SOTICS 2015

The Fifth International Conference on Social Media Technologies, Communication, and Informatics

ISBN: 978-1-61208-443-5

November 15 - 20, 2015

Barcelona, Spain

SOTICS 2015 Editors

Carla Merkle Westphall University of Santa Catarina, Brazil

Dumitru Roman, SINTEF, Norway

SOTICS 2015

Forward

The Fifth International Conference on Social Media Technologies, Communication, and Informatics (SOTICS 2015), held on November 15 - 20, 2015 in Barcelona, Spain, was an event on social eco-informatics, bridging different social and informatics concepts by considering digital domains, social metrics, social applications, services, and challenges.

We take here the opportunity to warmly thank all the members of the SOTICS 2014 technical program committee, as well as all of the reviewers. We also kindly thank all the authors who dedicated much of their time and effort to contribute to SOTICS 2014. We truly believe that, thanks to all these efforts, the final conference program consisted of top quality contributions.

We also gratefully thank the members of the SOTICS 2014 organizing committee for their help in handling the logistics and for their work that made this professional meeting a success.

We hope that SOTICS 2014 was a successful international forum for the exchange of ideas and results between academia and industry and to promote further progress in the area of social eco-informatics. We also hope Barcelona provided a pleasant environment during the conference and everyone saved some time for exploring this beautiful city.

SOTICS 2015 Chairs

SOTICS 2015 General Chairs

Philip Davis, Bournemouth and Poole College - Bournemouth, UK
David Newell, Bournemouth University - Bournemouth, UK

SOTICS Advisory Chairs

Petre Dini, Concordia University, Canada & IARIA, USA
Krzysztof Juszczyszyn, Wrocław University of Technology, Poland
Abdulrahman Yarali, Murray State University, USA

SOTICS Special Area Chairs on Social Networks

Feng Gao, Microsoft Corporation, USA
Lynne Hall, University of Sunderland, UK

SOTICS Special Area Chairs on eGovernment

Jennifer Watkins, Los Alamos National Laboratory, USA

SOTICS Special Area Chairs on Media

Claus Atzenbeck, Hof University, Germany / Aalborg University, Denmark

SOTICS Publicity Chairs

Nima Dokoohaki, Swedish Institute of Computer Science (SICS), Sweden
Christine Langeron, Hubert Curien Laboratory, Lyon University, France
Sandra Sendra Compte, Polytechnic University of Valencia, Spain

SOTICS 2015

Committee

SOTICS General Chairs

Philip Davis, Bournemouth and Poole College - Bournemouth, UK
David Newell, Bournemouth University - Bournemouth, UK

SOTICS Advisory Chairs

Petre Dini, Concordia University, Canada & IARIA, USA
Krzysztof Juszczyszyn, Wrocław University of Technology, Poland
Abdulrahman Yarali, Murray State University, USA

SOTICS Special Area Chairs on Social Networks

Feng Gao, Microsoft Corporation, USA
Lynne Hall, University of Sunderland, UK

SOTICS Special Area Chairs on eGovernment

Jennifer Watkins, Los Alamos National Laboratory, USA

SOTICS Special Area Chairs on Media

Claus Atzenbeck, Hof University, Germany / Aalborg University, Denmark

SOTICS Publicity Chairs

Nima Dokoohaki, Swedish Institute of Computer Science (SICS), Sweden
Christine Langeron, Hubert Curien Laboratory, Lyon University, France
Sandra Sendra Compte, Polytechnic University of Valencia, Spain

SOTICS 2015 Technical Program Committee

Witold Abramowicz, Poznan University of Economics, Poland
Fernando Albuquerque Costa, Universidade de Lisboa, Portugal
Mehdi Asgarkhani, CPIT – Christchurch, New Zealand
Simon Reay Atkinson, University of Sydney, Australia
Liz Bacon, University of Greenwich, UK
Thierry Badard, Centre for Research in Geomatics - Laval University, Canada
George Barnett, University of California, USA
Grigorios N. Beligiannis, University of Patras, Greece
Lasse Berntzen, Vestfold University College, Norway

Lukasz Bolikowski, Interdisciplinary Centre for Mathematical and Computational Modelling - University of Warsaw, Poland
Boyan Bontchev, Sofia University "St. Kl. Ohridski", Bulgaria
Christos Bouras, University of Patras / Computer Technology Institute & Press «Diophantus», Greece
Félix Brezo Fernández, University of Deusto, Spain
Piotr Bródka, Wroclaw University of Technology, Poland
Erik Buchmann, Karlsruhe Institute of Technology (KIT), Germnay
María Luisa Carrió-Pastor, Universitat Politècnica de València, Spain
Te-Shun Chou, East Carolina University, USA
Aspassia Daskalopulu, University of Thessaly, Greece
Guillermo De Ita, Universidad Politécnica de Puebla, Mexico
Gert-Jan de Vreede, University of Nebraska at Omaha, USA
Petra Deger, PH Heidelberg, Germany
Matteo Dell'Amico, EURECOM, France
Sebastian Dennerlein, Graz University of Technology, Austria
Giuliana Dettori, Institute for Educational Technology - National Research Council, Italy
Raffaele Di Natale, DIEEI - University of Catania, Italy
Nima Dokoohaki, Swedish Institute of Computer Science (SICS), Sweden
Prokopios Drogkaris,University of the Aegean - Karlovasi, Greece
Arianna D'Ulizia, Institute of Research on Population and Social Policies - National Research Council (IRPPS - CNR), Italy
Wael M. El-Medany, University Of Bahrain, Bahrain
Szilard Enyedi, Technical University of Cluj-Napoca, Romania
Larbi Esmahi, Athabasca University, Canada
M. Caremen Fernandez-Gago, University of Malaga, Spain
Michael Fire, Ben Gurion University of the Negev, Israel
Jean Vincent Fonou-Dombeu, Vaal University of Technology, South Africa
Schubert Foo, Nanyang Technological University, Singapore
Jean-Gabriel Ganascia, University Pierre and Marie Curie, France
Rimantas Gatautis, Kaunas University of Technology, Lithuania
Chris Geiger, Duesseldorf University of Applied Sciences, Germany
Christos K. Georgiadis, University of Macedonia, Greece
Lilia Georgieva, Heriot-Watt University, UK
Apostolos Gkamas, University Ecclesiastical Academy of Vella of Ioannina, Greece
Bogdan Gliwa, AGH University of Science and Technology, Poland
William I. Grosky, University of Michigan-Dearborn, USA
Adrian Groza, Technical University of Cluj-Napoca, Romania
Richard Gunstone, Bournemouth University, UK
Panos Hahamis, University of Westminster – London, UK
Lynne Hall, University of Sunderland, UK
R. Gy Hashim, Universiti Teknologi MARA (UiTM) - Shah Alam, Malaysia
Hana Horak, University of Zagreb, Croatia
Yuh-Jong Hu, National Chengchi University, Taiwan
Yanping Huang, Google, USA
Yun Huang, Northwestern University, USA
Omar S. Hujran, Princess Sumaya University for Technology – Amman, Jordan
Darko Huljenic, Ericsson Nikola Tesla d. d., Croatia
Sergio Ilarri, University of Zaragoza, Spain

Barbara Re, University of Camerino, Italy
Miguel Rebollo Pedruelo, Universidad Politecnica de Valencia, Spain
Robert Reynolds, Wayne State University, USA
Carla Rodríguez, Instituto de Educação da Universidade de Lisboa, Portugal
Jia (Jackie) Rong, Deakin University - Melbourne, Australia
Azim Roussanaly, LORIA/University of Lorraine, France
Luis Enrique Sánchez Crespo, Sicaman Nuevas Tecnologías S.L. | University of Castilla-la Mancha, Spain
Marcello Sarini, Università degli Studi Milano-Bicocca, Italy
Sonja Schmer-Galunder, Smart Information Flow Technologies (SIFT), USA
Heiko Schuldt, University of Basel, Switzerland
Peter Smith, University of Sunderland, UK
Günther Specht, University of Innsbruck, Austria
Sofia Stamou, Ionian University - Corfu, Greece
Johann Stan, Ecole Nationale Supérieure des Mines de Saint Etienne, France
Dalibor Stanimirovic, University of Ljubljana, Slovenia
Emmanouil Stiakakis, University of Macedonia, Greece
Jacqui Taylor, Bournemouth University, UK
Maurice.Tchuente, University of Yaounde I, Cameroon
Marten Teitsma, Hogeschool van Amsterdam / University for Applied Sciences, The Netherlands
Raquel Trillo, University of Zaragoza, Spain
Gabriel Valerio, Tecnologico de Monterrey, Mexico
Rudi Vansnick, Internet Society Belgium, Belgium
Nikos Vrakas, University of Piraeus, Greece
Chunyan Wang, Pinterest Inc., USA
Fang Wang, UMASS - Amherst, USA | Nankai University - Tianjin, China
Liqiang Wang, University of Wyoming, USA
Nan Wang, Linkedin.com, USA
Jiang Wei, Missouri University of Science and Technology - Rolla, USA
Huadong Xia, Microstrategy Inc. / Network Dynamics and Simulation Science Laboratory - Virginia Tech, USA
Levent Yilmaz, Auburn University, USA
Ryan Zammit, Middlesex University, UK
Weining Zhang, University of Texas at San Antonio, USA
Wenbing Zhao, Cleveland State University, USA
Xingquan Zhu, Florida Atlantic University, USA

Copyright Information

Table of Contents

Smart Navigation:

Using Artificial Intelligent Heuristics in Navigating Multiple Destinations

Hatem F. Halaoui
Computer Science
Haigazian University, Lebanon
Email: hhalaoui@haigazian.edu.lb

***Abstract*—Navigation applications are becoming an essential need in any mobile device. Finding the best path (time and distance) from an address to another is one of the most asked queries among driving users. Moreover, finding the best path with multiple destinations is a query that could be asked by many, including commercial companies' drivers (similar to the famous "Traveling Salesman Problem"). Google maps, Yahoo maps, and tens of other solutions are examples of such mobile applications. Calculating the best driving path between two addresses is subject to many factors including distance, road situation, road traffic, speed limitations and others. This paper presents the use of smart heuristic functions, as well as an efficient data structure to be used in finding efficient path between multiple points (addresses) rather than one destination. It presents spatial databases, current solutions, heuristics in Graph problems, and finally a smart solution (our new Algorithm A*Multiple) using a smart heuristic function to determine the best path between multiple destinations.**

Keywords: Smart Navigation, Artificial Intelligence, Heuristics, GIS.

I. INTRODUCTION

This section introduces the paper's main subject. First, the idea of heuristics is briefly presented. Second spatial databases are introduced as the underlying databases to store related data. Finally in this section, a brief introduction to Geographical Information Systems (GIS) and driving path application are presented.

A. Heuristics

Most of what we do in our daily lives involves heuristic solutions to real-time problems. As an adjective, heuristic pertains to the process of gaining knowledge by intelligent guess rather than by following some pre-established formula [1][2]. In map problems, when moving from one point to another to reach a certain destination, we have two options:

1- The algorithm tries all possible paths from all possible neighbors (next address on the way to destination). It keeps doing this until destination is reached. Finally it chooses the best path among all possibilities.

2- At each location, the algorithm chooses the next move smartly using some evaluation function (called the heuristic function)

The first option is very time consuming and does not match with real-time problems. As a result, a solution using the second is being adopted in this paper

B. Spatial Databases

Spatial databases are the main data warehouses used by Geographical Information Systems. Spatial databases are databases used to store information about geography like: geometries, positions, coordinates, and others [8]. Also, they might include operations to be applied on such data.

C. Geographical Information Systems and Driving Path Applications

Geographic Information System (GIS) is a collection of computer hardware, software for capturing, managing, analyzing, and displaying all forms of geographical information [8].

Finding the Directions (driving/walking) path is one of most asked queries in GIS applications. These are the most important factors that influence such criteria::

- Distance: Distance between the source and destination.
- Road situation: Is the road closed? Is it raining?
- Road traffic: How much traffic?
- Speed limitations: Is there many traffic lights? Is it a local road or highway? What is the average speed?

D. Navigating Using Heuristic Functions

In this paper, we present the issue of navigating multiple destinations in any order. Our main problem is to find the fastest path between a given source passing over all given destinations in any order. The importance of our approach is that existing solutions, like Google Maps [7], let the user choose his order of destinations rather than suggesting a fast path.

Moreover, calculating the fastest path with traditional Mathematical algorithms like Hamilton path has a high time complexity and hence time-expensive for real time problems like the one in this paper. As a result we use heuristic algorithms like A* to incredibly minimize the running time of such navigation real-time solutions.

Our approach offers the user a full path with an order of destinations claiming an efficient time. The main concern is that heuristic functions does not guarantee an optimal (best) solution. For this reason, choosing the heuristic function is an important factor for getting good results. Choosing a good heuristic function in order to choose our series of destination is an open research question. Moreover, choosing the heuristic function is highly dependent on the geography of surface in query.

The paper is organized as follows Section 2 presents some related work including widely used applications. Section 3 presents the main solution in this paper. Section

4 discusses some results and finally section 5 presents conclusions and future work.

II. STATE OF ART AND RELATED WORK

Most of current applications provide a one destination solution (users provide the application with source and destination). If users intend to visit multiple destinations, they have to decide the order of visits and make different queries each time.

This section presents an overview of related theoretical and applied related work. First, the use of Artificial Intelligence in path problems is presented, discussed and compared with traditional optimal solution algorithms. Moreover, one of the most used applied application (solutions) is presented.

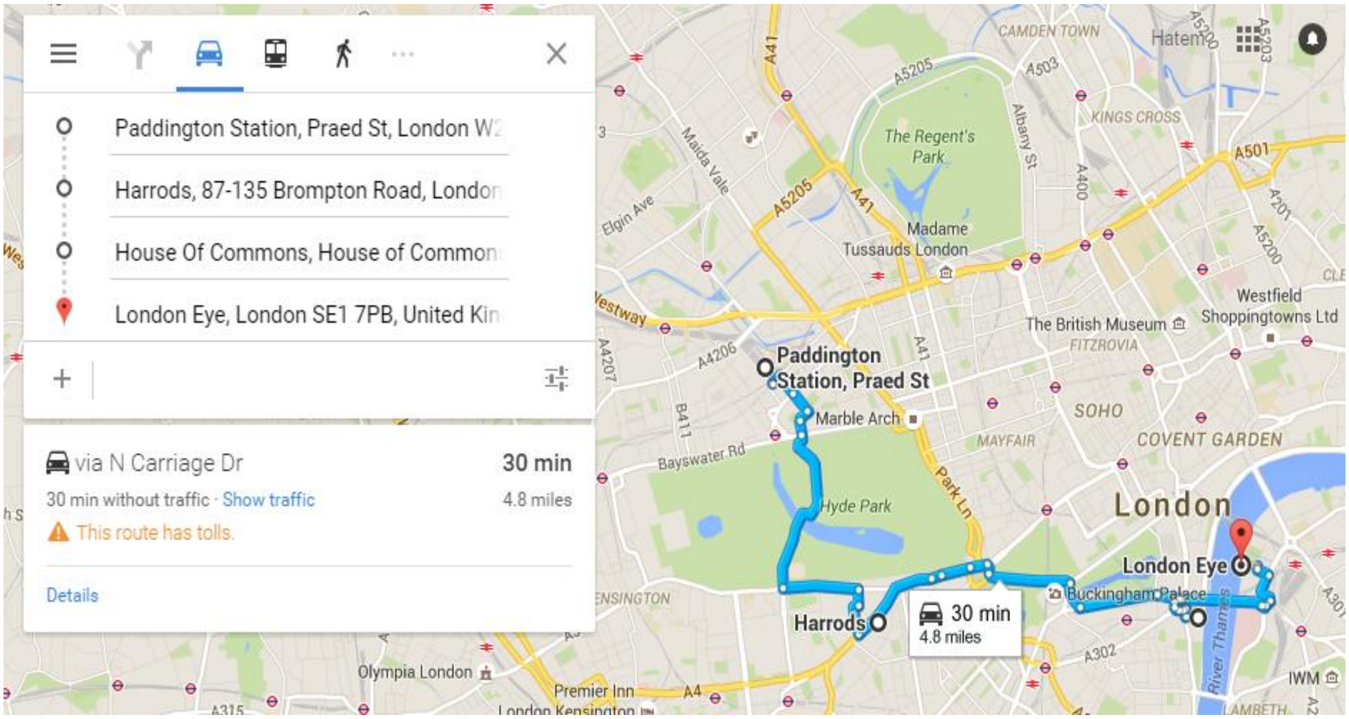
A. Driving Direction Applications: Google Maps as an example

This sub-section presents a widely used application for finding driving directions: Google Maps.

Google Maps [7] is a Web-based service that provides detailed information about geographical regions and sites around the world. In addition to conventional road maps, Google Maps offers aerial and satellite views of many places. In some cities, Google Maps offers street views comprising photographs taken from vehicles. Google Maps [7] offers several services as part of the larger Web application, as follows:

- A route planner offers directions for users of routes and public
- Google Maps for Mobile offers a location service for motorists that utilizes the Global Positioning System
- Google Street View enables users to view and navigate through horizontal and vertical panoramic street level images of various cities around the world.
- Provides user interaction.

Figure 1 shows an example a driving directions query using Google Maps [7]. The query is to get driving directions, over multiple destinations in London: Paddington station, Harrods, House of Commons, and London Eye. It also offers Real-time Traffic information. However Google Maps [7] does not suggest any order of visits. The user has to provide Google Maps with the order and he has to make multiple trials and look for the best sequence of destinations to be visited.



Paddington Station, Praed St, London W2

Harrods, 87-135 Brompton Road, London

House Of Commons, House of Common

London Eye, London SE1 7PB, United Kin

via N Carriage Dr 30 min

30 min without traffic · Show traffic 4.8 miles

⚠ This route has tolls.

Details

Figure 1. Path over multiple destinations in London

B. *Artificial Intelligence and Driving Directions*

Artificial intelligence is involved in graph searching algorithms. Russel and Norving [2] present many intelligent graph searching algorithms. Here are two important ones:

1. Greedy Best-First Search
2. A* Search

The main idea behind these algorithms is that they do not try all possible cases to give an answer. On the other hand, the algorithms use heuristic function to un-consider some of the paths. This issue saves huge amount of time but does not guarantee a best path. However, finding a good heuristic function could guarantee up to 95% finding the best path. Section 3 include the A* search algorithm which clears the idea in this section

C. *A* Traffic: Design, Algorithms and Implementation*

This section presents the application algorithms and the application of the intelligent driving path application used in our previous work, which is extended in this paper.

A* [2] is an Artificial Intelligent graph algorithm proposed by Pearl. The main goal of A* is to find a cheap cost graph path between two vertices in a graph using a heuristic function. The main goal of the heuristic function is to minimize the selection list at each step. In the graph example, finding the shortest path from a node to another has to be done by getting all possible paths and choosing the best, which is very expensive when having a huge number of nodes. On the other hand, using an evaluation function (heuristic) to minimize our choices according to intelligent and practical criterion would be much faster.

The heuristic function is not a constant static function. It is defined according to the problem in hand and passed to the A* as a parameter. In the case of A* search for a direction path, F is built up from two main factors:

H = Straight Line distance to destination.

G = Distance Traveled so far.

$$F = H + G \quad (1)$$

At each node n, we compute F (n) and we choose our next step accordingly.

A* Algorithm

```
A*(Graph, Source, Destination)
Task: takes a Graph (Vertices and Edges), Source and Destination (Vertices) and returns the Best path solution (stack of vertices) from Source to Destination

If Source = Destination then return solution (stack)
Else expand all neighbors Ni of Source
   Mark Source as Unvisited
   For each Neighbor Ni
       Get VNi = H(Ni, Destination)
       Add all (Ni, Vi) to the Fringe (list of all expanded Vertices)
       From the Fringe, Choose an Unvisited Vertex V with Least Vi
       If no more Unvisited return Failure
       Else Apply A*(V, Destination)

H(V, Destination)
Task: takes a vertex V and evaluate it using a heuristic function
Return: DistanceSoFar + StrightLineDistance (V, Destination)

Where
Distance_So_Far= Distance taken so far to reach the Node V
Stright_Line_Distance (V, Destination) = straight line distance to destination calculated by using the coordinated of V and destination
```

Figure 2 is an example of the A* algorithm behavior to find a path starting from “Arad” to “Bucharest”, cities in Romania [2]. First of all we start at Arad and go to the next neighbor with the best heuristic function (Sibiu). Second, explore all neighbor of Sibiu for the best heuristic function. The algorithm continues choosing the best next step (with the least value of heuristic function) until it reaches Bucharest. The interesting thing is that all vertices with values (calculated using the heuristic function) kept in the fringe in order to be considered at each step.

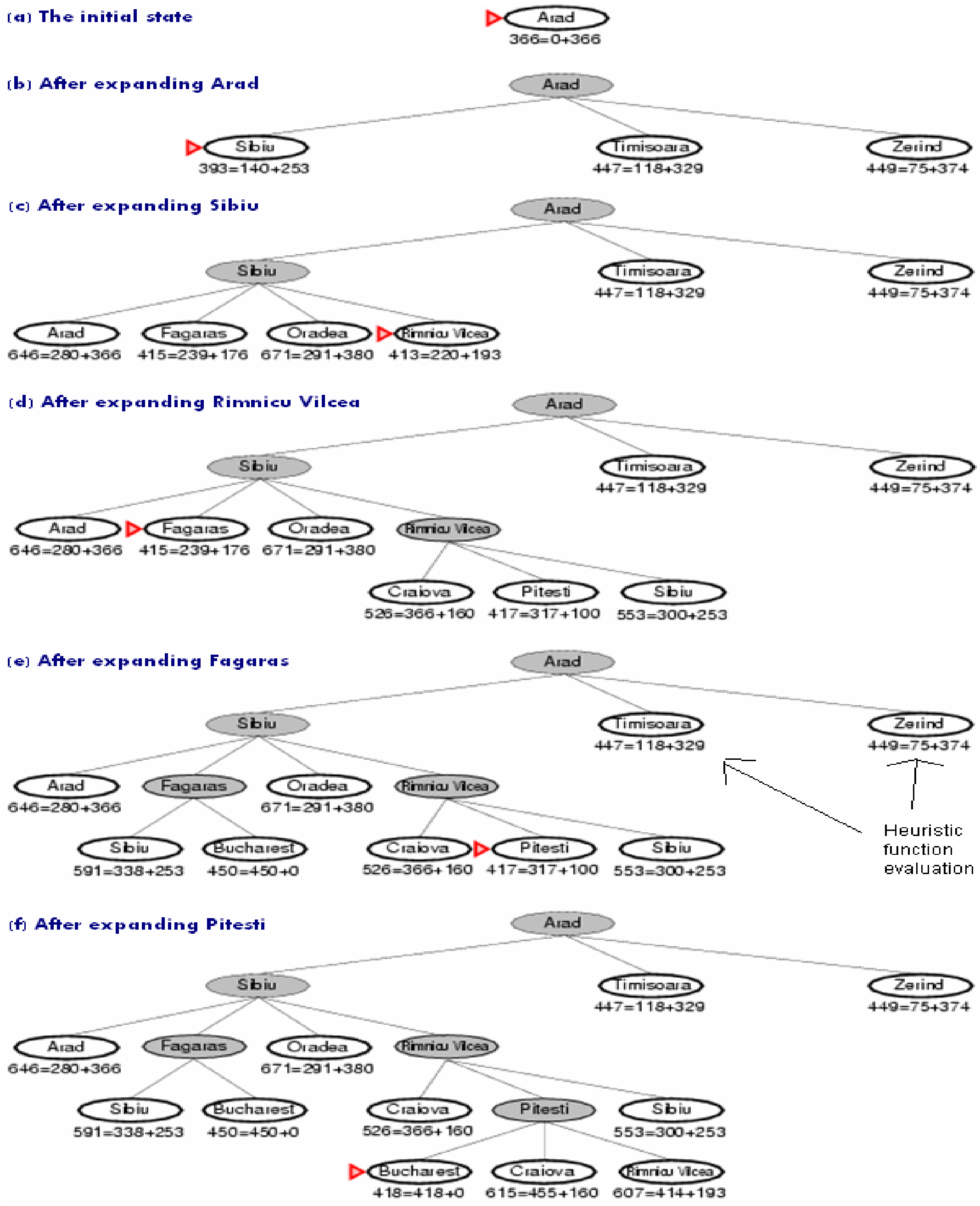


Figure 2. A* algorithm behavior to find a path starting from “Arad” “Bucharest city” [2]

D. *A*Traffic: A Variation of A* with Road Traffic as a Factor*

A*Traffic [5] (our previous approach) is a variation of A* with the ability to take Online traffic into consideration. The main job is done in the heuristic function where a new factor is used to choose the next step. The new factor is the average traffic value (got online from real time databases) represented in the following form time/distance (example: 3 min/km). The new Heuristic function will be:

$$F = H + G + T \qquad (2)$$

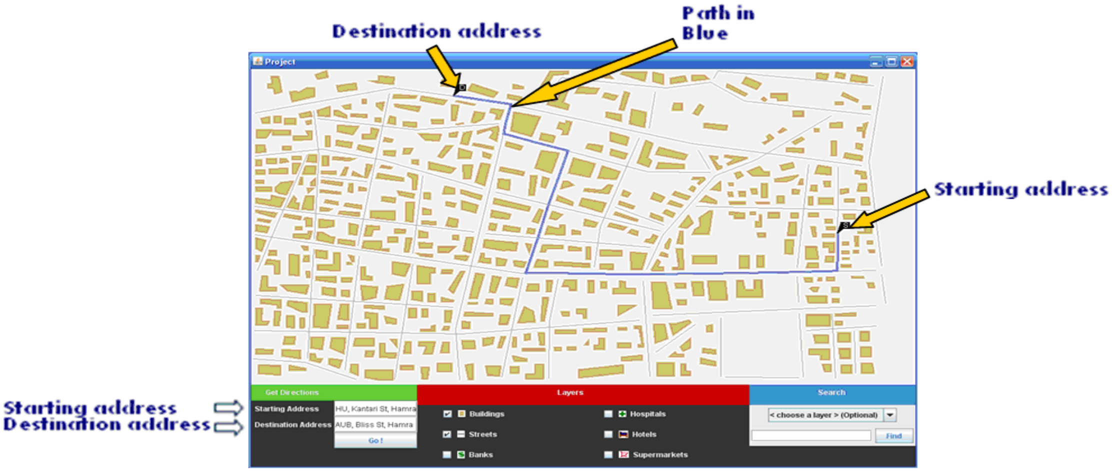
Where:
H = Straight Line distance to destination (distance between two coordinates).
G = Distance Traveled so far.
T= Average Traffic delay got from real online sensors.

E. *Testing Tool: Query Example*

This sub-section present the layout of the testing tool developed to test the algorithm proposal “A*Traffic”. For this purpose, an example query is presented.

A Query example

This example (Figure 3) demonstrates the main feature of the software. It provides the user with the driving directions between “HU, Kantari St, Hamra” (Haigazian University) and “AUB, Bliss St, Hamra” (American University in Beirut) in Beirut, Lebanon. The blue path generated is a short path (using A*Traffic) to follow in order to drive from the start address to the destination address.



Starting Address: HU, Kantari St, Hamra
Destination Address: AUB, Bliss St, Hamra

Figure 3. Path from “HU, Kantari St, Hamra” to “AUB, Bliss St, Hamra”

III. USING HEURISTICS AND TIME-WEIGHTED GRAPH IN MULTIPLE DESTINATIONS DIRECTION SOLUTIONS

This section, presents our new extension of the previous solutions to provide a good solution having multiple destinations to be visited.

In this section, we propose our smart solution for navigating multiple destinations. The section includes proposals for:

1. Time Weighted Graphs (TWG) as the main data structure used in our solution.
2. A*Multiple: the proposed smart algorithm
3. Execution examples of A*Multiple

A. *Time Weighted Graphs*

In our previous paper [6], we present TWG as our main data structure: a graph representing the map with edges weighted by numbers (minutes) representing the estimated time needed to drive the edge.

Distance is usually the main edge weight in Graphs. In TWG, time is used instead. A graph edge in the graph represents a road, street, highway, or part of any. Each of these has an average speed limit that is calculated using road speed sensors. The job of the real-time street sensor is to watch traffic and send the average speed during some period. The graph edge weight in terms of time (minutes) is computed as follows:

Average speed (AV) (minutes) =

Sum of speeds of n cars over a Period T (miles/minutes) ÷ n

The initial weight of the edge (minutes) (W) = (edge distance (miles) ÷ AV (miles/minutes))

B. *Example: Part of Manhattan in a Time-weighted Graph*

This Section shows the data structuring (transferring the map into TWG) of part of Manhattan (taken from Google Maps). Figures 4 and 5 show the location of vertices in the map (Figure 4) and digital graph after construction (Figure 5).

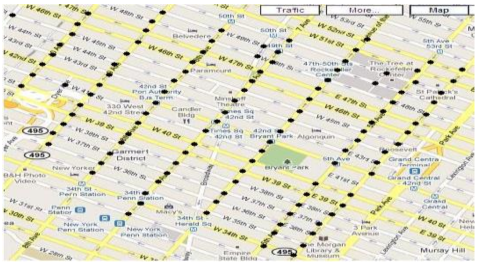


Figure 4. Modeling graph vertices

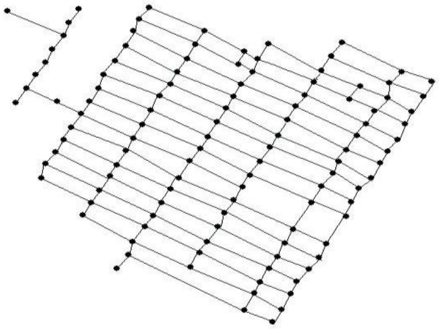


Figure 5. Building the graph edges (directed)

*The A*Multiple Algorithm*

The main idea behind A*Multiple is to find the best path (shortest in time) to visit multiple destinations in one tour. The algorithm uses a heuristic function to find the next destination and then uses the A*Traffic (which also use the same heuristic function) to travel to that destination.

A*Multiple (Source, Destinations)

Task: find an efficient path from source passing over all members in Destinations array.

Returns: 2 Lists

- VSL: The Vertices Solution List VSL, which is the vertices t visit in order
- PSL: Path Solution List PSL, which is the list of paths to take each time to each destination (vertex)

Pseudo code

```
If Destination is Empty return Done
For all Vertices Vi in Destinations
  Di=H(source, Vi)
  Get the Vs with the Minimum Di
  Remove Vs from Destinations
  Add Vs to the Vertices Solution List VSL
  Add A*Traffic (Source, Vs) to the Path Solution List PSL
 If A*Traffic fails return Failure.
A*Multiple (Vs, Destinations).
```

*How does A*Multiple Work?*

This section presents the execution of A*Multiple. To present our approach better, consider the following problem:

Suppose I am at Paddington station and want to visit the following destinations in London: “Eye of London”, “House of Commons”, and “Harrods”. If my only priority is time, means that I can visit them in any order with efficient time. In this case, I have to choose my next destination (at each step) smartly.

After creating the Time-Weighted graph (vertices shown in black in figure 7, over 5000 vertices) over the map of London (from Google Maps), the A*Multiple will return the following:

VSL: Harrods, House of Commons, Eye of London.

PSL: Path1, Path2, Path3.

Where VSL is the ordered list of destinations to be visited, PSL is the list of paths from each destination in VSL to the next one, Path1: Paddington – Harrods, Path2: Harrods – House of Commons, and Path3: House of Commons – Eye of London.

Each of these paths is calculated using A*Traffic.

Figure 6 shows these solutions in different colors: orange (Path1), Blue (Path2) and Pink (Path3). It also gives estimated time of each path according to current (at time of calculation) traffic situation.

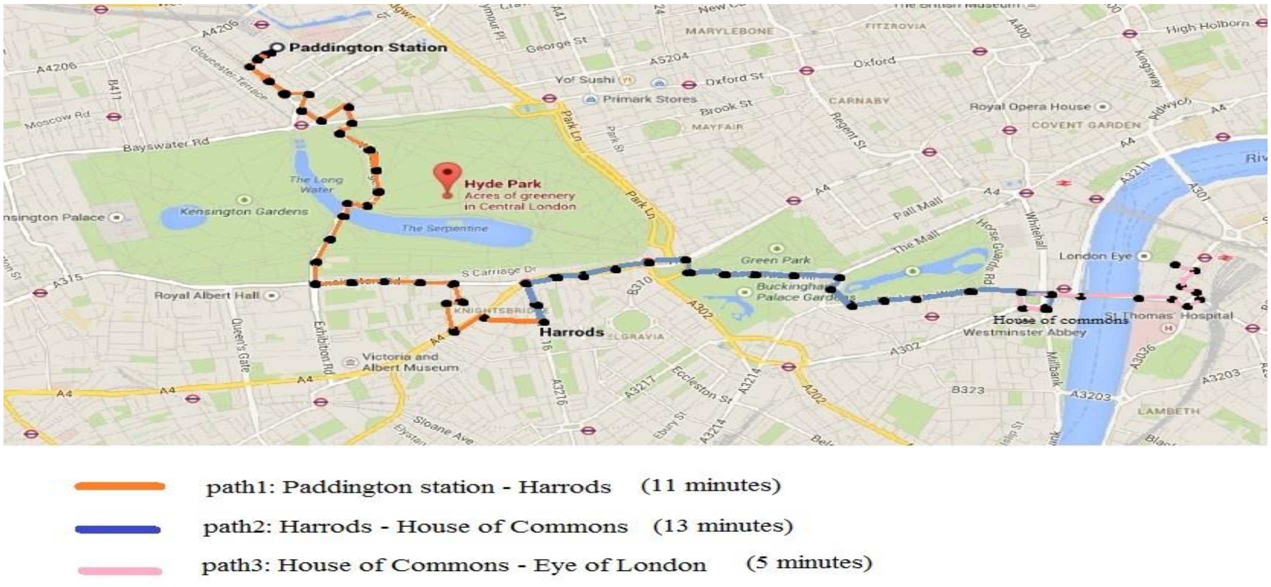


Figure 6. Paths for Multiple destinations (Paddington, Harrods, House of Commons, and Eye of London

IV. RESULTS

We have developed a testing tool (to test our approach) where 100 samples were tested in 3 groups. Our results showed that our solution is optimal in 81%. Table I presents the gathered results in each group/each case where:

- Optimal solution: Absolute best solution.
- Good solution: takes maximum of 20% more time than optimal solution.
- Bad solution: Takes more than 20% more time than optimal solution.

TABLE I. PERCENTAGES OF QUALITY OF SOLUTIONS

Distances	Optimal solution	Good Solution	Bad Solution
More than 10 destinations Over 5321 vertices	73 %	19%	8%
Less than 10 destinations Over 5321 vertices	88%	9%	3%
Average	81%		

Our approach is being evaluated according to optimal solutions (best solution). These optimal solutions were computed manually. Finding such solutions is time consuming and not applicable real time problems like navigation problems.

The Comparison of our solution to Google Maps [8], shows that ours suggests a smart order of visits over all destinations while we have to give the order of destinations to Google Maps In order to get the best solution we have to do try all possible orders of destinations, then we have to check for the best order (least time). This is not reliable when having a long list of destinations (over 10).

V. CONCLUSIONS AND FUTURE WORK

In brief, our approach, using A*Multiple algorithm with time-weighted graphs, has the following advantages:

- It uses time-weighted graphs, which takes distance and speed into consideration.
- It considers multiple destinations
- It saves a lot of execution time.

The reason this approach saves a lot of time is because if we do not use heuristics, we will have to find the path by getting all combinations (Hamilton Path). This will result into an exponential time algorithm. On the other hand, using heuristics considers options with best values when heuristic function is applied and we end up with polynomial time algorithm (A*Multiple and A*Traffic are in O(n^3) since they rely on A* [2]).

Our future work is focused on how to use heuristics combined with discrete structures algorithms like Hamilton path and Hamilton circuit. Moreover, finding the best heuristic function remains an open research question in the future.

REFERENCES

[1] J. Pearl, *Heuristics: Intelligent Search Strategies for computer Problem Solving*. Addison Wesley, Reading, Massachusetts, 1984.

[2] S. Russell and Peter Norving, *Artificial Intelligence a Modern Approach*. Prentice Hall, Upper Saddle River, New Jersey, 2003.

[3] H. Halaoui, Smart Traffic Online System (STOS): Presenting Road Networks with time-Weighted Graphs". IEEE International Conference on Information Society (i-Society 2010) London, UK. June 2010, pp. 349-356.

[4] Google Earth Blog Google Earth Data Size, Live Local, New languages coming Available: http://whatis.techtarget.com/definition/Google-Maps. Retrieved: September, 2015.

[5] H. Halaoui, "Smart Traffic Systems: Dynamic A*Traffic in GIS Driving Paths Applications". Proceeding of IEEE CSIE09. IEEE, Los Angeles, USA. March. 2009, pp. 626-630.

[6] H. Halaoui," *Intelligent Traffic System: Road Networks with Time-Weighted Graphs"*. International Journal for Infonomics (IJI), Volume 3, Issue 4, December 2010, pp. 350-359.

[7] Google Maps. Available: https:// Maps.google.com. Retrieved: September, 2015.

[8] H. Halaoui. "Spatial and Spatio-Temporal Databases Modeling: *Approaches for Modeling and Indexing Spatial and Spatio-Temporal Databases"*. VDM Verlag, 2009.

What Do People Affected by Cancer Talk about Online?

Text analysis of online cancer community usage in Bosnia and Herzegovina

Suncica Hadzidedic Bazdarevic, Alexandra I. Cristea
Department of Computer Science
University of Warwick
Coventry, United Kingdom
e-mail: s.hadzidedic@warwick.ac.uk, a.i.cristea@warwick.ac.uk

***Abstract*— Studies report that health information searching is among the top three activities on the Internet. Internet resources are a good alternative for initial information and health managing support, due to their accessibility and availability. However, *little is known about the type of support and information that people affected by cancer seek for and exchange via the Internet, in online communities, especially in non-English speaking environments*. To bring light to this important matter, we conducted a text analysis on an online cancer-related forum from Bosnia-Herzegovina. It revealed that the predominant topics of discussion are: *general cancer-related* and *time-wise discussions*, *various treatments*, *diet*, *doctors* and how to interpret *medical reports*, information exchange on *specific cancer-types*, *advice to caregivers*, *religious support*, and *community support* to members suffering from cancer. The most frequently exchanged word referred to the state of: being, being and doing something alone and by oneself. The majority of users of the online community were cancer patients' caregivers. This study represents a starting point for *identifying areas for improvement of online support and information for people affected by cancer, primarily in online communities, specifically for Bosnia-Herzegovina*, but can be generalized to countries with similar social, cultural, economic and public-health situation.**

Keywords-online health communities; text analysis; cancer; online health information; Bosnia-Herzegovina.

I. INTRODUCTION

Cancer is still recognized as the leading cause of death worldwide [1]. To this day the "number [of deaths globally caused by cancer] is far greater than the total number of deaths caused by HIV/AIDS, tuberculosis and malaria combined" [2]. The latest worldwide statistics show that 14.1 million adults had cancer [1]. Cancer does not only affect the person suffering from it. It impacts the life of the cancer patient's closest family, extended family and friends [3]. This research refers to *all* the people affected by cancer, which includes those on whose life cancer had a *direct* and *indirect* influence, and even those who are interested in informing themselves about cancer for prevention purposes. The global population affected by cancer is, thus, evidently much larger. Moreover, given cancer prevalence, the duration for which caregivers are needed, in most cases, expands over the period of several years or more [3].

Despite the fact that the health sector is one of the most resistant-to-change industries, *technology adoption for healthcare service improvements* has been witnessed in some countries, but progress and needs are not uniform, even in the Western developed countries [4].

Bosnia and Herzegovina (B&H) was selected as the environment to conduct the study in, due to the particular situation in the country. Public health services in B&H are still in the process of recovery and reform after the 1992-95 war; they are characterized by institutional fragmentation, existence of powerful interest groups, informal payments, unequal access to health care and low service quality and efficiency [5]. Support groups and cancer associations do not cover all types of cancer, are not vocal enough, and are not always easy to access offline.

The accessibility and availability of Internet resources could present a good alternative to people affected by cancer to find initial information and support. Electronic health (eHealth) contributes to disease management by facilitating information exchange between and among interest groups – medical professionals, patients and caregivers, and empowers patients by providing improved online support [6] [7]. However, a preliminary investigation of online health resources in B&H revealed that those from B&H institutions and in Bosnian language are but a few [8]. These are mainly endeavors by non-governmental cancer organizations (NGO) or private initiatives that aim at promoting their activities. Importantly for this study, the information and facilities presented are not always relevant. Information provided is mainly related to general health, and reused from online health services from other countries. Little consideration is given to reliability and comprehensiveness of cancer-specific online support and information. Moreover, the actual needs and interests of the country-specific population who are suffering from cancer are not addressed.

Even worldwide, patient input and opinions are often disregarded in health-related product adoption decision [4]. Nevertheless, when it comes to introducing improvements in eHealth, what is essential is a deeper understanding of how these applications are utilized by the target users, and what they consider useful and easy to use technology [6].

This study addresses an issue in eHealth - *little is known about the type of support and information that people affected by cancer seek for and exchange via the Internet, expressed by their behavior and activities, in online* *communities, especially in non-English speaking environments, as is B&H*. The goal was, therefore, to "listen" to the conversations led on the Internet by those members of the B&H population who are affected by cancer. In order to understand what it is they require, the usage of existing online cancer-related information and services was evaluated applying text analysis. This work extracted *topics most frequently discussed in a cancer-related online community*. The classification of topics of interest should serve in devising a set of content- and service-improvement recommendations for online communities for people affected by cancer, especially those from B&H.

This paper is structured around four remaining sections. Section II gives an overview of the research area and the related work. Analysis and results are presented in the third section, followed by a discussion of the findings. The last section concludes the paper and overviews future directions.

II. Background and Related Work

One of the primary purposes of Internet usage is to seek information about health [9]. While the traditional means for obtaining health information have stagnated [10], there has been a worldwide growth in online health information usage [11], but a particularly noticeable increase of around 50% in European countries in the period 2001 to 2009 [10][12]. The Pew Internet Project [13] concluded that 59% of adults in the U.S. have used the Internet to look for health information, usually (55% of the cases) about a specific illness or medical problem, and 43% searching for available treatments; followed by topics related to diet, particular doctor and experimental treatments [9][14]. Moreover, Internet health seekers look in equal percentage (39%) for health and medical information for themselves or for someone else [15].

The research presented here focuses on *online health information usage by people affected by cancer*. In a US based study, approximately 40% of the sample used the Internet to search for cancer-related information [9]. While exploring how Internet usage influences men and women affected by one of five different types of cancer in the UK in the period 2001 to 2002, [16] found that cancer patients used the Internet for three main reasons: the extent of resources available; to secretly verify what was told to them by their physicians; the need to show others and prove to themselves that despite their illness they are socially fit.

Health information available on the Internet is particularly appealing to people affected by cancer, as the sheer amount and variety of this type of information "which could be used by cancer patients for treatment decisions, medical consultations and social support, as well as by non-cancer patients for prevention, screening and risk evaluation" becomes a means to ease their physical and psychological ordeal with this illness [17].

There are various major health portals worldwide, such as the US-based PubMed [18], the UK's NHS [19], the German GoPubMed [20], as well as those that provide cancer-specific services and support, including the UK's Cancer Research UK [21] and Macmillan Cancer Support [22], and the US American Cancer Society [23]. The step forward in healthcare is home care, in particular with the aging of world population and increase in those requiring long term care [7]. Online health resources are the indispensable means for home care, and, thereby, patient empowerment. However, it is often the case that health consumers have to navigate through a sea of information, which is often difficult to understand [24], and sometimes irrelevant [25]. A user affected by cancer, whose needs are very specific, and who receives unsuitable, irritating information, might quickly terminate the visit to the health-website. To increase user engagement, online health services should address three main points: information quality, user interaction and tailoring information to user's needs [26].

Information better matched to user's health literacy and health situation, with a particular focus on patients with chronic medical conditions [25], is argued would lead to improved user engagement. Design of web tools for people with special health needs is another consideration. Including patients and caregivers in the decision making process [27] for eHealth design and information tailoring would result in more empowered patients, participating more actively in personal healthcare [14][25].

The lack of B&H online health services was mentioned in the previous section – the lack in number, but also in the provision of cancer-specific support. Some of the B&H health portals with an online presence are: *klix.ba* – a news portal offering health-related news and a forum feature used, among others, by cancer affected population; *port.org.ba* – a portal and community provided by an NGO for people with malignant diseases, mainly offering cancer-related articles from other sources; *source.ba* – a news portal; *srcezadjecu.ba* – a website, by a cancer association helping children suffering from cancer, used to present the NGOs projects and activities; *bhzdravlje.ba* – a health portal offering general health information, provided by a private organization in cooperation with one of the cantonal ministries of health. But often, online resources from the neighboring, similar-language-speaking, countries are used as an alternative, "the next best thing", solution.

Moreover, in order to evidence patient empowerment, via their improved online engagement, health service providers would have to show interest in exploring and understanding cancer patients' and caregivers' needs, to be able to match these with adequate information and support. However, to the best of our knowledge, studies on online health information seeking among people affect by cancer in B&H are lacking. We have addressed this in a previous pilot study [8], which showed that the target population in B&H does use the Internet to seek cancer information - the main reason for going online to seek for cancer information or advice is having someone close diagnosed with a medical condition.

However, this was an isolated study applied to the B&H population, and one that focused on collecting participants' perception on the matter. A gap still exists in *understanding the actual behavior and activities of cancer affected people in B&H*, and wider, when searching for cancer-related support and information on the Internet. To reduce the global burden of cancer, the specific needs of individual communities have to be addressed, primarily by the public health services, including those provided on the Internet.

III. Analysis and Results

A. Data Collection and Pre-processing

A health forum ‘Cancer and fighting cancer’ accessible from Klix.ba [28] - the most popular B&H online portal, with the most diverse forum that has the largest member base, was used. The forum content was in Bosnian language. *All* forum posts between December 2012 and April 2014, were extracted, as *semi-structured* text. Participants who made a minor number of posts (up to 3 posts) and had an insignificant amount of text posted compared to other forum participants, were excluded from the current analysis. The resulting sample was that of 38 users, 9597 unique words (length 1 character and above), and total word count of 38,387. Participation in the forum varied: membership ranged from < 1 to 12 years; some participated in a number of Klix.ba forum discussion boards and others participated specifically in the here studied cancer-related part of the forum; thus, the number of posts ranged from 3 to 28087.

Publicly available demographic data about the forum participants was collected from their user profiles, such as: username, number of posts, start of membership, age and location. Gender was automatically extracted from the text posted by the participants, given that in Bosnian language there is a gender difference in the third person tense.

B. Data Analysis

Word frequency analysis was performed using the QSR NVivo 10 tool [29] - as the available software to the researchers and the one popular for analyzing unstructured data, e.g., online textual content. The most frequent 1000 words of 3 or more characters were searched for in the sampled posts in Bosnian language; most stop words (including connector words: but, which, when, and similar) were excluded or manually removed from further analysis. Figure 1 shows a word cloud representation of the most frequent words encountered after this process.

Using the NVivo tool to analyze Bosnian text introduced a limitation to the study. While this tool has the functionality to identify stemmed words and synonyms for English, it was able to detect only slight changes in Bosnian words, and thus almost in no cases were stemmed words for the root word in Bosnian included in the frequency calculation.

C. Results

The demographics of the users showed 60.5% were female and 26% male; the remaining 13.5% was unknown. Those participants that did provide personal data were in their thirties and one a senior citizen, and lived in larger cities such as Sarajevo, Tuzla and Mostar.

The most frequent word in the sampled forum posts, as shown in Figure 1, is ‘**sam’** (weighted percentage 0.79). The word ‘sam’ occurs in 26 of the 38 sources, and in total 305 times in the sampled text. This is the highest count for a single word occurrence under the selected conditions. Translated into English the word ‘sam’ can be used as a verb ‘[I] am’ or ‘[I] have’, or otherwise it can mean ‘alone’ or ‘by him/her/your-self’. In all cases, it indicates that participants of the sampled forum posts were most frequently referring to *themselves*, that *the opinion they were expressing is theirs*, *to have done or gone through something related to cancer themselves*, or *to be affected by cancer themselves*. It can also be implied that they had to *go through something alone*, such as perform a cancer-related activity by themselves. The greatest number of references to the word ‘sam’ was found in the posts made by two participants (further anonymized here): *user VH* (86) and *user J* (37). Both of these forum members were cancer patients, as extracted from a more detailed analysis of their posts.

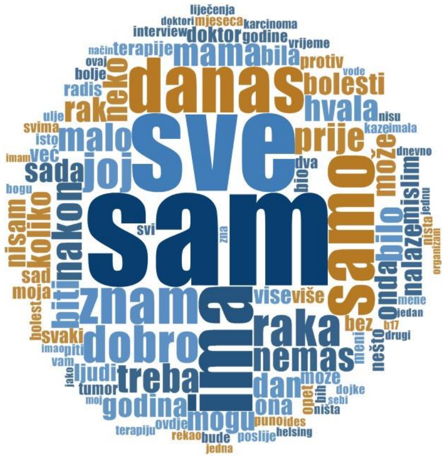


Figure 1. Word cloud representation of the most frequent words

Forum posts analysis further reveals other frequently occurring words, which can be indicative of the type of topics of greatest interest to people affected by cancer in B&H. A sample of highly ranked cancer-related words from the forum posts’ 1000 most frequent words is listed in Table 1. The table presents the Bosnian words, along with their respective translation into English language and the weighted percentage - representing the frequency of the word relative to all the words counted from the sampled forum posts. The word *‘mama’* (i.e., *mother* in English), for example, occurs in the posts made by 13 of the 38 sampled forum members. Analyzing the word in relation to other words within the posts, it is evident that ‘mother’ is the person mentioned as the one suffering from cancer. The indirectly affected forum participants mainly discussed: problems with treatments, the cause of positive and negative moods of their caretaker (mainly mothers), their health condition, the type of cancer they are fighting (mainly breast cancer, but not only limited to this cancer-type), and asking for treatment and diet advice.

The 1000 most frequent words (≥ 3 characters) were manually grouped into 9 categories of frequent discussions on the analyzed cancer forum. Words such as connectors, pronouns, and cancer discussions’ unrelated words were excluded from the categorization.

Figure 2 represents the relative frequencies of the 9 discussion categories. The most frequently exchanged words are those belonging to the category *General cancer-related discussions* (sum of words count is 1775). However, there was also great interest in time-related discussions - *Time* (sum of word counts: 1378), which was not one of the

TABLE I. FREQUENT WORDS IN THE KLIX.BA FORUM

Selected frequent words in Bosnian	*Selected frequent words – translation in English*	*Weighted percentage (%)*
sam	am / have / alone / by oneself	0.79
rak(a)	(of) cancer	0.47[a]
danas	today	0.40
mama	mom / mother	0.21
bolesti	disease, illness(es)	0.18
godina	year(s)	0.18
nalaze	find, finding, medical report	0.17
doktor	doctor	0.13
terapije	therapy(s), treatment(s)	0.12
tumor	tumor	0.11
piti	drink, drinks, drinking	0.10
vrijeme	time	0.09
bogu	God (to God)	0.09
dojke	breast(s)	0.09
ulje	oil	0.09
b17	Vitamin B17	0.09
user VH	*user VH*	0.09
lijek	treatment / cure / medication / medicine	0.08
abd	abbreviation for ‘God willing’	0.07
caj	tea	0.07
herceptin	Herceptin[b]	0.07
hemoterapije	chemotherapy(s)	0.06

a. The sum of weighted percentages for rak (cancer) and raka (of cancer), i.e., 0.19 and 0.28, respectively.

b. Treatment which can be used for breast and stomach cancer.

frequently sought topics found in previous research on target population’s perception. Categorization of forum discussions was self-developed by the authors, nevertheless, it was based on the findings of previous studies [9][13][14] that reported the most frequently searched online health information were the topics included in the 9 categories next presented.

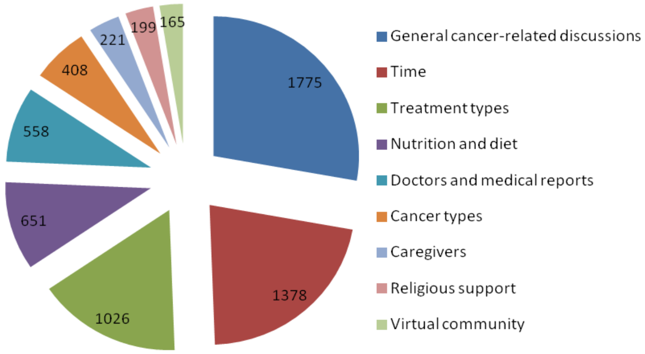


Figure 2. Categories of forum discussions and the sum of frequencies of the consisting words.

The 9 categories are listed below, sorted descending, based on the category level frequency (sum of counts of words making up the category). Each category is assigned the consisting total number of words, after excluding repeating occurrences of the same word (due to misspelling or missing symbols for affricates) and stemmed words.

- *General discussions* about cancer (76 words, e.g., *cancer, tumor, illness*)
- *Time* (45 words, e.g., *year, time*); which could be, but is not limited to, time since cancer has been diagnosed, duration of treatments and cancer prevalence, days stayed in hospital, or when a visit to the doctors is scheduled.
- Various *types of treatment* and medication (51 words, e.g., *therapies, b17 vitamin, cure, Herceptin, chemotherapy*); including alternative and experimental treatment, chemotherapies, and similar, and where to find it.
- *Nutrition and diet* (34 words, e.g., *oil, tea, drink*)
- *Doctors and medical reports* and findings (27 words, e.g., *doctor, find/finding/medical report*); who are the most renowned oncologists in the city, where they can be reached, experience others had with a specific doctor, doctor’s advice given, where specific medical analysis can be conducted, how to read medical findings and what they mean, are all questions asked within this topic category.
- *Type of cancer* (17 words, e.g., *breast, carcinoma, tumor*); narrowing down and specializing experience and information exchange to a specific type of cancer, and part of body where cancer occurred, rather than generic cancer information.
- *Caregivers* (9 words, e.g., *mother*); in this type of discussions, text analysis indicated that participants most often asked for support when a member of their closest family or a friend was affected by cancer, usually their mothers.
- *Religious support* (10 words, e.g., *God, God willing*); belief in God and help by a higher power.
- *Virtual community* (14 words, e.g., *user VH*); personal story sharing by forum participants who are suffering from cancer or are caregivers, and empathy and advice expression by their supporters.

IV. DISCUSSION

As cancer is not a local, but a global issue, all initiatives at supporting the needs of those affect by it resonate at a global level. The sample used for this study was geographically explicit, collected from an understudied population. Understanding what topics concern those who are affected by cancer in online communities around the world, not just in the Western developed countries, has various applications - from developing more geographically-tailored online support, to improved cancer prevention.

People affected by cancer in B&H are taking an action in improving their and their loved ones’ personal health, and/or doing this by themselves, as implied by the most frequent word - *'sam,'* in the analyzed cancer forum. This study indicates that women are predominant participants in B&H online health communities. It could be ascribed to the B&H culture, in which being a man still means hiding personal vulnerabilities. Nevertheless, worldwide studies also confirm that it is women who are more frequently online, seeking for health information and support [15]. What is remarkable in the B&H forum case is that the only two participants who were directly affected by cancer, and openly discussing their condition, were men. This occurrence also coincides with the B&H [30] and world statistics [31], which show that cancer is more frequent among the male population. On the other hand, the findings do suggest there is still a need to *encourage the male population affected by cancer to reach out for help and speak out about their health-related problems*. The majority of those using a cancer-related online community in B&H have joined these services in the desire to find *support for their loved one who was diagnosed with cancer*. The *members of the community mainly offer support – both emotional and information related* – to other members, who are cancer patients, or caregivers of cancer patients. Assistance is exchanged in *interpreting the diagnosis and approaching the specific type of cancer*.

The findings imply the primary type of conversation led in online health communities in B&H are those related to general discussions about cancer – admitting to be affected by cancer, the type of illness it is, finding out it metastasized, how it makes the person feel, what impact it has on the person's psyche and body, the pain it causes, help sought, the fight put into it, and similar. This type of online health information – about a specific medical problem, in addition to information on treatments, diet and doctors, has been reported the most frequently sought in Western countries [9][13][14]. However, this study is the first account, to the best of our knowledge, of Internet usage for health information seeking in B&H, based on actual user behavior on online health services, which, moreover, reports on cancer-specific information seeking. Furthermore, the findings show that people affected by cancer in B&H, when participating in online communities, are preoccupied with a type of discussion not commonly reported in previous studies on health information seeking on the Internet, this namely being the *time-related discussions*. They share information about when they had a medical appointment, the date of the next operation, how long they were receiving treatment, how long a loved one battled with cancer, up to a point of sharing how long their doctors said they are expected to live. Turning to online communities seems sensible for this type of personal story sharing; members want to exchange experiences, in an anonymous way, with others who have been in similar situations and, perhaps, to find consolation and a possibility for an alternative outcome. But, it is evident that this is a major concern of people in B&H who are affected by cancer, and thus, effort should be made by health professionals to inform this population about the time-related effects of cancer.

Similar conclusion can be drawn from turning to an online community to interpret medical reports. The *doctors and medical findings* category of discussions implies that health professionals are neglecting to explain the diagnosis to their patients, and are not cooperating with each other to inform their patients of the best specialists for their type of cancer. The result is that those affected by cancer have to turn to each other for advice from personal experience and for interpreting the doctor's notes on medical reports.

Another issue that was revealed is the variety of types of treatments (especially alternative and experimental) and diets suggested by the forum participants. These include: cytostatic drugs, Herceptin, Laetrile (or Vitamin B17), Zofran, Alimta, Avastin, Melatonin, Ecomer, sodium bicarbonate, cannabis, marihuana, plant roots, oil, petroleum, teas, etc. Many of these are very expensive, have side effects, some are illegal in B&H, and some diets suggested are even of questionable effect on human health. Health professionals and agencies in B&H, should invest into informing about all available treatments, raise awareness about the specific types of treatments and foods, verify or dispute the information exchanged by laymen, and assist in making available to the target population the needed effective treatments.

The frequently occurring categories of forum discussions related to virtual community, caregivers, religious support and time imply that the greatest benefit of online communities for cancer-affected people is, moreover, having a place to express oneself in an anonymous way; exchange experience, talk about the burden of cancer, knowing someone is listening and responding even if all they can offer is empathy and a kind word. Providing additional, *cancer-specific online communities*, assisted by health professionals for information verification, and raising awareness about these online services in the cancer-affected population, *is the type of support that could have a resonating effect*.

To extend this and the previous work [8], and confirm the findings of other related research [16], further ongoing work looks at the cancer patients' motivators for internet-based usage. In future studies, the intention is to explore the relationship between user-created content and their emotions with the application of different text analysis tools, e.g., LIWC: Linguistic Inquiry and Word Count.

V. Conclusion and Future Work

One of the approaches to reducing the global burden of cancer is diminishing the sense of helplessness and loneliness in those who are cancer sufferers, but also the sense of loss in those who are their caregivers. Electronic health can empower people affected by cancer, however, these online applications have to address the specific requirements of individual communities. This study evaluated the information exchange in an existing online community formed by people affected by cancer in B&H. Thereby, it identified the topics of discussions, type of support required, their concerns and information sought. It is primarily the obligation of public health institutions to take these findings into account to provide the target population the type of support they are evidently lacking. It is also up to online health service providers to offer people affected by cancer reliable information on the identified topics and the form of support this target population voiced a need for via their conversations. Thus, the future direction of this research is to enhance an existing major community platform with *personalized cancer-focused* services, and evaluate the end-user acceptance and satisfaction via both actual behavior and surveyed perception.

References

[1] International Agency for Research on Cancer. *GLOBOCAN database: Cancer Fact Sheets.* [Retrieved: September, 2015]. Available from: http://globocan.iarc.fr/Pages/fact_sheets_cancer.aspx

[2] Norwegian Cancer Society. *Reasons for our international involvement.* [Retrieved: September, 2015]. Available from: https://kreftforeningen.no/en/international-collaboration/

[3] American Cancer Society. *Who are caregivers, and what do they do?* [Retrieved: September, 2015]. Available from: http://www.cancer.org/treatment/caregivers/caregiving/whatyouneedtoknow/what-you-need-to-know-as-a-cancer-caregiver-who-and-what-are-caregivers

[4] S. Hengstler, "Wireless Health: Making Your Devices Talk - A Review, Solution, and Outlook for Wireless Health Connectivity," International journal on advances in life sciences, vol. 6, 2014, pp. 147-156.

[5] Federal Ministry of Health, Federation of Bosnia and Herzegovina. *Strategic basis for adopting and implementing the Project for Strengthening the Health Sector.* [Retrieved: September, 2015]. Available from: http://www.fmoh.gov.ba/index.php/projekt-jacanja-zdravstvenog-sektora

[6] M. Wentzel, d. N. Jong, L. Nijdam, v. R. Drie-Pierik, and v. J. Gemert-Pijnen, "Understanding eHealth use from a Persuasive System Design perspective: an Antibiotic Information Application for Nurses," International journal on advances in life sciences, vol. 6, 2014, pp. 210-219.

[7] F. Sieverink, *et al.*, "The Diffusion of a Personal Health Record for Patients with Type 2 Diabetes Mellitus in Primary Care," International journal on advances in life sciences, vol. 6, 2014, pp. 177-183.

[8] S. Hadzidedic Bazdarevic and A. I. Cristea, "Does the use of cancer-related websites depend on personalizaation services and user emotions?," unpublished.

[9] M. A. Bright, *et al.*, "Exploring e-Health usage and interest among cancer information service users: the need for personalized interactions and multiple channels remains," Journal of health communication, vol. 10, 2005, pp. 35-52.

[10] P. E. Kummervold, *et al.*, "eHealth trends in Europe 2005-2007: a population-based survey," Journal of medical Internet research, vol. 10, 2008, doi: 10.2196/jmir.1023.

[11] R. Siliquini, *et al.*, "Surfing the internet for health information: an italian survey on use and population choices," BMC medical informatics and decision making, vol. 11, 2011, p. 21, doi: 10.1186/1472-6947-11-21.

[12] S. Ek, K. Eriksson-Backa, and R. Niemelä, "Use of and trust in health information on the Internet: A nationwide eight-year follow-up survey," Informatics for Health and Social Care, vol. 38, 2013, pp. 236-245, doi: 10.3109/17538157.2013.764305.

[13] L. Rainie. *E-patients and their hunt for health information.* [Retrieved: September, 2015]. Available from: http://www.pewinternet.org/2013/10/10/e-patients-and-their-hunt-for-health-information/

[14] E. Huang, C.-c. A. Chang, and P. Khurana, "Users' preferred interactive e-health tools on hospital web sites," International Journal of Pharmaceutical and Healthcare Marketing, vol. 6, 2012, pp. 215-229, doi: http://dx.doi.org/10.1108/17506121211259395.

[15] S. Fox and M. Duggan. *Health Online 2013.* [Retrieved: September, 2015]. Available from: http://www.pewinternet.org/2013/01/15/health-online-2013/

[16] S. Ziebland, *et al.*, "How the internet affects patients' experience of cancer: a qualitative study," Bmj, vol. 328, 2004, p. 564, doi: http://dx.doi.org/10.1136/bmj.328.7439.564.

[17] N. Xiao, R. Sharman, H. R. Rao, and S. Upadhyaya, "Factors influencing online health information search: An empirical analysis of a national cancer-related survey," Decision Support Systems, vol. 57, 2014, pp. 417-427, doi: 10.1016/j.dss.2012.10.047.

[18] National Center for Biotechnology Information. *PubMed.* [Retrieved: September, 2015]. Available from: http://www.ncbi.nlm.nih.gov/pubmed

[19] National Health Sservice. *NHS Home Page.* [Retrieved: September, 2015]. Available from: http://www.nhs.uk

[20] GoPubMed. *GoPubMed Searching is now sorted.* [Retrieved: September, 2015]. Available from: http://gopubmed.org/web/gopubmed/

[21] Cancer Reasearch UK. *Let's beat cancer sooner.* [Retrieved: September, 2015]. Available from: http://www.cancerresearchuk.org/

[22] Macmillan Cancer Support. *We are Macmillan Cancer Support.* [Retrieved: September, 2015]. Available from: http://www.macmillan.org.uk

[23] American Cancer Society. *Cancer.org - Join the fight against cancer.* [Retrieved: September, 2015]. Available from: http://www.cancer.org/

[24] R. Jucks and R. Bromme, "Choice of words in doctor–patient communication: An analysis of health-related Internet sites," Health Communication, vol. 21, 2007, pp. 267-277.

[25] L. Alpay, J. Verhoef, B. Xie, D. Te'eni, and J. Zwetsloot-Schonk, "Current challenge in consumer health informatics: Bridging the gap between access to information and information understanding," Biomedical informatics insights, vol. 2, 2009, p. 1.

[26] E. Sillence, L. Little, and P. Briggs, "E-health," Proc. 22nd British HCI Group Annual Conference on People and Computers: Culture, Creativity, Interaction-Volume 2, 2008, pp. 179-180.

[27] M. Span, *et al.*, "An Interactive Web Tool to Facilitate Shared Decision Making in Dementia: Design Issues Perceived by Caregivers and Patients," International journal on advances in life sciences, vol. 6, 2014, pp. 107-121, doi: 10.3389/fnagi.2015.00128.

[28] Klix.ba. *Cancer and fighting it.* [Retrieved: September, 2015]. Available from: http://www.klix.ba/forum/karcinom-i-borba-sa-njim--t26209.html.

[29] QSR International. *NVivo 10 for Windows*. [Retrieved: September, 2015]. Available from: http://www.qsrinternational.com/products_nvivo.aspx

[30] International Agency for Research on Cancer. *Country: Bosnia, Most frequent cancers in both sexes 2012.* [Retrieved: September, 2015]. Available from: http://eco.iarc.fr/eucan/Country.aspx?ISOCountryCd=70

[31] WebMD. *Study Suggests Diagnosis of Cancer Is More Frequent for Men.* [Retrieved: September, 2015]. Available from: http://www.webmd.com/cancer/news/20110712/men-have-higher-cancer-death-rates-than-women

Towards Social Media Platform Integration with an Applied Gaming Ecosystem

Munir Salman
FernUniversität in Hagen
Faculty for Multimedia and Computer Science
Hagen, Germany
Email: munir.salman@studium.fernuni-hagen.de

Kam Star
PlayGen
42 - 46 Princelet Street
London, E1 5LP
Email: kam@playgen.com

Alexander Nussbaumer
Graz University of Technology
Graz, Austria
Email: alexander.nussbaumer@tugraz.at

Michael Fuchs, Holger Brocks, Duc Binh Vu
Research Institute for Telecommunication and Cooperation
Dortmund, Germany
Email: mfuchs@ftk.de
Email: hbrocks@ftk.de

Dominic Heutelbeck
Telecommunication and Cooperation
Dortmund, Germany
Email: dheutelbeck@ftk.de

Matthias Hemmje
FernUniversität in Hagen
Faculty for Multimedia and Computer Science
Hagen, Germany
Email: matthias.hemmje@fernuni-hagen.de

***Abstract*—The European (EU)-based industry for non-leisure games (so called Applied Games, AGs) is an emerging business. As such it is still fragmented and needs to achieve critical mass to compete globally. Nevertheless, its growth potential is widely recognized and even suggested to exceed the growth potential of the leisure games market. The European project Realizing an Applied Gaming Ecosystem (RAGE) is aiming at supporting this challenge. RAGE will help to seize these opportunities by making available an interoperable set of advanced Applied Game (AG) technology assets, as well as proven practices of using such AG assets in various real-world contexts. RAGE will finally provide a centralized access to a wide range of applied gaming software modules, relevant information, knowledge and community services, and related document, media, and educational resources within an online community portal called the RAGE Ecosystem. Besides this, an integration between the RAGE Ecosystem and relevant social network interaction spaces that arranges and facilitates collaboration that underlie research and development (R&D), as well as market-oriented innovation and exploitation will be created in order to support community building, as well as collaborative asset exploitation of User Generated Contents (UGCs) of the RAGE Ecosystem. In this paper, we will outline a conceptual approach exploring methods to first of all integrate Content Management- and Community Collaboration support including advanced portal features based on Digital Library (DL), Media Archive (MA), and Learning Management System (LMS) infrastructures with Social Network (SN) integration support technologies and on capturing support for Semantic Social Media (SSM) content. This will allow for a seamless integration of social network advantages within community portal operation. On the other hand it will support information, UGC, and knowledge sharing, as well as persistency of social interaction threads within Social Networking Sites (SNSs) that are connected to the RAGE Ecosystem. The paper reviews possible alternative architectural integration concepts, as well as related authentication, access, and information integration challenges. In this way, on the one hand a qualitative evaluation regarding an optimal technical integration approach is facilitated while on the other hand design approaches towards support features of resulting user interfaces are initiated.**

Keywords—Social Media; Applied Gaming; Digital Ecosystem; Access and Information Integration; Know-how transfer; Social Networking Site;

I. INTRODUCTION AND MOTIVATION

The EU-based industry for AGs is an emerging business. As such it is still fragmented and needs to achieve critical mass to compete globally. Nevertheless, its growth potential is widely recognized and even suggested to exceed the growth potential of the leisure games market. The RAGE project [6] is aiming at supporting this challenge. RAGE will help to seize these opportunities by making available an interoperable set of advanced technology assets, tuned to applied gaming, as well as proven practices of using asset-based applied games in various real-world contexts. This will be achieved by enabling a centralized access to a wide range of applied gaming software modules, information, knowledge and community services, as well as related document, media, and educational resources within the RAGE Ecosystem. Furthermore, the RAGE project aims to boost the collaboration of diverse actors in the AG environment. Therefore, the main objectives of the RAGE Ecosystem are to allow its participants to get hold of advanced, usable gaming assets (technology push), to get access to the associated business cases (commercial opportunity), to create bonds with peers, suppliers, and customers (alliance formation), to advocate their expertise and demands (publicity), to develop and publish their own assets (trade), and to contribute to creating a joint agenda and road-map (harmonization and focus).

This means that seen as a whole, the RAGE project is a technology and know-how driven research and innovation project. Its main driver is to be able to equip industry players (e.g., game developers) with a set of AG technology resources (so-called Assets) and strategies (i.e., know-how being provided by means of information services and knowledge resources) to strengthen their capacities to penetrate a market (non-leisure), which is new for most of them, and to consolidate a competitive position in it. Figure 1 represents the positioning of the project in the spectrum from 'theory to application'.

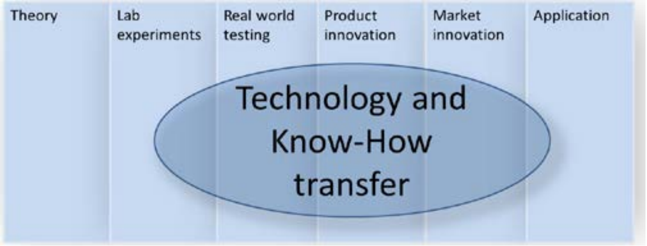


Figure 1. Technology and Know-How transfer [6]

In consequence, the RAGE Ecosystem and its integration with social networks of game-research-, game-developing- , gaming-, and AG communities will on the one hand become an enabler to harvest community knowledge and on the other hand it will support the access of such communities to the RAGE Ecosystem as an information and knowledge resource.

Building on the results of this SNS integration with the RAGE Ecosystem including corresponding SNS-enabled content and knowledge management, the RAGE Ecosystem will in the future also support Social Network Analysis

(SNA) by means of applying technologies for Natural Language Analysis (NLA) for discourse analysis, as well as Named Entity Recognition (NER) and Semantic Representation and Annotation (SRA) of its results. This will, e.g., enable users to utilize the envisioned Ecosystem with features of a social mediation engine going beyond content syndication, i.e., it will serve as a social space that mediates collaboration partners, while content remains the main attractor. Finally, an interactive map of supply- and demand-side stakeholders and resources will be provided for domain and community orientation, as well as visual access support.

Firstly, section II provides a brief introduction of a set of exemplar target communities that are present in SNS. Furthermore the section III describes related researches. The section IV is about state of the art in science and technology. Section V, more specifically, reviews the integration possibilities of social network technologies and their interfaces that could possibly support integration with the RAGE Ecosystem. Furthermore, it will investigate how to support access to resources and assets from such SNSs. Additionally, in section VI, it will outline design approaches towards supporting users in the target communities by services provided by the RAGE Ecosystem by means of outlining several use case scenarios for using Social Networking Features (SNFs**)** within the RAGE Ecosystem user interfaces. Finally, it will present conclusions and future work.

II. TARGET SNS USER COMMUNITIES AND CORRESPONDING EXEMPLAR USER STEREOTYPES

As outlined above, the EU-based industry for AG is an emerging business, which is still fragmented and needs to achieve critical mass for global competition. The AG industry and developer groups want to keep their developments innovative, i.e., attractive and technologically in good condition. These groups already have a very good understanding of their competitive advantage and corresponding assets (e.g., software, documents, and social media objects, etc.). However, they also need innovative ideas to develop innovative AGs in order to stay competitive. Therefore, they look for possibilities to cooperate with AG Research and Development (R&D) groups. Besides this, the AGs that researchers create within research projects produce a lot of AG research assets and prototypes, which need to be fully developed and deployed by AG software developers to become marketable. Apart from AG developers and researchers, there are also AG customers and players who on the one hand want to learn about or contract the development of AGs and on the other hand can also contribute to the development of AG usage scenarios. Many of these communities (AG developers, researchers, customers and players) are already present in a fragmented way within several groups in several SNSs. In [21], there are some examples of AG research, as well as industry and developer communities in, e.g., LinkedIn and Twitter. The Applied Games and Gamification (AGG) LinkedIn group, see [24] has over 4,500 members and has been running since 2011. The group claims to be one of the largest collective of creators, developers, researchers and users of applied games and gamification globally. The typical users can be divided approximately into those from the industry and those from the academia. From professors and recent graduates in gaming and related technologies, to CEOs, founders and directors of a wide variety of organisation that work or research the domain. The majority of discussion posts are promotions of products, methodologies for design, reposts of other interesting blogs on the topic and individuals' thoughts on implications of games and gamification for learning, training and behaviour change. The most prolific posters tend to be consultants and individuals representing organisation that are looking to showcase their abilities to a more business oriented community toward winning more business. Many posts do not garner comments or discussion as they are often pointing to other resources; however posts which pose interesting questions do receive attention and lead to interesting discussions from the more active members. Similarly the Serious Games Group on LinkedIn, see [25], has over 5100 members and has been running since 2008. The group memberships somewhat overlap with the applied games and gamification, however the audience tends to be more focused on the learning solutions and learning providers, with fewer CEOs and marketing directors, and more game designers as compared to the AGG, although the mode of use are very similar.

RAGE will help to overcome this fragmentation and aims to support the capturing, as well as the representation, management, sharing, and exchange of social media produced content and knowledge resources through its Ecosystem. Therefore, the integration of SNSs hosting such target communities with the RAGE-Ecosystem and at the same time enabling the connectivity between SNSs and the RAGE-Ecosystem will connect research-, gaming industry-, intermediary-, education provider-, policy maker- and end-user communities. Furthermore, it will facilitate the centralized access to the valuable assets beyond the SNSs.

As a whole, take up of RAGE results will generate impacts that will be visible through multiple enhancements in the performance of European Applied Game industries, especially in terms of reducing the current fragmentation, improving their innovation capacity and fostering their progress towards global technological leadership. By offering reusable Applied Games assets, the RAGE Ecosystem infrastructure and marketplace will play a key role in support of applied research and technology development, including demand driven research and productification activities, easing technology transfer and field validation of novel products and services, on a broad collaborative basis. The combined effects will allow end-to-end Applied Games value chain players to dramatically improve their competitive position.

III. RELATED WORK

The work presented in this paper is related to a number of topics in research. The RAGE Ecosystem will be built upon the Educational Portal (EP) technology and application solution, which was developed by the software company GLOBIT [1] that already was used in APARSEN [2].

APARSEN (Alliance Permanent Access to the Records of Science in Europe Network) was an EU-funded project within the digital preservation area with the goal to create a virtual research center in digital preservation in Europe. The so-called EP tool-suite offers a wide variety of tools. This includes a web based, user-friendly User and Community Management (UCM) including an advanced Contact and Role Management (CRM) based on MythCRM [20], as well as knowledge management support in the form of Taxonomy Management (TM) support and semi-automatic taxonomy-based Content Classification (CC) support [3][16], as well as a Learning Management System (LMS) based on Moodle [4] and an advanced Course Authoring Tool (CAT) [23]. In this way, the Content & Knowledge Management (C&KM) tools of the EP tool suite support the management of documents in a taxonomy-supported Digital Library, the management of multimedia objects in a taxonomy-supported Media Archive and the management of Learning Objects in a competence-based Learning Management System [19]. Furthermore, one of its additional purposes is to support Continuous Professional Education (CPE) and training of practitioners, experts, and scientists, which are members of professional communities of practice or scientific communities. Figure 2 displays the components and services in the EP tool suite as described in [16]. EP was built based on Typo3 [1] and, therefore, can be extended with the help of Typo3 extensions. Evgeny, Bogdanov et al [11] extend a social media platform in higher education with lightweight tools (widgets) aimed for collaborative learning and competence development. Our work will establish the new EP module Community & Social Network Support with a so-called Agile Application Interface (AAPI), which facilitates the connectivity to a wide range of SNSs.

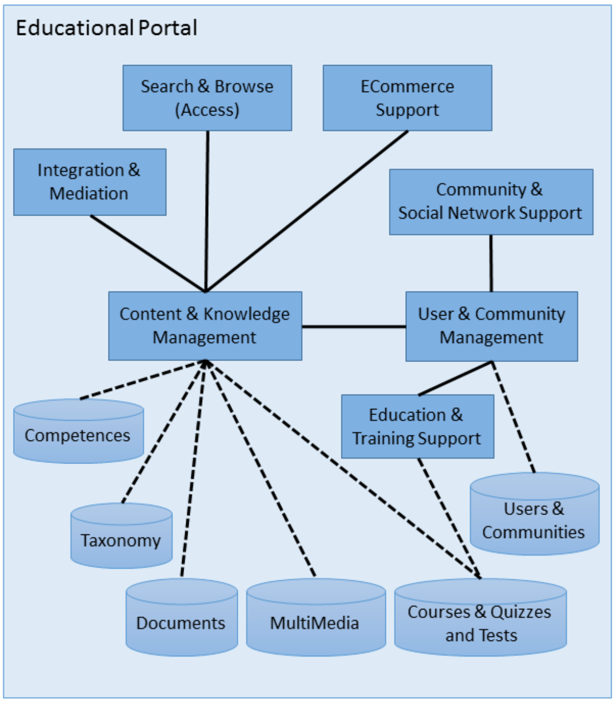

Educational Portal
Search & Browse (Access)
ECommerce Support
Integration & Mediation
Community & Social Network Support
Content & Knowledge Management
User & Community Management
Education & Training Support
Competences
Taxonomy
Users & Communities
Documents
MultiMedia
Courses & Quizzes and Tests

Figure 2. EP Tool Suite - Components and Services

IV. RELEVANT STARTING POINTS WITHIN THE STATE OF THE ART IN SCIENCE AND TECHNOLOGY

SNSs have changed the way of information sharing and learning processes by adding innovative features to social communication. SNS were defined as “*Internet or mobile-device based social spaces designed to facilitate communication, collaboration, and content sharing across networks of contacts. SNS allows its users to become content creators and content consumers at the same time, thus allowing instant participation, sharing of thoughts or information and personalised communication*” [7]. Therefore, SNSs are becoming increasingly important. This holds especially true for various SNFs like, e.g., rating, commenting, tagging, chatting, liking, posting new Social Media (SM) and UGC, following actors or celebrities, playing games etc. These SNFs are not only entertaining and exciting but also useful for learning and for information enrichment. Research has shown that distance education courses are often more successful when they develop communities of practice [10]. Besides, SM content becomes increasingly important in business and research. Kaplan [14] gave a clarification what the term SM means and how the concept of SM differs from the concept of related concepts such as Web 2.0 and UGC. Furthermore, he presents 10 pieces of advice to utilize SM. On the other hand, Agichtein et al [9] focus on the Semantic Social Media (SSM) and investigate methods exploiting community feedback, e.g., to automatically identify high quality content. Breslin et al. define in [12] “The Social Semantic Web as a vision of a Web where all of the different collaborative systems and social network services, are connected together through the addition of semantics, allowing people to traverse across these different types of systems, reusing and porting their data between systems as required.” RAGE will use Semantic Web technologies in order to describe in an interoperable way users’ profiles, social connections, and social media creation and sharing across different SNSs, as well as within the RAGE Ecosystem. Therefore, RAGE will be able to deliver well-grounded recommendation and mediation features AG R&D communities.

Today, most SNSs provide so-called Application Programming Interfaces (APIs) for developers to integrate the SNSs into their systems. Although, the SNSs are different in their functionality, i.e., their social networking feature support, their software architecture for the communication with distributed other systems is similar. Most of the SNSs offer REST APIs like [12][13][15], which can be used for integration with other systems. In the following, the description of the LinkedIn REST API software architecture as described in [7], as well as the Twitter Semantic REST API as described in [18] will be cited as an exemplary, illustrative, and at the same time representative example. The following features can be accomplished with the LinkedIn self-service APIs: Sign in with LinkedIn, Apply with LinkedIn, Share on LinkedIn and Manage Company Pages. One of the most important LinkedIn APIs is the share content. There are two methods for sharing content via the REST API [7].

- Post a comment that includes a URL to the content, which should be shared— LinkedIn analyzes the

TABLE I. TWITTERS REST API INSTANCES OF THE RESOURCETYPE CLASS [18]

ResourceType Instance	**Description**
Connections	Connections of the authenticating user to other users
DirectMessage	A private message sent from a user to another user
Entity	Metadata and additional contextual information about content posted on Twitter
EntityMedia	Media elements uploaded with a Tweet
FollowersIds	Numeric IDs of users that follow the authenticating user
FriendsIds	Numeric IDs of users that the authenticating user is following
TwitterList	A collection of tweets, posted by users belonging to a curated list
Trend	A popular topic in Twitter
UserCategory	Categeory of users
UsersLookup	Extended information about users

included URL and automatically identifies the title, description, image, etc.

- Share with specific values — developer should provide the title, description, image, etc., directly via the parameters of the API call.

Figure 3 displays a coding example for sharing content with specific values on LinkedIn [7].

The documentation of the Twitter API explicitly specifies four main response objects (Tweets, Users, Entities and Places) [22] but Togias et al. [18] identified with a more thorough study 44 of such resource types. Some of them can be directly accessed through actions provided by the API, while other can be accessed through the fields of other resource types. All of them defined the identified resource types as instances of *ResourceType* Class. Table I describes some of these instances in more detail.

```
{
"comment": "Check out
     developer.linkedin.com!", "content": {
   "title": "LinkedIn Developers Res.",
   "description": "Leverage LinkedIn's
      APIs to maximize engagement",
   "submitted-url":
   "https://developer.linkedin.com",
   "submitted-image-url":
   "https://example.com/logo.png"
  },
  "visibility": { "code": "anyone"}
}
```

Figure 3. Sharing content with specific values on LinkedIn

In summary, it is a big advantage to aim at supporting the integration of SNSs including relevant SNFs, as well as SSM content capturing, management, sharing, and dissemination support through their REST API into the RAGE Ecosystem. This will on the one hand facilitate to extend the envisioned RAGE Ecosystem with features of a social mediation engine going beyond content syndication, i.e., it can serve a social space that mediates collaboration partners, while content remains the main attractor. On the other hand it focuses on identifying collaboration opportunities between individuals and among groups, to support matchmaking and collaboration between stakeholders, and to identify and provide support for innovation opportunities and creativity efforts. That allows communities (such as technology providers, game developers and educators, game industries, researchers) to create their own assets and post them to the Ecosystem's repository without major effort. Besides this, the above approach enables follow-up work in the area of social network analysis and discourse analysis, which can then be conducted and used to provide feedback, recommendations, mediations, and relevant information to the communities. This feedback can e.g., help gaming companies to develop new markets in applied gaming.

V. INTEGRATION APPROACH AND METHODOLOGY

The following section presents the main technical integration possibilities in the backend, as well as in frontend. In this way, our integration approach and methodology is enabling us to differentiate between how to get access to resources and assets in the RAGE Ecosystem from external SNS communities and how to push contents from the RAGE Ecosystem to the external SNSs in order to improve user acceptance of services provided by the RAGE Ecosystem. Figure 4 displays the concept of a bi-directional integration approach of the RAGE Ecosystem with SNSs using a REST API.

Corresponding to this bi-directional integration approach, table II details scenarios following possible Tight and Loose Coupling methodologies that have to be considered for achieving an integration of SNS to RAGE and vice versa. In the following, the description of the As Another example, the so-called SlideShare API and its software architecture as described in [8] will be cited in

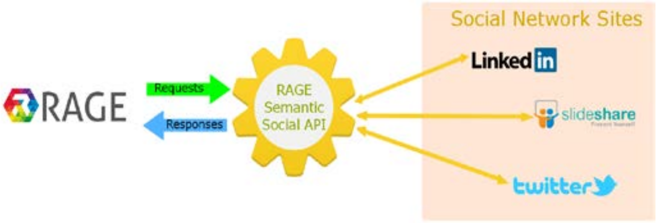


Figure 4. Integration Approach of RAGE Ecosystem with SNSs

TABLE II. LOOSE AND TIGHT COUPLING INTEGRATION METHODS BETWEEN SNS AND THE RAGE ECOSYSTEM

Method	FROM SNS TO RAGE ECOSYSTEM	FROM RAGE ECOSYSTEM TO SNS
Tight Coupling	Integration of RAGE-Interface within the SNS, user does not need to leave the SNS Environment (e.g., user posts a content to the RAGE Ecosystem without leaving the SN-Environment; user remains on the SNS)	Integration of SN-Interface within the RAGE Ecosystem, user doesn't leave the RAGE Environment (e.g., user posts, likes etc. a content without switching to the SNS; user remains on the RAGE Ecosystem)
Loose Coupling	SNFs are related to SNS (links from RAGE to the SNS) User lefts the RAGE Environment and switches to the SNS; user has to complete the action on the SNS, not on the RAGE Ecosystem	SNFs are only related to SNS (link from RAGE to the SNS) User leaves the RAGE Environment and switches to the SNS; user has to complete the action on the SNS, not on the RAGE Ecosystem

the following as an exemplary, illustrative, and at the same time representative example for the loose and tight coupling between the RAGE Ecosystem and SNSs.

The SlideShare API is based upon the REST model and supports the following functions:

- Upload, edit and delete slideshows
- Retrieving slideshow information by user, tag or group
- Retrieving groups, tags, and contacts by user
- Search slideshows

Those facilitate the loose coupling integration into the RAGE Ecosystem e.g., to get Slideshows by tag using the SlideShare REST API the Request should include followings parameter:

- Request Type: HTTPS GET
- Authorization: None
- URL: https://www.slideshare.net/api/2/get_slideshows_by_tag
- Tag: tag name
- [limit: specify number of items to return]
- [offset: specify offset]
- [detailed: Whether or not to include optional information. 1 to include, 0 (default) for basic information.]

The Slideshows REST API Response is in XML format and it looks like the following coding:

```
<Tag>
  <Name>{ Tag Name }</Name>
  <Count>{Number of Slideshows}</Count>
  <Slideshow>
  { as in get_slideshow }
  </Slideshow>
...
</Tag>
```

Furthermore, SlideShare provides an oEmbed API, which follows the oEmbed [5] standard. This standard facilitates the tight coupling integration of the Slideshows within the RAGE Ecosystem. The following codes show an example for making the embeddable media available through its oEmbed API endpoint and an API response example.

1. Example: XML Request

```
http://www.slideshare.net/api/oembed/2?
url=http://www.slideshare.net/
haraldf/business-quotes-for-
2011&format=json
```

2. Example: XML Response

```
{ The oEmbed version number }
{ Media type }
{ Embed media height }
{ Embed media width }
{ Embed content provider, SlideShare }
{ URL of the provider }
{ Thumbnail URL }
{ Thumbnail height}
{ Thumbnail width}
{ Author of embed content }
{ oEmbed version number }
{ Author SlideShare homepage }
{ ID of the slideshow }
{ Total number of slides in slideshow}
{ base URL of the slideshow images }
{ base URL suffix }
{ version number of the slideshow }
```

VI. SNF USAGE SCENARIOS AND DESIGN CONCEPT

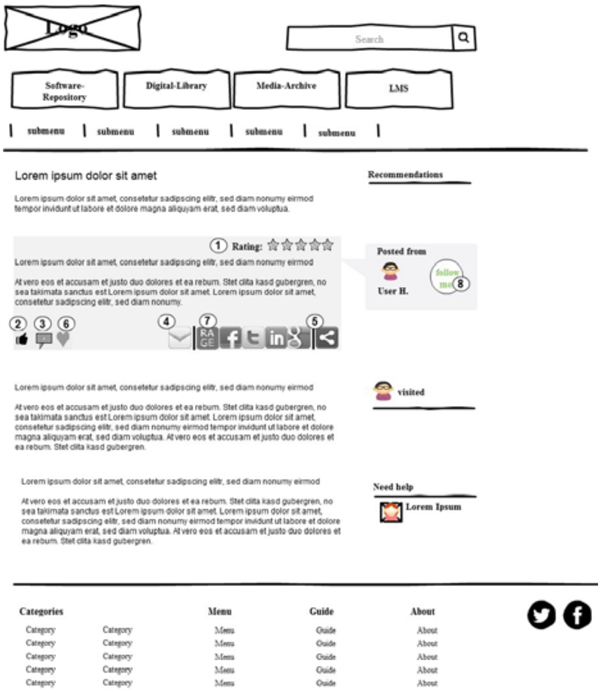
In addition to outlining our SNS integration approach and methodology, Figure 5 displays how the SNS usage scenarios can be integrated into the RAGE Ecosystem itself. RAGE Ecosystem users can visit content and knowledge management support within the RAGE Ecosystem's a Digital Library, Media Archive, Software Repository, which is currently under development based on [17], and Learning Management System. Here, users have the opportunity to:

a. Rate **(1)**, like **(2),** and Comment **(3)**: these Social Networking Features (SNFs) are e.g., important for the recommendation system (also currently under development, see [16] to get more useful suggestions.
b. Tell a friend **(4)**: users can send links to selected content (or the content itself) through email. Email addresses can be selected either from the RAGE address book or from users' address books, which are located in SNSs.
c. Share and post **(5)**: Users can share the selected content to one of their favourite SNSs or on the fly to more than one by selecting them from the share button.
   Users also have the possibility to publish content to a repository (e.g., GitHub's repository) or to cloud storage (e.g., into Dropbox).
d. Favourite **(6)**: Users can add content to their favourite lists, which facilitates to later, e.g., share/post their entire favourite list to a community.
e. Share and post to RAGE Communities **(7)** within the RAGE Ecosystem and also from any other platforms outside the RAGE Ecosystem. A RAGE Share-Button can be released and, e.g., be integrated by developers into other portals, homepages, ecosystems etc. to provide the possibility to Internet Users to share and post their content to the RAGE-Ecosystem.
f. RAGE Follow-me **(8)**: RAGE users can follow other users, groups or content in order to keep themselves up-to-date.

VII. CONCLUSION AND OUTLOOK

In this paper, we have introduced the RAGE Ecosystem supporting community-based content and knowledge management. In detail it will support the collection, sharing, access, and re-use of AG R&D assets, including SSM content and UGC resources, as well as academic, industry, and end user best practice knowhow represented in corresponding knowledge resources. In this way, the RAGE Ecosystem will provide AG communities, and therefore SNS communities, too, an opportunity to interact, share and re-use SSM content and UCG including corresponding knowledge resources, as well as communicate and collaborate using the RAGE Ecosystem as a back-end community content and knowledge management portal in addition to their favorite SNSs. Besides this introduction, we have presented how a technical integration between the RAGE Ecosystem and SNSs can be achieved to reduce the fragmentation and to increase the knowledge exchange among AG communities (such as AG developers, researchers, customers, and players). The RAGE Ecosystem and its SNS and SNF integration are currently under development. In the future, RAGE is aiming at increasing outreach and take-up of the RAGE Ecosystem through further SNS integration and SNF implementation. For example, the SNA and discourse analysis will be used for collecting, analyzing, and presenting data about various patterns of relationships among people, objects and knowledge flows within the RAGE Ecosystem and will provide additional functionality and sophisticated services for end-users, enhancing the emergence of communities. In particular, future developments will focus on identifying collaboration opportunities among individuals and groups, to support matchmaking and collaboration among main stakeholders, and to identify and provide support for innovation opportunities and creativity efforts. In this way, the RAGE project currently anticipates the following tools and services:

a) The RAGE Diagnostic tool based on various metrics for analyzing the usage of resources, the formation of different users groups, the level of social interactions, etc.,
b) the RAGE awareness tool can increase participation of different target groups in the Ecosystem,
c) the RAGE Knowledge Mapping tool builds and analyses knowledge maps for all kind of resources available in the Ecosystem.
d) the RAGE Professional support tool will support the users by letting them know whom or where to ask for support in different situations,
e) the RAGE Community detection tool will use available clustering algorithms (also called ''community detection algorithms'') that automatically identify and locate existing communities, in order to enhance the communication between gaming practitioners,
f) the RAGE Ecosystem analysis tool will apply network



Lorem ipsum dolor sit amet

Lorem ipsum dolor sit amet, consetetur sadipscing elitr, sed diam nonumy eirmod tempor invidunt ut labore et dolore magna aliquyam erat, sed diam voluptua.

Lorem ipsum dolor sit amet, consetetur sadipscing elitr, sed diam nonumy eirmod

At vero eos et accusam et justo duo dolores et ea rebum. Stet clita kasd gubergren, no sea takimata sanctus est Lorem ipsum dolor sit amet. Lorem ipsum dolor sit amet, consetetur sadipscing elitr, sed diam nonumy.

Lorem ipsum dolor sit amet, consetetur sadipscing elitr, sed diam nonumy eirmod

At vero eos et accusam et justo duo dolores et ea rebum. Stet clita kasd gubergren, no sea takimata sanctus est Lorem ipsum dolor sit amet. Lorem ipsum dolor sit amet, consetetur sadipscing elitr, sed diam nonumy eirmod tempor invidunt ut labore et dolore magna aliquyam erat, sed diam voluptua. At vero eos et accusam et justo duo dolores et ea rebum. Stet clita kasd gubergren.

Lorem ipsum dolor sit amet, consetetur sadipscing elitr, sed diam nonumy eirmod

At vero eos et accusam et justo duo dolores et ea rebum. Stet clita kasd gubergren, no sea takimata sanctus est Lorem ipsum dolor sit amet. Lorem ipsum dolor sit amet, consetetur sadipscing elitr, sed diam nonumy eirmod tempor invidunt ut labore et dolore magna aliquyam erat, sed diam voluptua. At vero eos et accusam et justo duo dolores et ea rebum. Stet clita kasd gubergren.

Categories		Menu	Guide	About
Category	Category	Menu	Guide	About
Category	Category	Menu	Guide	About
Category	Category	Menu	Guide	About
Category	Category	Menu	Guide	About
Category	Category	Menu	Guide	About

Figure 5. SNF Usage Scenario in the RAGE Ecosystem

analysis including many algorithms for identifying the most important, or central in some sense, nodes within a network,

g) the RAGE Recommendation may generate value interventions towards stimulating the participation of users. Such interventions include suggesting connections among users, setting up groups, closing the gaps in people’s knowledge of other members’ expertise and experience, and strengthening the cohesiveness within existing teams. Social media data such as tags, comments, purchasing patterns, and ratings can be used to link related gaming assets and users together into networks, the RAGE Social learning tool applies SNA to online learning environments, as well, focusing on the structural relationships between all learning objects and users, that support learning communities.

With the design and development of a comprehensive approach as pursued with the RAGE Ecosystem, ethical issues need to be taken into account. The integration of users’ SN profiles from different SNS, as well as the use of features carrying out analyses on top of Ecosystem user data have ethical implications in terms of privacy and data protection and require appropriate information and consent in the terms and conditions of use, as well as compliance to national and international data protection regulations. The same is for any use of log data for the purpose of system evaluation, or for UGC and user actions shared among different SNS and the RAGE Ecosystem. In addition, with UGC questions related to verification and validation of contributions, as well as to copyright ownership and infringement become relevant. The consideration of such ethical and legal requirements shall be incorporated in the system design and development process in terms of an ethics-by-design approach [26]. This means that data protection and privacy is already taken into account when the system is being designed. Design principles, such as purpose binding, would ensure that personal information is only accessible, if there is a need for it when performing a certain action. The system can also control data access by respecting personal settings which data should be available to others or the public. Other ethics-enabled features include the modification or deletion of personal data.

ACKNOWLEDGEMENTS AND DISCLAIMER

This publication has been produced in the context of the RAGE project. The project has received funding from the European Union’s Horizon 2020 research and innovation programme under grant agreement No 644187. However, this paper reflects only the author's view and the European Commission is not responsible for any use that may be made of the information it contains.

REFERENCES

[1] „Globit.com. Educational portal,“ [retrieved: Sep. 2015].

[2] S. Schrimpf, „Aparsen - alliance permanent access to the records of science in europe network.,“ Dialog mit Bibliotheken, Feb. 2014, pp. 52-53.

[3] T. Swoboda, Towards effectivity augmentation of automated scientific document classification by continuous feedback, Sep. 2014, Hagen.

[4] „moodle,“ [retrieved: Sep. 2015]. Available: http://moodle.de/.

[5] „oembed,“ [retrieved: Sep. 2015]. Available: http://oembed.com/.

[6] „RAGE,“ [retrieved: Sep. 2015]. Available: http://www.rageproject.eu.

[7] „LinkedIn,“ [retrieved: Sep. 2015]. Available: https://developer.linkedin.com/docs/share-on-linkedin.

[8] „SlideShare,“ [retrieved: Sep. 2015]. Available: http://de.slideshare.net/developers.

[9] E. Agichtein, C. Castillo, D. Donato, A. Gionis and G. Mishne, „Finding High-quality Content in Social Media,“ 2009, pp. 183-194.

[10] S. Barab and A. Duffy, „From practice fields to communities of practice. In D. Jonassen & S. M. Land Eds.,“ Theoretical foundations of learning environments, 2000, pp. 25-56.

[11] E. Bogdanov, F. Limpens, N. Li, S. El Helou, C. Salzmann and D. Gillet, „A Social Media Platform in Higher Education,“ IEEE Global Engineering Education Conference (EDUCON), Apr. 2012, pp. 1-8.

[12] J. Breslin and S. Decker, „The Future of Social Networks on the Internet: The Need for Semantics,“ IEEE Internet Computing, May 2008, pp. 86-90.

[13] G. Decker, A. Lüders, H. Overdick, K. Schlichting and M. Weske, „RESTful Petri Net Execution},,“Lecture Notes in Computer Science, Jul. 2009, pp. 73-87.

[14] A. M. Kaplan and M. Haenlein, „Users of the world, unite! The challenges and opportunities of Social Media,“ Business Horizons, Jan 2010, pp. 59-68.

[15] J. Mangler, P. Beran and E. Schikuta, „On the Origin of Services Using RIDDL for Description, Evolution and Composition of RESTful Services,“ May 2010, pp. 505-508.

[16] M. Schmedding, C. Nawroth and M. Hemmje, „Get along in the RAGE Ecosystem - Hybrid Domain Specific Knowledge Capturing in an Heterogeneous Software Development Environment,“ 2015. Manuscript submitted for publication.

[17] E. Stefanova, N. Nikolova, E. Peltekova, K. Stefanov, T. Zafirova-Malcheva and K. Kovatcheva, „Share.TEC: An innovative Solution for Teacher Educators,“ In proceedings of 4th International Conference of Education, Research and Innovation, Nov. 2011, pp. 1679-1688, ISBN: 978-84-615-3324-4, ISSN: 2340-1095.

[18] K. Togias and A. Kameas, „An ontology-based representation of the Twitter REST API,“ in Tools with Artificial Intelligence (ICTAI), IEEE 24th International Conference, Nov. 2012, pp. 998-1003.

[19] „TENCompetence,“ Building The European Network for Lifelong Competence Development, [retrieved: Sep. 2015]. Available: http://www.tencompetence.org.

[20] V. Duc Binh, „Customer Relationship Management based on Identity Management for Scientific Associations,“ Jul. 2015, pp. 43-48.

[21] M. Salman, M. Hemmje, D. Heutelbeck, M. Fuchs and H. Brocks, „Towards Social Network Support for an Applied Gaming Ecosystem,“ 2015. Manuscript submitted for publication.

[22] „TwitterDev,“ [retrieved: Sep. 2015]. Available: https://dev.twitter.com/.

[23] M. Then, B. Wallenborn, B. R. Ianniello and M. Hemmje, „Learning Design and Integration of Legacy Tools into Modern Learning Platforms,“ 6th Annual international Conference on Computer Science Education: Innovation and Technology, 2015. Manuscript submitted for publication.

[24] "The Applied Games and Gamification Group on linkedin", [retrieved: Sep. 2015]. Available: https://www.linkedin.com/grp/home?gid=3889283.

[25] "The Serious Games Group on linkedin", [retrieved: Sep. 2015]. Available: https://www.linkedin.com/grp/home?gid=137156.

[26] D. Gotterbarn, K. Miller, S. Rogerson, „Software Engineering Code of Ethics," Nov. 1997.

ISBN: 978-1-61208-443-5

A Real-Time Disaster-Related Information Sharing System Based on the Utilization of Twitter

Osamu Uchida*, Masafumi Kosugi†, Gaku Endo‡, Takamitsu Funayama§, Keisuke Utsu¶,
Sachi Tajima‖, Makoto Tomita**, Yoshitaka Kajita††, and Yoshiro Yamamoto‡‡
*Dept. of Human and Information Science, Tokai University, Japan
Email: o-uchida@tokai.ac.jp
†Dept. of Human and Information Science, Tokai University, Japan (Currently, Yahoo Japan Corporation)
‡Graduate School of Engineering, Tokai University, Japan
§Graduate School of Science and Technology, Tokai University, Japan
¶Dept. of Communication and Network Engineering, Tokai University, Japan
‖Student Project Center, Tokai University, Japan
**Dept. of Arts, Tokai University, Japan
††Dept. of Civil Engineering, Tokai University, Japan
‡‡Dept. of Mathematics, Tokai University, Japan

***Abstract*—In order to minimize the damage in case of a disaster, it is important to collect and spread accurate information quickly. Therefore, utilizing Twitter at the time of accidents has been gaining attention. In this paper, we propose a real-time disaster-related information sharing system based on the utilization of Twitter. The proposed system consists of two functions, the one is a disaster-related information tweeting function that automatically attaches user's accurate current geo-location information (address) and the hashtag of the form "#(municipality name) disaster," and the other is a disaster-related information mapping function that plots neighboring disaster-related tweets on a map in real time. We implemented the system on a linux server and conducted a verification experiment. The results of the experiment verify the usefulness of the proposed system.**

***Keywords*–Twitter, tweet mapping; disaster-related information; disaster mitigation.**

I. Introduction

In order to minimize the damage in case of a disaster, it is important to collect and spread accurate information quickly. Therefore, utilizing Twitter at the time of accidents has been gaining attention. The primary reasons for this are that Twitter is a social media having the characteristics of high immediacy, sharing information is very easy using Twitter, and the number of Twitter users is very large. In recent large scale natural disasters, such as the Grate East Japan Earthquake occurred in 2011, Hurricane Sandy, a Category 3 Superstorm which hit the U. S. East Coast in 2012, and Typhoon Haiyan which is called Typhoon Yolanda in the Philippines and caused severe damage to the islands of Leyte in 2013, Twitter was utilized as a communication tool by many people [1][2]. Due to the reasons mentioned above, lots of Japanese national and local governmental agencies have taken a great deal of utilizing Twitter in order to collect and distribute disaster-related information in recent years. For example, in Wako City, Saitama Prefecture, Japan, the government decided that the hashtag "#和光市災害" ("和光市" means Wako City and "災害" means disaster) is used as an official hashtag for reporting disaster situations in Wako City and conducted a disaster drill in order that citizens may get familiar with using this hashtag. Using the hashtag of the form "# (municipality name) disaster" has been spreading to other municipalities in Japan.

By the way, in order to optimize the utilization of disaster-related tweets in case of disasters, it is desirable that these tweets are geo-referenced accurately, that is, have geotags (longitude and latitude coordinates information). This is because geo-referenced tweets can be plotted on a map automatically, and both governmental agencies and disaster victims can obtain desired information by taking advantage of this map easily and promptly. However, it is well known that the percentage of Twitter users who permit the use of location services on Twitter is small, and then the ratio of geotagged tweets is quite small. For example, Cheng et al. found that only 0.42% of the tweets in their dataset were geotagged [3]. If tweets contain some geo-related information, such as building name or street address, it may be possible that obtaining longitude and latitude coordinates information by applying geoparsing technique [4][5] and geocoding. However, it is impossible to identify the location from tweet text by geoparsing completely because, for example, there are a dozen of "central park" all over the world.

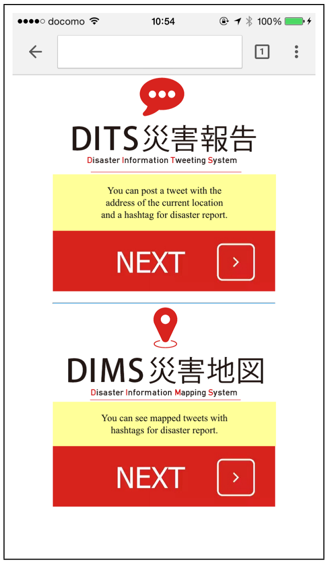
Moreover, considering the active utilization of tweets posted during disasters, these tweets should contain appropriate hashtags. During past disasters, the number of tweets increased dramatically, then hashtags played an important role in information retrieving from the stupendous tweets [6]. For example, if we access https://twitter.com/search?q=%23earthquake, we can obtain tweets that contain the hashtag "#earthquake."

In this study, based on these backgrounds, we have implemented a real-time disaster-related information sharing system utilizing Twitter that supports self-, mutual-, and public-help at the occurrence of a disaster (Figure 1 shows the top page of the proposed system). The system consists of the following two functions:

(1) a disaster-related information tweeting function that automatically attaches both the user's current geo-location information (address) and the hashtag of the form "#(市区町村名) 災害" (that means #(municipality name) disaster), and

(2) a disaster information mapping function that plots tweets posted using the function (1) on a map.

In this paper, we describe the outline of the proposed


●●●●○ docomo
10:54
100%
1
DITS災害報告
Disaster Information Tweeting System
NEXT
DIMS災害地図
Disaster Information Mapping System
NEXT

You can post a tweet with the address of the current location and a hashtag for disaster report.

You can see mapped tweets with hashtags for disaster report.

Figure 1. Top page of the proposed system

system, and report the results of the verification experiment that was conducted in June 2015.

II. Utilization of Twitter in Case of Disaster

In recent catastrophic natural disasters, Twitter was widely used for communication tool [1]. For example, it is widely known that during the Great East Japan Earthquake occurred on March 11, 2011, thousands of people took advantage of Twitter to retrieve information on the tsunami, shelter, the state of public transportation services, and so on [2][7][8][9][10]. More than 20 million tweets about Hurricane Sandy were posted between Oct. 27 and Nov. 1, 2012 [11]. Over a quarter-million tweets were posted during the first 72 hoours after Typhoon Haiyan destroyed large areas of the Philippines and crisis map was made with the aid of crowd sourcing [1][12]. When a flood disaster caused by a heavy rainfall was occurred on September 10, 2015, in Joso City, Japan, a victim who posted a tweet with address for a rescue request was saved by a rescue squad.

III. Real-Time Disaster-Related Information Sharing System

A. *Disaster-Related Information Tweeting Function*

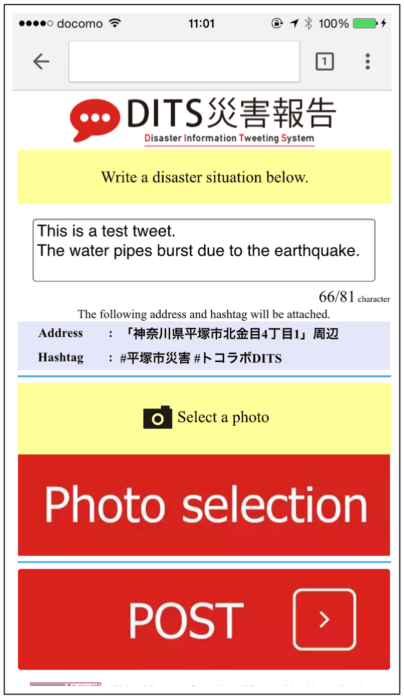
A disaster-related information tweeting function was implemented as a web-based application on a linux server using PHP and JavaScript (Figure 2). This function has the following features:

(1) Tweets are posted as tweets from the user's own Twitter account (Twitter authentication is conducted at the start of the use of the function).

(2) Information of the user's current geo-location information (the longitude and latitude coordinates) is acquired by utilizing location specification functions, such as the Global Positioning System (GPS), and based on the acquired location information, both address of the user's current location and a hashtag for disaster reports of the form "#(市区町村名) 災害" (#(municipality name) disaster) is automatically attached to the tweet. In the case where the user requires rescue, the hashtag to be attached becomes of the form "#(市区町村名) 災害_要救援" ("要救援" means rescue requirement (Figure 3).

(3) If the user enable location services on Twitter, the tweet is geotagged.

(4) An image can be attached (Figure 3).



Write a disaster situation below.

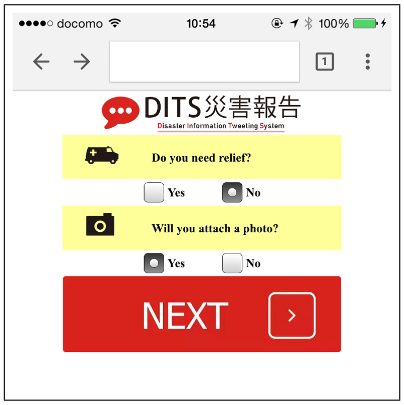
This is a test tweet.
The water pipes burst due to the earthquake.

66/81 character

The following address and hashtag will be attached.

Address : 「神奈川県平塚市北金目4丁目1」周辺
Hashtag : #平塚市災害 #トコラボDITS

Figure 2. Disaster-related information tweeting function

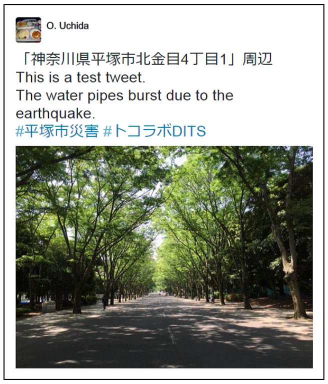


Do you need relief?
□ Yes ● No

Will you attach a photo?
● Yes □ No

Figure 3. Selection of the type of the hashtag to be attached and whether the user attaches a photo or not

Both address, which is determined by reverse geocoding (Yahoo! Japan Developer Network, Yahoo! Reverse Geocoder API [13] is used in this study) the location (the longitude and latitude) acquired by W3C Geolocation API [14], and hashtag for disaster reports are automatically attached to tweets. For example, the acquired location information is that the latitude and longitude are 35.361743 north and 139.273691 east, respectively, the address of the user's current location is determined as "神奈川県平塚市北金目 4 丁目 1" (4-1 Kitakaname, Hiratsuka City, Kanagawa Prefecture, Japan) and the hashtag is determined as "#平塚市災害" (#Hiratsuka City disaster) (if relief is need, it becomes "#平塚市災害_要救援" (#Hiratsuka City disaster rescue requirement)). This function



O. Uchida

「神奈川県平塚市北金目4丁目1」周辺
This is a test tweet.
The water pipes burst due to the earthquake.
#平塚市災害 #トコラボDITS

Figure 4. An example tweet posted by the proposed system

allows the user to attach images stored in their device. Users also can take a photos to be attached when they post tweets if their devices are smartphones or tablet PCs. The application uses Twitter API [15] to post texts and images. The maximum file size of the image that can be posted by Twitter API is about 3MB. Therefore, the application shrinks the file size of the image to be posted by decreasing the resolution of it before posting. Figure 4 shows an example Tweet posted by the proposed disaster information tweeting function. As you can see from the figure, the address of the current location "神奈川県平塚市北金目 4 丁目 1" is written at the beginning of the tweet and the hashtag "#平塚市災害" corresponding to the current location is attached.

B. Disaster-Related Information Mapping Function

The function to plot tweets posted using the disaster-related information tweeting function stated in the previous subsection, that is, tweets that are containing hashtags "#(市区町村名) 災害" (#(municipality name) disaster) or "#(市区町村名) 災害_要救援" (#(municipality name) disaster _ rescue requirement) on a map is implemented as a web-based application on the linux server using PHP and JavaScript (Figure 5). This application has the following features:

(1) It is available for a person without any twitter account to use.

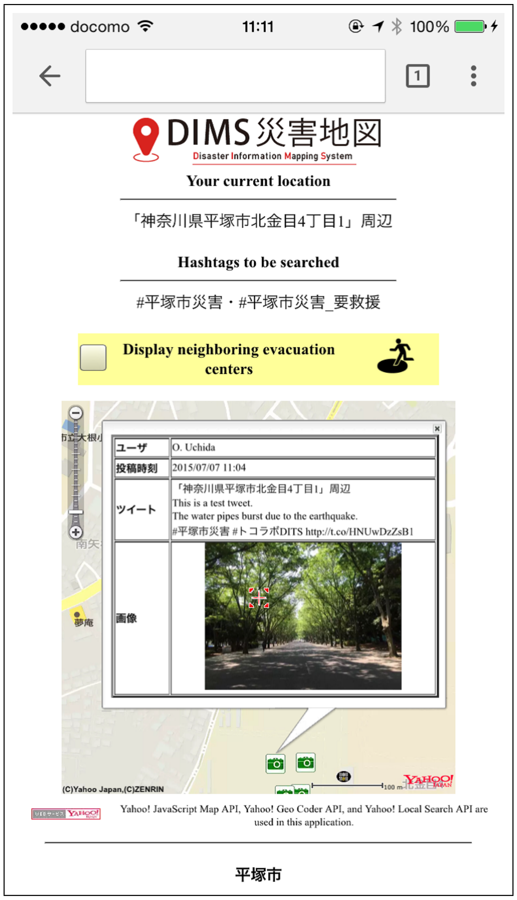
(2) It identifies the municipality name from the user's current location information (longitude and latitude) and displays tweets containing the municipality name in the hashtag on a map. For example, if the user's current location is 平塚市 (Hiratsuka City), tweets with "#平塚市災害" (#Hiratsuka CIty Disaster) or "#平塚市災害_要救援" (#Hiratsuka CIty Disaster _ rescue requirement) are pinned on the map.

(3) Tweets with geotags are mapped by using the latitude and longitude. Tweets without geotags are mapped by geocoding the address written at the head of the tweet (Yahoo! Japan Developer Network, Yahoo! Geocoder API [16] is used in this study).

(4) The shape and/or color of the icon that is used to indicate the position of the tweet change depending on the type of the attached hashtag (whether it is the rescue requirement hashtag or not) and whether an image is attached or not (Table 1).

(5) The places of shelters within 2km from the current location of the user are displayed on the map. If user click (tap) an icon indicating a shelter, the shortest route from the current location of the user to the shelter is displayed by using Yahoo! Japan route map API [17] (Figure 6).

(6) If the municipal government of the current location of the user has an official Twitter account, the latest 10 Tweets of the account will be also shown.



DIMS災害地図
Disaster Information Mapping System

Your current location

「神奈川県平塚市北金目4丁目1」周辺

Hashtags to be searched

#平塚市災害・#平塚市災害_要救援

□ Display neighboring evacuation centers

ユーザ	O. Uchida
投稿時刻	2015/07/07 11:04
ツイート	「神奈川県平塚市北金目4丁目1」周辺 This is a test tweet. The water pipes burst due to the earthquake. #平塚市災害 #トコラボDITS http://t.co/HNUwDzZsB1
画像	

Yahoo! JavaScript Map API, Yahoo! Geo Coder API, and Yahoo! Local Search API are used in this application.

平塚市

Figure 5. Disaster-related information mapping function

Using this function, we can get neighboring disaster-related information in nearly real time easily. We believe that the function is useful both governmental agencies and disaster victims, that is, supports self-, mutual-, and public-help at the occurrence of a disaster.

IV. Verification Experiment

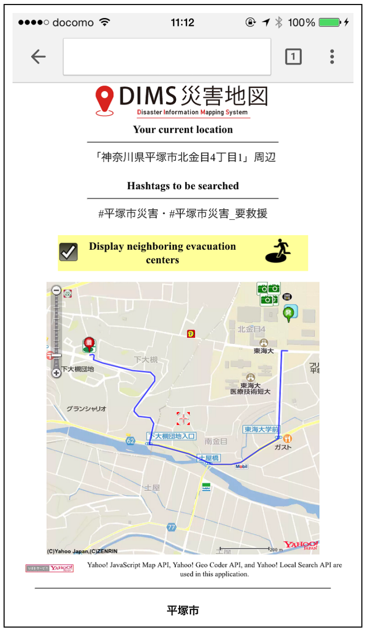
From 10:00 am to 4:00 pm, June 3, 2015, we conducted a verification experiment jointly with "The Study Group on Collection and Sharing of Disaster Volunteer Information Utilizing ICT, Kanagawa Prefectural Activity Support Center." There were 41 participants. The participants tweeted 217 Tweets in all. Generally, the system operated well and many of participants expressed their impression that by using the proposed system they were able to post tweets with address of their location and the corresponding hashtag of the form "#(市区町村名) 災害" (#(municipality name) disaster) very easily. However, the following issues were noticed:

(1) When the file size of the attached image was large, the uploading time was long depending on the connection. It sometime caused time-out error of the system.

(2) When it was used indoors, geotagging feature was in-accurate and the address attached to the tweet deviated from the user's actual address.

TABLE I. ICONS FOR TWEET PLOTTING

icon	image	rescue requirement	geotag
t			✓
t			
t		✓	✓
t		✓	
📷	✓		✓
📷	✓		
📷	✓	✓	✓
📷	✓	✓	


●●●●○ docomo 11:12 100%

DIMS 災害地図
Disaster Information Mapping System

Your current location

「神奈川県平塚市北金目4丁目1」周辺

Hashtags to be searched

#平塚市災害・#平塚市災害_要救援

☑ Display neighboring evacuation centers

(C)Yahoo Japan,(C)ZENRIN

Yahoo! JavaScript Map API, Yahoo! Geo Coder API, and Yahoo! Local Search API are used in this application.

平塚市

Figure 6. An example of displaying the shortest route from the current location of the user to the shelter the user selected

To solve the problem (1), we planned incorporating a feature that will reduce the file size of the image to be attached on the client-side device. For the problem (2), it is difficult to find a fundamental solution. However, we would like to consider that if the deviation estimated by the value of the accuracy obtained by W3C Geolocation API is large, only municipality name will be attached in the tweet.

V. CONCLUSION AND FUTURE WORK

We implemented a real-time disaster-related information sharing system with two functions: a disaster-related information tweeting function and a disaster-related information mapping function, and moreover conducted verification experiment. By using the proposed system, users can post disaster related tweets with geo-location information (address) and the corresponding hashtag of the form "#(市区町村名) 災害" (#(municipality name) disaster) easily even if they do not now where they are. In the future, we will try to improve the system by realizing features such as, for example, automatic useful tweets extraction [18] and tweet classification according to the contents of tweets, aggregating similar tweets, and the personalization of the tweet map according to the users' attributes, as well as improving and fixing the issues revealed by the verification experiment. The proposed system can be basically used only in Japan. Therefore, we will try to extend the system to so as to be used throughout the world.

ACKNOWLEDGMENT

The authors would like to thank the members of The Study Group on Collection and Sharing of Disaster Volunteer Information Utilizing ICT, Kanagawa Prefectural Activity Support Center for their cooperation in the verification experiment.

REFERENCES

[1] P. Meier, Digital Humanitarians: How Big Data Is Changing the Face of Humanitarian Response, CRC Press, 2015.

[2] B. D. M. Peary, R. Shaw, and Y. Takeuchi, "Utilization of Social Media in the East Japan Earthquake and Tsunami and its Effectiveness," Journal of Natural Disaster Science, Vol. 34, No. 1, 2012, pp. 3–18.

[3] Z. Cheng, J. Caverlee, and K. Lee, "You Are Where You Tweet: A Content-Based Approach to Geo-Locating Twitter Users," Proc. 19th ACM Int'l Conference on Information and Knowledge Management, 2010, pp. 759–768.

[4] X. Liu, F. Wei, S. Zhang, and M. Zhou, "Named Entity Recognition for Tweets," ACM Trans. Intelligent Systems and Technology, Vol. 4, No. 1, Article No. 3, 2013.

[5] J. Gelernter and S. Balaji, "An Algorithm fo Local Geoparshing of Microtext," GeoInformatica, Vol. 17, No. 4, 2013, pp. 635–667.

[6] L. Potts, J. Seitzinger, D. Jones, and A. Harrison, "Tweeting Disaster: Hashtag Constructions and Collisions," Proc. 29th ACM Int'l Conference on Design of communication, 2011, pp. 235–240.

[7] F. Toriumi, T. Sasaki, K. Shinoda, K. Kazama, S. Kurihara, and I. Nodao, "Information Sharing on Twitter during the 2011 Catastrophic Earthquake," Proc. 22nd Int'l Conference on World Wide Web Companion, 2013, pp. 1025–1028.

[8] S. Doan, B.-K. H. Vo, and N. Collier, "An Analysis of Twitter Messages in the 2011 Tohoku Earthquake," Lecture Notes of the Institute for Computer Sciences, Social Informatics and Telecommunications Engineering, Vol. 91, 2012, pp. 58–66.

[9] A. Acar and Y. Muraki, "Twitter for Crisis Communication: Lessons Learned from Japan's Tsunami Disaster," International Journal of Web Based Communities, Vol. 7, No.3, 2011, pp. 392–402.

[10] H. Wilensky, "Twitter as a Navigator for Stranded Commuters during the Great East Japan Earthquake," Proc. 11th Int'l ISCRAM Conference, 2014, pp. 695–704.

[11] A tweet from @twitter on Nov. 3, 2012, https://twitter.com/twitter/status/264408082958934016 [accessed: 2015-10-03].

[12] iRevolutions, "Live Crisis Map of Disaster Damage Reported on Social Media," http://irevolution.net/2013/11/11/live-crisis-map-of-disaster-damage-reported-on-social-media/ [accessed: 2015-10-03].

[13] Yahoo! Japan Developer Network, Yahoo! Reverse Geocoder API, http://developer.yahoo.co.jp/webapi/map/openlocalplatform/v1/reversegeocoder.html [accessed: 2015-10-03] (in Japanese).

[14] W3C Geolocation API Specification, http://dev.w3.org/geo/api/spec-source.html [accessed: 2015-10-03].

[15] Twitter Developers, https://dev.twitter.com/ [accessed: 2015-10-03].

[16] Yahoo! Japan Developer Network, Yahoo! Geocoder API, http://developer.yahoo.co.jp/webapi/map/openlocalplatform/v1/geocoder.html [accessed: 2015-10-03] (in Japanese).

[17] Yahoo! Japan Developer Network, Yahoo! Route Map API, http://developer.yahoo.co.jp/webapi/map/openlocalplatform/v1/routemap.html [accessed: 2015-10-03] (in Japanese).

[18] R. Kitajima, R. Kamimura, O. Uchida, and F. Toriumi, "Neural Potential Learning for Tweets Classification and Interpretation," 7th International Conference on Soft Computing and Pattern Recognition, 2015 (Accepted).

Making Sense of Large-Group Discussions using Rhetorically Structured Text

Ana Cristina Bicharra Garcia
Computer Science Department
Universidade Federal Fluminense
Niterói, Brazil
bicharra@ic.uff.br

Mark Klein
Sloan School, MIT
Boston, USA
University of Zurich
Zurich, Switzerland
m_klein@mit.edu

***Abstract*— Recent advances in social media technology have made it possible to involve large groups in online deliberations, using such tools as forums and argument maps. As discussions develop, however, making sense of their content can become a big challenge for newcomers, thus impeding their potential participation. We posited that rhetorically organized narratives can foster superior comprehensibility, and conducted an experimental evaluation that supports this claim. Human subjects were asked to answer a questionnaire about a discussion presented in one of three formats: forum, argument map, and rhetorically structured text. The rhetorical structures produced superior question-answering performance for complex questions. In this paper, we discuss these results, as well as their implications for the design of large group interaction tools.**

Keywords-Rhetorical structure theory; RST; online group deliberation; web forums; argument maps; crowd-computing; social computing

I. INTRODUCTION

Forums are virtual places for hosting online discussions among people on subjects of mutual interest. They have been an important source of knowledge in many fields, ranging from computer software development to education [1]. Typical forums have a *chronological* structure. Each new contribution (post) is appended at the end of the list of previous contributions, labeled by a time stamp and the name of its author. As a discussion develops, however, it becomes increasingly difficult for newcomers to understand the intertwined contributions from other participants, and this problem gets amplified as the group grows. Threaded discussions provide an additional layer of organization, based on capturing post reply structures [2], but this is of limited value for increasing comprehensibility because there is no clear relationship between reply structures and the semantics of a discussion.

Argument maps [3]-[7] provide an alternative, logic-based, structure where users organize their contributions using a pre-defined taxonomy of post types (e.g., issues, ideas, pros and cons). Such maps reveal the intentions of the posts, but at the cost of removing information about the chronological order of contributions, potentially impairing understandability [8].

In this paper, we explore whether a *narrative* organization for discussion content, one based on the principles of rhetorical structure theory (RST), can transcend the comprehensibility limitations of (chronologically-structured) forums and (logically-structured) argument maps. To test this idea, we conducted an experimental comparison with three groups of 16 people in each. Participants in the groups were demographically balanced by gender, age and background. Each group interacted with the same discussion content in one of these three formats:

- discussion forum
- argument map
- rhetorically-structured text

We measured how quickly and well the participants in each group could answer a range of questions about the content. We found that rhetorically structured text significantly improved the participant's abilities to answer complex questions relative to web forum and argument map structures ($p<0.05$), but did not have a statistically significant impact on answering simple questions.

Section 2 presents related work, followed by background explanation concerning RSTs presented in Section 3. Section 4 describes the experiments and finally Section 5 presents the conclusions including final remarks and future work.

II. RELATED WORK

The goal of crowd-scale deliberation is to allow communities to identify and evaluate possible solutions for a problem of shared concern [9][10]. A wide range of social computing technologies have emerged to address this challenge in the past few decades, including email, chat, wikis, web forums, open innovation systems, group decision support systems, as well as debate and argumentation systems.

How well do existing social computing technologies fare in terms of realizing these potentially powerful effects in the context of crowd-scale deliberation? There are several key types of applicable technology, each with their own strengths and weaknesses, including time-centric systems, question-centric systems, topic-centric systems, debate-centric systems, and argument-centric systems. We will review each type in the sections below.

A. *Time-Centric Systems*

Time-centric systems include tools - such as, email, chat rooms, blogs, micro-blogs like twitter, and web forums – in which content is organized based on when a post was contributed. Currently, time-centric systems are by far the dominant technology used for online deliberation. These systems enable large communities to weigh in on topics of interest, but face serious shortcomings that can deeply undercut the value of the deliberation engagements [11], such as:

- Low signal-to-noise ratios,
- Insular ideation,
- Balkanization,
- Non-comprehensive coverage,
- Dysfunctional argumentation and
- Opaque Processes

Because of all these issues, the content generated by time-centric deliberation tools is typically very sub-optimal from both a depth and breadth perspective.

B. *Question-Centric Systems*

Question-centric systems [12] are organized around questions: one or more questions are posted and the community is asked to contribute, rate, and comment on proposed solutions for these questions. These systems can be divided into two subtypes based on whether the questions are "close-ended" (there are only one or few correct answers, and the answers are relatively easy to verify), and "open-ended" (the system is soliciting ideas for large complex problems which have many possible solutions and where identifying the best answers is not straightforward). Close-ended question-centric sites such as stackoverflow.com, a programming Q&A site, have been remarkably successful [13], but are applicable only to a small subset of the entire scope of potentially important deliberation problems. Open-ended systems - such as group decision support systems as well as such open innovation platforms as IdeaScale and MindJet - can elicit huge levels of activity, and organize content better than time-centric tools. Like time-centric systems, however, they are prone to high levels of redundancy, wherein many of the ideas represent minor variations of each other. Also like time-centric systems, they tend to elicit many simple single-author ideas rather than a smaller number of collaborative efforts.

C. *Topic-Centric Systems*

Topic-centric systems, most notably wikis, organize content into collaboratively-authored articles that each focus on a single topic. A simple watchlist-rollback mechanism helps authors become aware of, and quickly repair, any damage caused by the work of other authors. Studies have shown that wiki content, despite often being contributed by non-experts, can have equivalent quality, greater currency, and much more comprehensive coverage than conventional, expert-curated sources [14]. Wikis, however, are deeply challenged by controversial topics [15][16]. They capture, by their nature, the “least-common-denominator” consensus between many authors (any non-consensus element presumably being edited out by those that do not agree with it), and the controversial core of deliberations are typically moved to massive talk pages for the article, which are essentially time-centric venues prone to all the limitations we noted above.

D. *Debate-Centric Systems*

Debate-centric systems, such as whysaurus.com, debatepedia.com, debatewise.org, and debate.org, have been designed to address the weakness of topic-centric systems around controversial topics. In such tools, a debate question is posed e.g. "Is the death penalty justified?", and users contribute arguments for and against that question, typically organized as two columns: one for pros, another for cons. Such tools, especially when curated to avoid duplication, provide an effective means for gathering a broad range of arguments on divisive topics, but are limited in several important ways. They are, to begin with, limited to "binary" debates where the question admits of only a "yes" or "no" answer. They are thus not suited to problems e.g. "how can we protect ourselves from climate change?" that have a large open-ended set of possible solutions. They also do not provide a systematic structure for supporting or rebutting arguments, since arguments can not be linked to other arguments. For both these reasons, the structure is not well suited for exploring open-ended deliberation problems in depth.

E. *Argument-Centric Systems*

Argument-centric systems [17][18] allow groups to systematically capture complex deliberations as tree structures made up of issues (questions to be answered), ideas (possible answers for a question), and arguments (statements that support or detract from an idea or argument) that define a space of possible solutions to a given problem:

Such tools have many advantages. Every unique point appears just once, radically increasing the signal-to-noise ratio, and all posts must appear under the posts they logically refer to, so all content on a given question is co-located in the tree, making it easy to find what has and has not been said on any topic, fostering more systematic and complete coverage, and counteracting balkanization by putting all competing ideas and arguments right next to each other. Careful critical thinking is encouraged, because users are implicitly encouraged to express the evidence and logic in favor of the options they prefer [19], and the community can rate each element of their arguments piece-by-piece.

Most argumentation systems have been used by individuals or in small-scale settings, relying in the latter case on a facilitator to capture the free-form interactions of a collocated group as a commonly-viewable argument map [20]. The Deliberatorium [21] is a web-based tool to allow crowd-scale online discussion and deliberation.

As we can see, argumentation systems offer much promise as a medium for enabling large-scale online deliberation. One key challenge for such systems, however, is that the logical structure of argument maps, while systematic, is not a good match with the narrative forms of knowledge communication that most people are much more familiar with. The project reported here has explored whether narratives generated *from* argument maps, using a technique called RST, can make the results of argument-centric deliberations more accessible and understandable to the average user

III. RHETORICAL STRUCTURE THEORY

RST is a theory of text organization where semantically related clauses are structured hierarchically [22]. An RST structure, more specifically, is a network made up of two basic units: the nucleus and the satellite. Nuclei represent the essence of the communication, while satellites contain additional information about the nucleus. The satellite is often incomprehensible without the nucleus, whereas nuclei without satellites can be understood to a certain extent.

RST relations are classified according to their expected effect on the reader. Mann and Thompson [22] originally proposed 24 semantic relations, including: Attribution, Cause, Circumstance, Contrast, Elaboration, Enablement and Solutionhood

Figure 1 presents a sample of a RST schema. In this figure, the central information is "I use sun protection SPF30". "to prevent skin cancer" is an enablement provided by the central idea. Additionally, the entire utterance is attributed to "My mother always says".

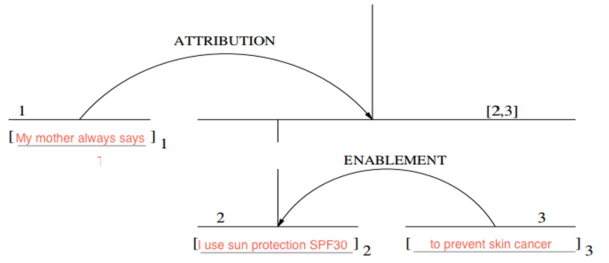

ATTRIBUTION
1
[2,3]
[My mother always says]1
ENABLEMENT
2
3
[I use sun protection SPF30]2
[to prevent skin cancer]3

Figure 1. RST example for the text: "My mother always says that I should use sun protection SPF30 in order to prevent skin cancer".

RST has been successfully applied to many different areas over the past 30 years, and it remains a research baseline for text analysis, parsing, summarization, essay scoring and natural language generation.

"RST is a theory of text organization resulting from exhaustive analysis of texts" [23]. It was meant to propose a guideline to computational text generation based on constructs that reflect how written text works. Every piece of text is included for a perceivable reason, dictated by the RST framework, leading to a coherent text that will foster readers' understanding.

The emphasis of our research is to propose a method for building coherent text over a discussion to improve newcomers' understanding. RST offers an interesting framework to organize large online discussion. Before building an automatic tool to generate RST-based explanation over a discussion, we developed an experiment to test its effectiveness on average users' understanding.

IV. EXPERIMENTAL EVALUATION

This section describes the experiment developed to confirm our hypothesis that RST-organized explanations positively impact understanding.

A. Subjects

An invitation email was sent to 70 people associated to the computer science department of a Brazilian University. Forty-eight people accepted to participate. Individuals were randomly assigned to each group, but considering their gender, age and education level to create three demographically homogeneous groups of sixteen people each, as illustrated in Table I. Participants were mostly young (between 20 and 30 years old), male and educated. Most were computer science undergrad and graduate students. They all had substantial experience with social media tools, as well as some previous experience with forums and argument maps.

The strong participation of young people suggests the need for further studies to test the generalizability of our conclusions.

B. The Task

The task consisted of reading a debate concerning the design of a virtual coin for a new computer game, and then answering questions about it.

We initially selected a hot discussion topic that was going on in the news for quite a long time. The discussion was in an open form in the Internet that attracted many posts from many different people, concerning the Brazilian post office efficiency. There was a corruption scandal and people were discussing the need to have a public post office.

The second experiment, reported in this paper, included a competitive ingredient as the incentive to participation and a topic that participants did not have a prior opinion. Among the options considered, we decided for a discussion presented in game designers' forums.

TABLE I. PRTICIPANTS' PROFILE DESCRIPTION OF THE THREE GROUPS PARTICIPATING IN THE EXPERIMENTS.

Group	**Age (years old)**				**Gender**		**Education**		
	20-30	30-40	40-50	>50	Female	Male	Undergrad Student	College	MSc/PhD
Group 1 (G1)	10	4	1	1	4	12	6	7	3
Group 2 (G2)	9	4	1	2	5	11	4	10	2
Group 3 (G3)	10	4	1	1	6	10	5	9	2

The discussion material concerned the design of a virtual coin for a new computer game and was taken from a known website forum. The game was inspired by the Lord of Rings tale with dwarves, elves, orcs, hobglobins, and drows societies. The debate was held in a computer game community forum [24] and lasted a few days in 2009. We selected this material for the high number of posts, 69 posts generated by 36 people in a 3-days discussion.

The task consisted in reading the material received in one of the three possible formats and answering a questionnaire of 13 questions. The material in all three scenarios were displayed in an online website. In the forum scenario, user interaction was constrained to search for words, scroll the document and copy&paste material from the document to the answer slot in the questionnaire area. Furthermore, RST and ArgMap scenarios allowed obtaining additional information by clicking in the paragraph and by clicking on the node, respectively. Detailed information concerned the author and the time stamp of the posts.

The task was performed in a controlled room with 10 computers. The experiment responsible received the participants, placed them in the computer stations and red the instructions aloud. Participants asked many questions, such as if the duration was a hard constraint, if they had to write complete sentences as answers, if they could copy parts of the original material and paste as answers and what would happen with the prize in case of ties. The experiment responsible answered the questions and stayed at the control room during the entire experiment.

The experiment website first displayed the experiment's instruction. This screen contained the same material read by the experiment responsible. After reading and agreeing in participating, a second screen appears with the material displayed according to the three possible scenarios. After reading the material they would click in the OK button to start answering the questionnaire. They could go back at any time to review the material when answering the questions.

No communication was allowed. We believe they obeyed this rule because they were competing among themselves. There was a small monetary payment for participating and a prize for the best three scores.

The material, in all three scenarios, was divided in chunks of information that received a number. These numbers worked as indexes to content, working as a set of discrete options from which users could select and assemble to compose an answer. Participants could answer questions using their own words, copying and pasting sentences from the original material or by writing the "chunks of information" indexing numbers.

The experiment lasted about 1 hour. After reviewing the material, participants could start the question and answer phase of the experiment. Questions were presented one at a time in a random order. After submitting an answer, a new question was presented. Participants could go back, at anytime, to the reading material, but not to a previously answered question.

There was an incentive for participants answering correctly. The best three scores would receive a monetary prize during a later workshop, so recognition from the community was also a reward.

The questionnaire had an answers' sheet prepared by a group of two graduate students and revised by one Linguist. Most questions had just one correct answer that contained from one to 10 segments of information. Precision was calculated as the relation between the number of corrected segments in the answer and the number of segments in the answer. Recall was calculated as the relation between the number of corrected segments in the answer and the number of segments in the expected answer. We also consider the F-measure metric because it is a balance between precision and recall metrics. F-measure is the harmonic mean Precision and Recall metrics.

C. *Question Types Used in Study*

We focused our research on generating answers to six frequent types of questions a newcomer might have concerning a discussion: What, Compare, Explain, Justify, Choose and Summarize. An answer is designed, as shown in Table II, according to the type of the question and the expected completeness of the answer. Optional information concerning social and temporal context can also be derived.

TABLE II. A SAMPLE OF RHETORICAL RELATIONS FOR GENERATING ANSWERS TO DIFFERENT TYPES OF QUESTIONS.

Question Type	Question's quantifier			Context	
	One	Some	All	Chronological Factor	Social Factor
What	Purpose	Sequence	Sequence	Motivation	Motivation
Compare	---	Contrast	Contrast	Antithesis	Antithesis
Explain	Interpretation & Evaluation; Relations of Cause	Interpretation & Evaluation; Relations of Cause	Interpretation & Evaluation; Relations of Cause	Enablement	Enablement
Justify	Condition &Otherwise	Condition &Otherwise	Condition &Otherwise	Justify	Justify
Choose	Purpose	Purpose	Purpose	Evidence	Evidence
Summarize	Restatement & Summary	Restatement & Summary	Restatement & Summary	Background	Background

The rhetoric structure guides the construction rules. There are explicit rules for generating the rhetoric answers, as described below.

```
1. Query←Get query from user
2. QueryType← Classify(Query)
3. QueryQuantifier←ClassifyQty(Query)
4. DecomposeQuery (QueryType, Query,
   QueryOrganization)
5. GetAnswerComponents(QueryQty,
   QueryOrganization, AnswerOrganization)
6. GenerateAnswer(Answer, AnswerOrganization)
```

For example, suppose a question over a discussion, as an argumentation map, concerning options for buying a car, such as, How does a Toyota Rav 4 compare to a BMW X1?

```
1. Query← Compare a ToyotaRav4 to a BMW_X1
2. QueryType←Contrast
3. QueryQuantifier←ALL
4. QueryOrganization← (COMPARE (EXPLAIN ToyotaRav4)
   (EXPLAIN BMW_X1))
5. AnswerOrganization←
   (ANTITHESIS (ISSUE  "Car Buying Options")
     ((ALTERNATIVE ToyotaRav4)
        (ADVANTAGE (CRITERION
                          "Quality" "Deluxe")
                  (CRITERION "Safety" "well-trusted
                              breaks on snow"))
        (DISADVANTAGE (CRITERION  "Cost" "Very high")))
   ((ALTERNATIVE BMW_X1)
        (ADVANTAGE (CRITERION "Quality" "Cool")
                   (CRITERION "Quality" "Beautiful" ))
        (DISADVANTAGE (CRITERION "Cost" "Very high"))))
6. Answer shown in Table III.
```

TABLE III. AN EXAMPLE OF A RST GENERATED ANSWER. » MEANS LINK TO INFORMATION CONCERNING AUTHOR, DATE AND SUPPORTERS.

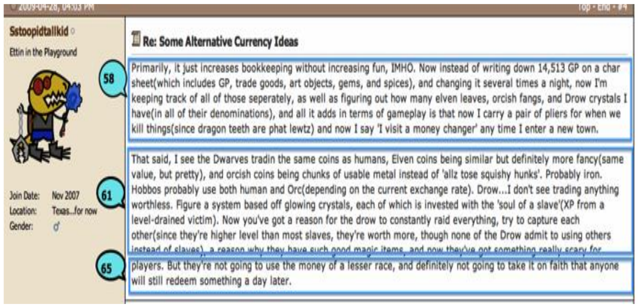
	Car Buying Options »			
	ToyotaRav4 »		BMW_X1 »	
Criterion	Pros	Cons	Pro	Cons
Quality	Deluxe »		Cool » Beautiful »	
Safety	well-trusted breaks on snow »			
Cost		Very high »		Very high »

D. Material and Apparatus

The task consisted of reading a debate concerning the design of a virtual coin for a new computer game.

Participants were divided into three groups. Each group received the reading material in one of the three formats:

- Forum format (scenario 1), as shown in Figure 2: a sequence of textual posts with date stamps and a nickname signatures.



2009-04-28, 04:03 PM Top - End - #4

Sstoopidtallkid
Ettin in the Playground

Join Date: Nov 2007
Location: Texas...for now
Gender: ♂

Re: Some Alternative Currency Ideas

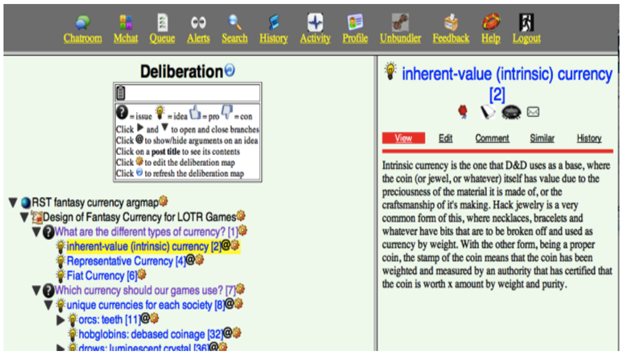
Primarily, it just increases bookkeeping without increasing fun, IMHO. Now instead of writing down 14,513 GP on a char sheet(which includes GP, trade goods, art objects, gems, and spices), and changing it several times a night, now I'm keeping track of all of those seperately, as well as figuring out how many elven leaves, orcish fangs, and Drow crystals I have(in all of their denominations), and all it adds in terms of gameplay is that now I carry a pair of pliers for when we kill things(since dragon teeth are phat lewtz) and now I say 'I visit a money changer' any time I enter a new town.

That said, I see the Dwarves tradin the same coins as humans, Elven coins being similar but definitely more fancy(same value, but pretty), and orcish coins being chunks of usable metal instead of 'allz tose squishy hunks'. Probably iron. Hobbos probably use both human and Orc(depending on the current exchange rate). Drow...I don't see trading anything worthless. Figure a system based off glowing crystals, each of which is invested with the 'soul of a slave'(XP from a level-drained victim). Now you've got a reason for the drow to constantly raid everything, try to capture each other(since they're higher level than most slaves, they're worth more, though none of the Drow admit to using others instead of slaves), a reason why they have such good magic items, and now they've got something really scary for

players. But they're not going to use the money of a lesser race, and definitely not going to take it on faith that anyone will still redeem something a day later.

Figure 2. Discussion sample presented in a Forum format. The blue balloons represent the "chunk of information" indexing number.

- Argument map (ArgMap--scenario 2), as shown in Figure 3: the discussion from the original forum was reread and logically organized into issues, ideas, and arguments. We used the Deliberatorium tool [4] to build the argument map and the same wording as used in the original forum.



Click ▶ and ▼ to open and close branches
Click @ to show/hide arguments on an idea
Click on a **post title** to see its contents
Click to edit the deliberation map
Click to refresh the deliberation map

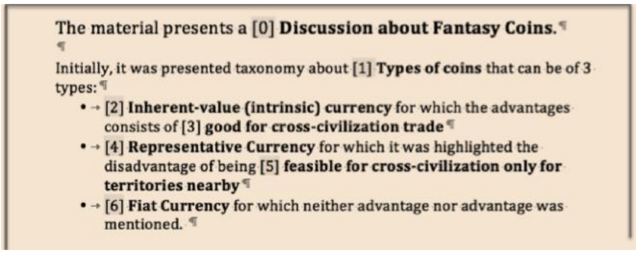
- ▼ RST fantasy currency argmap
  - ▼ Design of Fantasy Currency for LOTR Games
    - ▼ What are the different types of currency? [1]
      - inherent-value (intrinsic) currency [2]@
      - Representative Currency [4]@
      - Fiat Currency [6]
    - ▼ Which currency should our games use? [7]
      - ▼ unique currencies for each society [8]@
        - ▶ orcs: teeth [11]@
        - hobglobins: debased coinage [32]@
        - ▶ drows: luminescent crystal [36]@

inherent-value (intrinsic) currency [2]

Intrinsic currency is the one that D&D uses as a base, where the coin (or jewel, or whatever) itself has value due to the preciousness of the material it is made of, or the craftsmanship of it's making. Hack jewelry is a very common form of this, where necklaces, bracelets and whatever have bits that are to be broken off and used as currency by weight. With the other form, being a proper coin, the stamp of the coin means that the coin has been weighted and measured by an authority that has certified that the coin is worth x amount by weight and purity.

Figure 3. Discussion sample presented as an argument map. Numbers within [] represent the “chunk of information” indexing number.

- RST text (scenario 3), as shown in Figure 4: a rhetoric text that combines, temporal, logical and social aspect of the discussion, as proposed in our research.



The material presents a [0] **Discussion about Fantasy Coins.**¶

¶

Initially, it was presented taxonomy about [1] **Types of coins** that can be of 3 types:¶

- → [2] **Inherent-value (intrinsic) currency** for which the advantages consists of [3] **good for cross-civilization trade**¶
- → [4] **Representative Currency** for which it was highlighted the disadvantage of being [5] **feasible for cross-civilization only for territories nearby**¶
- → [6] **Fiat Currency** for which neither advantage nor advantage was mentioned. ¶

Figure 4. Discussion sample presented as a rhetoric text. Numbers within [] represent the “chunk of information” indexing number.

After reading the material, participants had to answer a questionnaire. As shown in Table IV, the questions were classified according to Bloom’s taxonomy [25], reflecting the cognitive skills expected to be triggered in the participant when answering a question, such as:

- Remembering: retrieving facts;
- Understanding: interpreting the meaning of facts by being able to Exemplify, Classify, Summarize, Infer, Compare and/or Explain;
- Analyzing: breaking content into parts and detecting how the parts relate to each other and to an overall structure or purpose, being able to differentiate and organize the answer;
- Evaluating: making judgment based on criteria;

Additionally, also presented in Table IV, we decoded a possible understanding of the question using a graph database query language. Participants only received the textual questions. The corresponding computational queries were developed to explain the difficulties users might have when answering questions, as if they were computer agents. The graph representation helped to visualize the indirection of the search when answering the questions. The objective of the two last columns of Table IV was to provide readers a notion of the complexity of answering questions. The query language representation suggests the cognitive activities and efforts an agent is required to perform to answer a question, as if the material was represented in a database. The graph representation is another approach to assess the cognitive effort. The greater the number of nodes and indirections, the more complex will be to answer the question.

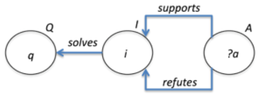
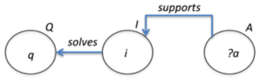
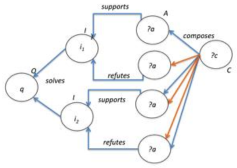
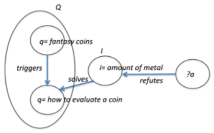
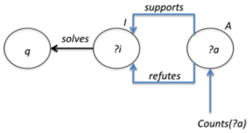
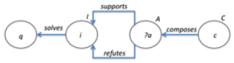
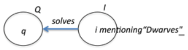
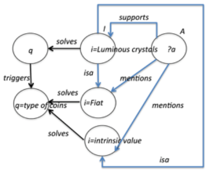
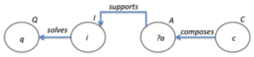
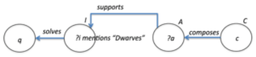
Both pieces of information were used to objectively measure the question’s complexity and check the correlation with precision and recall metrics.

E. The Procedure

The experiment took place in November of 2013, involving 48 participants, spending about 1 hour to answer questions concerning a previous discussion held by a computer game community that lasted a few days in 2009 [24]. Their task was to read the material concerning the discussion and to answer 13 questions about it. All the participants were assured that their information would remain anonymous.

The experiment took place in a controlled room with one computer per participant. A moderator read the experiment description, the permit form and the instructions. Participants were told they could quit at anytime. Actually, two participants quitted. The moderator remained in the room until the end of the experiment. There was no communication allowed among participants during the experiment started.

TABLE IV. QUESTIONNAIRE SAMPLE WITH A QUESTION PROPERLY CLASSIFIED.

Id	Bloom's Classification	Question	Analogous computational question	Graph Representation
1	Remember	What are all arguments for using "Teeth as coins" for the "Orchs civilization"?	$answer(?x,y) \leftarrow (?x\ isa\ Argument), (y\ isa\ Idea),$ $(y = \text{"Teeth as coins "}),$ $(?x\ supports\ y)\|(?x\ refutes\ y)$ ¶	 supports Q I A solves q i ?a refutes
2	Remember	Provide, at least 2 positive evidences, to use "Dungeon &Dragon (D&D) coins with exotic names for each civilization"?	$answer(?x,y) \leftarrow (?x\ isa\ Argument), (y\ isa\ Idea),$ $(y = \text{"Dungeon \& Dragon"}), (?x\ supports\ y),$ $count(?x) \geq 2$ ¶	
3	Understand	What are the similarities and differences of using "Teeth for the Orch civilization" and "Luminous Crystals for the Drows"? Provide at least two of each	$answer(\pi_1,\pi_2) \leftarrow (c\ isa\ Criterion),$ $(i_1\ isa\ Idea), (i_2\ isa\ Idea),$ $(i_1 = \text{"}Teeth\ for\ Orch\text{"}),$ $(i_2 = \text{"Luminous Crystals for Drows"}),$ $(i_1\ \pi_1 c), (c\ \pi_2 i_2)$ ¶ $where\ \pi_1 and\ \pi_2 are\ the\ actual\ paths\ that\ matched$	 supports A I i_1 ?a composes ?c ?a C Q refutes q solves I supports ?a i_2 refutes ?a
4	Understand	Provide an argument that weakens the option of using the amount of metal within a coin as a way to estimate the value of a coin	$answer(?a,i) \leftarrow (?a\ isa\ Argument), \quad (i\ isa\ Idea),$ $(q_1\ isa\ Question), (q_2 is\ a\ Question), (q_2 triggers\ q_1),$ $(i\ isa\ Argument), (q_1 = \text{ Fantasy coins}),$ $(q_2 = \text{"How to estimate the value of a coin"}),$ $(i\ solves\ q_2), (?a\ refutes\ i)$ ¶	 Q q= fantasy coins I triggers i= amount of metal ?a solves refutes q= how to evaluate a coin
5	Evaluate*	Which civilization produces coins with the best quality metal?	$asnwer\left(civilization, best\left(evaluate(a)\right)\right)$ $\longleftarrow (i\ isa\ Idea), (a\ isa\ Argument), (c\ isa\ Criterion),$ $(i\ mentions\ civilization), (a\ supports\|refutes\ i),$ $(c\ composes\ a), (c = foreign\ trading)$	
6	Evaluate *	Which coin seems the best for foreign trading?	$asnwer(i, best(evaluate(a))\ ) \longleftarrow$ $(a\ isa\ Argument), (c\ isa\ Criterion),$ $(a\ supports\|refutes\ i), (c\ composes\ a), (c =$ $foreign\ trading)$	
7	Evaluate *	Which coin was most discussed?	$answer\ (?i, \max\ (count(a)) \longleftarrow$ $\leftarrow (?i\ isa\ Idea), (a\ isa\ Argument), (a\ supports\ i)$	 supports I A solves q ?i ?a refutes Counts(?a)
8	Understand	What was said about "Luminous Crystals" for "foreign trading"?	$answer(?a,c) \leftarrow (?i\ isa\ Idea), \quad (c\ isa\ Criterin),$ $((?a\ supports\ i)\|(?a\ refutes\ i)),$ $(c\ composes\ ?a), (i = \text{"Luminous Crystals"}),$ $(c = \text{"}Foreign\ trading\text{"})$	 supports I A C solves q i ?a composes c refutes
9	Remember	What is the complete list of coins proposed for the Dwarves civilization?	$answer(?i,q) \leftarrow (?i\ isa\ Idea), (q\ isa\ Question),$ $(q = \text{"Fantasy coins"}),$ $(?i\ solves\ q),$ $(?i\ mention\ \text{"}Dwarves\text{"}))$	
10	Evaluate	Do you think, according to the text, that the "Luminous Crystals" are best classified as a fiat or as an intrinsic value coin? Explain your choice	$answer(?i, best((i_1, evaluate(?a_1)), (i_2, evaluate(\ a_2))))$ $\leftarrow (?i\ isa\ Idea), (q_1\ is\ a\ Question), (q_2 isa\ Question),$ $(q_1\ triggers\ q_2), (q_1 = \text{ Fantasy Coins}), (q_2 = \text{ Type of Coins}),$ $(i_1 isa\ Idea), (i_2 isa\ Idea), (i\ isa\ Idea), (i_1 = Fiat), \quad (i_2$ $= Intrinsic\ Value), \quad (i$ $= Luminous\ Crystals), (?a_1\ supports\ i), (?a_1 mentions\ i_1),$ $(?a_2 supports\ i), (?a_2 mentions\ i_2)$	 supports I A solves q i=Luminous crystals ?a triggers isa mentions solves q=type of coins i=Fiat mentions solves i=intrinsic value isa
11	Analyze	What are all arguments supporting "Teeth as coins" related to the "intrinsic value" of a coin?	$answer(x,y) \leftarrow (x\ isa\ Argument), (y\ isa\ Idea),$ $(c\ isa\ Criterion), (x\ supports\ y), (c\ composes\ x),$ $(y = \text{"Teeth as coins"}), (c = \text{"Intrinsic value"})$	 supports Q I A C solves q i ?a composes c
12	Remember	What is the main discussion all about?	$answer(x) \leftarrow (x\ isa\ Question), (x\ Mention\ \text{"Main"})$	
13	Analyze	What are all arguments supporting the claim that Dwarves coins are good for foreign trading?	$answer(?a,c) \leftarrow (q\ isa\ Question), (?i\ isa\ Idea),$ $(?a\ isa\ Argument), (c\ isa\ Criterion),$ $(?i\ solves\ q), (?i\ mentions\ \text{"}Dwarves\text{"}),$ $(?a\ supports\ ?i), (c\ composes\ ?a),$ $(q = \text{Fantasy Coins}), (c = \text{"}Foreign\ trading\text{"})$	 supports I A C solves q ?i mentions "Dwarves" ?a composes c

ISBN: 978-1-61208-443-5

F. *The Metrics*

We considered a set of 26 variables, organized in four groups, as described in Table V, to select the statistically significant ones that might affect the results.

TABLE V. THE SET OF INDEPENDENT VARIABLES.

Variable Type	Variable ID	Description
Material	MT	Forum, argumentation map or RST text
	MNumLetters	Number of letters in the displayed material
	MWC	Number of words in the displayed material
	MBC	Number of blocks in the material. Blocks are posts in forum, nodes in argumentation maps and paragraphs in RST text.
	MIdentation	Maximum indentation of displayed material
Question	QT	Question type according to Bloom's taxonomy
	QNumLetters	Number of letters in the question
	QWC	Number of words in the question
	QQ	Question quantifier: all, some or one
	QClauses	Question's number of clauses
	QNodes	Question's number of nodes
	LLinks	Question's number of links
Expected Answer	EANodes	Number of nodes in the expected answer
	EAFirstNode	Smallest node number in the expected answer
	EALastNode	The greatest node number in the expected answer
	EAMaxSpam	EASpam=EALastNode - EAFirstNode
Participant 's Answer	PANodes	Number of nodes in the participant's answer
	PAFirstNode	The smallest node number in the answer
	PALastNode	The greatest node number in the answer
	PAMaxSpam	PASpam=PALastNode - PAFirstNode
	PADFirstLast	Number of letter from PAFirstNode to PALastNode
	PANumLetters	Number of letters in the participant's answer
	PAWC	Number of words in the participant's answer
	PACNodes	Number of corrected nodes in the participant's answer
	TRM	Participant's time to read the material
	TUAQ	Participant's time to answer a question

The dependent variable included the classic metrics of document retrieval domain, as described in Table VI.

TABLE VI. THE SET OF DEPENDENT VARIABLES.

Variable Type	Variable name	Description
Participant's Answer	Precision	$\frac{PACNodes}{EANodes}$
	Recall	$\frac{PANodes}{EANodes}$
	F-Measure	$\frac{(2 * Precision * Recall)}{(Precision + Recall)}$
	PrecisionHit	$If\ Precision = 1,$ $Then\ PrecisionHit = 1$ $Else\ PrecisionHit = 0$
	RecallHit	$If\ Recall = 1,$ $Then\ RecallHit = 1$ $Else\ RecallHit = 0$
	F-MeasureHit	$If\ F\_measure = 1,$ $Then\ F\_measureHit = 1$ $Else\ F\_measureHitHit = 0$

G. *Statistical Analysis*

The comparison of means test was performed for each of the 13 questions considering the F-measure metric for being a balance between precision and recall. We considered three comparison scenarios:

- Test1: RST and Forum, the null hypothesis is that the F-measure for the RST scenario is not significantly higher than in the Forum scenario;
- Test2: RST and Arg. Map, the null hypothesis is that the F-measure for the RST scenario is not significantly higher than in the Arg. Map scenario;
- Test3: Arg. Map and Forum, the null hypothesis is that the F-measure for the Arg. Map scenario is not significantly higher than in the Forum scenario.

The T-test [26] assumes that samples are randomly drawn from normally distributed populations with unknown population means. For this reason, before performing each of the t-tests, the Kolmogorov-Smirnov test [26] was performed to check the hypothesis of normality. The hypothesis of normally distribution data was only observed for questions 1, 3, 8, 11 and 13. Table VII presents p-values for the three tests. P-value reflects the probability of proving the null hypothesis, i.e., the probability that our hypothesis is false [26]. For questions that did not pass the normality distribution, it was possible to evaluate the proportion of hits for precision and recall, as shown in Table VIII.

TABLE VII. P-VALUES FOR F-MEASURE METRIC. GREEN CELLS HIGHLIGHT P-VALUE < 0.05.

Question	Test 1: RST and Forum	Test 2: RST and Arg. Map	Test 3: Arg. Map and Forum
Q1	0.004639	0.8278	0.0001833
Q3	0.01968	0.006144	0.786
Q8	0.02396	0.05142	0.3286
Q11	0.1363	0.2848	0.2274
Q13	0.1124	0.3507	0.1585

ISBN: 978-1-61208-443-5

TABLE VIII. PRECISION AND RECALL "HIT" P-VALUES. GREEN CELLS REPRESENT P-VALUE <0,1.

	HitPrecision			HitRecall		
Question	Test 1	Test 2	Test 3	Test 1	Test 2	Test 3
Q1	0.00013037	0.2326044	0.001140098	0.07206352	0.9002284	0.00745985
Q2	0.05123522	0.5	0.05123522	0.03137432	0.6868228	0.01105973
Q3	0.1439504	0.03826125	0.7673956	0.1548145	0.1548145	0.5
Q4	0.5	0.5	0.5	0.1548145	0.1548145	0.5
Q5	0.1548145	0.5	0.1548145	0.1548145	0.5	0.1548145
Q6	0.2720985	0.2720985	0.5	0.2720985	0.2720985	0.5
Q7	0.2720985	0.07206352	0.8174849	0.2720985	0.07206352	0.8174849
Q8	0.00114010	0.01625472	0.1439504	0.3131772	0.3131772	0.5
Q9	0.1548145	0.03442252	0.8574753	0.07206352	0.07206352	0.5
Q10	0.00328921	0.5	0.003289207	0.1425247	0.03442252	0.8451855
Q11	0.06356221	0.5	0.06356221	0.5	0.8451855	0.1548145
Q12	0.5	0.5	0.5	0.5	0.5	0.5
Q13	0.07206352	0.8451855	0.01625472	0.06356221	0.2326044	0.2071081

We had to discharge task duration from our analysis, since there was too much noise in this measurement because some, but not all, participants did a great deal of copying from the original material and pasting as an answer. This answering method was fair, but the time spent typing an answer and the time spent copying&pasting could dramatically mask the results.

F-measure for questions 1, 3 and 8, as well as PrecisionHit and RecallHit for questions 1, 2, 8, 9, 10, 11 and 13 support our claim that RST-organized text improves newcomers understanding of a discussion. Two questions were raised from these results: "Why were not good the results for questions 4, 5, 6, 7 and 12?" and "What may explain the good performance of RST?".

First of all, let us investigate possible reasons for the bad results. Questions 5, 6 and 7 concern providing the "best perceived value" for an entity: "civilization with best quality coin", "the best coin for foreign trade" and "the most discussed coin", respectively. Although intuitively these questions should have triggered a search, evaluation, comparison and selection processes, participants perform as in a pattern matching style. They answered very fast and their answers presented very high precision and recall values, no matter the method. The task becomes just a matter of retrieving from the text the information that caused the highest impact in the reader. Thus, no matter the method, it will be a matter of being impacted by the information. When we are asked for explaining our selection, as triggered in question 10, probably the 4-stage process is called to take place. In this later case, RST text method played an important role facilitating newcomers' understanding.

Question 12 was also a question to check the minimum attention of the participants. It was a pretty easy question with very high precision values no matter the method. Actually, we could use this question as a filter: only consider respondents with high values in this question.

We are still investigating the possible answers for the bad results of question 4

Although we had very good results for an exploratory research, we wanted to understand why the RST format was causing such positive impact. We investigated the 26 independent variables searching for single or group correlations that would explain the results on the 624 answers.

We used the LASSO statistic method [27] to select the relevant variable to run a linear regression model. The method penalizes models with higher number of variables. It does a balance between quality and complexity.

V. CONCLUSION AND FUTURE WORK

The contribution of this research was to show that the presentation format of a group discussion impacts comprehension. Furthermore, a rhetorically organized text improves understanding, especially for answering cognitively complex questions, over classic sequential forum's organization. We are currently implementing a tool developing according to our RST answer generator method. We did a small experiment to explore the idea.

The results, as presented in Table VIII, were very promising and interesting, even when p-value was not small enough to refute the null hypotheses.

From the 26 variables, results indicated HitF-measure (0 or 1 value) was mostly affected by EAMaxSpam parameter. The results indicate, with p-value<0.05, that F-measure is inversely affected by the size of the spam of the expected answer (EASpam).

Inspecting our automatic RST text generation, we realize that this is exactly what the method is meant to do: grab the relevant information pieces and organize them in a concise text, bringing together time, logic and social information to provide context to the message.

This is still a first, but promising step towards specifying the design of large-scale collaboration environments. Although we did a comparison study, we believe RST texts can be used as a storyteller, guiding participants through logical (argument map representation), temporal (forum) and social aspects of a discussion.

The main objective of our research was accomplished. We could show that making sense of an ongoing discussion can be sensibly improved by using RST-based textual explanation generated from argument map organized discussion. Additionally, our observations on participants' behavior answering questions about the discussion raised the possibility of integrating this Q&A feature to crowd source answers that would be generated exploring the material of a discussion done by experts.

While RST was developed as a descriptive technique for analyzing natural language text, it can also be used prescriptively to describe how logical points can be structured in order to be persuasive and clear. Given that RST structures demonstrably increase comprehensibility of complex content, our next step is to explore how we can generate RST structures automatically for real-world online discussions. Our strategy for this will include:

- generating argument maps - natively or by argument mining [28][29] from web forums
- developing algorithms to automatically generate RST-structured responses to queries from these argument maps, building upon on a taxonomy of canonical query types which each have an RST template plus rules describing how to harvest the argument map information needed to fill in the empty template slots.

REFERENCES

[1] M. A. Andersen, “Asynchronous discussion forums: success factors, outcomes, assessments, and limitations,” Educational Technology & Society, vol. 12 (1), 2009, pp. 249–257.

[2] D. Feng, E. Shaw, J. Kim, and E. H. Hovy, “An intelligent discussion-bot for answering student queries in threaded discussions,” Proc. 11th Intelligent User Interface Conference, 2006, pp. 171-177.

[3] J. Conklin, A. Selvin, S. B. Shum, and M. Sierhuis, “Facilitated hypertext for collective sensemaking: 15 years on from Gibis,” Proc. 8th International Working Conference on the Language Action Perspective on Communication Modelling (LAP'03), July 2003, pp. 1-22.

[4] M. Klein, “The MIT Deliberatorium: Enabling Large-Scale Deliberation About Complex Systemic Problems,” Proc. International Conference on Agents and Artificial Intelligence, 2011, pp. 15-24.

[5] W. Kunz and H. Rittel, “Issues as Elements of Information Systems”, Working Paper 131, Center for Planning and Development Research, University of California, Berkely, CA, 1970.

[6] S. B. Shum and A. M. Selvin, “Structuring discourse for collective interpretation,” Proc. Distributed Collective Practices: Conference on Collective Cognition and Memory Practices, 2000, pp. 1-16.

[7] V. Uren, S. B. Shum, G. Li and M. Bachler, “Sensemaking Tools for Understanding Research Literatures: Design, Implementation and User Evaluation,” International Journal of Human Computer Studies, vol. 64(5), 2006, pp. 420–445.

[8] U. Hermjakob, “Parsing and question classification for question answering,” Proc. ACL Workshop on Open-Domain Question Answering, vol. 12, 2001, pp. 1-6.

[9] F. H. Eemeren and R. Grootendorst, A Systematic Theory of Argumentation: The Pragma-dialectical Approach. Cambridge, MA: Cambridge University Press, 2003.

[10] D. N. Walton and E. C. W. Krabbe, Commitment in dialogue: Basic concepts of interpersonal reasoning. Albany, NY: State University of New York Press, 1995.

[11] M. Klein and G. Convertino, “A Roadmap for Open Innovation Systems,” Journal of Social Media for Organizations, vol. 1 (2), 2015, pp. 1-16.

[12] M. Klein and G. Convertino, “An Embarrassment of Riches: A Critical Review of Open Innovation Systems,” Communications of the ACM, vol. 57(11), 2014, pp. 40-42.

[13] F. Calefato, F. Lanubile, F., M. Raffaella and N. N. Merolla, “Success Factors for Effective Knowledge Sharing,” Proc. 10th International Forum on Knowledge Asset Dynamics (IFKAD'15), Jun 2015, pp. 1-11.

[14] J. Giles, “Internet encyclopaedias go head to head,” Nature, vol. 438(7070), 2005, pp. 900-901.

[15] A. Kittur, B. Suh, B. A. Pendleton and E. H. Chi, “He says, she says: conflict and coordination in Wikipedia,” Proc. SIGCHI Conference on Human Factors in Computing Systems, ACM Press, 2007, pp. 453-462

[16] F. B. Viegas, M. Wattenberg and K. Dave, “Studying cooperation and conflict between authors with history flow visualizations,” Proc. SIGCHI conference on Human factors in computing systems, ACM Press, 2004, pp. 575–582.

[17] P. A. Kirschner, S. J. B. Shum and C. S. Carr, Visualizing Argumentation: Software tools for collaborative and educational sense-making, Springer, 2003.

[18] A. D. Moor and M. Aakhus, “Argumentation Support: From Technologies to Tools,” Communications of the ACM, vol. 49(3), 2006, pp. 93-98.

[19] C. S. Carr, “Using computer supported argument visualization to teach legal argumentation,” in Visualizing argumentation: software tools for collaborative and educational sense-making, P. A. Kirschner, S. J. B. Shum and C. S. Carr, Eds. Berlin: Springer-Verlag, 2003, pp. 75-96.

[20] S. J. B. Shum, A. M. Selvin, M. Sierhuis, J. Conklin and C. B. Haley, “Hypermedia Support for Argumentation-Based Rationale: 15 Years on from gIBIS and QOC,” in Rationale Management in Software Engineering, A. H. Dutoit, R. McCall, I. Mistrik and B. Paech, Eds. Berlin: Springer-Verlag, 2006, pp. 111-132.

[21] M. Klein, “Enabling Large-Scale Deliberation Using Attention-Mediation Metrics,” Computer-Supported Collaborative Work, vol. 21(4), 2011, pp. 449-473.

[22] W. C. Mann and S. A. Thompson, “Rhetorical structure theory: towards a functional theory of text organization”. Text, 8 (3), 1988, pp. 243-281.

[23] M. Taboada and W. C. Mann, Rhetorical Structure Theory: looking back and moving ahead, Discourse Studies, London: SAGE, 2006, pp. 423-459.

[24] Giant in the Playground Forum: http://www.giantitp.com/forums/showthread.php?110342-Some-Alternative-Currency-Ideas. Accessed September, 8th, 2015.

[25] L. W. Anderson, D. R. Krathwohl, P. W. Airasian, K. A. Cruikshank, R. E. Mayer, P. R. Pintrich, J. Raths, and M. C. Wittrock, A Taxonomy for Learning, Teaching, and Assessing: A revision of Bloom's Taxonomy of Educational Objectives, New York: Pearson, Allyn & Bacon, 2001.

[26] R. S. Witte and J. S. Witte, Statistics, Wiley, 10th edition, 2013.

[27] R. Tibshirani, “Regression Shrinkage and Selection via the Lasso,” Journal of the Royal Statistical Society, 58(1), 1996, pp. 267-288.

[28] B. Pang and L. Lee, “Opinion mining and sentiment analysis”. Foundations and trends in information retrieval, vol 2(1-2):1-135, (2008) .

[29] C. Reed and G. Rowe, “Araucaria: software for argument analysis, diagramming and representation,” International Journal of Artificial Intelligenc Tools, vol. 13(4), 2004, pp. 961-979.

Virtual Activity Ontology Modeling: the Case of Sociocultural Knowledge Sharing

Kaladzavi Guidedi
University of Maroua
Maroua, Cameroon
kaladzavi@ymail.com

Papa Fary Diallo
University of Gaston Berger
Saint-Louis, Senegal
diallo.papa-fary@ugb.edu.sn

Kolyang
University of Maroua
Maroua, Cameroon
dtaiwe@gmail.com

Moussa LO
University of Gaston Berger
Saint-Louis, Senegal
moussa.lo@ugb.edu.sn

***Abstract*-In the light of the rhythm of the current cultural mixing, we believe that in the long term, culture of African people in particular may disappear. Some new computational techniques (semantic web technologies) are needed to manage the large repositories of sociocultural data and to discover useful patterns and knowledge from them. This paper presents a virtual activity ontology modeling approach, in the case of sociocultural knowledge sharing and co-construction named Ontoshare. Our modeling approach is based on Engeström's Human Activity Theory (HAT). With Ontoshare we designed how Internet users could build the content of a sociocultural Knowledge Management System (KMS); this vocabulary also organises data, facilitates information retrieval by introducing a semantic layer in social web platform, we plan to implement. The platform could be considered as a « collective memory » which will allow communities to share, co-construct and discover sociocultural knowledge in the Cameroonian context.**

Keywords-sociocultural knowledge; sharing; Human Activity Theory.

I. Introduction

Since 2005, UNESCO has projected that local knowledge will increasingly become the main point of social mutations; it leads economic, political, and sociocultural projects. Emerging societies must avoid being mere components of the "Global Information Society". Effective participation of African Countries in "Societies of knowledge" is required [1]. While, local knowledge is considered in the conception of different project development, B. Z. Deli deplores the fact that, in Africa, cultures are deteriorating and emptying of their meanings, their mellow content and values [2]. To promote indigenous knowledge, some media have been proposed: a permanent (re)education, radio broadcasting, and of course Internet. On the Web, any topics may be discussed. As a result, the Web constitutes the source of global information. However, when we consider how Internet works by focusing on data flow, social values such as equality, freedom, democracy which are supposed to characterise the "Global Information Society" appear to be an illusion. The unbalance between these social values is described by "*digital divide*" issue between Northern and Southern countries. The "*digital divide*" has many views. Out of these views, there are "*divide by access*", "*divide by use*", "*divide by decision*", and "*divide by content*". The "*divide by content*" is defined as the gap between Western culture over-represented and African culture under-represented. This gap is considerable [3]. To reduce this gap and promote the African culture, to refresh the memory of our citizens and show the transparent view of opportunities (unknown infrastructures, spaces, etc.) and challenges (investments unequally distributed, marginalised communities, etc.) through endogenous information (from involved actors), rather than external analysis, we propose, the implementation of ontologies-based web platform for sociocultural knowledge sharing and co-construction in the Cameroonian context.

By Web 2.0, Internet users are no longer just consumers but also authors of information. Semantic Web is a Web evolution, through ontologies; it improves data organisation, and information retrieval. Ontology is an explicit specification of a shared conceptualisation [4]. As such, ontologies are considered to be knowledge representation tools, transforming data into information and information into knowledge. They are recommended tools for knowledge reuse, organisation, interoperability, integration, and valorisation on the Web [5].

The objective of this paper is to present a virtual activity ontology modeling approach in the case of sociocultural knowledge sharing and co-construction named Ontoshare. Ontoshare modeling approach is based on Engeström's Human Activity Theory (HAT). HAT is a conceptual framework which, the foundational concept is "activity", understood as purposeful, transformative, and developing interaction between actors ("subjects") and the world ("objects") [6].

This paper is organised as follows: In Section 2, we present research work related to our domain. In Section 3, we shall explain our modeling approach, which is based on Engeström's Human Activity Theory and how we reused the related ontologies to solve interoperability issue. Moreover, we will describe in Section 4, how Ontoshare has been populated and improved through inference services provided by reasoning engine in Protégé. In the same section, we will

ISBN: 978-1-61208-443-5

point out some SPARQL queries on a Knowledge Base related to the use cases to illustrate how the “collective memory” could be built. This will end with, a conclusion and future work.

II. RELATED WORK

To the best of our knowledge, only the Sociocultural Ontology [7][8] covers the sociocultural domain which has been developed in the Senegalese context. The modeling approach is based on Vygotsky framework [9]. This framework is considered as the first Generation of HAT. It is organised around the "*mediation*" concept and based on the idea that, human actions are mediated by cultural, symbolic or physical artifacts that enable man to act on his environment.

The main limitation of Vygotsky's framework is the fact that, it focuses solely on individual actions and not on individual actions within the community. The modeling approach proposed in [7][8] tried to solve this limitation by substituting *subject* by *community*. In this model, there are three main axes, which are:

1. *Community*: which is a group of people sharing a common interest in a sociocultural domain;
2. *Object*: considered as a locality or infrastructure where community evolves;
3. *Artifact*: it mediates the actions of community with object.

Translating *subject* by *community* has just solved the problem concerning community action but has failed to model the dynamics within the community. The authors of this model consider the community as an atomic entity. The modeled ontology in [7] hides some knowledge on the internal dynamics (collaborations, interactions, actors, roles, etc.) of the community and the contextual nature (regulations) while organising activities considering the fact that knowledge on internal dynamic within community will enrich our Knowledge Base and allow deep analysis of communities and activities.

Due to the limitations of sociocultural ontology [7], we proposed in our previous paper an improved sociocultural ontology named OntoSOC [10]. Its modeling approach considers the dynamics within communities and the contextual nature for organising activities. This vocabulary enabled us to semantically circumscribe the content of sociocultural knowledge.

However, on the Web, the flows have been reversed: Internet user is no longer passive (reader) but active (author). The transition from Web 1.0 to Web 2.0 done in 2004 was a decisive transition to social media paradigm. Social media includes all tools and applications that allow interaction between Internet users. Within that “*galaxy*” of social media, there are several "*planets*". Out of them, there are texts publishing tools (wikis, blogs, etc.), exchanging and sharing tools (YouTube for videos, Slideshare for sharing presentations, etc.), tools for discussion (Skype, Messenger, etc.), and networking tools (Twitter, Facebook, MySpace, etc.), etc., are in continuous supply.

Unfortunately, most of them are not sharing and discovering local knowledge oriented, except Wikipedia [11], the online encyclopedia which is inter-domain and inter-society. The second limitation and not of lesser importance, is the fact that they are not ontologies-based. In the case of Wikipedia, DBpedia [12] project has been launched to address this issue. DBpedia is an academic and community project for automatic data exploration from Wikipedia to propose a structured version in semantic web format (RDF) of data. Due to the previous limitations and the unknown of a sociocultural sharing ontology, designing ontology of how Internet users could interact to share knowledge is needed.

III. ONTOSHARE MODELING

In this section, we present, why and how we used Human Activity Theory to model Ontoshare.

A. Methodology

Practically, ontological engineering does not propose a standardised methodology for designing ontologies [13]. Our concern is to implement a “*collective memory*” system helping people to have a holistic view on local changes, while considering culture and historicity in our localities. Otherwise, we agree with P. Berger and T. Luckmann in [14] with the fact that reality is a social construction and that, the universe is evolving. These changes are driven by groups of individuals through their various activities. For analysing and understanding these changes and how they transform reality, some activity models have been proposed. There are three generations of this model: the first and the second are characterised by focusing on subject and its actions; the third generation is known as collective view framework making distinction between individual actions and collective activities. For these purposes, Engeström’s model is more adapted and has been chosen as framework of our modeling process.

Engeström’s model clearly points out distinction between individual action and collective activity. It is produced according to the historical and cultural view of activity. The model has six poles (*Subject, Object, Tools, Rules, Community and Division of Labour*) as shown in Figure 1.

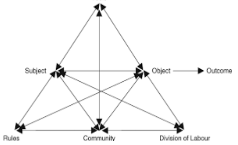


Figure 1. Engeström’s Human Activity triangle

Copyright (c) IARIA, 2015. ISBN: 978-1-61208-443-5

- *Subject*: represents the chosen individual to analyse;
- *Object*: environment transformation by activity (task to be performed, objective to be achieved);
- *Tools*: materials or symbolic tools that mediate activity ;
- *Comm*unity: set of individuals that share the same interests and thereby differ from other communities;
- *Division of Labour:* it considers, the horizontal distribution of actions among subjects, community members and the vertical hierarchy or responsibilities and status;
- *Rules*: they refer to implicit and explicit standards, conventions, habits, etc. that maintain and regulate actions and interactions within the community.

Many Human-Computer-Interaction (HCI) researchers were eager to move beyond the confines of traditional cognitive science to HAT direction [16]. Activity Theory is a powerful and clarifying descriptive tool rather than a strongly predictive theory. It incorporates strong notions of intentionality, history, mediation, collaboration and development in constructing conscientiousness. HCI research needs to understand and describe "*context*", "*situation*" and "*practice*". Considering these previous features of HCI, its objectives are related to HAT objectives. Although, the HAT is developed for offline communities, the concept of online community is still the set of people (Internet users) who share the same values and interests. Thus, to analyse how internet users will interact with computer to co-construct knowledge in our context, Human Activity Theory constitutes the powerful tool [17].

B. Concepts and Relationships identification process

In this section, we focus on how Engeström' HAT has been used to construct our ontology (Ontoshare). In other words, we want to show how we deduced concepts and relationships from HAT. To do that, due to "collaborative persona" approach, we simulated HAT in three use cases of online knowledge sharing. "*Persona method*" is a modeling strategy used by software architects. This idea was introduced by A. Cooper, software designer [18]. In Software Engineering, this approach is called "*Goal-directed design*". It represents patterns of users' behavior, goals and motivations, compiled in a fictional description of a single individual. It also contains made-up personal details, in order to make the person more "*tangible and alive*" for the development team. In our case, we use "*collaborative persona*" suitable for collaborative, participative and interactive context as community [19].

To simulate the online shared information, the following activities were chosen: a cultural event organised by *NakoSenda* community in Mokolo (locality), rural library building activity conducted by CDE-SAARE [20] in Kolara (locality) and a soccer tournament holidays organised by *Club 2-0-UMa.*

There are two types of Internet users: passive and active users. Active user is the one who edits (shares) information. For the purpose of this paper, active user is named contributor. Any contributor can be a member of a community which organises activity, or a witness who attends the activity or just anyone who has information (partial or complete) to share on an activity. For example, in the first case, a contributor named D, created in our platform a page named "*cultural event in Mokolo",* in which he edited information about the location (weather forecast, vegetation, etc.) of Mokolo. The second contributor named MP who is Nakosenda's member, shared information on the same page relative to *NakoSenda* (the founder, headquarter, regulations, type). The third contributor called M shared his impression on the event. All these contributors shared and co-constructed sociocultural knowledge on cultural event organised by *NakoSenda* in the respect of platform regulations. Thus, for each use case and for each contributor, we simulated separately, twelve triads within the overall triangle as shown in Figure 2.

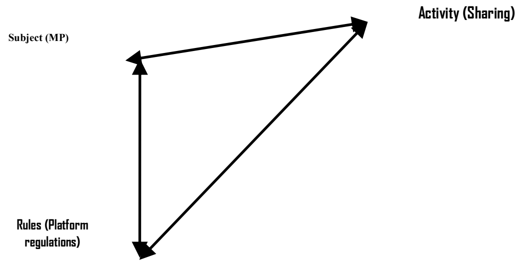


Figure 2. Subject-regulations-Activity triad

As seen in Figure 2, the above triad gave rise to the following triples:

- ***isRespectedBy (****contributor**,** Paltform regulations**) ;***
- ***isParticipedby (****contributor**,** sharing**) ;***
- ***isRegulatedBy (****sharing**,** Paltform regulations**).***

Thirty triples in all were deduced from twelve triads. The connexity of some triads gave in some cases identical triples. For example: *Subject-Community-Division of labour* and *Community-Division of labour-Activi*ty producing identical triple, ***isCreatedBy*** (*Community*, *Division of labour*). After identifying and eliminating redundant triples, the following results were obtained:

- ***isUsedBy****(sociocultural Knowledge, Contributor);* ***isMemberOf*** *(Contibutor, community of Contributors);*
- ***isRegulatedBy (****Sharing**,** Paltform regulations**)**;*
- ***isAllowedBy*** *(Contributions, Administrators);*
- ***isEditedBy*** *(Contibution,Contributor);*
- ***isConcernedBy*** *(Sharing, Contribution);*
- *is**Monitored*** *(Sharing, Administrators);*

- ***isRespectedBy*** ***(****Contributor****,*** *Paltform regulations****);***
- ***isParticipedby*** ***(****Contributor****,*** *sharing****).***

Subjects and objects of these triples model fundamental concepts and predicates representing relationships between them. Table 1 shows mapping carried out between poles of Engeström's model and upper-level concepts of our ontology.

TABLE I. MAPPING BETWEEN ENGESTRÖM'S MODEL POLES AND ONTOSHARE UPPER- LEVEL CONCEPTS

Engeström model poles	**Ontoshare Upper-level concepts**
Object	Sharing
Subject	Contributor
Rules	Platform regulations
Community	Community of Contributors
Division of Labour	Contributions
Tools	Sociocultural Knowledge

- ***Community of Contributors****:* set of contributors;
- ***Sharing:*** information co-construction activity ;
- ***Contributor:*** active Internet user;
- ***Platform Regulations****:* rules defined for guiding sharing activity and contributors;
- ***Contributions:*** shared information for describing any *Sociocultural Knowledge* on the platform. In others words, this concept models each piece of information edited (shared) by any contributor into the knowledge construction in platform;
- ***Sociocultural Knowledge:*** This concept models information about society and culture. According to [21], sociocultural knowledge concept, concerns all forms of human knowledge: objects that compound the real world, facts and events. This concept is complex. As a result, sociocultural ontology named OntoSOC proposed in our previous work [9] was reused. This vocabulary helped us to semantically circumscribe the content of sociocultural knowledge. To reuse it, *Sociocultural Knowledge* concept was used as a "bridge" concept.

C. Concepts and Relationships

Different concepts obtained in Table 1 represent fundamental classes of Ontoshare. There are seven upper-level concepts in all. Figure 3 illustrates these concepts and relationships between them. In fact, classes alone are not enough to define ontology; we need to define also relations between them and attributes to characterise classes. These two notions add semantics to ontologies. In our study, use cases were required in identifying the following relationships:

- ***isUsedBy****(Sociocultural Knowledge, Contributor) ;*
- ***MemberOf*** *(Contibutor, Community of Contributors) ;*
- ***isRegulatedBy*** *(Community of Contributors, Plateform-Regulations) ;*
- ***isAllowedBy*** *(Contribution, Administrators) ;*
- ***isEditedBy*** *(Contibution, Contributor) ;*
- ***isConcernedBy*** *(Sharing, Contribution) ;*
- ***isMonitored*** *(Sharing, Administrators).*

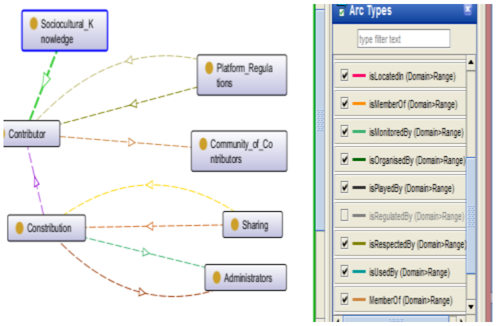

Sociocultural_K nowledge
Platform_Regula tions
Contributor
Community_of_Co ntributors
Constribution
Sharing
Administrators
Arc Types
type filter text
☑ isLocatedIn (Domain>Range)
☑ isMemberOf (Domain>Range)
☑ isMonitoredBy (Domain>Range)
☑ isOrganisedBy (Domain>Range)
☑ isPlayedBy (Domain>Range)
☐ isRegulatedBy (Domain>Range)
☑ isRespectedBy (Domain>Range)
☑ isUsedBy (Domain>Range)
☑ MemberOf (Domain>Range)

Figure 3. Ontoshare concepts and relationships

D. Hierarchy of classes

Out of seven concepts, three have variant depth, going from one to level six. There are many approaches to define hierarchy of classes: top-down, bottom-up and hybrid approach. The use of HAT not only enabled us to generate fundamental concepts but also marked the beginning of top-down method. Thereafter, to better define hierarchy, we intended to "*think up*" before making specifications. This is a top-down development process that begins with definition of the most general concepts in the domain and continues with sub-concepts specialisation. Figure 4 shows an overview of possible articulation between various generality levels of Ontoshare.

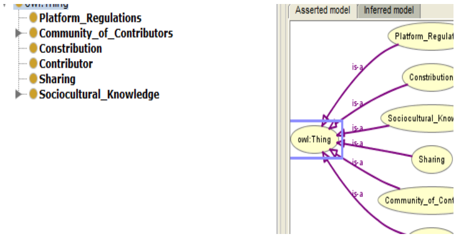


Figure 4. Ontoshare hierarchy extract

For example, *community* of *contributors* consisting of *Administrators*, *Community members* and *not-Community members* gives the overview of those who can share information on the platform. *Administrators* are particular members. In addition to their contributions, they have to work on monitoring edited contents by implementing and setting platform regulations. It clearly appears from this that, hierarchy of *sociocultural knowledge* provides information on different views (task, tools, community, actors, regulations, etc.) of data that each contributor can share through *sharing* concept.

E. Ontoshare alignment

For semantic web, alignment is a solution to the interoperability issue. It helps not to recreate those that exist but only to improve them. Ontoshare is an inter-domain vocabulary. It reuses some concepts of related ontologies (FOAF, Schema.org, DBpedia, and wai). Figure 5 presents alignment (manually) done between related vocabularies and some corresponding Ontoshare concepts.

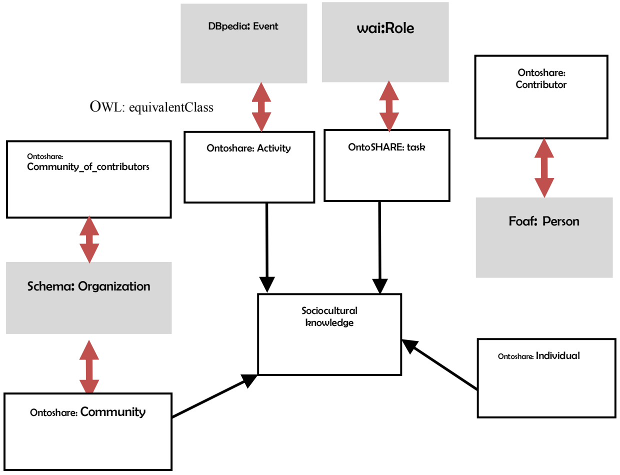


Figure 5. Ontoshare alignment

In recent years, the concept of Linked Open Data (LOD), and the so-called Web of Linked Data, has attracted tremendous attention from both the academic and real application world. The idea is, if we start to publish machine-readable data, such as RDF documents on the Web, and somehow make all these documents connected to each other, then we will be creating a Linked Data Web that can be processed by machines. Alignment technic participates to build the LOD project at two levels:

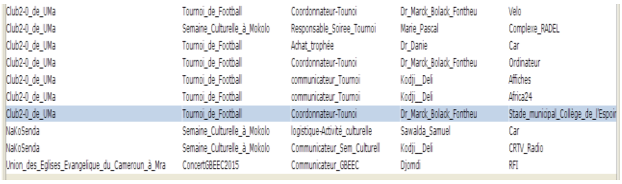
- Using *OWL:equivalentClass* property to connect Ontoshare concepts to standard vocabularies concepts and share semantics between these concepts. It is the case of *ontoshare:task* which is semantically equivalent to *wai:role and ontoshare:contributor* which is semantically equivalent to *foaf:person*;
- Using *is-a* property to add news concepts to LOD through standard vocabularies (FOAF, Schema.org, DBpedia, and wai). In our case, we added to LOD, a local concept as « *tontine* » which is the sub-concept of *ontoshare:community* class representing the local social network.

IV. ONTOSHARE POPULATING AND VALIDATION

For editing, we used "Protégé 5.0", an ontology development tool [22]. Ontoshare populating was done with data related to some activities (cultural events, sport events, and religious events). Three of them have been used previously for simulating HAT.

Protégé offers a number of reasoning engines and SPARQL endpoint in its standard distribution. A reasoner checks for consistency of description of class, subsumption between classes, taxonomy of class names (classification) and finds classes that match known instances. The performance of the proposed ontology has been evaluated at the following levels: classification, consistency checking using a reasoner, and competence question checking by SPARQL queries. According to classification checking, we tried to identify by classify function; if instances are automatically classified in a defined class. we did it for all concepts. For consistency checking, we aimed to verify, if there is any class which could never have an instance due to its definition. The competence question checking allowed by SPARQL endpoint enables to verify, if ontoshare can answer a competency question that guided its design. Thus, we focused on various activities organised in a specific locality to evaluate consistency of the Knowledge Base, and check for infinite query (eventually). The following query enabled to extract the relevant information on organised activities, resource used and tasks realised by subjects for any given community from our Knowledge Base.

```
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
PREFIX OntoShare: <http://maroua-univ/ns/ OntoShare #>
SELECT  ?Communuties ?Activité ?task ?person ?tools
WHERE { ?task OntoShare:isUsedBy ?tools
OPTIONAL { ?Activité OntoShare:isRealizeBy ?task }
OPTIONAL { ?task OntoShare:isPlayedBy ?person }
OPTIONAL { ?task OntoShare:isCreatedBy ?Communuties }
} ORDER BY  ?Communuties
```



Club2-0_de_UMa	Tournoi_de_Football	Coordonnateur-Tounoi	Dr_Marck_Bolack_Fontheu	Velo
Club2-0_de_UMa	Semaine_Culturelle_à_Mokolo	Responsable_Soiree_Tournoi	Marie_Pascal	Complexe_RADEL
Club2-0_de_UMa	Tournoi_de_Football	Achat_trophée	Dr_Danie	Car
Club2-0_de_UMa	Tournoi_de_Football	Coordonnateur-Tounoi	Dr_Marck_Bolack_Fontheu	Ordinateur
Club2-0_de_UMa	Tournoi_de_Football	communicateur_Tournoi	Kodji__Deli	Affiches
Club2-0_de_UMa	Tournoi_de_Football	communicateur_Tournoi	Kodji__Deli	Africa24
Club2-0_de_UMa	Tournoi_de_Football	Coordonnateur-Tounoi	Dr_Marck_Bolack_Fontheu	Stade_municipal_Collège_de_l'Espoir
NaKoSenda	Semaine_Culturelle_à_Mokolo	logistique-Activité_culturelle	Sawalda_Samuel	Car
NaKoSenda	Semaine_Culturelle_à_Mokolo	Communicateur_Sem_Culturell	Kodji__Deli	CRTV_Radio
Union_des_Eglises_Evangelique_du_Cameroun_à_Mra	ConcertGBEEC2015	Communicateur_GBEEC	Djomdi	RFI

Figure 6. SPARQL query overview

The result of the SPARQL query is presented in Figure 6. It illustrates how the content of our KMS has been enriched by Internet users by editing some pieces of information related to sociocultural activities.

We would like to point out that, access features to ICT such as "*divide by access*", "*divide by use*", "*divide by decision*" in Cameroon must be improved, even if, sharing knowledge and discovering on that platform will not only be done by Cameroonians, but all Internet users worldwide. In fact, according to [23] "*digital divide*", statistics are very low. As a result, Cameroon would be the seventieth country over fifty four in Africa with about one million of Internet users. This number represents only 5% of its population. The same survey, recorded that the Internet penetration rate would be around 0.01 %.

V. CONCLUSION AND FUTURE WORK

In this paper, we presented Ontoshare modeling approach. Ontoshare is online activity ontology in the case of sociocultural knowledge sharing. To get there, we used Engeström's Human Activity Theory (HAT). To fix the interoperability problem, we have established mapping between Ontoshare concepts and those related. We populated our ontology with use cases and applied some SPARQL tests. Certainly, our use cases are far to be representative, but, their data helped us to eliminate or explain some inconsistencies and demonstrate how our "collective memory" could be co-constructed. It should be noted that redundant triples elimination process was done empirically. We have no guarantee of reaching minimal coverage. Nevertheless, reduction rate is considerable, about 60%. In addition, due to the fact that sociocultural knowledge is complex, we reused OntoSOC to circumscribe data type to be shared and co-constructed into our platform through sociocultural knowledge concept.

In perspective, we will focus on domain ontology and design the platform's architecture.

ACKNOWLEDGMENT

The authors would like to thank the University of Maroua which partially funded this work. Special thanks to LANI (Laboratoire d'Analyse Numérique et d'Informatique) of the University of Gaston Berger, Saint-Louis, in Senegal to our research position in the Laboratory.

REFERENCES

[1] UNESCO, "Toward the knowledge societies", world report of UNESCO, editions UNESCO, 2005.

[2] B. Z. Deli, "Western culture imperialism and the future of African culture: Challenges and Perspectives" GSSA, Maroua, Cameroon, 2008, pp15-20.

[3] http://www.enssib.fr/bibliotheque-numerique/notices/1948-de-la-fracture-numerique-a-la-fracture-cognitive-pour-une-nouvelle-approche-de-la-societe-de-l-information, [retrieved: 10, 2015].

[4] T. R. Gruber, "Toward Principles for the Design of Ontologies used for Knowledge Sharing", International Journal of Human Computer Studies, Special issue: the Role of Formal Ontology in the Information Technology, Vol. 43, No. 5, , 1995, pp.907-928.

[5] G. Kaladzavi, P. F. Diallo, Kolyang, and M. LO, "The role of ontologies in valuing African sociocultural knowledge on the Web ", 2nd Computer Science Reasearch Conference, Yaoundé, Cameroon, 2015, In press.

[6] https://www.interaction-design.org/literature/book/the-encyclopedia-of-human-computer-interaction-2nd-ed/activity-theory, [retrieved: 10, 2015].

[7] P. F. Diallo, S. M. Ndiaye, and M. LO, "Study of sociocultural ontology", In the First International Conference on Social Economics and Informatics, Spain, October 2011, pp. 43-53.

[8] P. F. Diallo, O. Corby, M. LO, I. Mirbel, and S. M. Ndiaye, "Sociocultural Ontology: Upper-level and Domain Ontologies" in Actes JFO, Hammamet, Tunisia, november 2014, pp. 15-27.

[9] L. S. Vygotsky, "Mind in society: the development of higher psychological processes, "Cambridge: Harvard University Press.

[10] G. Kaladzavi, P. F. Diallo, Kolyang, and M. LO, "OntoSOC: Sociocultural knowledge ontology", International Journal of Web & Semantic Technology, vol 6, No 2, April 2015, DOI: 10.5121/ijwest.2015.6201.

[11] https://en.wikipedia.org/wiki/Main_Page, [retrieved: 09, 2015].

[12] http://mappings.dbpedia.org/server/ontology/classes/, [retrieved: 10, 2015].

[13] A. Gangemi, A. Prisco, M. T. Sagry, G. Steve, and D. Tiscornia,"Some Ontological Tools to Support Legal Regulatory Compliance, with a Case Study", In OTM Confederated International Workshops, 2003, pp. 1-2.

[14] http://mip-ms.cnam.fr/servlet/com.univ.collaboratif.utils.LectureFichiergw?ID_FICHIER=1295877017861, [retrieved: 09, 2015].

[15] L. S. Vygotsky, "Mind in society: the development of higher psychological processes." Cambridge: Harvard University Press.

[16] Y. Engeström, "Activity theory as a framework for analyzing and redesigning work", Ergonomics, 2000, pp. 960-974

[17] K. Kuutti, "Activity Theory as a potential framework for human-computer interaction research", Cambridge: MIT Press, 1995, pp. 17-44, [retrieved: 10, 2015].

[18] S. Blomkvis, "The User as a personality. Using Personas as a tool for design", Theoretical perspectives in Human-Computer Interaction, IPLab, KTH, September 3, 2002.

[19] A. Giboin, "From Individual to Collective Personas Modeling Realistic Groups and Com munities of Users (and not Only Realistic Individual Users) ", The Fourth International Conference on Advances in Computer-Human Interaction, 2011. pp. 32-35.

[20] J. P. Haton and M. C. Haton, "Knowledge Representation and Reasonings", in Artificial Intelligence, 15e ed. Paris, France: PUF, pp 39-67, 1989.

[21] cde-saare.de, [retrieved: 09, 2015].

[22] protege.stanford.edu/, [retrieved: 10, 2015].

[23] http://www.journaldunet.com/web-tech/chiffres-internet/cameroun/pays-cmr, [retrieved: 17, 2015].

Influence of Couple Patterns on Entropy in Multilayer Networks

Xiaoling Nian
School of Computing Science & Engineering
University of Electronic Science and Technology of China
Chengdu, China.
e-mail: xiaoling.nian@gmail.com

***Abstract*—Many real-world systems are networks coupled with other networks, and research on these multilayer networks about their structure properties and functions recently produced significant and remarkable findings. There is one type of multilayer network in which an individual has more than one counterpart on other layer network. For instance, important station on an infrastructure network usually has more than one supporter on its counterpart network for optional access or risk diversification. In this paper, we investigate the influence of couple patterns on information entropy and energy of two layer coupled networks with community structures. Couple patterns refer to the allocation of counterpart numbers according to nodes' degrees and the establishment of interconnections between nodes on different layer networks according to their degree of assortativity. Nodes' degrees and counterpart numbers are assorted based on the tendency of large degree nodes with more counterparts or the reverse. Additional, a pair of nodes on different layer networks can be interconnected according to their degree of assortativity. Under the scenario of a heterogeneous distribution of counterpart numbers, we have found that the influences of couple patterns on entropy and energy are negative. That is, entropy and energy of the two layer coupled network decrease when counterpart number assortativity and/or degree assortativity positively increase, while increase when the two assortativity are negatively enhanced. Moreover, networks with weak community structures extend these influences compared to networks with obvious community structures.**

Keywords-multilayer network; couple pattern; influence; entropy; energy.

I. INTRODUCTION

A large quantity of real-world complex systems are networks coupled with other networks, and recent studies on these complex systems are fruitful [1]-[13]. For instance, an infrastructure network[14][15], such as a power grid, is often coupled with water, gas, or other resource supply networks to turn various resources into power. Diseases spread in face-to-face networks, while the information of contagion spreads in the coupled online social networks[16]-[18]. The multilayer point of view can help find important layer network to improve the robustness of the whole bank networks. In[19], bank networks are composed of several layer networks which represent different types of exposure, such as unsecured overnight, unsecured short-term, unsecured long-term, secured short-term, and secured long-term. The authors found there is a high heterogeneity between these layer networks, and the unsecured overnight layer is the especially important layer network related to the stability of the overall network in a financial crisis. Researchers define the multidegrees, multilinks, cluster coefficients, and degree correlations to describe the properties of these multi-layer networks, and give deep insights into the dynamic processes of these coupled networks, such as percolation, spreading, and growth. In [14], authors investigated cascade failures on two layer coupled networks and modelled a true coupled system composed of a grid and internet. The effects of assortativity and cluster coefficient on the robustness of two layer coupled networks were clarified in [7]. These researches demonstrate that the dynamics of multi-layer networks relates directly to the topologies of these coupled networks. Among these cases, there is one kind of multilayer network in which nodes on one layer network have more than one counterpart on the other layer network. A node with multiple counterparts naturally brings advantage of optional access and adequate supply, and could disperse risk avoiding cascade of failures. The deep exploration of structures and functions of this kind of multilayer network appears in[20]. The author mathematically analysed percolation based on generating functions and concluded that multi-interconnections can significantly lower the percolation threshold. Besides several related studies, research work on this kind of multilayer network is still a little. More explorations should be carried out to look into the properties and dynamics of these complex systems. Spreading information across multilayer networks, however, always sparks interests of researchers in how to maximize information entropy [21]-[23]. Entropy in information theory characterizes uncertain sources of information. The larger the entropy is, the more random information sources are. In [24], we investigated how the overall entropy of a two layer coupled network varies based on the assumption that a group of nodes on one layer network have the same number of counterparts on the other layer network. In this paper, we explore the impacts of couple patterns on information entropy and energy in two layer coupled networks with community structures, especially in heterogeneous interconnection scenarios. Couple patterns refer to the allocation of counterpart numbers according to nodes' degrees and the establishment of interconnections between nodes on different layer networks according to their degree of assortativity. Counter number couple pattern couples nodes that have various numbers of counterparts. The counterpart number of a node is allocated according to node's degree. Nodes' degrees and counterpart numbers are assorted based on the tendency of large degree nodes with more counterparts or the reverse. Degree couple pattern interconnects a pair of nodes on different layer networks according to their degree of assortativity. In this work, the distribution of node's counterpart on the other network follows power law, and nodes interconnect with their counterparts according to the assortativity or disassortativity between them. We will find out how couple patterns affect entropy and energy when information flows on two layer coupled networks.

The outline of the paper is as follows. In Section II, we describe couple patterns in detail and establish the two layer interconnected network model. In Section III, we present the influence of couple patterns on entropy and energy. In Section IV, we give our conclusions.

II. The Model

First, we use [25]-[27] to construct two standalone networks and set both of them to have the same number of nodes, number of communities, community size, and modularity strength. Then, we choose one random community in each network and establish interconnected links between nodes in the community. Figure 1 shows the illustration of this two layer couple network. As we can see, nodes in a community of the top network can interconnect several counterparts of the lower network. Their counterparts may be totally different or may partly overlap.

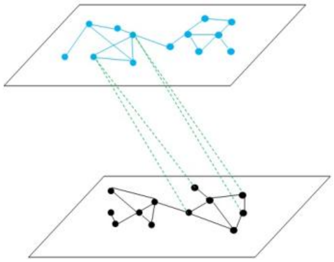


Figure 1. Illustration of a two layer coupled network model. Nodes in a community on one layer network interconnect with their counterparts on the other layer network.

Due to the fact that distributions of the wide variety of phenomenon in nature and man-made world follow power laws[28], a series of counterpart numbers that are distributed heterogeneously are generated. The tendency of assorting counterpart number with node's degree indicates couple strength between nodes of different layer networks. Positive assortativity means that large degree nodes have more counterparts than small degree nodes, while negative assortativity means that small degree nodes prefer to link more nodes on the other layer network. After the above two steps, we get a sequence of nodes' degrees and a sequence of nodes' counterpart numbers. Each node in the community will be allocated a certain number of counterparts from the counterpart number sequence according to the method proposed in [29]. In the simulation, we range assortativity from -0.05 to 0.1, which can significantly affect entropy.

On the other hand, the tendencies of interconnecting nodes in different layer networks are various. Large degree nodes may tend to couple with nodes similar to themselves, while the opposite is also possible. Two layer coupled networks are established after nodes of different layer networks interconnect with each other according to their degree assortativity. Couple patterns are made of counterpart number assortativity and/or degree assortativity. Then we use the method introduced in [24] to compute the entropy and energy expended in these two layer coupled networks in order to find out how they are affected by couple patterns. To describe our model completely, we retell this method. Nodes collect information most from their first neighbour nodes, second neighbour nodes, and their couple nodes on the other layer network. q_i is the frequency of node i spreading information, then we use

$$H(x) = -\sum_{i=1}^{n} q_i \log q_i \qquad (1)$$

to compute entropy [30]. In order to determine about how much energy is expended, we first interpret asymmetry and similarity between two nodes. Asymmetry is defined according to the covariance of degree between two nodes, given by

$$A(x_{i,}x_j) = \left|(K_{x_i} - \mu)(K_{x_j} - \mu)\right| \qquad (2)$$

where μ is the mean value of nodes' degrees of community x_i belongs to. Similarity describes common properties between a pair of nodes. In this paper, node similarity is quantified by the portion of common friends to total friends when two nodes are on the same network. If two nodes belong to different layer networks, then we regard their common friends as the portion of their mutual coupled neighbours to their total friends. For a pair of nodes that are asymmetrical and dissimilar to each other, it could be deduced that they need more energy in order to share information. Hence, we calculate energy by

$$E = \sum_{j \in N(i), Cou(i)} A_{i,j}(1 - S_{i,j}) \qquad (3)$$

$N(i)$ and $Cou(i)$ are the neighbour set and counterpart set of node i . $S_{i,j}$ is the similarity between node i and j , defined by

$$S_{i,j} = \frac{N(i) \cap N(j)}{N(i) \cup N(j)} \qquad (4)$$

We expect to find out how the allocation of counterparts and the establishment of interconnections among nodes affect information entropy and energy.

III. Results

We construct a two layer coupled network with the first couple pattern in which there is only counterpart number assortativity (the assortativity between node's degree and its counterpart number) and observe how entropy varies with this couple pattern. Because correlations of entropy and couple pattern have no significant difference under values of counterpart number assortativity larger than 0.5, in the simulation, the counterpart number assortativity ranged from 0.05 to 0.5. Each node interconnects its counterparts on the other layer network regardless of degree assortative mixing between the mutually connecting nodes. Pearson correlation coefficients of entropy and the counterpart number assortativity are calculated as influences of couple pattern. As shown in the top of Figure 2, when counterpart number assortativity is positive, the influence is negative, and the absolute value increases with the absolute value of counterpart number assortativity. When counterpart number assortativity is negative, there is the same tendency. Hence, the overall influence of this couple pattern on entropy is negative. That is to say, if more counterparts are allocated to nodes with larger degrees, then the entropy of whole system tends to decrease. Conversely, if nodes with small degrees have more counterparts, then the entropy of whole system tends to increase. We also notice that weak community structure on each layer network stretches this effect when compared to networks with obvious community structures. The range of correlation is from 0.21 to -0.14 when modularity is 0.2, which is wider than the range from 0.16 to -0.06 if modularity is 0.8.

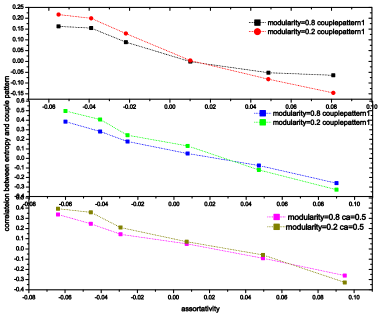


Figure 2. Pearson correlation coefficient of entropy and couple patterns. Degree assortativity between node pairs ranged from -0.08 to 0.1. Ca refers to counterpart number assortativity for short.

Next, we coupled two layer networks with a more complex couple pattern in which not only nodes' degrees and nodes' counterpart numbers are assorted but also nodes interconnect their counterparts by degree assortativity. The influence of this couple pattern on entropy when counterpart number assortativity is 0.05 is shown in the middle of Figure 2. Analogously, the influence is negative and entropy decreases when the degree assortativity between nodes is positive. Therefore, if nodes prefer to interconnect to nodes similar to themselves, then the entropy of two layer network system tends to reduce. The interconnection between similar nodes cannot produce large entropy, even if both of them have large degrees. However, the interconnection between dissimilar nodes can increase entropy. The influence also is stronger in networks with weak community structures than those with obvious community structures. This is similar to the couple pattern discussed previously. The correlations of entropy and the couple pattern cover broader ranges than those with the first couple pattern.

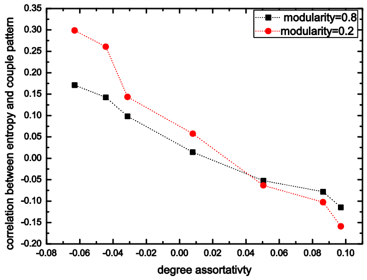


Figure 3. Pearson correlation coefficient of entropy and couple pattern as a function of degree assortativity. Network with weak community structure stretch the influence especially in disassortative degree mixing.

The influence of degree assortativity couple pattern on entropy when counterpart number assortativity is 0.5 is presented at the bottom of Figure 2. It matches our first result that counterpart assortativity has a negative effect on entropy. There is a positive correlation between a node's degree and its counterpart number, which lowers entropy; however, information entropy of a two layer coupled network is remarkably augmented as the result of adding degree assortative mixing between nodes into the couple pattern. Consequently, we can conclude that degree assortative mixing among interconnecting nodes is the dominant couple pattern affecting the entropy of the system. Therefore, we wonder what will happen if we extract the degree assortative mixing effect and allocate counterparts equally among nodes. It is straightforward that each node has only one counterpart on the other layer network. However, a node chooses its counterpart according to degree assortativity. In Figure 3, the influence holds the same tendency. Degree assortative mixing has a significant influence on entropy, and networks with weak community structures gain more entropy when nodes are disassortatively mixed.

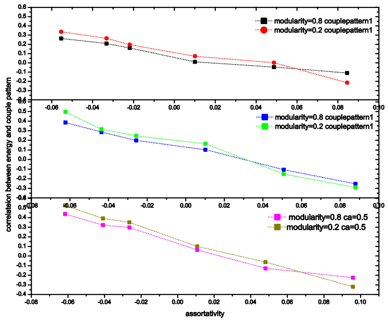


Figure 4. Pearson correlation coefficient of energy and couple pattern. The couple patterns are the same as Figure 2. The tendency is not the same but similar to that of entropy.

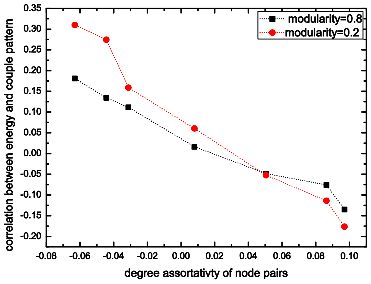


Figure 5. Pearson correlation coefficient of energy and couple pattern as a function of degree assortativity. The couple pattern is the same as Figure 3. The tendency is similar to that of entropy.

Consuming the least energy to gain information is often expected. Figure 4 shows energy expended under the above two couple patterns. It follows a similar tendency as entropy. In Figure 5, influences of couple pattern on energy are presented when each node has only one counterpart but there are degree assortative mixing between a pair of nodes. Compared to entropy, the Pearson correlation coefficients of energy and couple patterns are larger. Therefore, if we want to obtain much more diverse information, then the energy we consume is greatly affected by these couple patterns.

IV. CONCLUSION AND FUTURE WORK

In this work, we established two layer coupled network models and investigated the influences of couple patterns on entropy and energy from the aspect of counterpart number assortative mixing and degree assortative mixing. Under counterpart numbers of nodes are heterogeneously distributed scenario, a node’s degree is assorted with its counterpart number in order to allocate counterparts according to node’s degree. And nodes on one layer network interconnect their counterparts on the other layer networks according to degree assortative mixing among them. Couple patterns are made of one or both of the two assortative mixing couplings. We found that the influences of these couple patterns on entropy and energy are negative. Entropy and energy of the two layer coupled network decrease when counterpart number assortativity or/and degree assortativity positively increase, while increase when they are negatively enhanced. Furthermore, weak community structures stretch the influences that nodes have on networks that obtain more diverse information. Compared to the counterpart assortativity couple pattern, the degree assortativity couple pattern exerts more significant influences. We verified this phenomenon through extracting the effect of degree assortative mixing and allocating one counterpart to each node. Therefore, the degree assortativity couple pattern is the dominant pattern that affects the entropy and energy of the system greatly. Specifically, under degree assortativity couple pattern, networks with weak community structures gain much more entropy than those with obvious community structures if nodes are disassortatively mixed. Future deeper endeavours on this research should analysis available large data sets to explore more interesting new findings and propose a mathematical framework with which to theoretically support and predict the tendency of influence of corresponding couple patterns. Assortative or disassortative mixing in other nodes’ properties needs to pay more attention. And particular assortative or disassortative mixing hiding only in special local structures is waiting to be discovered.

REFERENCES

[1] S. Boccaletti, et al. “The structure and dynamics of multilayer networks,” Physics Reports, vol. 544.1, 2014, pp. 1-122.

[2] M. Kivelä, A. Arenas, M. Barthélemy, J. P. Gleeson, Y. Moreno and M. A. Porter, “Multilayer networks,” Journal of Complex Networks, vol. 2.3, pp. 203-271.

[3] R. Gallotti and M. Barthelemy, “Anatomy and efficiency of urban multimodal mobility,” Scientific Reports, 2015, 4, pp. 6911.

[4] O. Yagan, D. Qian, J. Zhang and D. Cochran, “Conjoining Speeds up Information Diffusion in Overlaying Social-Physical Networks,” Selected Areas in Communications, 2013 vol. 31. 6, pp. 1038 - 1048

[5] L. Lotero, A. Cardillo, R. Hurtado and J. Gómez-Gardeñes, “Several Multiplexes in the same City: The role of wealth differences in urban mobility,” arXiv preprint arXiv: 1408.2484, 2014.

[6] M. De Domenico , M. A. Porter and A. Arenas. “Multilayer Analysis and Visualization of Networks,” arXiv preprint arXiv: 1405.0843, 2014.

[7] D. Zhou, G. D'Agostino, A. Scala and H. E. Stanley, “Assortativity decreases the robustness of interdependent networks,” Physical Review E, 86.6, 2012.

[8] E. Ernesto and J. Gómez-Gardeñes, “Communicability reveals a transition to coordinated behavior in multiplex networks,” Physical Review E, 89.4, 2014.

[9] M. Magnani and L. Rossi, “Formation of Multiple Networks," Social Computing, 2013, vol. 7812, pp. 257-264.

[10] F. Battiston,V. Nicosia and V. Latora, “Structural measures for multiplex networks,” Physical Review. E, 2014, vol. 89.3:032804.

[11] V. Nicosia, P. S. Skardal, V. Latora and A. Arenas, “Spontaneous synchronization driven by energy transport in interconnected networks,” arXiv preprint arXiv:1405.5855, 2014.

[12] S. Grauwin, S. Sobolevsky, S. Moritz, I. Gódor and C. Ratti, “Towards a comparative science of cities-using mobile traffic records in New York, London and Hong Kong,” Springer International Publishing, 2015, pp. 363-387.

[13] M. De Domenico, A. Lancichinetti, A. Arenas and M. Rosvall, “Identifying modular flows on multilayer networks reveals highly overlapping organization in social systems,” Physical Review X, 2015, Vol. 5. 1, 011027.

[14] S. Havlin, et al. “Catastrophic cascade of failures in interdependent networks,” Nature, vol. 464.7291, 2010, pp. 1025-1028.

[15] M. Chertkov, V. Lebedev and S. Backhaus, “Cascading of Fluctuations in Interdependent Energy Infrastructures: Gas-Grid Coupling,” arXiv preprint arXiv: 1411.2111, 2014.

[16] G. Clara, S. Gómez and A. Arenas. “Competing spreading processes on multiplex networks: awareness and epidemics,” Physical Review. E, 2014.

[17] B. Franco and E. Massaro. “Epidemic spreading and risk perception in multiplex networks: a self-organized percolation method,” Physical Review E, 2014.

[18] L. G. A. Zuzek, H. E. Stanley and L. A. Braunstein, “Epidemic model with isolation in multilayer networks,” arXiv preprint arXiv: 1412.1430, 2014.

[19] L. Bargigli, G. Di Iasio, L. Infante, F. Lillo and F. Pierobon, “The multiplex structure of interbank networks,” Quantitative Finance, vol. 15.4, April 2015, pp. 673-691.

[20] E. A. Leicht and R. M. D'Souza, "Percolation on interacting networks," arXiv preprint arXiv:0907.0894, 2009.

[21] B. Ginestra, “Statistical mechanics of multiplex networks: Entropy and overlap,” Physical Review E, vol. 87.6, 2013: 062806.

[22] M. De Domenico, et al. “Mathematical Formulation of Multilayer Networks,” Physical Review X, vol. 3.4, 2013: 041022

[23] G. Menichetti, D. Remondini, P. Panzarasa, R. J. Mondragón and G. Bianconi, “Weighted Multiplex Networks,” PLoS ONE, 2014.

[24] X. Nian and H. Fu, “Maximization of entropy in a two layer asymmetry-coupled network,” Computational Aspects of Social Networks (CASoN), 2014.

[25] M. Molloy and B. Reed, “A critical point for random graphs with a given degree sequence,” Random structures & algorithms, vol. 6.2‐3, 1995, pp. 161-180.

[26] R. Milo, N. Kashtan, S. Itzkovitz, M. E. J. Newman and U. Alon, “On the uniform generation of random graphs with prescribed degree sequences,” arXiv preprint cond-mat/0312028, 2003.

[27] B. Karrer and M. E. J. Newman, “Stochastic blockmodels and community structure in networks,” Physical Review E, vol. 83.1, 2011: 016107.

[28] A. L. Barabási, and R. Albert, “Emergence of scaling in random networks,” Science, vol. 286.5439, 1999, pp. 509-512.

[29] M. E. J. Newman, “Assortative mixing in networks,” Physical review letters, vol. 89.20, 2002: 208701.

[30] T. M. Cover and J. A. Thomas, “Elements of information theory,” John Wiley & Sons, 2012.

Social Network Analysis Tools for Career Advancement

Timothy Arndt
Department of Electrical Engineering and Computer Science
Cleveland State University
Cleveland, OH 44115, U.S.A.
t.arndt@csuohio.edu

***Abstract*—Social networks, which have been studied by sociologists for many decades, have risen greatly in visibility due to the widespread growth of online social networks like Facebook, which facilitate the analysis of social network structures by special purpose software tools. Sociologists have long recognized that special advantage can accrue to individuals who occupy certain strategic locations in their social networks, or whose local neighborhood in their social network exhibits certain characteristics. In this paper describing our initial research directions in this area, we propose to develop social network analysis tools which will allow individuals to analyze their social networks with the express purpose of cultivating or pruning social ties which will enhance their career advancement prospects within an organization. This paper presents our work in progress in this area.**

Keywords-social networks; software tools; social simulation; legal aspects.

I. Introduction

Social networks have been the object of study by sociologists for many decades. In recent years, with the availability of software for analysis of the social network structures, as well as with the growth of online social networks like Facebook, LinkedIn, and Google+, the computer science community has become interested in studying, analyzing and categorizing social networks of various types, both online information networks as well as the more traditional informal social networks which sociologists have studied [1].

Social networks can be modelled by a graph structure where the nodes represent individuals and the links (ties, in social network terminology) represent some relationship between individuals. An illustration of part of an example social network is shown in figure 1. The social network has eight individuals (A through H) and 13 ties among those individuals.

Depending on the amount or degree of interactions between two individuals, we may characterize a tie between them as either a strong tie (much interaction) or a weak tie (little interaction) [1]. One famous result in the field of social network analysis is that reported by Granovetter [2], who found, surprisingly, that most persons reported that they found a new job not through persons with whom they had strong ties, but rather among those with whom they had weak ties. The reason for such a result is that those colleagues with which we have strong ties are likely to have the same knowledge which we have, so new opportunities, such as job openings are likely to come from those individuals with which we have only weak ties.

For example, in figure 1, maybe all of the ties are strong ties except for that between C and E, which is a weak tie. Individuals A through D might be members of one organization while E through H might be members of a different organization. C might learn of an opening in E's department through his (weak) tie with E. Furthermore, the theory of triadic closure suggests that if E had a strong tie to C, then ties would develop between E and A, B, and C (either weak or strong) as well.

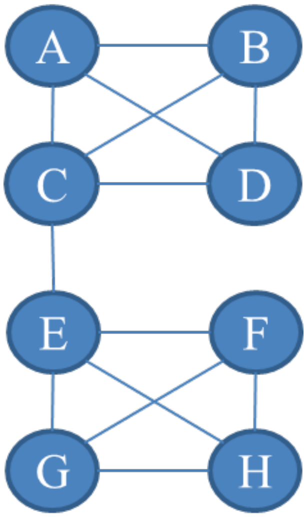


Figure 1 – Part of a social network

The theory of social networks suggest that social capital exists when people have an advantage over their rivals because of their position in some social network. Normally, social structures have a dense structure of strong ties among the participants in the network. This dense structure of strong ties tends to engender trust among the participants, and thus is usually something to be cultivated. On the other hand, in large scale networks, not all of the individuals are members of the same cluster. There may thus exist multiple clusters. When two clusters contain non-redundant information . there is said to be a structural hole between them [6]. A node which is the only (local) connection between two such clusters as said to be a (local) bridge. The person who is the bridge between two clusters gains much social capital, since he can act as a broker for the flow of information between two clusters.

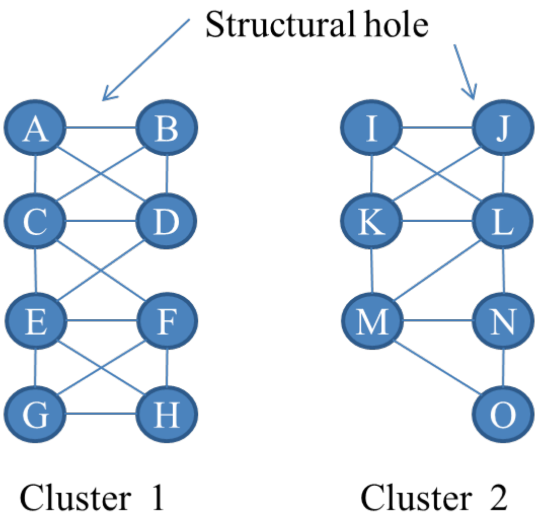


Figure 2 – A structural hole

It has long been recognized that one's position in a social network can play a very important part in one's career success [3]. However, on the part of the participants themselves (as opposed to those who study them), this understanding is more or less inchoate. As these concepts enter more and more into the mainstream due to the success of online social networking platforms such as Facebook and LinkedIn, we expect that individuals will become more aware of the role such networks and their place within the network can play in their career success. Further, we expect that they will purposefully cultivate those ties which will benefit them, and attempt to suppress ties among others which would hurt them. Since cultivating ties which might be advantageous is time consuming, any one individual can only cultivate a limited number of ties, thus part of the active process of improving his social capital might be the downgrading (strong tie to weak) or jettisoning those ties which are less advantageous. The ultimate goal of the research described in this paper (preliminary work) is to develop software tools to support this purposeful manipulation of one's place in social networks in order to support career advancement.

The rest of this paper is structured as follows. Section 2 sketches our main research goals in the area of social network analysis and section 3 presents future research and conclusions.

II. Purposeful Manipulation of Social Networks

Much research has been done on the analysis of social network structures as they exist at a moment in time or as they evolve over time. Less study has been done, however, on the purposeful cultivation and manipulation of the social network structure of an individual in order to enhance the career opportunities of that individual and/or maximize his opportunities in an organization.

Podolny and Baron [3] examined how the structure and content of individuals' networks in the workplace affect intraorganizational mobility. They found that an individual's mobility is enhanced by having a large, sparse network of informal ties for acquiring information and resources and that well-defined performance expectations are more likely to arise from a small, dense network of individuals. For the purposes of the present research, the first of these findings is most important. We would like to empower individuals to develop and cultivate such large, sparse networks of informal ties.

With the rise of online social networks, software tools to manipulate those networks have been developed. For example, Vizster [9] is a tool for end-user visualization, navigation, and exploration of large-scale online social networks. It builds upon familiar node-link network layouts to contribute techniques for exploring connectivity in large graph structures, supporting visual search and analysis, and automatically identifying and visualizing community structures. In addition many GUI (graphical user interface) based software packages for online social network analysis have been developed such as Pajek [10] and UCINet [11].

In order to describe the social network being analyzed, it is necessary to classify the ties in the network as being either weak or strong. That this can be done in an automated way is demonstrated by [5] in which a predictive model was used to map social media data to tie strength. The model was exercised on a data set of two thousand social media ties from Facebook and achieved 85% accuracy in classifying the ties as either weak or strong. Another work on Facebook data is described in [4] in which distinctions are drawn between the different types of interactions users have on Facebook. One to one communications (receiving messages from friends) is associated with bridging social capital, but other uses are not. However, even using the site passively to consume news assists those with lower social fluency draw value from their connections.

In order to meet our requirements, the software we are developing should incorporate the functionality of the above-mentioned packages such as GUI and extensive visualization, navigation and exploration of large-scale online social networks, but it should do more to enable purposeful manipulation of an individual's social network for career advancement. First, it should attempt to construct the user's social network from multiple sources, both online social networks such as Facebook, LinkedIn and Google+, as well as informal social networks using information gathered from such sources as email logs, phone call records, organization charts, etc. Notice that except in exceptional circumstances, the software will not have complete access to data such as email or phone records, due to privacy and legal requirements. For these sources, it is more likely that the software will have access only to the Ego network (or a portion thereof) of the user – that is, that part of the social network consisting of the user (the "ego"), those nodes to whom the ego is directly connected (the “alters”) plus the ties, if any, among the alters. See figure 3 for an example of an ego network. Incorporating these multiple, possibly contradictory, likely incomplete networks into a complete network representing the individual’s context in an organization is a major part of the research.

It will also be a part of the required functionality of the software we are developing to provide hints for the user, beyond just allowing for freeform navigation and exploration of an online social network. So, for example, after having integrated the various social networks from the sources available, the software should analyze both the social network as a whole (representing the individual’s organization) as well as the user’s place in the network. It should then point out, for example, structural holes in the network, offering hints on how the user can fill that structural hole by cultivating (weak) ties with other employees. Since there are likely to be many such possibilities in a large organization, the opportunities will have to be prioritized by the system, possibly involving user interaction to clarify information such as the roles that the user and others play in an organization. A plan for cultivating ties might be generated.

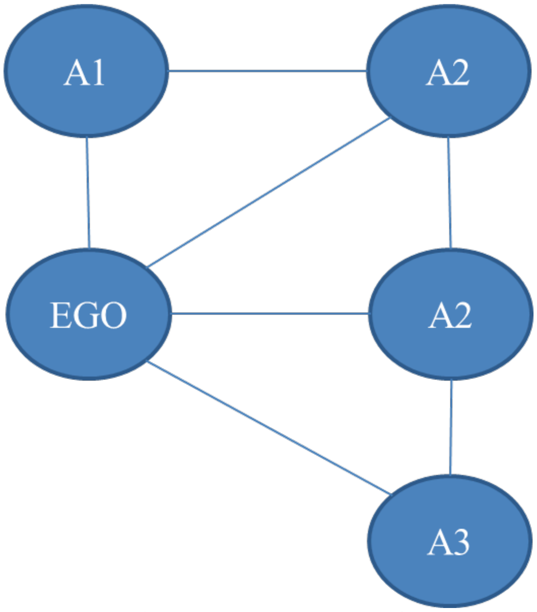


Figure 3 – An ego network

The overall data flow of the proposed system is shown in Figure 4. In the following paragraphs we give a very hiogh level, conceptual description of the system. The algorithms and data structures which will be used in the actual implementation of the system are actively under investigation and will be incorporated as the research proceeds. The input to the system consists of all of those pieces of information which will be used to construct the user’s social network in the context of the organization in which he works. This will include online social networks such as Facebook, LinkedIn and Google+ (probably delimited to those contacts who are relevant for the career advancement objective), ego networks for the user which might be constructed by scraping data from the user’s contact list, email client, phone records, etc., and other types of documents such as organization charts which can help to fill in those parts of the social network which the other two collections of information might miss. It is expected that this information will be incomplete, overlapping and contradictory. The first module of the system attempts to disambiguate and flesh out the network as much as possible automatically. The output of the first module is a preliminary aggregate social network for the user (preliminary because it is not expected that the fully automatic system will be able complete the network without manual intervention).

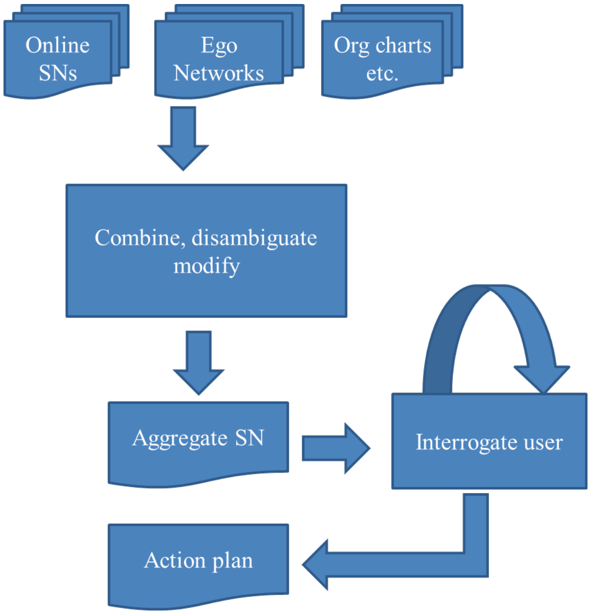


Figure 4 – High-level system data flow

The next module of the system will carry out a number of rounds of interrogation of the user. It will ask the user to correct any errors in the preliminary social network, performing such tasks as aggregating nodes which represent the same actor, if such has not been correctly performed by the automated portion of the system. The user will also be asked to fill out any missing information – such as roles for individuals which can help with the network classification algorithm and prioritization of equally useful ties in the social network. The user may also enter his career goals and any other information which may help the system produce the desired output. This process is an interactive one, and proceeds in a loop until a desired level of refinement is reached.

The final output of the system is an action plan (or a set of action plans) which can be employed by the user to enhance his social capital (for the advancement of his career), for example by filling structural holes, cultivating a denser network of strong ties in a department in order to increase trust, and hence productivity, in the department, etc. Note that the action plan (or plans) which are the final result of the process will normally be acceptable to the user, since if they weren't, the previous stage would just go through one or more additional rounds of refinement.

III. Conclusions and Future Research

This paper has introduced my current work in progress research in the area of social network analysis for career advancement. I am currently at a very early stage in this research. The theme of the future research has been chosen, and currently available social network analysis software has been chosen, installed and tested for suitability. Sample data sets have been installed and tested.

The immediate next stage of the research will be to identify a set of social network data that can be used for an empirical study. Possibilities include Facebook or LinkedIn data sets, email data for an organization, cell phone trace data, etc. The use of at least some of the data sets will raise privacy and legal issues, so these will need to be studied as well. In the extreme case, simulated data may need to be used.

The overall structure of the system has been constructed, as outlined in the previous section, however specific algorithms and data structures need to be chosen. An algorithm for integrating diverse social networks, some of which may be incomplete and which may contain contradictory information, will be an interesting topic of future research. We are currently investigating the possibility of using an ontology-based approach in the part of the system.

An interesting question which might be a topic of future research is how the widespread adaption of tools such as those proposed in this paper would affect the social networks of organizations in the long term. Several researchers have looked at similar questions for the evolution of social networks. Burt, Merluzzi, and Burrows [7] analyzed network volatility as something akin to the hum of a running engine. People active in a network produce vibration and wiggle where the connections and the network structure around these people changes frequently. They distinguish four dimensions to network volatility (churn, variation, trend, and reversals), measure them with panel data on a population of bankers, and then add them to analysis predicting compensation from status and structural-hole measures of network advantage. They find that volatility creates a slope adjustment that enhances the returns to network advantage. They identify two stability traps that destroy advantage, but the key is not to avoid the traps so much as to avoid them in a particular way. The volatility that enhances is reversal. Bankers who go through reversals were shown to enjoy significantly higher returns to their network advantage.

Even more pertinent for our research, Buskens and van de Rijt [8] examine the question of whether those who strive to fill structural holes can gain and maintain an advantage over time. Burt's informal treatment and economic models of information network evolution suggest as equilibrium network the star, in which a single broker acquires all of the access and control benefits. The work of the authors, on the other hand, shows that if everyone is seeking the same type of advantage, adding beneficial links and removing costly ones, the predominant equilibrium turns out to be the "balanced complete bipartite network." Paradoxically, this network,– in stark contrast to the star – distributes benefits evenly, so no one has a structural advantage. So, if the tools we propose were adopted universally, they would not result in an advantage for anyone! Luckily, such a prospect is far in the future (if it is ever achieved).

Acknowledgements

Timothy Arndt's research was supported by a Dean's Faculty Development Grant from the Monte Ahuja College of Business of Cleveland State University. Cleveland State University MCIS student Pranjal Shah helped with the installation and testing of the social networking software.

References

[1] D. Easley and J. Kleinberg, Networks, Crowds, and Markets: Reasoning about a Highly Connected World, New York: Cambridge University Press, 2010.

[2] M. Granovetter, Getting a Job: A Study of Contacts and Careers, University of Chicago Press, 1974.

[3] J. Podolny and J. Baron, "Resources and Relationships: Social Networks and Mobility in the Workplace," American Sociological Review, vol. 62, no. 5, July 1996.

[4] M. Burke, R. Kraut, and C. Marlow, "Social capital on facebook: differentiating uses and users," Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '11), ACM, 2011, pp. 571-580. doi:10.1145/1978942.1979023

[5] E. Gilbert and K. Karahalios, "Predicting tie strength with social media," Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '09), ACM, 2009, pp. 211-220.

[6] R. Burt, Structural Holes: The Social Structure of Competition, Harvard University Press, 1992.

[7] R. Burt, J. Merluzzi, and J. Burrows, "Path dependent network advantage," Proceedings of the 2013 conference on Computer supported cooperative work (CSCW '13), ACM, 2013, pp. 1-2. DOI=10.1145/2441776.2441778

[8] V. Buskens and A. van de Rijt, "Dynamics of Networks if Everyone Strives for Structural Holes,"American Journal of Sociology, Vol. 114, No. 2 (September 2008), pp. 371-407

[9] J. Heer, and D. Boyd, "Vizster: visualizing online social networks," Proceedings of IEEE Symposium on Information Visualization, 2005. INFOVIS 2005, pp. 32-39, 2005.

[10] W. De Nooy, A. Mrvar, and V. Batagelj, Exploratory social network analysis with Pajek. Vol. 27. Cambridge University Press, 2011.

[11] S. Borgatti, Ucinet for Windows: Software for Social Network Analysis. Harvard, MA: Analytic Technologies, 2002.

Mapping Competences of Distance Education Students

Ketia Kellen Araújo da Silva, Patricia Alejandra Behar
Universidade Federal do Rio Grande do Sul
UFRGS
Porto Alegre, Brasil
e-mail: ketiakellen@gmail.com, pbhear@terra.com

***Abstract*— This article aims to present a mapping of the competences needed by students of Distance Education. The study establishes a relationship between competences and students of distance education, highlighting knowledge, competences and attitudes linked to this mode of learning. The qualitative-quantitative approach employed presents an explorative and unique case study. As part of the study, a Learning Object was developed and subsequently validated through an extension course. The course was one of the strategies used to map competences and was complimented by interviews and questionnaires. Analysis consisted of an evaluation of the data and a description of the competences. The final competence map was organised using a twelve-part diagram presenting the following competences: digital fluency, autonomy, organization, planning, time management, communication, reflection, virtual presence, self-evaluation, self-motivation, flexibility, and teamwork.**

Keywords - Distance Education; Competences; Distance Education students

I. Introduction

The advance that Distance Education has made in the Brazilian context in recent years is undeniable. One of the key factors in this phenomenon is the development of technology, in particular Information and Communication Technology (ICT). The introduction of ICT in education has meant a reduction in distances, which in turn has favoured Distance Education and thus new opportunities for creating, storing and transmitting data.

Technology has generated a profound social change whereby, increasingly, the generation born and living in this new age of technology, develops new ways of acting, thinking, learning and being [1][2]. All of these transformations have had a huge impact on education, modifying learning environments and the resources used for teaching and molding the profile of the students themselves. These peculiarities of Distance Education require those being taught to demonstrate unique knowledge, abilities and attitudes, which can be identified as specific competences. Thus, both the teacher and student of Distance Education must possess knowledge of technology and its potential [3]. In the Brazilian education system, Distance Education was regulated in law in 1996 through the so called *Lei de Diretrizes e Bases da Educação National* (LDB) nº 9.394 [4]. In 2005, with the updating of the LDB, Distance Education was defined as: “[...] a type of education whose pedagogical didactic is characterized by the use of information and communication technology where students are found in diverse places or times.”

In 2007, the Brazilian Education Ministry (MEC) released guiding principles of the process of Distance Education in tertiary education systems, with the intention of informing the evaluation, organization, supervision and wider role of this mode of education. With this in mind, Distance Education can be seen as having significant potential to democratize and elevate the standard and quality of Education in Brazil.

Students, the focus of this study, possess characteristics in transformation, most apparent in the way they behave and act through technology [5]. There exists a broad spectrum of individuals who have had different experiences and relations with ICT, while not all Distance Education students will have a natural affinity with the use of technological resources [6]. For students to drive their learning, it is necessary to understand what competences, knowledge and attitudes are needed to undertake study in Distance Education, as opposed to in traditional classroom learning. Paloff and Pratt, state that “[...] with the freedom and flexibility of the on-line environment comes responsibility” [7]. In other words, the student becomes responsible for his individual study plan as Distance Education course planning is more open and flexible regarding the distribution of reading hours, assignments and other activities. In classroom learning, these considerations are strongly guided by the classes attended and by direct engagement with the teacher.

It is in this context that discussion surrounding the basic skills essential for students in Distance Education is set. Sections II and III of this article contain a brief bibliographic revision, focusing on themes relating to competences, their elements and students of distance education. Following this, the research itself is presented in Section IV, while Section V covers the conclusions of the study.

II. Competences

The term competence initially originated in the business world, where it denotes someone capable of completing certain activities efficiently. Etymologically, competence shares its Latin roots with competition, both words coming from the Indo-European expression pot (pet), meaning to put oneself against and compete, to find oneself at the same point, to be adequate, to bring together circumstances. There is, therefore, a relation to the professional world in which there is the need for competitiveness. As such, people are not resources, which the organization consumes, uses and which

ISBN: 978-1-61208-443-5

produce costs, but rather they constitute a competitive factor in the same way as do the market and technology [8].

Being professional, competitive and competent is linked to those most capable and efficient - those who achieve the most in professional terms. Competence is knowing how to act responsibly and to be recognized by others. It implies knowing how to mobilize, integrate and transfer knowledge, resources and skills in a given professional context [9]. For Fleury and Fleury it is an underlying characteristic of someone who is related to a higher performance in the realization of a task or in a specific situation. The job market is based on a process of preparation for operational roles and functions. Thus, those that have greater formal education achieve management positions or higher hierarchical levels [10].

From an educational perspective Gaspar [11] states that the concept of competence arises from studies carried out in Canada, Switzerland and Belgium at the start of the 1990s, where actions go beyond knowledge, aptitude or ability. Competence is understood as the mobilization of these resources depending on personal experience, one's psychological, cognitive and affective background and the situation in which the individual is placed.

Even today, there exists a great deal of uncertainty about the nature of competences and how they are being applied in education, as at times they have different meanings and show contradictions. For Perrenoud [12] competence is the aptitude to effectively confront a range of situations, mobilizing the conscience more and more quickly and creatively. For Zabala and Arnau [13] present the concept of competence that is adopted in this study as being the capacity or ability to carry out tasks or act in a range of situations in an effective way in a given context. It is necessary to mobilize attitudes, abilities and knowledge at the same time and in an interrelated way.

In Brazil, the concept focused on education was incorporated in 1996 through the *Lei de Diretrizes e Bases (LDB)* nº 9.394 [4], which states that the curriculum of secondary schools should be aimed at the development of competences for citizenship.

The CNE/CEB 16/99 report, which deals with curricular directives for professional education, presents curricular reform in a professional way. Here, the concept of competence is understood as "[...] the capacity to articulate, mobilize and put into action values, knowledge and skills necessary for the efficient and effective development of the required activities in a given working environment." [14].

In 2001, the CNE/CP 9/2001 report surrounding teacher training, has the development of competences as its central focus. The report states that: "It is not enough for a professional to possess knowledge about their job. It is fundamental that they know how to summon this knowledge and transform it into action." [15].

The Brazilian national secondary school examination known as ENEM (*Exame Nacional do Ensino Médio*), administered by the Brazilian Ministry of Education (MEC) since 1998 is applied with the aim of evaluating the performance of students. This evaluation is based on five competences: 1 - Mastering languages; 2 - Understanding phenomena; 3 - Problem solving; 4 - Constructing arguments; 5 - Elaborating proposals.

There is an exaggerated use of the term competence, which can lead to incorrect use and confusion. It is therefore necessary to be able to truly understand the changes and new perspectives brought by the concept of competence in education.

As such, with the aim of contextualizing and using the term coherently in education, one must understand the context of the competence. This is particularly salient given that, naturally, the great challenge in education remains the association of competences with the final performance of the student. In education, one must consider the entire process of development and mobilization of competences and not only the result.

With these different definitions of competences in mind, it is possible to see commonalities such as:

- The behaviour of students in new and complex situations;
- The mobilization of resources, depending on the willingness or not of the student to solve problems, i.e., with specific attitudes and intention;
- The command of the procedures in the action being carried out;
- The action must be inter-related as it depends on the gathering of resources or command of the student, not only in terms of knowledge but also in experience and attitude etc.;
- The resources are, therefore, composed of three fundamental elements: knowledge, ability and attitudes.

The analysis of these points forms the concept of competence, as it is necessary to understand the composing elements: knowledge, abilities and attitudes. It is thus not sufficient to merely understand what a competence is but also to understand the wider process, beginning with its elements as described below.

III. Elements

The elements of competences correspond to the joining of the elements that an individual has at their disposal. "[...] competences presuppose the existence of resources [...]. No resource belongs exclusively to a competence in that it can be mobilized by others." [16]

The majority of our resources can be used or reused in different contexts, i.e., they are at the service of a range of different intentions. The dynamic character of competences relates to the elements, which modify or transform them according to socio-cultural changes. Thus, it is important to understand each element and their characteristics.

A. Knowledge

Knowledge is constructed through contact with one's environment and is not synonymous with information or wisdom. This study understands knowledge in terms of Piaget's constructivist vision [17]: "[...] the essential point of our theory is that knowledge is the result of interactions between the subject and the object, which is richer than what objects can supply by themselves." This perspective holds that the construction of knowledge is attained through the interaction between the student (subject) and the environment (object) and its structures. As such, the acquisition of knowledge depends on both the cognitive structures of the student and their relation with the object. The construction of the subject's knowledge of the object is therefore achieved through construction and reconstruction in constant spiral movement.

B. Ability

Ability is the element of competence which demonstrates what the student knows and can learn. It is related to the application of knowledge, can be constructed through practice and can suffer alterations according to the socio-cultural and cognitive context of the student.

The concept of ability also has different perspectives. Perreound [16] states that "[...] in general, an ability is not as broad as a competence, that is why it is understood by many authors to be one of the elements of competence." Thus different abilities form one or more competences, i.e., they are used in different situations. In this way, abilities are both those which present mental/cognitive processes and those which present motor and technical processes. Indeed, different abilities form one or more competences in that they are used in different situations [16].

C. Attitude

Attitudes determine how individuals position themselves in relation to others and to wider circumstances. They also serve to evaluate feelings, behaviour and choices. A number of studies have exhaustively demonstrated that attitudes are behind behaviour. An attitude is a state of readiness organised by experience which exercises a guiding and dynamic influence over the responses of an individual *vis-a-vis* given objects or situations. Thus, attitude can be understood as the motivation behind an action.

Following an understanding of the link between competences and their elements, it is necessary to understand the characteristics of distance learning students.

IV. The Distance Learning Student

The range of changes generated by technology have had a significant impact on education, modifying learning spaces, educational environments and teaching resources, as well as the profile of the student arriving at school. As such, distance education has also been reorganized to include technology and has redefined its structure.

The profile of the student has also changed, given the nature of the moment of transition and the fact that not all are born and raised in contact with technology. Pozo and Monereo [6] call this phenomenon the *digital divide*, as, just as there exist young people who remain distant from ICT, there also exist those in the older generation who have had close contact with the latest technology from the outset and whose current modes of working, communicating and thinking are guided by computer systems. Thus, there is a great diversity of students with different profiles, tastes, knowledge, background and ideas [6].

According to the Quality References for Higher Distance Education (*Referenciais de Qualidade para Educação Superior a Distância*) [18], the student is the centre of the educational process. When distance education first emerged, all attention was focused on the teacher and technological resources while the student was marginalized. Today the student is regarded as the centre and focus of on-line learning [19].

The course, materials and structure are created with the virtual student in mind. Paloff and Pratt [7] present the following as necessary resources for students: connection with technology; training and support on the use of course technology; access to services, such as those found on the university campus; support services and feedback and evaluations.

The virtual student cannot always adapt to this mode of education. In addition to personal issues, experience in using technology can also influence and present difficulties for students. [3]

Prensky [1] describes the relationship that people have with technology by characterizing two types of people: Digital Natives and Digital Migrants. These terms distinguish those who came into contact late with digital technology, migrating from technology based on conventional texts, from those that have been raised with these technologies as their "natural" environment of development. The profile of distance learning students dealt with in this study is largely that of the digital migrant, previously used to printed linear and statistical texts.

In Brazil, data from the 2008 Higher Education Survey published by the Brazilian National Institute of Studies and Research (Inep) show that there were 115 institutions with 647 undergraduate distance courses, a total of 727.961 students enrolled and 70.068 graduates of these courses.

Research released in 2010 by the Brazilian internet regulator CGI (*Comitê Gestor da Internet no Brasil*) shows that more than seven million Brazilians have already taken distance courses using the internet. This study excludes those who have completed courses using other technological means such as video and radio.

This suggests that the number of distance education students and distance courses has risen rapidly. Thus, it is important to attract students, present the innovations and opportunities of distance education and identify the necessary skills in this process. In this way, it will be possible to anticipate and reduce future conflicts and problem situations, increasing motivation – so essential in the process of teaching and learning. The focus of this study is on adult learners undergoing post-graduate study in a distance learning format.

The following section discusses the methodology used in this study to carry out the mapping process.

V. Delineating The Mapping

This research was carried out over a two year period in which time, during the first year, a Learning Object was constructed. This Learning Object was called CompMap –

Competence Mapping of Distance Education students (*Mapeamento de Competências dos alunos da EAD*). The main function of the object was to act as a digital resource with a content specially developed for the mapping of competences and with a focus on Distance Education students. The content and activities of the Learning Object were developed with a focus on the mapping of competences and the Object was employed with post-graduate students in an extension course; the intention being to map the competences of these students.

During the second year, the process of data collection was completed using Questionnaires, an Extension Course and Interviews. The Questionnaire was given to distance learning students, teachers and tutors and was developed in an on-line form using Google Docs. Through this source, it was possible to obtain information on the profile of the distance education students and their skills, tying in with relevant studies and literature. In total, 17 tutors, 2 teachers and 7 students who participated in the extension course responded to the questionnaire The extension course was given to post-graduate distance learning students who were doing teacher training. The theme of the course was distance learning competences and consisted of 40 hours, over 7 weeks with 3 traditional face-to-face classes (at the start, middle and end of the course) and 4 distance classes. The extension course used the UFRGS AVA Moodle tool as well as the OA CompMap. During this time, participating students completed activities aimed at mapping their competences. In order to obtain the final mapping, comparisons were made between competences noted in relevant academic literature, the answers given by teachers and tutors and the result of the mapping carried out during the course. Finally at the end of the course, some of the participating tutors and students were interviewed. The interviews were performed with students who participated in the extension course, and with tutors who did not answer the initial questionnaire. A total of five tutors and three students were interviewed for more than 30 minutes each and the interviews were transcribed.

VI. COMPETENCE MAPPING

Mapping was carried out using an analysis of two categories created from the data collected. The categories were: 1. Distance Education students and 2. Competences of Distance Education students. Mapping in the former category was noteworthy in demonstrating issues with technology, time management and communication. Moreover, adults enrolled in this class work 40 to 60 hours a week. It was noted that, just as students in the classroom require competences which allow them to act like students, distance education students also require skills to face their difficulties and discover the opportunities presented by technology. The quotes below were taken from responses to the questionnaire and reveal some of the principal characteristics of the of the profile of the students.

> "For a distance learning student … who works from 40 to 60 hours is difficult to take traditional classes," said a student 4.
>
> "People who work for 40 hours don't have the time to go to a University. It is much more comfortable to study at home," a student 5 commented.

Today's distance learning students were formed through the traditional classroom model throughout their learning process. As adults, they are not very familiar with technology and have trouble feeling responsible for their own learning. Moreover, for a long time they had only been content reproducers rather than producers. This is illustrated in the following quotes from students:

> "The distance education student depends on a physical professor, has some difficulty to interpret the tasks proposed, as well as to be able to use the Virtual Learning Environment well," student 1 stated.
>
> "My biggest challenge as a distance learner was the lack of physical contact, the exchange of looks, facial expressions, and other non-verbal communication that just does not happen in distance learning," student 2 said.

It is therefore necessary for students to construct new identities – that of a virtual student. In order to do so, one must remodel what has already been elaborated in the face-to-face classroom environment. This process is not only a cause or product of interaction but must be a constant transformation.

In order to develop a new identity, three fundamental points are necessary. The first is strategic action: time management, ways of communicating, disposition and motivation. The second is an understanding of the characteristics of the group and of the tasks, objectives and wider context of the course. The third is the technological conditions available to the student, such as reliable Internet connection, the use of tools and familiarity with technology, shown in "Fig. 1".

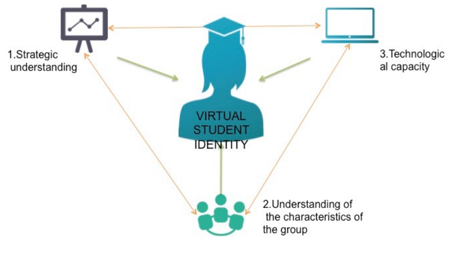

1.Strategic understanding
3.Technologic al capacity
VIRTUAL STUDENT IDENTITY
2.Understanding of the characteristics of the group

Figure 1. Representation of the Identity of the Virtual Student

To begin with, students can take their real world experiences to the virtual world along with the understanding of strategies, characteristics and technology. Students can also start to create a type of hybrid or blended learning [20].

This convergence of real world and virtual experiences will unlock the students' style and behaviour in each situation, be it face to face or virtual. It can be concluded that the profile of distance education students is composed of three interlinked contexts/dimensions, those being: social/family related, professional and academic. Technology is transversal and permeates all other contexts owing to its daily presence in the life of the student. However, technology is not a resource with which this student is familiar, being a context under construction and which, owing to experiences and time, will go through transformations with different results for each student.

Having raised and organized the profile, it was possible to present and list the competences of these students. This is the subject of the next section.

VII. Competences of Distance Learning Students

Basic competences of students of distance education were identified through analyzed data. These competences derive from an initial mapping of the relevant literature, which was subsequently compared with the data obtained from students taking the course and statements given by teachers and tutors. From this, text extracts were taken in which the research subjects indicated the competences that a distance learning student should have. The extracts below are the tutors' opinions about the students.

> "The competences required are: Autonomy, discipline, reflective reading, suitable production of writing, proper reference of other's texts,," said tutor 1.
>
> "The student has to be autonomous, be able to manage his time, have basic knowledge of computing, know the virtual environment that he will use," tutor 3 said.
>
> "[...] the student should have knowledge about computer tools and computing itself, and be autonomous," stated tutor 7.
>
> "The student must be organized and disciplined, and should seek a theoretical base from the classes' suggested readings.," tutor 20 said.

The following extracts reveal the students and tutors answers about the competences the students had during the course.

> "I think that the group I classified as knowing more about technology had technological competences but no communicative competences. Yet, the other group had more difficulties with technological issues. Both groups lacked good time management strategies," tutor 4 stated.
>
> "There are cases in which students didn't have any of this competence [technological]. In this case, the focus has to be to invested in the student's digital literacy, which I think is a very important," tutor 5 said.
>
> "I believe I have the main competences, though due to the everyday rush, I end up falling behind on my homework/readings," student 1 stated.

The mapping itself, shown in "Fig. 2", demonstrates the organization of the twelve competences listed in this study together with the profile of these students. While many other competences of course exist, the students in this study showed little familiarity with distance learning, thus generating a list of basic competences.

The results obtained show that there exists a gradual transition, beginning with competences of digital fluency and ending with team work. This analysis, based on levels of difficulty, was made possible via the statements made by statements as well as by activities undertaken on the course.

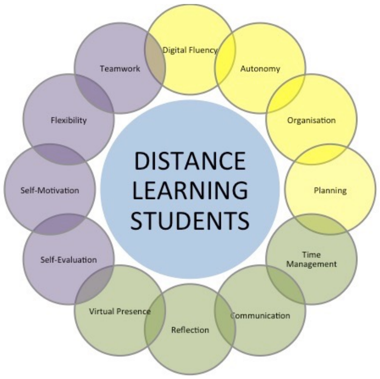


Figure 2. Twelve Competences of the Distance Learning student

The full competence mapping, and their elements (Knowledge, Abilities and Attitudes), will be presented below.

A. Digital Fluency

Digital Fluency is linked to the use of technology whereby the student feels an active participant in advances in digital technology. Fluency makes possible not only the use of technology but also the creation of content and materials. Knowledge: Theoretical and technological knowledge of the tools. Abilities: The ability to use, search for, select and produce. Attitudes: Having the initiative to seek innovations and keep up to speed with technological advances at all times.

B. Autonomy

For Piaget, autonomy is being governed by oneself. It is the opposite of heteronomy which is where one individual is governed by another. Knowledge: Knowledge of social and cultural norms, moral values and ethics. Abilities: The ability to analyse and interpret data and situations, making complex choices, anticipating situations, selecting, systemising, relating and interpreting data and information and making decisions. Attitudes: Having self-control and being responsible, self-critical, proactive, committed and ethical.

C. Organization

Organization is related to the ordering and structuring of activities, materials and groups. Knowledge: Having self-knowledge, being able to plan and being aware of deadlines. Abilities: Creating strategies, systemising, ordering and classifying. Attitudes: Being engaged, involved and proactive, taking decisions and being persistent.

D. Planning

Planning is based on the establishment of priorities, goals and objectives. In education, planning is also having the necessary conditions to create learning situations and apply learning strategies. Knowledge: Knowledge of types of planning, context, opportunities, weaknesses and audience (if applicable). Abilities: Systemising, evaluating and analysing. Attitudes: Being proactive, objective and methodical.

E. Time management

Time management is necessary in fulfilling diary commitments, organizing and managing activities and fulfilling priorities, goals and objectives. Knowledge: Knowledge of deadlines, methods of organization and self-knowledge. Abilities: Using time efficiently, establishing limits, deadlines and priorities, ordering actions and identifying objectives. Attitudes: Being proactive, objective and focused.

F. Communication

Communication is founded in the clarity and objectivity of oral, gestural and written expression. Knowledge: Linguistic knowledge, understanding rules of behaviour, forms of communication and different audiences. Abilities: The ability to write clearly, objectively and coherently, to interpret messages received, and knowing how to use one's voice, articulate words and use appropriate language. Attitudes: Being expressive, empathetic, cautious and articulate.

G. Reflection

Reflection is based on being able to reflect on and critically analyse situations, activities and ways of behaving. Knowledge: Knowing the object in question and its different aspects. Abilities: This consists of analysing and interpreting data/facts and situations. Attitudes: Being proactive, critical and reflective and having self-control and the ability to self-teach.

H. Virtual presence

Virtual Presence is the concept of being present in the virtual environment through interaction with classmates and the carrying out of activities. Knowledge: Knowledge of the virtual environment and its tools, forms of communication and deadlines. Abilities: Using virtual environment tools efficiently for communication and the sending of activities. Attitudes: Being proactive, analytical, having insight and being willing to participate.

I. Self Evalaution

Self-evaluation is having knowledge about one's own learning processes and thus being able to colaborate in or evaluate the activities proposed. Knowledge: Knowing one's learning needs, one's learning processes and the ways of evaluating. Abilities: Being able to analyse the learning process, systemise activities, mediate and take into consideration one's individual characteristics. Attitudes: Having self-control, being critical, being up to date with developments and being receptive.

J. Self-motivation

Self motivation is establishing the conditions to be able to maintain motivation amongst peers and with oneself and being a facilitator in the process. It is being able to be receptive to the difficulties being faced by others and encouraging peers to continue and conclude an activity, being active and participating. Likewise, it is being able to deal with one's own difficulties. Knowledge: Self-knowledge, knowledge about others and about motivational mechanisms. Abilities: Insight and ability to criticise, analyse and face obstacles. Attitudes: Having self-esteem and self-confidence, being open, engaged, participative, receptive, empathetic, open to exchanges and being able to put oneself in the place of others.

K. Flexibility

Flexibility is being able to deal with a range of needs, examining and interpreting the opportunities presented by actions as well as by changes of opinion and attitude. Knowledge: Knowledge of interpersonal relations, knowing how to deal with socio-cultural differences. Abilities: Being able to identify situations, analyse possible solutions and mold situations. Attitudes: Being ethical and responsible and knowing how to change one's posture.

L. Teamwork

Teamwork relates to intra and interpersonal relations which allow individuals to effectively express and communicate their feelings, desires, opinions and expectations. Teamwork also requires interpersonal conduct and dexterity to interact with others in a socially acceptable way so as to bring benefits to participants in moments of interaction. These elements can also be complimented by an affective perspective as the complexity of social relations also demand the capacity to notice and make distinctions in mood, intentions, motivations and in the feelings of others. Knowledge: Knowledge of types of teams and the parts which compose a team. Abilities: Adapting intra and interpersonal actions, creating strategies and articulating communication with others. Identifying the profile and needs of the team in which one is placed, knowing how to work in a fair environment, articulating conflicts, negotiating, communicating, collaborating, cooperating, being able to adapt to new situations and deal with different situations. Attitudes: Being concerned to reach team objectives, being flexible and open to criticisms and suggestions, knowing how to listen to others and being collaborative and cooperative.

VIII. Conclusions

The focus of this study was to relate studies of distance education competences with the aim of identifying competences, which may be able to help students in the learning process. The final objective was to map these competences and their elements. By analyzing the student of this education method, the necessity for new educational opportunities was apparent. The competences relate to these opportunities seeking, via the input of students and their trainers, to uncover solutions, which can provide action and change, above all, in the challenges faced by students who begin their education in the distance education format.

In this context, it is more necessary than ever for students to have adequate skills to manage their own learning, aiming to learn through autonomy.

Distance courses demand a great deal of organisation and flexibility on the part of the student. Thus, understanding the competences and elements that can facilitate the learning process for students would appear essential for those participating in this process.

It is therefore believed that the results of this study can bring about reflections relating to best practice in distance education as well as to new ways of teaching and learning.

A continuation of this research will therefore map the digital competences of these students, focusing on technological questions, so as development of these competences at the beginning of distance learning course.

REFERENCES

[1] M. Prensky, “Digital Natives, Digital Immigrants,” In: On the Horizon. NCB University Press, n. 5, v. 9. 2001, pp.1-6.

[2] C. Coll, C. Monereo, “The studente in virtual environments: conditions, profile and competences,” Porto Alegre: Artmed, 2010.

[3] M. Moore and G. Kearsley, “Learning distance education: an integrated view,” São Paulo: Thomson Learning, 2008.

[4] BRASIL. Law n. 9.394, “Guidelines and Bases of National Education,” <http://www.planalto.gov.br/ccivil_03/Leis/L9394.htm>. [retrieved: september, 2015].

[5] S.D. Gilbert, “How to Be a Successful Online Student” New York: McGraw-Hill, 2001.

[6] C. Monereo, “Internet and basic competences”, Barcelona: Graó, 2005.

[7] R. Palloff and K. Pratt, “The Virtual Studente,” Porto Alegre: Artmed, 2004.

[8] I. Chiavenato, “Introduction to general theory of administration,” 6. ed. Rio de Janeiro: Campus, 2000.

[9] G. Le Boterf, “The competence,” Paris: Quatrième Tirage, 1995.

[10] A. Fleury and M. T. Fleury, “Building the concept of competence,” 2001: <http://www.scielo.br/pdf/rac/v5nspe/v5nspea10.pdf>. [retrieved: september, 2015].

[11] I. M. Gaspar, “Competences in questions,” Portugal, 2004. <http://repositorioaberto.univ-ab.pt/bitstream/10400.2/158/1/Discursos%E2%80%93Forma%C3%A7%C3%A3o%20de%20Professores55-71.pdf> [retrieved: september, 2015].

[12] P. Perrenoud and M. G. Thuller, “The competences to teach in the XXI century,” Porto Alegre: Artmed, 2002.

[13] A. Zaballa and L. Arnau, “How to learn and to teach competences”, Porto Alegre: Artmed, 2010.

[14] BRASIL, Notion CNE/CEB n ° 16/99, “National Guidelines for Professional Education Technical Level” <http://portal.mec.gov.br/setec/arquivos/pdf/PCNE_CEB16_99.pdf>. [retrieved: september, 2015].

[15] BRASIL, ______. Notion CNE/CP 9/2001, “National Curricular Guidelines for Teacher Training of Basic Education at a higher level”, Brasília: http://portal.mec.gov.br/cne/arquivos/pdf/009.pdf>. [retrieved: september, 2015].

[16] P. Perrenoud, “Build competenes from scholl,” Porto Alegre: Artmed, 1999a.

[17] J. Piaget, “The intelligence born in children,” Suíça: Guanabara, 1987.

[18] BRASIL, “Quality benchmarks for Higher Distance Education, Brasília: Ministério da Educação, Secretaria de Educação a Distância,” 2007. <http://portal.mec.gov.br/seed/arquivos/pdf/referenciaisead.pdf>. [retrieved: september, 2015].

[19] M. R. Notare and P. A. Behar, “The mathematical communication online through ROODA Exact”, Porto Alegre: Artmed, 2009.

[20] R. Tori, “Without distance education:interactive technologies to reduce in teaching and learning,” São Paulo: Senac São Paulo, 2

Open Data from Social Media as Tool for Better Understanding Complex Territory

Application through Photos Data in Calabria

Alexandra Middea, Silvia Paldino, Sara Maria Serafini
Università della Calabria
Arcavacata di Rende (Cs), Italy
email: alexandra.middea@me.com, silvia.paldino@unical.it, saramariaserafini@gmail.com

***Abstract*—Data are becoming increasingly important nowadays, because they represent a concrete and inexhaustible source of information, which could be transformed into knowledge. Then, knowledge is synonymous with resource, because it represents a source of personal and community enrichment and, at the same time, it allows to accomplish more aware actions and to take advantage of every moment of freedom that comes from it. Thanks to today's technology, that creates services in response to the growing needs of citizens, public administrations and companies, the value of these data is finally disclosable in simple and immediate ways. An important tool, particularly with regard to more precise issues, such as security, economy or quality of life, is Geographic Information System, since it allows not only to represent the collected data in an immediate way, but above all to provide it with georeferencing, and, then, to spread it through mappings, easily interpreted also by non-experts.**

In particular, research is always more interested in big and Open Data; for this reason, the present work aims to analyze a significant amount of information on geo-referenced data in order to provide a broad view on issues related to tourism flows, in a reality such as Calabria, an Italian region, which is configured as a territory characterized by complex interactions and dynamics.

Keywords-Big Open Data; collective sensing; community; territory.

I. INTRODUCTION

The meaning of Open Data can be clarified by using one of the commonly accepted definitions provided by the Open Data Manual, the "Bible" for anyone who wants to embrace this philosophy, which describes it as:

"[...] data that can be freely used, reused and redistributed, with the only limitation - at most - of the request for allocation and the redistribution of the author in the same way, so without any change."

As highlighted above, we talk about "open" data, i.e., freely transmitted and distributed information, that is exchanged in the network in ways that provide for the total absence of forms of control (such as copyright and patents) and other restrictions that may limit the use, integration and reuse.

Starting from the concept of open knowledge as outlined by the Open Knowledge Foundation (a no-profit foundation founded on 24 May 2004 in Cambridge with the aim to promote open content and Open Data), Open Data can also be characterized by the same principles:

- availability and access: data must be available in a convenient and modifiable form, preferably by downloading it from the Internet. Data must be available in a useful and editable format;
- reuse and redistribution: data must be provided so as to allow its reuse and redistribution, this includes the ability to combine it with other databases;
- Universal participation: everyone should be able to use, redistribute and reuse data, without any discrimination towards application areas or people or groups.

Open knowledge is a prerequisite for collective intelligence, through which it is possible to implement the main practical advantage of the opening that is to exponentially increase the ability to control, certify, explore and combine different databases and then develop new products and services [1].

The use of Open Data is also connected to the tools used for its cataloguing, processing and representation. Since most data are equipped with a system of coordinates that make the data itself georeferenced, it is logical to connect the subject with the Geographic Information System (GIS), a system designed to receive, store, process, analyze, manage and display data geographically; GIS, in fact, allows to work on maps and to show, through an endless series of layers, all the features that are highlighted in a given territory.

But why should research about urban planning be so much interested in these technologies?

1. Because, in the short period, the distance between the digital and the real world becomes shorter and shorter;
2. Because researchers have been very good in the last few years at talking about new technologies, but not as good at understanding how these technologies can actually improve our lives.

It is time to explore these issues, so, the question should be: How can we extract added value from these datasets that are constantly increasing? Everything comes together for creating abundance of data.

Collective sensing is focused on the human aspect that can be drawn from this data, and it would be quite interesting to understand how data can be representative of some collective phenomena, such as mobility, transportation, tourism, etc.

One area that has much interest in Big Data and Collective Sensing is, in fact, tourism. As it will better explained later, tourism in Italy represents 10% of the national Gross Domestic Product (GDP). However, nobody knows how many tourists are present at a given time in a specific area of our nation. These details can be retrieved only with great expense in terms of time and costs, and after a certain period; it is the same case of a company that does not know who its users are, it is unfamiliar with its production cycles, and that, by working in this way, is bound to fail [2]. Big Data can help in this situation, providing knowledge about what happens and also analyzing it.

The work presented is divided as follows: Section II explains the importance of Big Data to analyze complex cities and to help decision makers to invest in and improve this kind of cities; Section III describes how GIS can be used as a tool to mapping Open Data, Section IV provides an example of the application is presented, and Section V concludes the work with some considerations.

II. The use of Big Data for the complex cities

“Big Data” is a huge collection of such complex datasets as to require the use of different tools as compared to traditional ones in all phases of the process: acquisition, also through sharing, analysis and visualization. The increasing size of the dataset is related to the need to analyze a single dataset, with the aim of extracting additional information as compared to what could be obtained by analyzing just small series of data, for example, the analysis to gauge the “mood” of the markets and trade, and thus, the overall trend of the company and the flow of information travelling and passing through the Internet.

Big Data also represents the interrelationship of data from potentially disparate sources; these are structured set of data, such as databases, but also unstructured ones, such as pictures, emails, Global Positioning System (GPS) data, as well as information taken from social networks. So, we can talk about “Big Data” when we have a large dataset which requires unconventional tools to extract, manage and process information within a reasonable time [3]. This issue is ever changing because machines are getting faster and datasets are getting bigger. According to a 2001 study [4], the analyst Doug Laney defined the growth model as a three-dimensional one (model of “3V” [5]):

- *volume*: is the size of the data set;
- *velocity*: is the velocity of generation of the data; there is a tendency inherent in making analysis of the data in real time or nearly so;
- *variety*: refers to the various types of data from different sources (structured and unstructured);

This model, summarized in Figure 1, is still valid, although in 2012 the model was extended to a fourth variable V to indicate the *veracity* of the data [6], i.e., the informative value that you can extract.

Over time the model was extended, adding the following features:

- *variability*: this feature can be a problem because it refers to the possibility of inconsistency of data;
- *complexity*: the huge size of the dataset increases the complexity of the data to be managed; the most difficult task is to link the information to obtain interesting outputs.

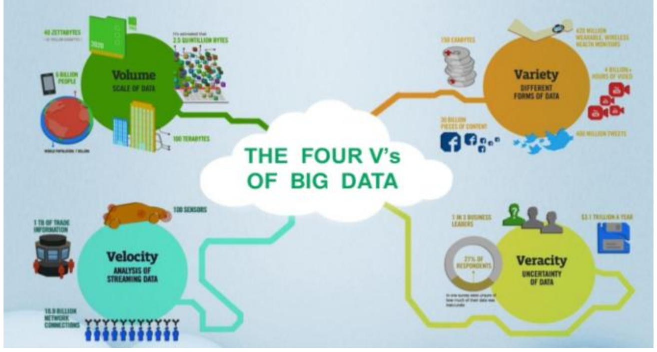


Figure 1. The four V’s of Big Data, source IBM.

With 7 billion people on the planet, who access about 1.2 billion personal computers and 1.5 billion smart phones, growing at a rate of about 30% annually, the scale of data being generated by these devices is daunting [7], but it is important that Big Data are not turned into Bad Data [8]. In fact, the possibility to collect digital traces on a massive scale could be transformed from a tool of potential liberation — the fuel that drives Open Data initiatives in cities and states across the world — into an instrument of abuse, surveillance and asymmetrical control.

Yet, Big Data still hold many promises, not only for the private but also for the public welfare. In cities, Big Data is making a tremendous impact across a broad spectrum: it is helping to imagine a more efficient mobility [9], reducing pollution [10], showing humanity patterns [11]-[13], from energy [14] to waste [15]; moreover, understanding city patterns is a useful instrument for urban planning in general [7]. It is a silent tool that can promote new forms of civil engagement. Nonetheless, a new way to frame the relationship between individuals and Big Data is urgently needed, to move beyond today's pseudo-feudal system of trading personal information for a service.

The challenge, of course, is that Big Data will shortly provide new ways to analyze topical issues of the world, offering new immediate solutions.

Big Data is certainly enriching our experiences of how cities function, and it is offering many new opportunities for social interaction and more informed decision-making with respect to our knowledge on how to better interact in cities. However, it is important to use it properly and respecting everything, keeping in mind that citizens are essentially people and not only data.

III. Mapping Open Data with GIS

GIS is a system designed to capture, store, manipulate, analyze, manage, and present all types of spatial or geographical data. The acronym GIS is sometimes used for geographical information science, or geospatial information studies to refer to the academic discipline, or career of working with geographic information systems, and it is a large domain within the broader academic discipline of Geoinformatics [16]. In a general sense, the term describes any information system that integrates, stores, edits, analyzes, shares, and displays geographic information. GIS applications are tools that allow users to create interactive queries (user-created searches), analyze spatial information, edit data in maps, and present the results of all these operations [17]-[19]. Geographic information science is the science underlying geographic concepts, applications, and systems [20]. GIS is a broad term that can be referred to a number of different technologies, processes, and methods. It is attached to many operations and has many applications related to engineering, planning, management, transport/logistics, insurance, telecommunications, and business [18]. For this reason, GIS and location intelligence applications can be the foundation for many location-enabled services that rely on analysis and visualization. GIS can relate unrelated information by using location as a key index variable. Locations or extents in the Earth space–time may be recorded as dates/times of occurrence, where x, y, and z coordinates represent, longitude, latitude, and elevation, respectively. All Earth-based spatial–temporal location and extent references should, ideally, be relatable to one another and ultimately to a “real” physical location or extent. This key characteristic of GIS has begun to open new avenues of scientific inquiry. The GIS tool allows to map the territory and to have a complete visualization of the overall situation of a determined phenomenon.

The importance of the localization component, i.e., the possibility to have geographical knowledge and information detailed up to the urban level, is often an essential element of the knowledge base of businesses, institutions, local administrative bodies, and public and private operators, providing services in many areas; moreover, with the greater interest in economic, social, political and environmental issues, the availability of data and information, that could be traced on a geographical basis, is increasing.

For example, GIS is a fundamental tool for smart cities, as explained in the book “Geographic information systems for Smart Cities” by Professor Vinod [21], focusing on how the future development of GIS will be triggered by Smart City challenges. In order to explain how GIS system performs, GIS experiences in conjunction with Smart Cities from many countries are shared by GIS experts who have designed and maintained it for several years. GIS is employed for sea erosion issues, urban resilience issues, slum rehabilitation (for example in India where slums represent one of the most important social aspects to be considered in the urban planning) and state perspective; GIS is also used for smart growth and transport planning, for land use allocation and also for community planning and so on.

The investigation about smart cities requires an integrated approach through innovative, sustainable and inclusive dimensions with knowledge across green energy, sustainable transportation, quality environment and smart building, risk and resilience and many other different domains in which geospatial data and GIS are fundamental elements.

IV. The use of Open Data for the analysis of tourist flows

As discussed in Section III, the availability of digital geographic information at different scales (national, regional and urban) has produced a crucial transformation in the use of spatial data in recent years, with important benefits for organizations, institutions, governments, public and private operators in the different sectors of economy and services.

It is sensible to assume that the possibility of operating in a relatively simple way with geographically related information, i.e., “geo” information linked directly to the territory, and “graphic” information based on the effective graphical representations of digital mapping, can produce benefits in several application areas, with research or operational purposes, resulting from enhanced possibilities to integrate databases and from their use “on site” (by their direct acquisition), and/or “on line” (through the construction of databases that make information available on the web immediately accessible).

GIS applications are those traditionally used for the production and use of digital maps, such as monitoring and mapping territorial, environmental protection, urban planning, design and operation of road networks, stations and, more generally, technological networks.

Recently, the use of GIS and spatial data has become more popular and widespread; therefore, GIS applications, often integrated with simulation and visualization tools, are covering new areas, such as telecommunications, but little has been done in the cultural and tourist sector so far.

Our country is universally recognized for its great cultural heritage: 3.609 museums, nearly 5.000 cultural sites (including monuments, museums and archaeological sites), 4.000 entertainment places, 49 UNESCO sites, hundreds of festivals, traditions and cultural events.

Tourism is a key sector for our economy (10% of GDP), but beyond figures and statements about our tangible and intangible heritage, the truth is that culture is not considered as a priority in the political choices for the development of the country [22].

For several years, the cultural sector has been suffering due to a serious decrease in resources, that was the consequence of a substantial absence of active investment policies for the development of cultural, creative and artistic activities, and of a renunciation to an effective protection and enhancement of our heritage. Moreover, the interest people have in heritage and cultural activities in general, is progressively increasing, surely thanks to a medium level of culture, that has improved considerably as compared to the past, but also thanks to an instinctive attraction to beauty, which leads us to approach, look at and understand it.

These statements are supported by data provided by the National Statistical System (SISTAN) and Statistical Office of the Ministry of Heritage and Culture (MIBAC); from the analysis, it is clear that the number of museum visitors (including also monuments, archaeological sites, etc. in this category) is steadily increasing. In fact, cultural tourism remains a key segment of the tourism industry, which accounts for about 35%; moreover, 17.6% of the Italian and foreign expenditure in our country in 2012 (i.e, 12.6 billion euro), was represented by expenses made for cultural activities.

The increasing diffusion of social networks, internet, the use of new technologies and strategies of digital communication, have generated deep changes and imposed new rules, new speed and new spaces. Essentially, we have created new ways of interaction and relationship with end-users, users themselves and between users and cultural institutions; a new communication space, made not only of exclusive content, created ad hoc, but mainly based on sharing, discussions, constant feedback and interaction with users, before, during and after the experience of enjoyment [23]. Therefore, we should reconsider everything. And we should do it fast.

Based on the data provided by Globalwebindex 2014, Figure 2 shows the daily hours spent on social media by the people who use them in different countries. Italy, with 2.5 hours / day, perfectly ranks in the middle between the minimum value of 0.7 hours / day in Japan, and the maximum value of 4.3 hours / day in Argentina and the Philippines.

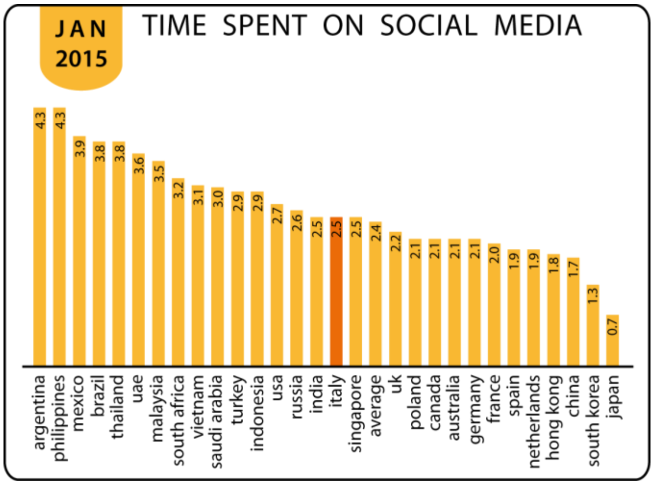


Figure 2. Average number of hours / day spent by users on social media (this figure only applies to people who use social media, not who do not use them), source Globalwebindex 2014.

Although our country has the highest concentration of cultural heritage, certainly it does not stand out for a promotion activity able to communicate with the new generations, exploiting the full potential of digital channels, starting with the name of the museum that we seek in the web. The *Uffizi Gallery*, for example, is one of the most famous museums in the world, but on the search engine Google it ranks third among the research results; first of all, the name of the website is *polomuseale.firenze.it*, the website is translated in English only, the graphical interface is not attractive, the Facebook page has only 28.794 “likes” and only 117 people who “are talking about it”, nothing as compared to the *British Museum*, that has 468.747 “likes”, and 12.057 people who “are talking about it” [24].

Heritage and cultural activities, by their nature, are perfect candidates to support an effective endogenous development, however, art can help the economy grow if a proper strategy [25] is conceived.

The convenience to invest in the cultural field, therefore, lies not in an immediate economic advantage, but in the utilities flow generated by the use, research, and propagation of the heritage and territory where cultural heritage is located.

It is clear how much social media can help promote a territory rather than another. In fact, people buy through the social and choose clothes, shoes and accessories for the car through Facebook and Twitter, and of course, they also plan their vacations by using the same media. So, we can say that by studying social network information, it is possible to have a general idea of people perceptions.

The big novelty of this study does not only lie in mapping Big Data in GIS, since there are many examples of that in the literature, but in the particular kind of data considered.

In the large variety of social networks, we have chosen Flickr, the most famous social for picture sharing. We focused on geo-tagged photos because photography is a disciplined way of viewing and investigating landscapes, able to inform about design and planning in a more “qualitative” way. Residents and visitors take photos in particular places they consider important for some reasons. It could be very useful to understand what we like in our cities, in our territories, what we are interested in, or also where residents or tourists want to go. In turn, understanding this could provide important indications for urban innovation. For this reason, photography is already considered a good mean of inquiry in architecture and urban planning, being it quite useful for understanding the landscape [26].

As we can see in Figure 3, for Globalwebindex, in 2014, Flickr has been the first social network dedicated to sharing photographs.

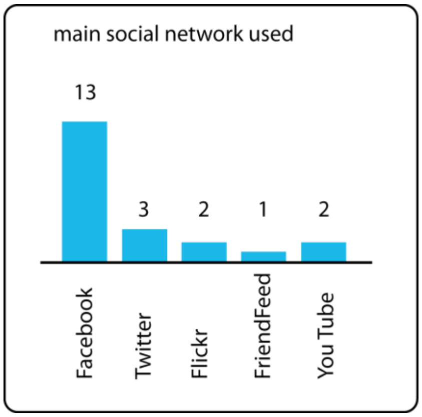


Figure 3. Main social networks used in Italy, source Globalwebindex 2014.

Flickr photos are publicly available, and there are many "Applications Programming Interfaces" (APIs) allowing the download of picture-related data. To carry out this study a tool was prepared, connected with flickr.photos.search API. Through this, we had the possibility to insert the name of the place we were interested in and download the information related to the pictures shared for that place.

Regarding the "spatial" issue, we have chosen to consider the poorest Italian region, Calabria: a land of seas, of mountains, of emigrants, of social, economical and geological problems, nevertheless a wonderful land. Regarding the "temporal" issue, we have chosen to consider data related to the last 5 years, i.e. the time-span from 2010 to 2014. Then, collecting data on pictures in Calabria related to these 5 years, we obtained a large amount of information. We have chosen to consider only spatial coordinates, i.e., the longitude and latitude of places where the picture was taken, year by year, and, for each year, month by month. Thanks to this approach, we were able to geo-localize our Big Data on a map, using a GIS tool, which can read and elaborate our files. Thanks to the GIS tool, the interaction between our cleaned Big Data from Flickr and the effective GIS in a map become possible.

The idea is to be able to describe the tourist potential of an area, and consequently its weaknesses in terms of attractiveness, through the study and revision of data collected from users who visited the territory, and to represent them through the use of GIS. In order to do it, as explained, we elaborated maps year by year, using different colors, and then we overlapped all of them in one single map (Figure 4).

Until now, in the literature, we can find studies about a particular event in a particular area of a city [27], or in three consecutive years in ten of the most famous and attractive cities of the world [13]. In this study Big and Open Data from photography is used for the first time to understand people perceptions in a poor region, so different in each of its own areas, including relatively small cities, little events, sea and mountains, and in such a large temporal range.

Referred to the Map of Attractiveness [13], we created the Map of Attractiveness of Calabria for each year, and at the end, instead of different categories of users, we overlapped the years, comparing what happened each year. and the attractiveness evolution.

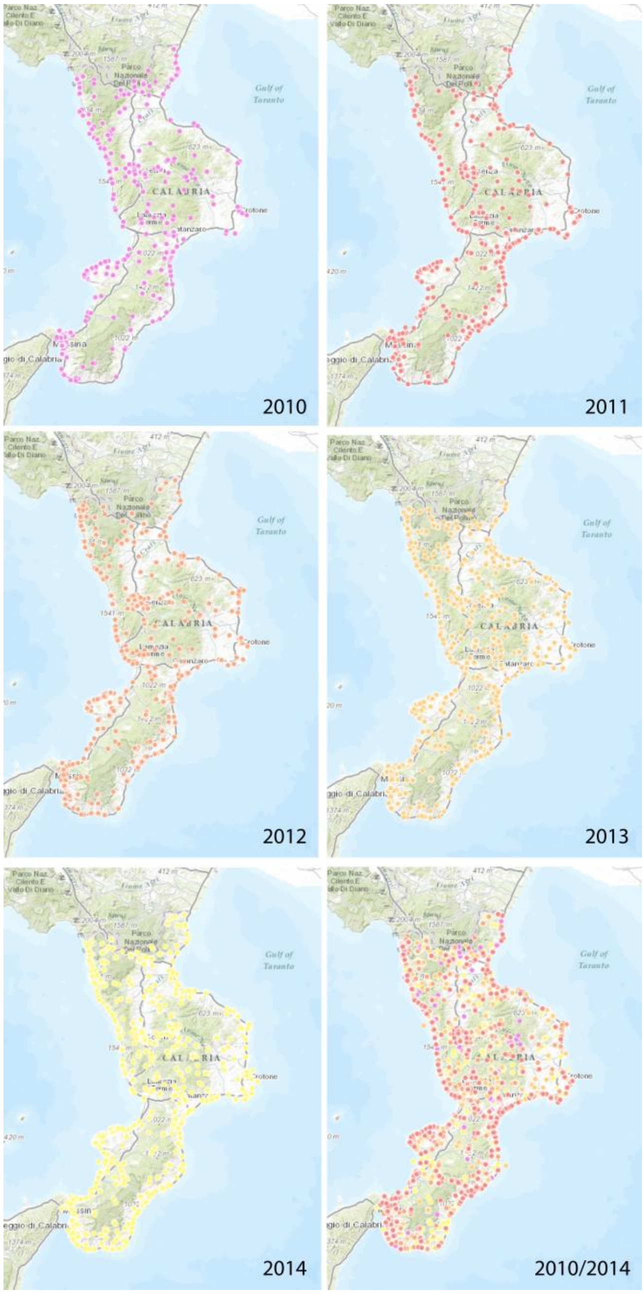


Figure 4. Calabria (Italy), representation of Open Data from Flickr - series 2010/2014, and overlapping of several years, processing by GIS.

We think that this visualization can be considered as representative of the popularity rating of people, who, satisfied after visiting a place, want to preserve the memory, and then take a picture and share it on the social.

Through the downloaded Data, the origin of the users can be traced, information for year can be classified, and the geographical coordinates of the place where it was taken can be associated to each photograph; thanks to this approach it was simple to recognize the most popular tourist sites, the months of the year when tourist flows are greater, and to try to create attractions that enhance tourism itself; but above all, it becomes clear the need to analyze the deficiencies of the territory, which are immediately visible by the mapping of the area, and suggest strategies to overcome the gaps found.

In particular, by the analysis carried out in the years 2010/2014, shown in Figure 4, a repeated pattern for each year stands out: the predominance of the coast respect to the hinterland. This is not related to a particular period or month, even if it is clear, through the temporal analysis, that the most attractive period is summer.

The hinterland is almost completely unknown, despite the presence of three National Parks (Aspromonte, Pollino, Sila), a Regional Park (Serre) and numerous reserves. Also, as far as the coast is considered, each year the Tyrrhenian coast is more attractive than the Ionian one. These are the macro considerations relative to what happens in the whole region, considering coast and hinterland. We also want to underline what happens, at regional scale, in the different cities.

The experiment showed that out of 409 towns divided into five provinces (Catanzaro, Cosenza, Crotone, Reggio Calabria, and Vibo Valentia), the popular and well-known places are less than 70. These information indicate an isolation of some towns, a tourism flow which is limited to certain nationally known and well advertised places, probably included in the catalogues of travel agencies. But Calabria is also made of smaller and typical towns, of characteristic places full of traditions, which should be valued and handed down.

The idea was to compare the Data from the last five years to assess the possible evolution of the interest and the attractiveness of the Calabria region; overall the results indicated it as an interesting place and with a good tourism potential, in fact the tourists' interest is generally constant, although the region faces some problems, which we have already mentioned above, probably related to poor accessibility of the hinterland, and to the inadequacy of the infrastructure network and of the media system.

V. Conclusion

In this study, we introduced and explained two new tools that are becoming fundamental in the new approach to observation of cities and territories in general: Big Data and GIS. These tools are also mutually related, because the term Big Data also indicates Open Data, the latter are most of all georeferred data, particularly in the cases we are interested in, like cities. With them it is possible to conduct analyses applying the complex theory that better responds to city descriptions, revealing patterns and predicting phenomena that can allow to improve everyone's life. These geotagged data can be spatially visualized on maps to have a complete and immediate vision of the situation we are talking about, thanks to the GIS.

Both tools are rapidly becoming necessary in space and territory related researches. However, we are quite at the beginning of such explorations, therefore research has to advance more and more in order to improve and maximize the huge potential of these instruments minimizing the controversies and the problems that a wrong use of the same can generate. Therefore, it is correct to affirm that GIS and Big Data suggest a future in which experiences of common citizens, and of true tourists, can be used for better understanding people's taste. Data give a "safe surprise", i.e., information about a reality away from home. Predicting that consumers will prefer a place instead of another one, or eat more in a restaurant rather than a pub, does not sound like a Big Data issue; nevertheless, by giving exact information about the favorite places, seasons, or hours of the day, it represents a new and powerful tool for fine-tuning and maximizing the tactical brand decisions. For instance, the capacity to adjust prices, mobility, services, in a quick and competitive way and in response to an analytically predictable change in tourist vocations, is clearly a significant tactical power. Thus, also through these predictions, Big Data will greatly enhance the capabilities of travel companies, tour operators and urban planners.

Previous studies on tourism, retrieved from the Internet or from scientific articles, are mostly based on surveys and interviews with experts carried out by the Ministry of Industry, Energy and Tourism, the main public organizations, or private companies; in this particular case this indicates that the industry does not have real data about tourists and it can only take samples from the population as a whole. In contrast, the innovative approach achieved and proposed through this study, specially through the use of photos and GIS for localization, is to introduce data based on real actions of users and not on surveys. In other words, real actions have been analyzed instead of stated intentions or answers to questions, that can be interpreted through a subjective vision and can be, therefore, less useful for business and development.

This study has involved many different indicators, useful to carry out more precise contributions, such as:

- Visitors' main country of origin;
- Geographical position.

Based on the conclusions drawn from data analysis, the study ends with a series of tactical and strategic recommendations for managers. These recommendations focus on:

- Attracting more customers and pinpointing the countries where it is recommended to focus on marketing;
- Detecting areas of the city in which commercial transactions are carried out, specially, those referring to accommodation;
- Ensuring an attractive product suited to customers' true needs (ideal length of package offers, information about complementary services demanded by nationalities, etc.).

It is realistic to believe that a new frontier of social, territorial and economic retraining, could be based on open source data and geographic information, along with the use of information technology-based GIS.

The analysis of the vast amount of data produced by digital activities, opens up a wide range of opportunities for companies, for enhancing the services they offer and for the management of their business. This study is a first step for understanding the possibilities of Big Data, especially in Calabria region, where the potential is enormous. In this specific case, we are trying to contribute and add value to what is a key sector for the Calabrian economy, underlying that there are places already discovered and appreciated by tourists, such as coasts and the Sila area, but that there are also suggestive and unexplored towns. Moreover, this kind of study could become a replicable model, useful for analyzing other economic and social sectors, or other territories.

For sure digital traces we leave every day on social media will increase, providing a very accurate representation of what we do. Social media are a kind of expression of people and crowds participation because they reveal interests, tastes and perceptions of the community. Therefore, we can use these data to know, understand and analyze our collective behavior and, as a function of them, we can imagine better spatial planning in the territories where we live, by taking into account what each of us spontaneously shares and expresses on social media every day.

ACKNOWLEDGMENT

The authors want to acknowledge Stefano Vena – PhD - for kindly offering his technical support in downloading Flickr data.

REFERENCES

[1] S. Aliprandi, “Il Fenomeno Open Data: Indicazioni e Norme per un Mondo di Dati Aperti”, Ledizioni, 2014.

[2] Euro Beinat speech, Academy dell’Innovazione Urbana, https://www.youtube.com/watch?v=PXSjlGFPYzg, [retrieved: January, 2013].

[3] C. Snijders, U. Matzat, and U. D. Reips, “Big Data: Big gaps of knowledge in the field of Internet”, International Journal of Internet Science, 7, 1-5. International Journal of Internet Science, Volume 7, Issue 1, 2012, pp. 1-5.

[4] D. Laney, “3D Data Management: Controlling Data Volume, Velocity and Variety”, META Group Research Note 6, 2001, p. 70.

[5] M. Beyer, “Gartner Says Solving Big Data Challenge Involves More Than Just Managing Volumes of Data”. *URL http://www. gartner. com/newsroom/id/1731916* (2011).

[6] Villanova University, “What is Big Data?”

[7] L. M. A. Bettencourt, “The Uses of Big Data in Cities Big Data”, March 2014, 2(1), pp. 12-22.

[8] Conference “Engaging Data 2013 - Big Data or Bad Data”, MIT SenseABLE City Lab, 15 November 2013

[9] M. Batty, “Big Data, smart cities and city planning”, Dialogues in Human Geography n.3, 2013, pp. 274-279.

[10] C. Ratti, S. Di Sabatino, R. Britter, M. Brown, and F. Caton, “Analysis of 3-D urban databases with respect to pollution dispersion for a number of European and American cities”, Water, Air and Soil Pollution: Focus, no. 5-6, 2002, pp. 459-469.

[11] M. C. Gonzalez, C. A. Hidalgo, and A. L. Barabasi, “Understanding individual human mobility patterns”. Nature 453, no. 7196, 2008, pp. 779-782.

[12] C. Ratti, S. Williams, D. Frenchman, and R. M. Pulselli, “Mobile landscapes: using location data from cell phones for urban analysis”, Environment and Planning B Planning and Design 33, no. 5, 2006, p. 727.

[13] S. Paldino, I. Bojic, S. Sobolevsky, C. Ratti, and M. C. Gonzalez, “Urban magnetism through the lens of geo-tagged photography”. *EPJ Data Science* 4, no. 1, 2015, pp. 1-17.

[14] C. Ratti, N. Baker and K. Steemers, “Energy consumption and urban texture”, Energy and buildings 37, no. 7, 2005, pp. 762-776.

[15] H. Shahrokni, B. Van Der Heijde, D. Lazarevic, and N. Brandt, “Big Data GIS Analytics Towards Efficient Waste Management in Stockholm”. In *Proceedings of the 2014 conference ICT for Sustainability*, 2014, pp. 140-147.

[16] Fully Web-Based Geographic Information System (GIS) and Mapping, PolicyMap – Retrieved, 11 Feb 2015.

[17] H. K. Clarke, “Advances in geographic information systems, computers, environment and urban systems”, Vol. 10, 1986, pp. 175–184.

[18] V. Maliene, V. Grigonis, V. Palevičius, and S. Griffiths, “Geographic information system: Old principles with new capabilities”, Urban Design International 16 (1), 2011, pp. 1–6. doi:10.1057/udi.2010.25.

[19] A. Asgary and A. Middea, “Modelling Community Resilience: An Agent Based Approach”, Proceedings of Design for Urban Disaster Conference, Harvard – May 2014.

[20] M. F. Goodchild, “Twenty years of progress: GIScience in 2010”, Journal of Spatial Information Science 1, 2015, pp. 3-20.

[21] T. M. Vinod Kumar, “Geographic Information System for Smart Cities”, Vol.1, Copal Publishing Group, 2013.

[22] S. M. Serafini, “Il paesaggio agrario come bene da tutelare e risorsa economica. Il museo del territorio e le politiche di crescita legate al luogo”, Una politica per le città italiane - Urbanistica Informazioni, pp. 162-165, INU Edizioni, ISSN: 0392-5005, 2014.

[23] M. Gerosa, Milano, R., “Viaggi in rete. Dal nuovo marketing turistico ai viaggi nei mondi virtuali”, Franco Angeli Edizioni, Milano, 2011.

[24] F. Dallari and S. Grandi, “Economia e geografia del turismo: l'occasione dei Geographical Information System”, Pàtron, 2005.

[25] A. Pschera, edizione italiana a cura di Villani Lubelli, U., “Dataismo. Verso i Big Data: Critica della morale anonima”, goWare, 2013.

[26] A. Spirn, The language of landscape, Yale University Press, 1998, pp 3-81.

[27] F. Girardin, A. Vaccari, A. Gerber, A. Biderman, and C. Ratti, Quantifying urban attractiveness from the distribution and density of digital footprints. *International Journal* 4, 2009, pp. 175-200.

The Influence of Social TV Strategies and Contents on TV Online Engagement

Angela Fortunato, Michele Gorgoglione, Umberto Panniello
Dipartimento di Meccanica, Matematica e Management
Politecnico di Bari
Bari, Italy
e-mail: angela.fortunato@poliba.it

***Abstract*—The phenomenon of social TV is gaining importance in both industry and research. TV broadcasters are increasingly adopting social TV strategies both on first and second screen to increase the viewers' online engagement. The research done so far suggests quite simple models of the phenomenon, identifying and studying separately the effects of different variables on online engagement. This research represents one of the first attempts to develop a better research model. We analyzed a large dataset related to a popular Italian TV show using social strategies to engage viewers on Twitter. Through hierarchical linear regression models we studied the relationships among social strategies, TV contents, viewership, time and different kinds of viewers' online behavior. We demonstrate that (i) different factors play different roles in affecting the viewers' online engagement and (ii) the phenomenon can be better explained if we look at different kinds of online behavior that represent online engagement, such as posting original comments, sharing or replying to them. Despite some limitations, we think that this work's findings may be important for researchers to develop a holistic research model of social TV, and for practitioners to realize how to balance the factors affecting the viewers' online engagement in an effective way.**

Keywords-Social TV; social networks; viewers' online engagement.

I. INTRODUCTION

The phenomenon of Social TV has gained a striking importance in the last years. Increasing the amount of discussion associated with TV show's contents, i.e., the online engagement of TV viewers, has become a major goal of the companies in this industry. TV broadcasters encourage viewers interacting in real time with the TV shows and sharing online comments through "second screen" devices (smartphones and tablets). They do this through several social strategies, which can be delivered on both screens. Examples of strategies on the first screen include showing Twitter hashtags and recent viewers' comments on the TV screen during a show, or let the show hosts reply to some of the viewers' tweets. Second screen strategies include posting comments on a social network or delivering online messages to invite viewers polling a show's contestants. The main reason behind the use of these strategies is the idea that they can increase the viewers' online engagement in online communities, such as Twitter, which is the most popular social network in both research and industry domains. In the recent past, scholars have studied this phenomenon [34][35] examining the changes in the relationship between the broadcasting industry and its audience, thus considering the new ways of audience engagement through the integration of the technology. Particularly, recent research investigated the relationship between online engagement and social strategies. They showed that using social TV strategies on the first screen can be predictive of various types of online engagement [15]. Other works [5][10][12][32][33] have highlighted the existence of several variables, which can play an important role in driving the TV viewers' online engagement, such as viewership, time and the show contents. Moreover, recent studies [3][7][15][27][29][32] have identified different kinds of behavior, which can represent online engagement, such as posting tweets or sharing them. However, the approach used in these works suggests a quite simple model of the phenomenon, where the use of a certain social strategy can have an effect, potentially positive, on the online engagement of viewers. This research represents the first step in the development of an appropriate research model to study the phenomenon of social TV online engagement (OE). The goal of this research is to demonstrate that a change in the OE of TV viewers cannot be simply explained by the use of social TV strategies, rather different variables play different roles in the phenomenon. Based on prior research we have identified several variables, which can be included in a model. Firstly, three main factors can affect OE: the use of *social strategies*, the type of TV *content* on air, the *online behavior* of other viewers. Secondly, three variables may influence OE and have to be controlled: *viewership*, time during an *episode*, time during the *season*. Finally, OE can be represented by three different kinds of behavior: generation of *original* tweets, sharing tweets (*retweets*), *replies* to tweets. We studied the relationships among these variables by analyzing the behavior of viewers on Twitter during "L'Isola dei Famosi", a popular Italian TV show. We used hierarchical linear regression models. We have found that the phenomenon of OE can be better explained if we look at different kinds of behavior separately, namely generating original tweets, sharing tweets (retweets), replying to tweets. In this paper, we show the following main results: (1a) viewership affects the overall number of original tweets, as expectable; however, it does not affect retweets and replies; (1b) online engagement decreases with time during each episode, while it increases during the season; (2a) the effect of *social strategies* is different for different kinds of behavior: they increase generation while decrease retweets and replies; (2b) the effect of *TV contents* depends on the kind of content: commercial breaks decrease generation while increase retweets and replies, “challenges” always decrease OE. We also found that (3) an increase in the number of original tweets generated is correlated to an increase in retweets and replies. The contributions of this work are the following. First, we demonstrate that different factors play different roles in the phenomenon and their effect to online engagement is more complex than what research has shown so far. Second, we show that these effects depend of the type of online behavior we use to represent online engagement. Third, to our best knowledge this is one of the first attempts to develop a complex research model of the phenomenon of online engagement in the context of social TV. Despite some limitations in our research, we think that its findings may be important from the business viewpoint because they contribute to clarify the real factors affecting the phenomenon and help managers properly design TV shows and social strategies in order to drive viewers online and keep them engaged. The paper is structured as follows. Section II provides a description of the existing literature on social television. Section III depicts the methodology of our research, in terms of dataset, variables’ description and the propositions we explored in our research. Section IV illustrates the results and Section V describes their implications from research and managerial viewpoints and limitations of this work.

II. Prior work

Research has shown that television is a facilitator of social interactions, bringing people together and giving them a broad variety of topics to discuss [20][21][30]. Moreover, it affects viewers’ behavior [26] in terms of shaping, reinforcing or changing their reactions [16]. This depends on factors, such as the type of contents or messages [22][26]. In recent years, the television domain has been interested by the phenomenon of “Social TV”, which refers to the variety of systems that support social practices associated with TV viewing [13]. Social networks have gained a relevant role, since they allow viewers to share online their real-time viewing experiences [6], by interacting around the TV contents. Therefore, viewers can be affected by social interactions in online networks during a viewing experience [16][28]. Research has studied several aspects of Social TV. Some authors studied viewers’ motivations and the different ways to interact on social network sites during viewing [9][18][32]. Notably, some scholars analyzed the use of hashtags. For instance, [4] stated that hashtags are used to group tweets by topic, thus allowing people to follow and contribute to conversations on topics of interest, e.g., during televised political debates in the U.S. presidential primaries. Others studied how viewers’ messages are related to what viewers are watching and observed particular behavioral patterns focusing on specific TV programs and contents [5][12][32]. Moreover, viewers can deliver different types of messages (tweets) when using Twitter, which are generally studied separately: original tweets, when posting a message for the first time; replies, when responding to an existing message, and retweets, when sharing an existing message [3][7][15][27][29][32]. A major topic is the analysis of the Social TV Strategies that broadcasters often use during a show. These strategies may consist in adopting program’s official hashtags [8], displaying on the first screen social media elements such as viewers’ tweets [13][15] or in leveraging the use of second screen applications dedicated to the program to deliver several types of trigger [2][5][19]. These strategies aim at attracting people’s interest towards the live shows [24] and increase the viewers’ involvement towards the programs [31]. They are used to prompt real-time viewers to interact online with the TV programs and to share online messages [13][14][15]. A few studies have reported the effects of the Social TV Strategies on viewers’ behaviors. For instance, reference [15] analyzed the effect of the social TV strategies in the American show “The Voice”. They studied the effects of different social strategies displayed on the first screen: showing a show-related tweet on the screen increases the number of retweet, while showing a hashtag on the screen increases the viewers’ online engagement, i.e., the amount of discussions associated with the TV show’s contents [15], during commercial breaks. Another study [10] showed that in certain circumstances online engagement can be predicted by the show contents rather than by the use of social strategies and that the overall number of tweets is highly correlated to the number of viewers. Another work has examined viewers’ visual attention while interacting with synchronized second-screen applications and found that the presence of the second screen dramatically decreased the attention towards TV contents [17]. Each one of these works has studied the influence of one single type of variable on OE. This entails the use of very simple models, which, in our opinion, cannot explain a complex social phenomenon properly. Researchers, indeed, have not built yet a complex model to explain the whole phenomenon of the viewers’ online behavior. Building and testing such models is important because it would clarify the factors affecting online engagement and, in turn, the source of business value related to social TV and second screen applications. Our research represents the very first step in the attempt of filling this gap.

III. Methodology

In this section, we report the methodology followed for our experiments. In particular, we discuss the dataset used to perform analyses, the variables measured during our study and the propositions to explore.

A. Dataset

We collected data from the 2015 edition of the Italian TV show “L’Isola dei Famosi”, a reality show where celebrities have to survive on a desert island. The show had one episode a week and lasted 7 weeks, from the 2nd of February to the 23th of March, each episode lasted around 170 minutes. The TV broadcaster delivered several social strategies during the show on the second screen app dedicated to the program (see next section for the list). Viewers interacted on Twitter during the show using the official TV show hashtag. First of all, we acquired approximately 500,000 tweets, including the official TV show hashtag. The tweets in our dataset included original tweets, retweets and replies. We used the tweet time-stamp to associate each tweet with a minute during the show. Second, we structured in our dataset the type of TV content that viewers were watching minute by minute for the 1,242 minutes of show (including commercial breaks). Third, we collected the type of social strategy that the broadcaster was delivering on the second screen app minute by minute. Fourth, we gathered the number of viewers who were tuned on the show also minute by minute (viewership), provided by Auditel s.r.l.. In conclusion, we collected 1,242 observations corresponding to as many minutes and each observation included information on tweets, TV contents, social strategies and viewership.

B. *Measurements*

Based on the literature reviewed in the previous section, we have identified several variables, which can be included in a model of the Social TV phenomenon. Three main factors can generate OE, (1) the use of social strategies, (2) the type of TV content, (3) the online behavior of other viewers (the tweets generated). Three variables may influence OE and have to be controlled: (4) viewership (the number of TV viewers), (5) time (measured both during each episode and during the whole season). OE can be represented by three different kinds of behavior: (7) posting original comments, (8) sharing comments, (9) replying to comments. The dependent variable in our models is the online engagement (OE). According to the research we have reviewed, we have defined OE as the amount of discussions around the TV show’s content, measured by the total number of tweets including the TV show’s official hashtag. In addition we classified the tweets as follows: original tweets (OT), measured by the total number of tweets posted for the first time; original tweets from app (AT), measured by the total number of tweets posted for the first time, but delivered through the second screen application (AT represents a subset of OT); retweets (RT), measured by the total number of retweets, i.e., the share of existing tweets; replies (RP), measured by the total number of replies to existing tweets. According to prior research [15] the measurement of dependent variables is shifted by a time delay of one minute with respect to the measurement of independent variables. The independent variables in our models are the TV content and the social strategy. TV content is defined as a nominal variable that describes what type of contents viewers are watching on the TV screen and takes nine values: (1) general contents, (2) challenge, (3) nomination, (4) week summary, (5) contestant’s elimination, (6) appearance of eliminated contestant in studio, (7) visit in “Playa Desnuda”, (8) start of voting, (9) commercial break. Social strategy is defined as a nominal variable describing what message the broadcaster delivered to viewers through the second screen app. It takes eight values: (1) call to comment, (2) survey/quiz, (3) call to predict, (4) photo gallery, (5) call for appreciation, (6) call to vote, (7) displaying related information, (8) absence of strategy. Finally, we considered viewership and time (episode and season) as control variables. We measured viewership as the number of viewers in each minute. Since we observe the aggregate phenomenon of online interaction, a change in the number of viewers may affect the overall number of tweets. In addition, since a variation in OE may be caused by time [15], we used two measures of time: the minute in each episode and the number of episode. We included these measures of time to test the existence of a relationship between independent variables and dependent variables by excluding the effect of the control variables, but also to identify possible online engagement’s trends in time.

C. *Propositions development*

According to prior research, we developed three main propositions that we want to explore:

P1: *Viewership and time affect OE*.

P2: *TV contents and social strategies affect OE*.

P3: *OT affect RT and RP*.

We explored these propositions through hierarchical multiple linear regression models [1][11], using SPSS. In general, we first built a linear regression considering one dependent variable a time, checked the significance, then added more dependent variables and checked again the significance of the model. We repeated this procedure for each dependent variable. Dependent and control variables are metric, while we codified the nominal independent variables as dummy (binary) variables.

IV. Results

In order to explore P1, we built models having viewership, episode and minute as independent variables: first, we built simple linear regressions considering only one of the described variables, then we used all variables together. We iteratively repeated this procedure one dependent variable a time, i.e., OT, AT, RT and RP. Table I reports the results. We also ran regressions measuring viewership in a log scale. For lack of space we do not report these results. As expected, we found that viewership positively affects the online engagement. Looking at the three kinds of behavior, we observed that viewership positively affects the number of original tweets (both OT and AT), while it does not affect the number of RT and RP. We also found that the online engagement decreases minute by minute during each episode. Looking at the three kinds of behavior, the decrease holds for AT, RT and RP while OT increase. In addition, only the number of RT and RP increases during the season (episode by episode), while there is no effect for OT. Therefore, we found that P1 is valid, however the relationship between viewership, time and OE changes depending on what kind of online behavior we measure. In order to explore P2 we ran the models with

TV contents and social strategies as independent variables. First, we built the models considering only one independent variable and then all variables together, including viewership and time variables as control variables. We iteratively repeated this procedure one dependent variable a time. Table II reports the results of the final model. As expected, we found that TV contents and social strategies affect the online engagement. In particular, some contents (challenge), strategies (call to vote) and the absence of a strategy negatively affects the OE, while some contents (contestant’s elimination, commercial break, displaying related information) positively affects the OE. Looking at each kind of online behavior, we found more clear patterns. If we consider the TV contents, the “challenge” decreases all kinds of behaviors (OT, RT and RP); the “contestant’s elimination” increases the original tweets; the “appearance of eliminated contestant in studio” increases the number of RP. Interestingly, during “commercial breaks” the number of OT increases while the number of RT and RP decreases. If we consider the social strategies, we found that the “call to comment” has a positive effect only on the number of tweets from app (which is expectable because strategies were only shown on the second screen app) while it has a negative effect on the number of RT and RP. The “call for appreciation” and the “call to vote” have a negative effect on the number of RT, while the “displaying related information” has a positive effect on the AT. Finally, the “absence of strategy” has a negative effect on all kind of behaviors. Again, our models show that P2 is valid, however the relationships change depending on the way OE is measured. In order to explore P3, we ran the models with OT and original AT as independent variables. We included viewership and time as control variables. We used RT and RP as dependent variables. Table III reports the results of the final model, with all variables together. We found that the number of OT and the number of AT positively affects both the number of RT and RP. In this case, the models confirm the validity of P3.

V. Discussion and Conclusions

In this section, we present and discuss the findings and their implications from research and managerial viewpoints. First, we have shown that the phenomenon of OE can be better explained if we look at different kinds of online behavior separately (namely generating original tweets, sharing tweets and replying to tweets). In fact, we confirmed the influence of viewership on OE: the higher the number of viewers tuned into the program, the higher the number of tweets posted. However, looking at the different kinds of online behaviors, we found that the viewership affects the number of OT while it does not affect the number of RT and RP. We also found that the OE changes during each episode and during the season. During each episode, the number of RT and RP decreases while the number of OT increases. Finally, the number of RT and RP increases during the season, while the OT do not. Second, we have shown that TV contents and social strategies affect OE. Also in this case, the relationship between contents and OE as well as the relationship between the use of social strategies and OE changes depending on what online behavior is used as measure of OE. On the one hand, we found that some strategies have a positive effect on the generation of OT and a negative effect on RT and RP, while the absence of a strategy has a negative effect on all kind of behaviors. On the other hand, we found that different TV contents have different effects on the online behavior. During challenges, all kinds of OE decrease. During commercial breaks the generation of OT decreases while RT and RP increase. During the elimination, OT and RP increase. Finally, we found that the generation of OT positively affects sharing and replying behaviors: the number of OT generated is correlated with the number of RT and RP. This research has clearly some limitations. The main are the use of one dataset only, the fact that the social strategies in our setting were only delivered by the second screen app, and the use of “simple” regression models. Despite this, our findings may have interesting implications from both a research and a managerial viewpoint. For researchers, it is interesting to develop a consistent interpretation of viewers’ behavior to build a holistic research model. For instance, when viewers are watching a challenge they decrease each form of OE, but this is offset by an increase in the willing to post new comments when contestants are eliminated. Viewers seem to be willing to share comments and reply to comments during breaks. Challenges are likely to attract more attention towards the first screen, while commercial breaks attract viewers to the second screen: however, since viewers are not provided with any content to tweet about, they tend to read and react to existing tweets. From a managerial viewpoint, our results suggest that TV contents remain a major factor affecting OE. However, social strategies may increase the generation of OT, which, in turn, are a factor affecting RT and RP. Broadcaster should then carefully balance the use of social strategies with TV contents in order to drive the viewers’ online behavior effectively. In the next research steps, we plan to develop a behavioral interpretation of the phenomenon and test hypotheses through the use of more datasets and the use of more sophisticated statistical models. Moreover, we plan to analyze the relationship between contents and strategies and understand the effect of *combinations* of contents and strategies.

Acknowledgment

We wish to thank Telecom Italia for funding this research, Auditel for allowing the authors to access viewership data, Media Consultants for providing the viewership datasets.

References

[1] R. M. Baron and D. A. Kenny, “The Moderator-Mediator Variable Distinction in Social Psychological Research: Conceptual, Strategic, and Statistical Considerations,” Journal of Personality and Social Psychology, vol. 51, no. 6, Dec. 1986, pp. 1173-1182, doi: http://dx.doi.org/10.1037/0022-3514.51.6.1173.

[2] S. Basapur, G. Harboe, H. Mandalia, A. Novak, V. Voung, and E.C. Metcalf, “Field trial of a dual device user experience for iTV,” The 9th International Interactive Conference on

Interactive Television (EuroITV '11), ACM, June 2011, pp. 127-136, ISBN: 978-1-4503-0602-7

[3] D. Boyd, S. Golder, and G. Lotan, “Tweet, Tweet, Retweet: Conversational Aspects of Retweeting on Twitter,” The 43th Hawaii International Conference on System Sciences (HICSS), Jan. 2010, pp. 1-10, ISSN: 1530-1605

[4] A. Bruns and J. E. Burgess, “The use of Twitter hashtags in the formation of ad hoc publics,” The 6th European Consortium for Political Research (ECPR) General Conference, 2011.

[5] C. Buschow, B. Schneider, and S. Ueberheide, “Tweeting Television: Exploring communication activities on Twitter while watching TV,” Communications - The European Journal of Communication Research, vol. 39, no. 2, Jun. 2014, pp. 129-149, doi: 10.1515/commun-2014-0009.

[6] P. Cesar and D. Geerts, “Past, Present and Future of Social TV: A Categorization,” The IEEE Consumer Communications and Networking Conference, Jan. 2011, pp. 347-351, ISBN: 978-1-4244-8789-9

[7] G. M. Chen, “Tweet this: A uses and gratifications perspective on how active Twitter use gratifies a need to connect with others,” Computers in Human Behavior, vol. 27, no. 2, Mar. 2011, pp. 755-762, doi: 10.1016/j.chb.2010.10.023.

[8] R. Deller, “Twittering On: Audience Research and Participation Using Twitter,” Participations: Journal of Audience &Reception Studies, vol. 8, no. 2, May 2011.

[9] M. Doughty, D. Rowland, and S. Lawson, “Who is on your sofa?: TV audience communities and second screening social networks,” The 10th European conference on Interactive tv and video, July 2012, pp. 79-86, ISBN: 978-1-4503-1107-6

[10] A. Fortunato, M. Gorgoglione, and U. Panniello, “What Drives TV Online Engagement? The Influence of Social Strategies and Contents,” The 12th International Conference on Web Based Communities and Social Media, July 2015.

[11] P. A. Frazier, A. P. Tix, and K. E. Barron, “Testing Moderator and Mediator Effects in Counseling Psychology Research,” Journal of Counseling Psychology, vol. 51, no. 1, Jan. 2004, pp. 115-134, doi: http://dx.doi.org/10.1037/0022-0167.51.1.115.

[12] F. Giglietto and D. Selva, “Second Screen and Participation: A Content Analysis on a Full Season Dataset of Tweets,” Journal of Communication, vol. 64, no. 2, Apr. 2014, pp. 260-277, doi: 10.1111/jcom.12085.

[13] G. Harboe, “In Search of Social Television,” in Social Interactive Television: Immersive Shared Experiences and Perspectives, P. Cesar, D. Geerts and K. Chorianopoulos, IGI Global, USA, 2009, pp. 1-13.

[14] G. Harboe, C. Metcalf, F. Bentley, J. Tullio, N. Massey, and G. Romano, “Ambient Social TV: Drawing People Into a Shared Experience,” The 26th Annual SIGCHI Conference on Human Factors in Computing Systems, Apr. 2008, pp 1-10, ISBN: 978-1-60558-011-1

[15] S. Hill and A. Benton, “Social TV: Linking TV Content to Buzz and Sales,” The International Conference on Information Systems (ICIS), Dec. 2012.

[16] R.L. Holbertand and J.M. Tchernev, “Media Influence as Persuasion,” in The SAGE Handbook of Persuasion: Developments in Theory and Practice, J. Dillard and L. Shen, Eds. SAGE Publications Inc., 2012, pp. 36-52.

[17] M. E. Holmes, S. Josephson, and R. E. Carney, “Visual attention to television programs with a second-screen application,” The Symposium on Eye Tracking Research and Applications, Mar. 2012, pp. 397-400, ISBN: 978-1-4503-1221-9

[18] N. C. Krämer, S. Winter, B. Benninghoff, and C. Gallus, “How «social» is Social TV? The influence of social motives and expected outcomes on the usage of Social TV applications,” Computers in Human Behavior, vol. 51, no. A, Oct. 2015, pp. 255-262, doi: 10.1016/j.chb.2015.05.005

[19] M. Lochrie and P. Coulton, “Sharing the Viewer Experience through Second Screens,” The 10th European Interactive TV Conference (EuroITV), July 2012, pp. 199-202, ISBN: 978-1-4503-1107-6

[20] M. Morrison and D. Krugman, “A look at mass and computer mediated technologies: Understanding the roles of television and computers in the home,” Journal of Broadcasting & Electronic Media, vol. 45, no. 1, pp. 135–161, 2001, doi: 10.1207/s15506878jobem4501_9

[21] H. Newcomb, Television: The critical view, Fifth edition. New York: Oxford University Press, 1994.

[22] W. J. Potter, Media Effects. SAGE Publications Inc., 2012.

[23] S. Schirra, H. Sun, and F. Bentley, “Together Alone: Motivations for Live-Tweeting a. Television Series,” The SIGCHI Conference on Human Factors in Computing Systems, Apr. 2014, pp. 2441-2450, ISSN: 2308-4138, ISBN: 978-1-4503-2473-1

[24] M. Proulx and S. Shepatin, Social Tv: How Marketers Can Reach and Engage Audiences by Connecting Television to the Web, Social Media, and Mobile. Wiley, 2012.

[25] L. Rossi and M. Magnani, “Conversation Practices and Network Structure in Twitter,” The Sixth International AAAI Conference on Weblogs and Social Media, AAAI Press June 2012, pp. 563-566.

[26] L. J. Shrum, “Television and Persuasion: Effects of the Programs between the Ads,” Psychology and Marketing, vol. 16, no. 2, Mar. 1999, pp. 119-140, doi: 10.1002/(SICI)1520-6793(199903)16:2<119::AID-MAR4>3.0.CO;2-R

[27] D. Sousa, L. Sarmento, and E. M. Rodrigues, “Characterization of the Twitter @replies Network: Are User Ties Social or Topical?,” The 2nd International Workshop on Search and Mining User-generated Contents (SMUC), Oct. 2010, pp. 63-70, ISBN: 978-1-4503-0386-6

[28] B.G. Southwell and M.C. Yzer, “The roles of interpersonal communication in mass media campaigns,” in Communication Yearbook, C. Beck Ed., Lawrence Erlbaum Associates, New York, 2007, pp. 419-462.

[29] B. Suh, L. Hong, P. Pirolli, and E. H. Chi, “Want to be retweeted? Large scale analytics on factors impacting retweet in Twitter network,” The IEEE Second International Conference on Social Computing, Aug. 2010, pp. 177-184, ISBN: 978-1-4244-8439-3

[30] C. Tichi, “Electronic hearth: Creating an American television culture,” New York: Oxford University Press, 1991.

[31] J. Torrez-Riley, “The Social TV Phenomenon: New Technologies Look to Enhance Television’s Role as an Enabler of Social Interaction,” 2011, available at: http://www.torrezriley.com/projects/researchpaper/SocialTVpaper.pdf, [retrieved: September, 2015].

[32] D. Y. Wohn and E.-K. Na, “Tweeting about TV: Sharing television viewing experiences via social media message stream,” First Monday, vol. 16, 2011, pp. 3 – 7.

[33] S.Hill and A. Benton, “The Spoiler Effect?: Designing Social TV Content That Promotes Ongoing WOM,” The Conference on Information Systems and Technology (INFORMS), Oct. 2012.

[34] J. Jones, “Emerging platform identities: Big Brother UK and interactive multi-platform usage,” in Big Brother International: Format, Critics, and Publics, E. Mathijs and J. Jones, London: Wallflower Press, 2005, pp. 210-231, ISBN: 9781904764199

[35] J. Roscoe, “Multi-platform event television: Reconceptualizing our relationship with television,” The Communication Review, vol. 7, no. 4, 2004, pp. 363-369.

TABLE I. EFFECT OF VIEWERSHIP AND TIME ON ONLINE ENGAGEMENT: UNSTANDARDIZED COEFFICIENTS (STANDARD ERRORS).

	OE	**OT**	**AT**	**RT**	**RP**
Episode	3.110 (3.055)	-1.962 (1.934)	0.027 (0.048)	5.072 (1.754)**	0.297 (0.061)***
Minute	-0.218 (0.120) (*)[a]	0.163 (0.076)*	-0.005 (0.002)**	-0.381 (0.069)***	-0.014 (0.002)***
Viewership	6.076e-5 (0.000)***	5.522e-5(0.000)***	2.368e-7 (0.000)*	5.542e-6 (0.000)	1.508e-7 (0.000)
Constant	62.089 (39.298)	-105.969 (24.880)***	0.791 (0.616)	168.058 (22.564)***	4.285 (0.787)***
R^2	0.173	0.201	0.038	0.077	0.091
Adjusted R^2	0.171	0.200	0.036	0.075	0.089

a. Statistical significance: “***”*p<0.001; “**”p<0.01; “*”p<0.05; “(*)”p<0.1*.

TABLE II. EFFECT OF SOCIAL STRATEGIES AND CONTENTS ON OE: UNSTANDARDIZED COEFFICIENTS (STANDARD ERRORS).

	OE	**OT**	**AT**	**RT**	**RP**
Episode	1.679 (3.107)	0.469 (1.992)	0.140 (0.045)**	1.211 (1.691)	0.194 (0.061)**
Minute	-0.257 (0.129)*[b]	0.031 (0.083)	-0.010 (0.002)***	-0.288 (0.070)***	-0.013 (0.003)***
Viewership	7.181e-5 (0.000)***	4.815e-5 (0.000)***	-1.098e-7 (0.000)	2.366e-5 (0.000)***	4.842e-7 (0.000)**
Content 1 +					
Content 2	-89.090 (18.573)***	-47.754 (11.906)***	-0.224 (0.266)	-41.335 (10.109)***	-1.155 (0.366)**
Content 3	19.524 (18.890)	14.164 (12.109)	-0.100 (0.271)	5.360 (10.282)	-0.196 (0.373)
Content 4	-77.901 (61.385)	-40.600 (39.348)	-0.203 (0.880)	-37.301 (33.411)	-0.674 (1.211)
Content 5	151.100 (56.434)**	174.917 (36.175)***	0.887 (0.809)	-23.817 (30.717)	0.054 (1.113)
Content 6	13.422 (21.886)	17.141 (14.029)	0.369 (0.314)	-3.719 (11.913)	2.036 (0.432)***
Content 7	80.358 (60.334)	35.494 (38.675)	-0.580 (0.865)	44.865 (32.840)	0.499 (1.190)
Content 8	39.721 (80.547)	44.395 (51.632)	0.839 (1.154)	-4.674 (43.841)	0.789 (1.589)
Content 9	50.111 (15.708)**	-25.972 (10.069)*	0.061 (0.225)	76.083 (8.550)***	2.246 (0.310)***
Strategy 1	-14.593 (17.560)	3.858 (11.256)	4.317 (0.252)***	-18.451 (9.558) (*)	-0.889 (0.346)*
Strategy 2 +					
Strategy 3	14.274 (17.216)	14.213 (11.035)	0.321 (0.247)	0.061 (9.370)	0.150 (0.340)
Strategy 4	11.410 (14.599)	-3.317 (9.358)	-0.153 (0.209)	14.727 (7.946) (*)	-0.119 (0.288)
Strategy 5	-27.977 (14.578) (*)	-7.523 (9.345)	0.059 (0.209)	-20.455 (7.935)*	-0.460 (0.288)
Strategy 6	-63.054 (27.231)*	-27.674 (17.455)	-0.404 (0.390)	-35.379 (14.822)*	-0.522 (0.537)
Strategy 7	40.869 (24.504) (*)	25.961 (15.707) (*)	0.885 (0.351)*	14.908 (13.337)	0.518 (0.483)
Strategy 8	-130.203 (54.290)*	-38.319 (34.800)	-1.519 (0.778) (*)	-91.885 (29.550)**	-2.295 (1.071)*
Constant	12.204 (44.225)	-62.773 (28.349)*	2.117 (0.634)**	74.978 (24.071)**	2.736 (0.872)**
R^2	0.239	0.246	0.260	0.237	0.189
Adjusted R^2	0.228	0.235	0.249	0.226	0.177

b. Statistical significance: “***”*p<0.001; “**”p<0.01; “*”p<0.05; “(*)”p<0.1. +omitted*

TABLE III. EFFECT OF ORIGINAL TWEETS AND APP TWEETS ON RETWEETS AND REPLIES: UNSTANDARDIZED COEFFICIENTS (STANDARD ERRORS)

	RT	**RP**
Episode	5.740 (1.490)***[c]	0.315 (0.056)***
Minute	-0.456 (0.059)***	-0.016 (0.002)***
Viewership	-1.950e-5 (0.000)***	-5.006e-7 (0.000)***
OT	0.474 (0.022)***	0.012 (0.001)***
Constant	213.270 (19.279)***	5.461 (0.728)***
R^2	0.335	0.232
Adjusted R^2	0.333	0.229

	RT	**RP**
Episode	4.791 (1.727)**	0.289 (0.061)***
Minute	-0.343 (0.068)***	-0.013 (0.002)***
Viewership	4.393e-6 (0.000)	1.162e-7 (0.000)
AT	6.533 (1.023)***	0.196 (0.036)***
Constant	161.012 (22.237)***	4.073 (0.779)***
R^2	0.107	0.113
Adjusted R^2	0.104	0.110

c. Statistical significance: “***”*p<0.001; “**”p<0.01; “*”p<0.05; “(*)”p<0.1*.

Rumor Detection and Classification for Twitter Data

Sardar Hamidian and Mona Diab
Department of Computer Science
The George Washington University
Washington DC, USA
Email: sardar@gwu.edu, mtdiab@gwu.edu

***Abstract*—With the pervasiveness of online media data as a source of information, verifying the validity of this information is becoming even more important yet quite challenging. Rumors spread a large quantity of misinformation on microblogs. In this study we address two common issues within the context of microblog social media. First, we detect rumors as a type of misinformation propagation, and next, we go beyond detection to perform the task of rumor classification (RDC). We explore the problem using a standard data set. We devise novel features and study their impact on the task. We experiment with various levels of preprocessing as a precursor to the classification as well as grouping of features. We achieve an F-Measure of over 0.82 in the RDC task in a mixed rumors data set and 84% in a single rumor data set using a two step classification approach.**

***Keywords*–Rumor Detection and Classification; Supervised Machine Learning; Feature-based model.**

I. Introduction

Social media is currently a place where massive data is generated continuously. Nowadays, novel breaking news appear first on microblogs, before making it through to traditional media outlets. Hence, microblogging websites are rich sources of information which have been successfully leveraged for the analysis of sociopragmatic phenomena, such as belief, opinion, and sentiment in online communication. Twitter [27] is one of the most popular microblogging platforms. It serves as one of the foremost goto media for research in natural language processing (NLP), where practitioners rely on deriving various sets of features leveraging content, network structure, and memes of users within these networks. However, the unprecedented existence of such massive data acts as a double edged sword, one can easily get unreliable information from such sources, and it is a challenge to control the spread of false information either maliciously or even inadvertently. The information seeker is inundated with an influx of data. Most importantly, it is hard to distinguish reliable information from false information, especially if the data appears to be formatted and well structured [9] [24]. The problem is exacerbated by the fact that many information seekers believe that anything online in digital form is true and that the information is accurate and trustworthy; although, it is well known that a lot of the information on the web could be false or untrue. This is especially crucial in cases of emergencies. For example, by simply hitting the Re-tweet button on Twitter, within a fraction of a second, a piece of information becomes viral almost instantly. There are widely varying definitions of the term "rumor". We adopt the following definition of rumor: a rumor could be both true or false. A rumor is a claim whose truthfulness is in doubt and has no clear source, even if its ideological or partisan origins and intents are clear [2].

In verifying the accuracy of claims or events online, there are four major aspects that could be checked: *Provenance*, the original piece of content; *Source*, who uploaded the content; *Date-and-location*, when and where the content was created [22]. Analyzing each of these items individually plays a key role in verifying the trustworthiness of the data.

In this paper, we address the problem of detecting rumors in Twitter data. We start with the motivation behind this research, and then the history of different studies about rumors is overviewed in Section 2. Next, in Section 3, the overall pipeline is exposed, in which we adopt a supervised machine learning framework with several feature sets, and finally in Section 4, we compare our results to the current state of the art performance on the task. We show that our approach yields comparable and even superior results to the work to date.

II. Related Work

Psychologists studied the phenomenon of rumors from various angles. First studies were carried out in 1902 by German psychologist and philosopher, William Stern, and later in 1947 by his student Gordon Allport, who studied how stories get affected in their lifecycle [10]. In 1994, Robert Knapp published "A Psychol-

Copyright (c) IARIA, 2015. ISBN: 978-1-61208-443-5

ogy of Rumors", which comprised of a collection of more than a thousand rumors propagated during World War II. In his work, the rumor is what was transmitted by word of mouth and bore information about a person, an event, or a condition, which fulfilled the emotional desires of the public [11]. In 1948, Allport and Postman [12] studied the behavior of rumors and how one rumor reflects leveling, sharpening, and assimilation behavior in its propagation. Related studies in political communication conducted by Harsin [2] presented the idea of the "Rumor Bomb". For Harsin, a "Rumor Bomb" spreads the notion of the rumor into a political communication concept. In other research, Kumar and Geethakumari [5] explore the use of theories in cognitive psychology to estimate the spread of disinformation, misinformation, and propaganda across social networks. There are several studies about the behavior of misinformation and how they are distinguished in a microblog network. For example, Budak took a step further [4] investigating how to overcome the spread of misinformation by applying an optimized limitation campaign to counteract the effect of misinformation.

From an NLP perspective, researchers have studied numerous aspects of credibility of online information. For example, Ghaoui [3] detects rumors within specialized domains, such as trustworthiness, credibility assessment and information verification in online communication. Modeling and monitoring the social network as a connected graph is another approach. Seo [6] identifies rumors and their corresponding sources by observing which of the monitoring nodes receive the given information and which do not. Another relevant work, Castillo [23], applied the time-sensitive supervised approach by relying on the tweet content to address the credibility of a tweet in different situations. The most relevant related work to ours is that reported in [1], which addresses rumor detection in Twitter using content-based as well as microblog-specific meme features. However, differences in data set size and number of classes (rumor types) render their results not comparable to ours. Moreover, Qazvinian et al. [1] suggest label-dependent features in creating their User-based (USR) and URL features, which is only possible by having the input data labeled for being a rumor or not. In other words, labeled data is used for creating the language model (LM) with USR and URL features, and the trained LM is then used for extracting the value of each feature. In our study, we propose a totally label-independent method for feature generation that relies on the tweet content, and boosts our model in a realtime environment.

TABLE I. LIST OF ANNOTATED RUMORS [1]

Rumor	Rumor Reference	# of tweets
Obama	Is Barack Obama muslim?	4975
Michele	Michelle Obama hired many staff members?	299
Cellphone	Cell phone numbers going public?	215
Palin	Sarah Palin getting divorced?	4423
AirFrance	Air France mid-air crash photos?	505

III. APPROACH

We addressed the problem of rumor detection and classification (RDC) within the context of microblog social media. We focused our research on Twitter data due to the availability of annotated data in this genre, in addition to the above mentioned interesting characteristics of microblogging, and their specific relevance to rumor proliferation.

A. Data

Qazvinian et al. [1] published an annotated Twitter data set for five different 'established' rumors as listed in Table I. The general annotation guidelines are presented in Table II.

TABLE II. RUMOR DETECTION ANNOTATION GUIDELINES

0	If the tweet is not about the rumor
11	If the tweet endorses the rumor
12	If the tweet denies the rumor
13	If the tweet questions the rumor
14	If the tweet is neutral
2	If the annotator is undetermined

The following examples illustrate each of the annotation labels from the Obama rumor collection.

0: 2010-09-24 15:12:32 , nina1236 , Obama: Muslims 2019 Right To Build A Manhattan Mosque: While celebrating Ramadan with Muslims at the White House, Presi... http://bit.ly/c0J2aI

11: 2010-09-28 18:36:47 , Phanti , RT @IPlantSeeds: Obama Admits He Is A Muslim http://post.ly/10Sf7 - I thought he did that before he was elected.

12: 2010-10-01 05:00:28 , secksaddict , barack obama was raised a christian he attended a church with jeremiah wright yet people still beleive hes a muslim

13: 2010-10-09 06:54:18 , affiliateforce1 , Obama, Muslim Or Christian? (Part 3) http://goo.gl/fb/GJtsJ

14: 2010-09-28 22:22:40 , OTOOLEFAN , @JoeNBC The more Obama says he's a Christian, the more right wingers will say he's a Muslim."

2: 2010-10-05 17:37:04 , zolqarnain , Peaceful Islam- Muslims Burn CHURCH in Serbia: http://wp.me/p121oH-1ir OBAMA SILENT #politics #AACONS #acon #alvedaking #women #news #tcot

Table III shows statistics for the annotated tweets corresponding to each of the five rumors. The original data set as obtained from [1] did not contain the actual tweets for both Obama and Cellphone rumors, but they only contained the tweet IDs. Hence, we used the Twitter API for downloading the specific tweets using the tweet ID. Accordingly, the size of our data set is different from that of [1] amounting to 9000 tweet in total for our experimentation.

TABLE III. LIST OF ANNOTATED TWEETS PER LABEL PER RUMOR

Rumor	**0**	**11**	**12**	**13**	**14**	**2**	**Total**
Obama	945	689	410	160	224	1232	3666
Michelle	83	191	24	1	0	0	299
Palin	86	1709	1895	639	94	0	4423
Cellphone	92	65	3	3	3	0	166
Air France	306	71	114	14	0	0	505
Mix	1512	2725	2452	817	321	1232	9059

B. *Experimental Conditions*

We approached RDC in a supervised manner and investigated the effectiveness of multi step classification with various sets of features and preprocessing tasks versus a single step detection and classification approach. In the single-step classification for RDC, we performed detection and classification simultaneously as a 6-way classification task among the six classes in the labeled data, as shown in Table II, by retrieving the tweets as Not Rumor(0), Endorses Rumor(11), Denies Rumor(12), Questions Rumor(13), Neutral(14), and Undetermined tweets(2). In the two-step classification set up, an initial 3-way classification task is performed among the following groups of fine grained labels (0, Not Rumor), (2, Undetermined tweet), and the compound (11-14, Rumor) labels. This is followed by a 4-way classification step for the singleton labels, (11, Endorsing the Rumor), (12 Denys the Rumor), (13, Questions the Rumor), and (14, Neutral about the Rumor). In the second step, we took out class 0 and 2 tweets from the training data set and only classified the tweets from the test data set, which had been classified as rumor in the first step. The underlying motivation of our effort in designing the single-step and two-step classification is to investigate the performance of each technique in order to solve two problems. First, classifying tweets as ’Rumor’ and ’Not Rumor’, which can assist users to distinguish the type of tweets. Second, classifying the rumor type that the tweet endorses, denies, questions or is neutral. Although in both problems we investigated the rumor, these two problems are different. Our two-step model pipeline is dynamic in a way that the output of the the first step (Rumor Detection) is the input data set for the next step (Rumor Type Classification). We also designed a new set of pragmatic features along with updating the set of features in Twitter and network-specific category, which could boost the overall performance in our pipeline.

C. *Machine Learning Frameworks*

For our experiment we applied J48, a discriminative classifier that utilizes decision trees and supports various types of attributes. WEKA platform [25] is used for training and testing the proposed models in our pipeline.

TABLE IV. FINAL LIST OF USED FEATURES. ’*’ MARKED FEATURES ARE THE APPENDED SET OF FEATURES

	ID	Value
Twitter and Network Specific	* Time	Binary
	* Hashtag	Binary
	Hashtag Content	String
	URL	Binary
	Re-tweet	Binary
	*Reply	Binary
	User ID	Binary
Content	Content Unigram	String
	Content Bigram	String
	Pos Unigram	String
	Pos Bigram	String
Pragmatic	*NER	String
	*Event	String
	*Sentiment	String
	*Emoticon	Binary

D. *Feature Sets*

We experimented with content, network, and social meme features. We extended the number of features by including the pragmatic attributes. We employed all the features proposed in [1] in addition to developing more pragmatic attributes as well as additional network features. For network and meme features, we explicitly modeled source and timestamped information and for pragmatic features we proposed NER, Event, Sentiment, and Emoticon. Table IV lists all the features for the RDC task and marked the new features with “*”.

1) Content Features: This set of features is developed using tweet content. We applied various preprocessing granularity levels to measure the impact of preprocessing on the RDC task.

a) Unigram-Bigram Bag of Words (BOW): Similar to the content lexical features proposed in [1], we used a bag of words feature set comprising word unigrams and bigrams. We employed the WEKA's String To Word Vector along with N-gram tokenizer for creating this feature set with the TF-IDF weighting factor as the matrix cell content corresponding to each feature. We also generated the lemma form of the words in the tweets using WordNet [19] lemmatization capability. Accordingly, we created four feature sets: unigram tokenized word form, unigram lemma form, bigram tokenized word form, and bigram lemma form.

b) Part of Speech (POS): POS tagging for social media is challenging since the text genre is informal and quite noisy. We relied on the CMU Twitter POS tagger [7]. The feature values are set to a binary 0 or 1, corresponding to unseen or observed.

2) Pragmatic Features: In an extension to the features proposed by [1], we further explored the explicit modeling of pragmatic features to detect favorable and unfavorable opinions toward specific subjects (such as people, organizations). Applying this set of features offers enormous opportunities for detecting the type of rumors [21].

a) Sentiment: There are a wide variety of features for sentiment classification on Twitter data sets that have been studied in various publications. We believe that polarity of a tweet could be an informative factor to extract user's opinion about each rumor. For tagging the sentiment polarity of a tweet we applied the Stanford Sentiment system [18]. We preprocessed the data by removing punctuations, URL, "RT", and lowercased the content. Each tweet is tagged with one of the following sentiment labels; ***Very Positive, Positive, Neutral, Negative, or Very Negative***.

b) Emoticon: Another pragmatic cue is Emoticon. Studies on modeling and analyzing microblogs, which explicitly use emoticon as a feature, show its impact on classification [17]. We used the list of popular emotions described in Wikipedia [26]. We manually designated and labeled the list of entries as either expressing Positive (2), Negative (1), or Neutral (0) emotions.

c) Named-Entity Recognition (NER): We employed Twitter NLP tools [20] to explicitly extract information about named-entities, such as Location, Person, Organization, etc. In this paper we show how modeling NER has an explicitly positive impact on performance.

d) Event: Extracting the entity *Obama* and the event phrase *praises in connection with Muslims* is much more informative than simply extracting *Obama*. We utilized the same Twitter NLP tools [20] for tagging event labels.

3) Network and Twitter Specific Features: Relying on Twitter specific memes, we expanded features listed in [1] by adding time and network behavior features, such as Reply.

a) Time: It is quite remarkable that social networks spread news so fast. In a similar task to [13] we analyzed the process of rumor expansion on Twitter in our data set. Both the structure of social networks and the process that distributes the news lead to a piece of news becoming viral instantaneously. We labeled and ranked all the days based on the number of tweets posted in a day. We modeled the tweet creation-time attribute. We also observed that more than 90% of rumors are posted during the five most busiest days in the collected data set. Figure 1 shows the results of tabulating time frequency of the rumors in the Palin rumor data set and how the number of rumors changed within a six month period. Accordingly, we designated two labels for the time feature: Busy Day or Regular Day, depending on what type of day tweets were (re)tweeted.

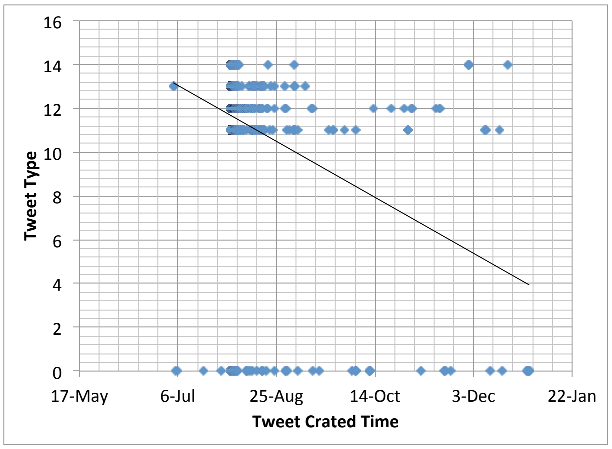


Figure 1. tweet Distribution for the Palin Rumor collection within a 6 month period

b) Reply, Re-tweet, User ID: Replying and retweeting in microblogs are revealing factors in judging user's trustworthiness when it comes to relaying information [14]. For example, *User A* is more likely to post a rumor than *User B* if *User A* has a history of retweeting or replying to *User C* who also has a rumor spreading history. Investigating the credibility of users is an expensive and almost impossible task, but it is doable when we only want to investigate a specific story. For example, knowing the limited number of users who have

TABLE V. NUMBER OF FEATURES AND LABELS USED IN SINGLE STEP AND TWO STEP CLASSIFICATIONS

1st Step		2nd Step
Method	Labels	Labels
(SRDC) 6-way classification	(0)(11)(12)(13)(14)(2)	
(TRDC) 3-way (1st step) — 4-way (2nd step) classification	(0)(2)(11-14)	(11)(12)(13)(14)

a history of posting rumors could be a hint to detect the large number of users that follow, retweet or reply to those tweets.

c) Hashtag: Hashtags serve as brief explanations of the tweet content [16]. We extracted hashtags in the labeled data set. Tweets with no hashtags are assigned a value 0 to their hashtag feature dimension, and tweets containing hashtag(s), received a 1 in the hashtag dimension. Additionally, we added all the observed hashtags as feature dimensions, thereby effectively identifying the tweets that share the same hashtag. For compound hashtags, we used a simple heuristic. If the hashtag contained an uppercase character in the middle of the hashtag word, then we split it before the uppercase letter. For instance, ***#SarahDivorce*** is separated into two hashtags and converted to **Sarah** and **Divorce**. We then modeled both compound and separated hashtags as hashtag feature dimensions.

d) URL: Twitter users share URLs in their tweets to refer to external sources as an authentic proof (a source of grounding) to what they share. All URLs posted in tweets are shortened into 22 characters using the Twitter t.co service. Analyzing the URL is an expensive task and requires a huge source of information to verify the content of the shared URL. We excluded all URLs but we modeled their presence as a binary feature.

IV. Experimental design

All the experiments are designed, performed, and evaluated based on various experimental settings and conditions, all elaborated in this section.

A. Data

We experimented with three data sets: the two largest rumor sets, Obama and Palin, and a mixed data set (MIX) which comprises all the data from the five rumors. We splited each of the three data sets into 80% train, 10% development, and 10% test.

B. Experimentation Platform

All experimentations were carried out using the WEKA-3-6 platform [25].

C. Baselines

We adopted two baselines: Majority and limiting the features to the set of features proposed in [1], which are Content, Hashtag-Content, URL, Re-tweet, and User ID. As the name indicates, the Majority baseline assigns the majority label from the training data set to all the test data.

D. Experimental Conditions and Evaluation Metrics

We had two main experimental conditions: single-step RDC (SRDC) and a two-step RDC (TRDC). We employed the set of 15 features listed in Table IV. Information about SRDC and TRDC is illustrated in Table V. In the development phase multiple settings and configurations were performed on the development data set for tuning, then the models that achieved the highest performance were used on the test set. Evaluating the performance of the proposed technique in rumor detection should rely upon both the number of relevant rumors that are selected (recall) and the number of selected rumors that are relevant (precision). Hence, we calculated F-measure, a harmonic mean of precision and recall due to its bias as an evaluation metric. TableVI shows the F-measure value for the different settings on the test set.

V. Results and Analysis

In this section the impacts of different experimental conditions are investigated.

A. SRDC and TRDC

By studying the results in Table VI, it can be observed that TRDC significantly outperforms SRDC, since TRDC achieves an F measure of 82.9% compared to 74% in SRDC for the MIX data set, and 85.4% for the Obama data set compared to a 71.7% in SRDC. By comparing the F-Measure with the Majority baseline and the features proposed in [1](VAR11) as the second baseline, we could explicitly see how applying the proposed methodology and set of features enhances the overall performance in certain rumors, and also leads to acceptable performance in the MIX data set.

TABLE VI. F-MEASURE RESULTS OF SRDC AND TRDC METHODS EMPLOYING 15 FEATURES AND VAR11 FEATURES

data set	Method	Our 15 Feat.	VAR11 Feat.
	Majority	0.30	
	SRDC	**0.743**	**0.748**
MIX	TRDC	**0.83**	**0.83**
	Majority	0.33	
	SRDC	**0.717**	0.705
1-3 Obama	TRDC	**0.854**	0.844
	Majority	0.46	
	SRDC	**0.754**	0.748
Palin	TRDC	**0.79**	0.70

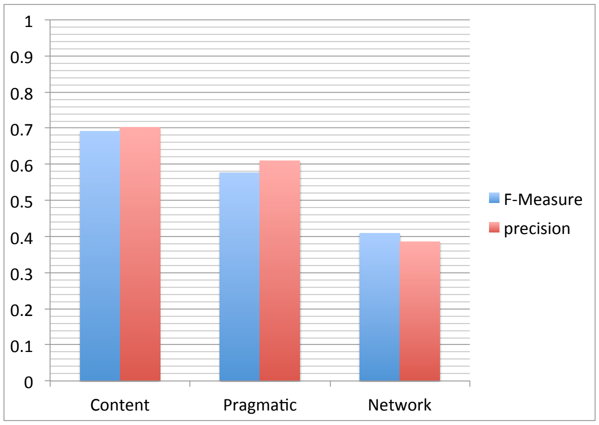


Figure 2. The Average F-measure and precision of SRDC and TRDC classifications employing each group of features: Content, Pragmatic, Network

B. *Impact of Feature Set*

In this experiment, we assessed the performance of different groups of features individually. Figure 2 shows the average F-measure and precision of SRDC and TRDC by employing the Content, Pragmatic and Network sets of features. As shown in Figure 2, employing the Content set of features yields the overall best precision. In contrast to the other features, the network feature set had the minimum impact on our classification.

C. *Impact of Preprocessing*

As mentioned above, we applied various levels of preprocessing to the content of tweets such as stemming, lemmatization, punctuation removal, lowercasing, and stop words removal. We measured the impact of applying such preprocessing versus no preprocessing. Figure 3 illustrates that accuracy doesn't benefit from preprocessing and results in the loss of valuable information.

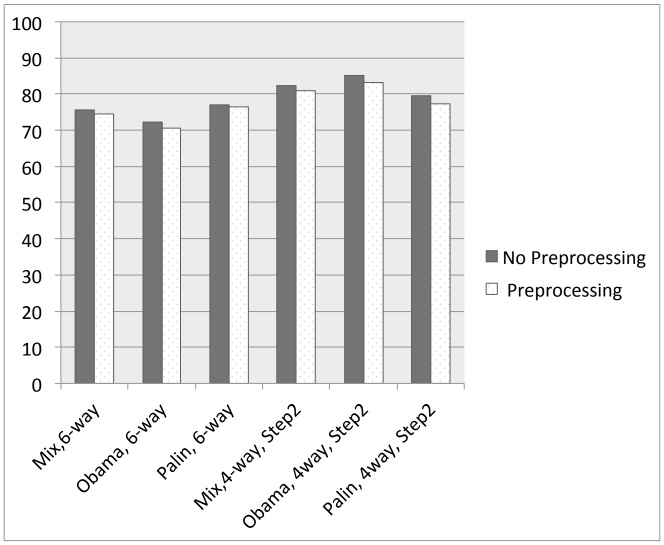


Figure 3. The overall accuracy in different experiments with and without preprocessing

VI. CONCLUSIONS AND FUTURE WORK

In this paper, we study the impact of a single-step (SRDC) 6 way classification versus a two-step classification (TRDC). Our contributions in this paper are two-fold: (1) We boosted the pipeline by decoupling the rumor detection from the classification task. We proposed an automated TRDC pipeline that employs the results from the rumor detection step and performs the classification task upon data and leads to promising results in comparison to SRDC. (2) We employed a new set of meta linguistic and pragmatic features, which leads and performs the experiments with and without preprocessing on the textual content. We achieved the F-Measure of more than 0.82 and 0.85 on a mixed and the Obama rumor data sets, respectively. Our proposed features achieved better performance compared to the state of the art features proposed in [1]. Our study however suggests that our pipeline does not benefit from preprocessing which might be attributed to the weakness of the tools used for processing twitter content at this stage. We are planning to expand the proposed methodology to streaming tweets. Having a limited amount of labeled data, we are investigating means of augmenting the training data with noisy data in a semisupervised framework.

ACKNOWLEDGEMENT

This paper is based upon work supported by the DARPA DEFT Program.

REFERENCES

[1] V. Qazvinian, E. Rosengren, D. R. Radev, and Q. Mei, “Rumor has it: identifying misinformation in microblogs,” In Proceedings of the Conference on Empirical Methods in Natural Language Processing EMNLP, Association for Computational Linguistics, Stroudsburg, PA, USA, 2011, pp. 1589-1599.

[2] J. Harsin, “The Rumour Bomb: Theorising the Convergence of New and Old Trends in Mediated US Politics,” Southern Review: Communication, Politics and Culture. Issue 1, 2006, pp. 84-110.

[3] C. Ghaoui, “Encyclopedia of Human-Computer Interaction,” Encyclopedia of human computer interaction. IGI Global, 2005.

[4] C. Budak, D. Agrawal, and A. E. Abbadi, “Limiting the spread of misinformation in social networks,” Proceedings of the 20th international conference on World wide web, ACM. 2011, pp. 665-674.

[5] C. Budak, D. Agrawal, and A. E. Abbadi, “Detecting misinformation in online social networks using cognitive psychology,” Human-centric Computing and Information Sciences. 2014, pp. 4-14.

[6] E. Seo, P. Mohapatra, and T. Abdelzaher, “Identifying rumors and their sources in social networks, SPIE Defense, Security, and Sensing,” International Society for Optics and Photonics, 2012.

[7] O. Owoputi, O. Connor, B. Dyer, C. Gimpel, and K. Schneider, “Part-of-speech tagging for Twitter: Word clusters and other advances,” School of Computer Science, Carnegie Mellon University, Tech, 2012.

[8] C. Silverman, “Verification Handbook: A Definitive Guide to Verifying Digital Content for Emergency Coverage,” European Journalism Centre, 2014.

[9] C. N. Wathen and J. Burkell, “Believe it or not: Factors influencing credibility on the Web. Journal of the American Society for Information Science and Technology,” 53, no 2, 2002, pp. 134-144.

[10] T. Takahashi and N. Igata, “Rumor detection on Twitter,” Soft Computing and Intelligent Systems (SCIS) and 13th International Symposium on Advanced Intelligent Systems (ISIS), 2012, pp. 452-457.

[11] R. H. Knapp, “A Psychology Of Rumor,” Public Opinion Quarterly, 1944, pp. 22-37.

[12] G. Allport and J. Postman, “Psychology of Rumor,” Russell and Russell. 1951, p. 750.

[13] K. Starbird, J. Maddock, M. Orand, P. Achterman, and R. Mason, “Rumors, false flags, and digital vigilantes: misinformation on Twitter after the 2013 Boston marathon bombing,” iSchools, 2014.

[14] T. Kurt, G. Chris, S. Dawn, and P. Vern, “Suspended Accounts in Retrospect: An Analysis of Twitter Spam,” Proceedings of the ACM SIGCOMM Conference on Internet Measurement Conference, 2011, pp. 243-258.

[15] A. Hassan, V. Qazvinian, and D. Radev, “What’s with the attitude?: identifying sentences with attitude in online discussions,” In Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing Association for Computational Linguistics, Oct. 2010.

[16] R. Deveaud and F. Boudin, “Effective tweet Contextualization with Hashtags Performance Prediction and Multi-Document Summarization,” Initiative for the Evaluation of XML Retrieval (INEX), July. 2013, pp. 1245-1255.

[17] A. Agarwal, B. Xie, I. Vovsha, O. Rambow, and R. Passonneau, “Sentiment analysis of Twitter data,” In Proceedings of the Workshop on Languages in Social Media. Association for Computational Linguistics, June. 2011. pp. 30-38.

[18] S. Richard, A. Perelygin, J. Y. Wu, J. Chuang, C. D. Manning, A. Y. Ng, and C. Potts, “Recursive deep models for semantic compositionality over a sentiment treebank,” Proceedings of the conference on empirical methods in natural language processing (EMNLP). vol. 1631, 2013, pp. 1642.

[19] G. A. Miller, WordNet: A Lexical Database for English, Cambridge, MA: MIT Press, Communications of the ACM Vol. 38, No. 11, pp. 39-41.

[20] A. Ritter, S. C. Mausam, and O. Etzioni, Named entity recognition in tweets: an experimental study. In Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP 11). Association for Computational Linguistics, Stroudsburg, PA, USA, 2011, pp. 1524-1534.

[21] T. Nasukawa and J. Yi, “Sentiment analysis: capturing favorability using natural language processing,” In Proceedings of the 2nd international conference on Knowledge capture, ACM. 2003, pp. 70-77.

[22] C. Silverman, The Poynter Institute, “Verification Handbook,” European Journalism Centre.

[23] C. Castillo, M. Mendoza, and B. Poblete, “Predicting information credibility in time-sensitive social media”, Internet Research. 23 (5), 2013, pp. 560-588.

[24] R. S. Young and N. J. Belkin, “Understanding judgment of information quality and cognitive authority in the WWW,” In Proceedings of the 61st Annual Meeting of the American Society for Information Science. ASIS, 1998, pp. 279-289.

[25] H. Mark, F. Eibe, H. Geoffrey, P. Bernhard, P. Reutemann, and I. H. Witten, The WEKA data mining software: an update. SIGKDD Explor. Newsl. 11, 1 (November 2009), 10-18. DOI=10.1145/1656274.1656278 http://doi.acm.org/10.1145/1656274.1656278. Sep. 2015.

[26] Wikipedia contributors. “List of emoticons,” Wikipedia, The Free Encyclopedia. Sep. 2015.

[27] www.twitter.com, Sep. 2015.

Helping Hands of Autism Blogger Community in Social Media Platforms

Amit Saha
Center for Distance Health
University of Arkansas for Medical Science
Little Rock, Arkansas, U.S.A.
email: asaha@uams.edu

Nitin Agarwal
Maulden-Entergy Chair Professor of Information Science
University of Arkansas at Little Rock
Little Rock, Arkansas, U.S.A.
email: nxagarwal@ualr.edu

Abstract— **With a high prevalence of autism among children, there is a shortage of autism support facilities around the world. Families dealing with autism use online social media to share experiences with other members of the community. Systematic analysis of the vast interaction between autism community members in blogs and Twitter can be used to build a learning tool for others who are dealing with autism. The study found that the autism blogger community provides substantial social support to other community members. Differences across various groups (autistic bloggers, mother bloggers with autistic kids, father bloggers with autistic children, and autism support group blogs) and different social media platforms (blogs and Twitter) were reviewed in context of social support. We found that the families dealing with autism have a better quality of life when facilitated with social support by community members.**

Keywords-Autism; ASD; Social Support; Twitter; Blogger; Community.

I. INTRODUCTION

According to estimates by the Centers for Disease Control and Prevention (CDC) approximately 1 in 68 in the USA are diagnosed with Autism Spectrum Disorder (ASD). ASD occurs in all the ethnicities and boys are four to five times more likely to have autism as compared to girls [1]. Intervention and efficient treatment such as specialty services and early detection can help people with Autism to lead a better quality of life [2].

Social media has provided Internet users an open platform for discussions, communication, and information exchange for various health related topics. Families with a member diagnosed with autism share their life experiences in social media (almost on a daily basis). This exchange of information, which is by default archived, has become an immense source of knowledge for others dealing with the same situation. Organizations working on spreading autism awareness encourage creation of an open social media platform on autism where members can share their experiences and get advice from others. Shared experience by an individual dealing with autism in the social media platform, especially blogs, Twitter and Facebook shed light on various issues of autism. To raise awareness among autism community members, premier non-profit organization like Autism Speaks [3] recognizes top autism bloggers based on feedback from families with autism. Shared know-how about autism helps to find a better way of life for the families dealing with autism. Social support in online platform can be defined as "information leading the subject to believe that he is cared for and loved, esteemed, and is a member of a network of mutual obligations" [4].

The purpose of this study is to offer a research-based understanding of the conversations in social media platforms especially blogs and Twitter among families dealing with autism. The study aims to shed light on characteristics of social support provided by autism blogger community towards other members of the community.

This paper is organized as follows: the prior related works are described in Section II, Section III depicts the methodology and data collection, Section IV shows the result, Section V discusses about the inferences drawn from the study; finally, in Section VII, we draw conclusions and possible future works.

II. RELATED WORK

Many clinical studies are ongoing to get an in-depth knowledge of causes and effective interventional strategies for ASD. These studies provide understanding of the outcome of various available therapy options for autism. With the high cost involved in clinical trials, the use of social media content in research analysis to assess the effectiveness of different intervention strategies for autism could be an economically viable option. Caregiver's preference for using social media platform as compared to any other communication platform was also established by Hamm, et al., [5]. Our study does not intend to provide a substitute for clinical tests of the intervention strategies. On the contrary, our methodology would provide the perceived effectiveness of the intervention strategies or the therapies from a practitioner’s perspective. This would include clinically evaluated as well as unevaluated strategies. This in turn would help prioritizing resources on the testing procedures of intervention strategies. In this study, however, we address a tiny part of this bigger research agenda, which is does autism blogger community provide social support to other community members? Are there differences in the offered social support across various groups of bloggers (i.e., autistic bloggers, mother bloggers with autistic kids, father bloggers with autistic kids, and autism support group blogs) and different social media platforms (e.g., blogs and Twitter)?.

Answers to these questions will help conduct a more systematic evaluation of interactions occurring on various online platforms, especially the social media, for evaluating the efficacy of intervention strategies from the perspective of the practitioners.

Sociologists published many research works on the social support concept. The link between social support and health is addressed by two different hypothesis: the buffering hypothesis and the direct effects hypothesis. In the buffering effect hypothesis, social support enhance good heath by reducing the effect of stressful life events [6] while in direct effect hypothesis, better health is provided by high social support [7].

Hamm et al., [5] found in their study that caregivers and patients started using social media to gather health information and exchange information related to health informatics. The mostly publicly available social media data facilitated by various healthcare communities can be analyzed effectively to build a knowledge-based source.

III. Methodology and Data Collection

Social support concept, although widely studied in the social science literature, there lacks a formal mathematical definition. Hence, we leverage various empirical definitions available in computational science literature that overlap with or application domain, i.e., healthcare.

This study evaluates the social support using sentiments expressed by the interactions of the autism bloggers community. The social network analysis features are used to evaluate the structural aspect of the generation of the social support within the autism community. Our methodology consists of the following steps:

1) Collect data from Autism bloggers in different social media platform (blogs and Twitter).
2) Pre-process and filter noise.
3) Perform topic and word analysis to ensure the subject of discussion is autism.
4) Construct networks for autism community for each social media platform.
5) Analyze the sentiment of the content of the interaction of the members of autism blogger community.
6) Calculate the degree of social support provided by the interaction of the autism community members.

Web search on autism keyword shows, presently there are more than a thousand active autism bloggers. For the initial phase of the study, we selected the top 40 autism bloggers based on the recommended list of popular bloggers by the Autism Speaks organization. The content and metadata of blogs by the 40 autism was extracted and analyzed. Further, we cross-referenced their blogger profile and Twitter profile (wherever the blogger had provided a link to his/her Twitter profile) and collected their tweets, and other network information, including friends and followers.

We retrieved the most recent permissible tweets (up to 3,200 each) for the 40 autism bloggers, resulting in 118,531 tweets. All the tweets are in the English language. Some of the tweets by autism bloggers are as follows, *"Autism is part of what we are. Neither good nor bad. Being unable to talk is a problem to be solved. Absolutely ", "I am a strong believer in developing therapies and tools to help people severely disabled by autism. Always believed that..".*

Profile analysis of the bloggers led to the classification of bloggers based on different characteristics. Classification of autism bloggers into different categories is done to deduce different capacities of social support based on defined blogger categories. Of the 40 autism bloggers, 13 were female bloggers with autistic kids who are termed as mothers. Male bloggers with autistic children termed as fathers are 10 in our database. Number of bloggers who blogged as groups to create autism awareness termed as autism support group are 13 and rest 4 termed as self-autistic bloggers who are diagnosed with autism and blogs for themselves.

To infer social support from the text content, the psycholinguistic analysis was used. De Choudhury [8] studied the online exchange of social support for health communities on depression using the psycholinguistic analysis technique. Blog content and tweets are categorized into psychological groups utilizing Linguistic Inquiry and Word Count (LIWC) program [9]. In LIWC, a word can belong to more than one groups. LIWC has been used by many researchers for text analysis, and promising results have been reported. Tov and Ng [10] found a consistent correlation between emotion rating values of LIWC with self-reported values.

LIWC categories used to infer social support were selected based on resemblance with social support concepts like the social process, which signifies feeling of solidarity. The definition proposed by Cobb [4] for online social support is used as the reference. The definition is quoted in Section 1. Social support in the text content is deduced primarily using the scale in the spoken category of assent along with positive emotion and social processes.

IV. RESULTS

To get an insight into social support characteristics of autism blogger community, various social network analysis techniques were used during the study. To infer network dimension of the autism blogger community, the activities like tweets, friends, followers, mentions, the hashtag of autism bloggers in Twitter and blogs were analyzed.

The autism blogger community interaction shows tightly linked community. The friend and follower Twitter network of autism blogger community is shown in Figure 1. The autism bloggers are annotated based on the classification defined in Table 1. Their real identity is anonymized. Different colors indicate various communities based on network modularity. For complex network structure, modularity is one of the effective function in community detection [11]. Figure 1 shows distinct characteristics of the autism blogger network on Twitter where any member of the community can reach a colleague on average 3.4 hops (average geodesic distance), as compared to the widely known 4.74 degrees of separation in Facebook network of active users [12].

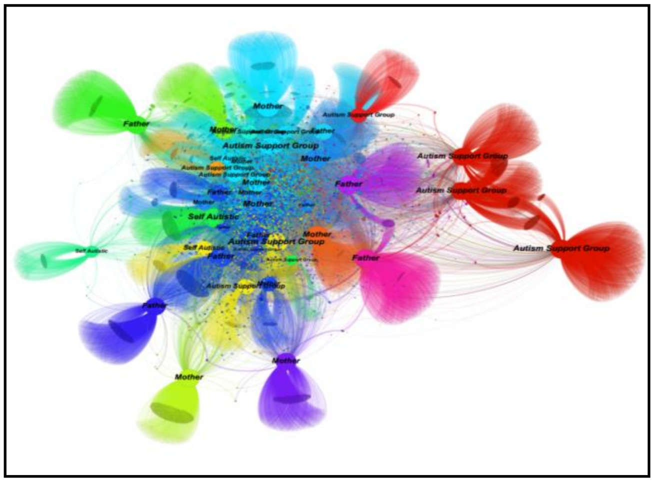


Figure 1. Friend and follower network of autism blogger community. Colors indicate different communities based on the network modularity.

The modularity of the Twitter network of autism bloggers is 0.623, which indicates that the community is well connected. The network is partitioned into 22 communities using Chinese whisper clustering algorithm [13]. The top hashtags of the autism blogger network are '*autism*', '*autismawareness*', '*autismhoops*' and '*specialneeds*' that indicate the network is highly focused on autism-based discussions. Wordpress, a very well-known platform for blogging, along with blogspot.com are found to be the top domains shared among tweets. Overall analyzed metrics of the Twitter friends and followers network is shown in Table 1.The tweets content of the autism bloggers found to be involved in many topics related to autism and the aim of the autism community bloggers' Twitter network seems to be spreading autism awareness.

TABLE 1. OVERALL TWITTER DATA CHARACTERISTICS OF THE AUTISM BLOGGER COMMUNITY NETWORK.

Metric	**Friends and Followers Network**	**Metric**	**Friends and Followers Network**
Number of users	48030	Out-Degree	2011 (Max), 1.137 (Average)
Total Edges	98099	Connected Components	1 with 35787 Maximum Vertices
In-Degree	2018 (Max), 1.88 (Average)	Geodesic Distance (Diameter)	5 (Max), 3.408 (Average)

Tweets collected and analyzed for the study shows signs of sentiments associated. "*I also really liked this positive post from about when and how to tell kids about their autism diagnoses*", "*The study means we are one step closer to understanding how one of the key components of autism happens in the brain*" are some of the examples of tweets with positive sentiments.

Based on the author characteristics of the autism bloggers and choice of social media platform, the positive sentiment in the text varies. LIWC provides the baseline values for the control writing, science articles and conversion medium. Chuang and Yang [14] in their study on online alcohol community found the presence of social support. Based on Chung and Yang [14] work we used the social support provided by the alcohol support forum as the baseline value to infer the presence of social support.

Major findings: Our study found average positive feeling of autism blogs is much higher as compared to another support forum like alcohol support community or by the general conversation medium. Moreover, mother of an autistic kid shows much more positive emotion on Twitter as compared to the father of an autistic child or autistic bloggers. For the autism support forum, the negative sentiment was evaluated to be a lot lower as compared to control writing and science articles. The negative attitude was quite higher for the father of an autistic kid, as compared to other categories of bloggers (i.e., mother and autistic bloggers).

Social support provided by the textual interaction was estimated using the scores of assent, positive sentiment, and social process in the text content. Our study found that in Twitter, mothers of autistic kids provide the maximum amount of social support as compared to other autism blogger categories. The autism bloggers' community in Twitter and blogs as a whole provided high social support as compared to other forms of text writing such as emotional writing, scientific articles, etc. Further, social support provided by talking or verbal communication is quite higher as compared to all other modes of communication, including online forums and social media. This finding demonstrates that there is no substitute for the verbal communication medium. However, social media and other online communication media could possibly fill the gaps, wherever verbal communication with an expert is instantly unavailable.

V. DISCUSSION

The study sheds light on characteristics of social support provided by online support community of autism on the different social media platform. The social support provided by interactions within autism blogger community by identifying the bloggers and the community members was unfolded in the study.

The tightly knit interaction within the autism blogging community was revealed in our study. Members of the autism community provide extensive social support to its members, by sharing information and extending emotional support. Members of the autism bloggers community in Twitter and blogs provide high social support as compared to other health groups like alcohol support forum.

For the tweets of the mothers, the amount of social support provided is higher than fathers or autistic bloggers with a given amount of positive emotion, but the ratio is highest in verbal communication as compared to any written text interaction. Figure 2 shows the variation of social support with positive emotion. Verbal communication provides the maximum amount of social support for a given value of positive emotion in contrast to other communication modes.

Statistical analysis of our model for social support determination based on various psychological groups of LIWC scores shows impressive results. For the mother and father with autistic kids, the value of social support provided given the score of positive emotion came out to be statistically significant (p <=0.001). Positive emotion is highly correlated with social support. Figure 2 also depicts the high correlation between positive emotion and social support. The correlation, however, is not monotonically increasing, which means beyond a certain degree of positive emotion, social support provided is unchanged.

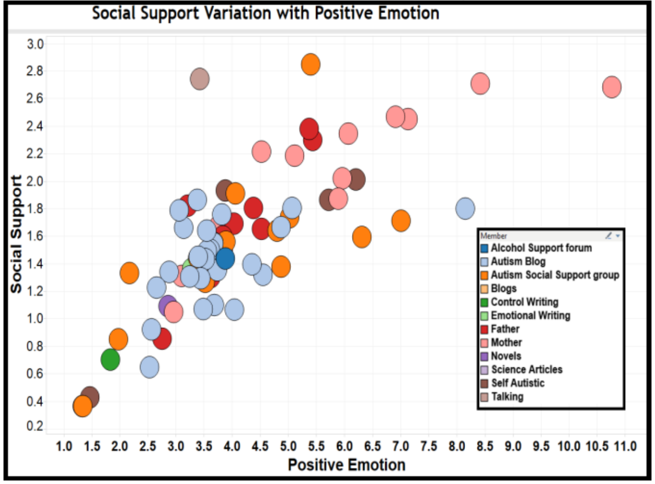


Figure 2. Social support variation with positive emotion. Vertical Axis shows the amount of Social Support and Horizontal axis represents positive emotion. Each circle represents a data point.

VI. CONCLUSION

In this research, we study the online social support provide by interaction among members of the autism blogger community in different social media platforms. The study extracts blogging activity of popular autism blogger and their Twitter activity including their friends, followers, tweets, retweets, mentions, and hashtags information. The tightly knit interaction within the autism blogging community was identified in our study. Our study found that Autism blogger community provides extensive social support to its community members in different social media platforms, especially on Twitter. The autism community members share a feeling of solidarity by providing support to other community members empathetically on social media. While negative sentiments are reflected in some tweets, the social support contributed by the autism blogger community overwhelmingly outweighs the negativity.

Social support provided by autism bloggers varies based on blogging characteristics and social media platform. Whether bloggers influence played a role in evoking social support within the readers/audience is an interesting question, which could be studied. Agarwal et al. [15] in their research suggested that influential bloggers are more likely to encourage discussions among the community members through comments in blogs or retweets in Twitter.

We envision our study will provide a mechanism to access social support in online health communities. However, the fact that autism bloggers also use other social media platforms, such as Facebook presents a limitation in our study. The findings of this study lay the groundwork to study our bigger research agenda, i.e., evaluating the efficacy several of therapies for ASD as perceived by the caregivers through the experiences they have shared in online forums and social media. This will help build a knowledge base for interventions and experiences, which in turn could assist the clinical research in better understanding of behavioral interventions for various health disorders.

ACKNOWLEDGEMENTS

This research was funded in part by the Jerry L. Maulden/Entergy Fund at the University of Arkansas at Little Rock. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the funding organization. The researchers gratefully acknowledge the support.

REFERENCES

[1] Centers for Disease Control and Prevention (CDC).. “Community Report on Autism”, http://www.cdc.gov/ncbddd/autism/states/comm_report_autism_2014.pdf/. [Online; accessed 10-June-2015].

[2] V. B. Gupta, et al., “Identifying children with autism early?” Pediatrics 119(1):152–153, 2007.

[3] Autism Speaks, http://www.autismspeaks.org (Online; accessed July 14, 2015).

[4] S. Cobb, "Presidential Address-1976. Social support as a moderator of life stress," Psychosomatic Medicine, vol. 38, 1976, pp.300-314.

[5] M. P. Hamm, et al., “Social media use among patients and caregivers: a scoping review”, BMJ open 3(5),2013.

[6] J. E. Wallace, "Job stress, depression and work-to-family conflict: A test of the strain and buffer hypotheses." Relations Industrielles/Industrial Relations (2005): 510-539.

[7] L. F. Berkman and L. Breslow, ”Health and Ways of Living: The Alameda County Study”, Oxford University Press, New York, 1983.

[8] M. De. Choudhury, S. Counts, E. J. Horvitz, and A. Hoff, “Characterizing and Predicting Postpartum Depression from Facebook Data”, ICWSM 2014, Baltimore, MD, 2014.

[9] Linguistic Inquiry and Word Count (LIWC) program http://www.LIWC.net (last visited May 14, 2015).

[10] W. Tov, K. L. Ng, H. Lin, and L. Qiu, “Detecting Well-Being via Computerized Content Analysis of Brief Diary Entries”, Journal of Personality Assessment, vol 25(4), 1069-1078, 2013. doi: 10.1037/a0033007.

[11] M. E. Newman, "Modularity and community structure in networks." Proceedings of the National Academy of Sciences 103.23 (2006): 8577-8582.

[12] L. Backstrom, P. Boldi, M. Rosa, J. Ugander, and S. Vigna, “Four degrees of separation”, In Proceedings of the 4th Annual ACM Web Science Conference (WebSci '12). ACM, New York, NY, USA, 2012, pp.33-42.

[13] C. Biemann, “Chinese Whispers - An Efficient Graph Clustering Algorithm and its Application to NLP Problems”. In Proceedings of TextGraphs, New York, USA ,2006, pp.73–80.

[14] K. Chuang and C. Yang, “Informational Support Exchanges on Different Computer-Mediated Communication Formats in a Social Media Community of Alcoholism,” Journal of the American Society for Information Science and Tech., vol.65, no.1, 2014, pp.37-52.

[15] N. Agarwal, H. Liu, L. Tang, and S. Y. Philip, “Modeling blogger influence in a community”, Social Network Analysis and Mining, 2(2), 2012, pp.139-162.

Sharing Our Heritage to Shape Our Future

How Effective Are Multi-User Sharing Platforms in Supporting Collaborative Visioning for the Future, and Why is Heritage Centre-Stage?

Andrea Nanetti
School of Art, Design and Media
Nanyang Technological University
Singapore
e-mail: andrea.nanetti@ntu.edu.sg

Anna Simpson
Futures Centre
Forum for the Future
Singapore
e-mail: a.simpson@forumforthefuture.org

***Abstract*—This is a position paper about heritage and futures that the authors present for discussion to set the stage for a subsequent analysis of the topic supported by statistics and formal proofs. In social eco-informatics, a significant, challenging, not yet solved problem comes from the fact that the global Internet community is entering the age of collaborative and generative design: not of a central object, but of complex systems. People are looking around them, spotting patterns of connection, extrapolating as trends converge, questioning the implications, and coming together in groups to reimagine the future. The recognition that we need to come together to shape our common futures does not just coincide with the rise of social media: it is thanks to social media that this consciousness has reached such scale. Now, social networking sites are being created with the intention to help people across the world talk about the future they want and co-design it. In this process, heritage, seen as the treasure trove of human experience, is found to be centre-stage. We do not know what the next generation will value and like, but we can display and discuss what humans valued and liked and why, using the results to make better decisions. The practice of sharing heritage can become the context and the method for collaborative design. This paper inquires how effective these new platforms are in shaping visions for the future, and what difference an emphasis on sharing heritage makes to their success. And, in conclusion, the paper underlines how comprehensive and systematic use of heritage (seen as the treasure of human experiences) becomes the key science for sustainable and dynamic innovation in the coming anthropogenic era, during which human activity is becoming the dominant influence – not only in climate and the environment, but also in the human genetic and epigenetic heritages evolutionary processes.**

Keywords—multi-user sharing platforms; collaborative visioning for the future; heritage science

I. INTRODUCTION

We have entered the age of collaborative design, not of a central object, but of complex systems. People are looking around them, spotting patterns of connection, extrapolating as trends converge, questioning the implications, and coming together in groups to reimagine the whole, 'a crude look at the whole', as Murray Gell-Mann would say [1].

There is an on-going shift in consciousness beyond an individualist mindset towards greater awareness of the collective, connected nature of our lives in our global Internet community. The understanding that “no man is an island” (John Donne, Meditation XVII) [2] is expressed in the ancient philosophy of the Tao – where human life is seen as “an integral feature of the world process, and not as something alien or opposed to it” [3]. But with the rise of liberalism, recognition of our interdependence became more an abstract theory than a lived experience: you might have known you were in tune with high-street trends, for instance, but believed it was a matter of personal choice, intoxicated with the idea of self-determination [4].

Today, things are different: we are confronted with the influence of wider forces shaping our lives on a minute-to-minute basis, thanks to the omnipresence of accessible and useful data, targeted advertising, and social media. Targeted communications and advertising run on the irony that people can be made to feel intimately known, *precisely because* their behaviours and interests can be mapped to wider trends and thereby anticipated. Behavioural researchers are asking “how modern collective behaviour may be changing in the digital age, including whether behaviour is becoming more individualistic, as people seek out exactly what they want, or more social, as people become more inextricably linked, even “herd like,” in their decision making” [5].

Some find the shift in our collective consciousness unnerving, as Cameron Tonkinwise, Director of Design Studies at Carnegie Mellon University, explains [6]:

> I think everyone feels the disjuncture between what big data is telling us – “I discover that I am part of a data set through the ads in my buzz feed and the way they have been tailored to me" or literally an article saying "You are a Gen Xer and you have no healthcare" – and our phenomenological experience of the world.

In response to this rising awareness of the forces at play, communities are coming together to shape their future. It might be interpreted as a reassertion of collective agency [7]. They are going beyond the traditional work of consumer-focused design, which focuses on the relation of subject to object. Their focus is how communities relate to commonalities [8].

Similarly, designers—from brand strategists to web programmers—are putting user experience at the heart of their processes. They recognise that we cannot recreate systems—such as cities, food cycles, and community energy schemes—without the volition and participation of their future users. The same goes for pensions, health plans, transport, and schools [9].

To reiterate, the recognition that we need to come together to shape our common futures does not just coincide with the rise of social media: it is thanks to social media that this consciousness has reached such scale. And so it is no coincidence that social sites are enabling people across the world to co-design their future spaces. The transition began with innovative companies recognising the opportunities in social media for co-creation [10]. But increasingly, communities of interest are collaborating on social platforms to take the design of their own futures in hand.

Take Visionmaker [11]: a New York-based platform that allows communities to rethink precise urban areas, projecting into the future. Individuals can select an area of the metropolis and a timeframe (for instance, Manhattan in 2050), and specify factors such as land use (from vegetable garden to street trees to retail complex), energy sources (e.g., waste-to-energy power plant), water source, storage and usage, and transport types. They can also select various 'precipitation events'—such as drought, or showers—and observe how their system fares. Crucially, designs can be shared and discussed with other visionaries or contributors.

Only some urban planners are thinking about what might social design sites, such as this, actually change. For an analogy, take the design of an energy efficient building. An architect can optimise natural light and ventilation all they like, but if the inhabitants choose to live with the curtains drawn for privacy and the lights on, the features are wasted. Similarly, our future will depend only in part on our external conditions: we must also design our future lifestyles.

This is something the creators of Visionmaker have worked into the fabric of their site: “Beyond changing the ecosystems of an area of interest, users are able to change the Lifestyle and Climate scenarios that inform their visions.” A lifestyle scenario—the site explains—will affect the consumption and waste generation patterns of the people living in the area, as well as factors such as take-up of various modes of transportation, and subsequent energy use.

Communities are also establishing themselves around social platforms to design global systems, beyond the bounds of single geographies or industries. We might even paraphrase the Latin author Cicero in his famous quote—dated June 46 BCE—"if you have a garden and a library too, you are not missing anything" (*Si hortum in bibliotheca habes, deerit nihil*), saying "if you have a garden and a 'connection to the internet' too, you are not missing anything" [12].

One example of such a social platform, where humanity's garden meets its vital knowledge-sharing infrastructure, is 'The Future of Protein'. This online sharing site is at the heart of The Protein Challenge 2040: a consortium convened by Forum for the Future and involving the World Wildlife Fund, the Global Alliance for Improved Nutrition, Firmenich, Hersheys, Quorn, Target, Volac and Waitrose.

This is the first international innovation partnership to explore how we balance supply and demand of protein for a growing population, in a way that is affordable, healthy and good for the environment. The stakeholders are now sharing their perceptions of change and their long-term implications via Forum's public trends-monitoring platform, the Futures Centre, in a topic hub called 'The Future of Protein' [13]. Users spot 'signals of change', from 'Nutrient shakes replace meal breaks in Silicon Valley' to 'Industry stakeholders call for debate on insects as food and feed'. They're also raising questions, such as 'What will it take to meet China's growing demand for pork?'

These questions converge around the theme of intangible cultural heritage. Can Westerners overcome their aversion to eating grubs? Will the attraction of meat continue to rise among Asia's growing middle class? Why are social mealtimes losing their value among tech entrepreneurs? Our assumptions about the world we want to create—its founding myths—are shaped by our cultures. In designing the future, we are always choosing the elements of our current culture that we want to carry forward with us. In this conscious or subconscious selection process, we are also designing what might be inherited by the future generations in terms of both active cultural politics and digitally archived latent legacy factors.

The paper is structured in nine sections (including this introductory one and the conclusion) followed by an acknowledgment of an epiphany and bibliographical references. Section 2 presents heritage as the treasure of human experience, which can pioneer a new science in the social media era. Then—after having introduced Singapore as a privileged view point in Section 3—some exemplary less privileged case studies are presented in the three following sections (Greece in Section 4, Syria in Section 5, and Rwanda in Section 6) to open the stage for Section 7 and Section 8 that propose a complex-network approach to heritage data analysis as the core methodology, with which heritage science can support multi-user sharing platforms in developing effective and sustainable collaborative visioning for the future. In the paper, it has been chosen to avoid a systematic review of online social media and platforms that could relate to the topic, because other conference papers are focusing on them in detail both in CENTRIC 2015 - The Eighth International Conference on Advances in Human-oriented and Personalized Mechanisms, Technologies, and Services and SOTICS 2015 - The Fifth International Conference on Social Media Technologies, Communication, and Informatics (Barcelona, Spain, November 15-20, 2015) [14].

II. Heritage: Treasure and Foundation of Human Experience

For the first time ever, our society has the technological capacity to organize and retrieve myriad records of human experience at any time and from any digitally connected place, and apply these elements of this treasure trove to any kind of future design process. We are able to select from the past and apply to the future in full consciousness that responsibility belongs only to those present, in their shaping of the past, and is heavily interdependent of the news flow. [15].

"Heritage poses the challenge of innovation in a new way: how does the new integrate with the old?" This was the key question raised by Helga Nowotny (co-founder and former president of the European Research Council) in her keynote address at the 1st Singapore Heritage Science Conference on *Heritage science as a complex system*: *The embarrassment of complexity: A phase of transition?*. The conference was organized and chaired by Andrea Nanetti and Siew Ann Cheong at Nanyang Technological University Singapore for the Complexity Program and the School of Art, Design and Media (6-7 January 2014), to pioneer a new science of heritage, as a state-of-the-art multidisciplinary domain able to investigate and discover integrated action plans and solutions in response to, and in anticipation of, the challenges arising from cultural heritage issues in society. In this way, heritage is closely linked to the history and identity of communities [16].

This new science of heritage focuses on accessing, interpreting, conserving and managing cultural heritage. It takes into account knowledge and values acquired in all relevant disciplines, from arts and humanities (philosophy, ethics, (art-)history, economics, sociology and anthropology) to fundamental sciences (chemistry, physics, mathematics and biology), as well as computer sciences and engineering, and media studies. In fact, the 2nd Singapore Heritage Conference "Heritage and the Creative Industry", held on 15-16 January 2015, wrestled with the tensions between age-old practices and our modern digital lifestyles. In particular, there was a sense that we might be losing our humanity, as our lives become more and more digital. In hearing experts—like Harold Thwaites in his keynote—talk about their past experiences, and draw from them creative inspirations for the future, one could realize that human qualities like ethics, empathy, identity, and spirituality are connective qualities that serve to bind people together. In short, to be human is to be connected to other humans, to our environments, and for some, to cosmic significance [17]. On that same note, significantly, Steve Dixon—in his talk given for the online international symposium *Art of the Networked Practice* chaired by Vibeke Sorensen and Randall Packer at the Nanyang Technological University Singapore for the School of Art, Design and Media on April 1, 2015—used the metaphor of existentialism to speak about commitment and engagement in contemporary networked practices [18].

The ethics of reporting current affairs has received much attention, but the emphasis has been on the present-day implications and the need to assume responsibility for past events: a stance shaped by notions of justice rather than an acknowledgement of any design agency in the practice of reporting itself. More attention is now due to the ethics of shaping the future, and the politics of heritage-selection as part of this. In this way, the social media experience is entering that same field of 'intangible cultural heritage' that UNESCO (and with it 153 countries) defines as [19]:

> … the practices, representations, expressions, knowledge, and skills—as well as the instruments, objects, artifacts and cultural spaces associated therewith—that communities, groups and, in some cases, individuals recognize as part of their cultural heritage. This intangible cultural heritage, transmitted from generation to generation, is constantly recreated by communities and groups in response to their environment, their interaction with nature and their history, and provides them with a sense of identity and continuity, thus promoting respect for cultural diversity and human creativity.

Thus, according to the UNESCO charter, the preservation of intangible cultural heritage requires the active collaboration of the people or community within which the heritage resides. This in turn requires protection of the processes that allow traditions and shared knowledge to be passed on to future generations, along with arts, science, problem solving and invention.

The 4th Singapore Heritage Science Conference (25-26 January 2016) will approach the interpretation of the nature of changes in intangible cultural heritage through the lens of complexity science: despite the fact that the roles of innovation and creativity in shaping the intangible cultural heritage of complex societies are not yet described in the literature, we believe that, in essence, these changes are voluntary human expressions of ingenuity. To build up a complexity science toolkit for incorporating innovation and creativity into the core processes of a living heritage, the role of multi-user sharing platforms will be crucial: these will both enable and build awareness of collaborative visioning for the future [20].

III. Singapore as a Privileged View Point

In Singapore this year (2015), the decision process for its future direction is very much a conscious one. Singapore was 'born' as a nation state 50 years ago, and has advanced largely (although not exclusively) to the plans of its first Prime Minister, the late Lee Kuan Yew. Because of the cultural dimension of urban planning—as addressed by Liu Thai Ker, former director in Singapore of the Urban Redevelopment Authority and now Chairman of the Centre for Liveable Cities in his lecture given on September 18, 2015, in the Chinese Heritage Centre at Nanyang Technological University Singapore—in a careful and attentive design, Singapore's urban development story offers not only a unique view point for an integrated engineered use of the built heritage but also an advantaged perspective for intangible cultural heritage appreciation [21]. In fact, many are now taking the opportunity to look another 50 years into the future. One sharing platform, Imagine2065, invites communities to redesign key social spaces. Among three spaces showcased, two are to be found on the map: Telok Ayer community centre and the youth centre *Scape. The third—the 'mothership'—invites designs for online shared spaces, with various aims: to channel funding for social projects, to offer a safe space for pre-formed thoughts, to raise difficult questions, such as 'Should new college graduates immediately get corporate jobs?'

Visions for both community centres illustrate the active process of heritage-design. Members of the Telok Ayer community call for "a living museum that reflects the organic and dynamic stories of the community" and "a gallery where community groups can showcase themselves", while the youth centre is imagined as "a blank canvas for youths to express themselves".

Imagine 2065 is the creation of The Thought Collective, a group of social enterprises with the shared aim to "build up

Singapore’s social and emotional capital” – beginning with young people. Founder Tong Yee explains [22]:

> We truly believe that in any form of solution, if you do not understand the context, you are not going very far. If you look at a young person, you have to ask, ‘How do I give you the context that will allow you to be a much more strategic thinker and partner in the future?’

The need for young people to be at the heart of future developments has been powerfully advocated by one of Singapore’s most respected strategists, the former Foreign Minister, George Yeo, an early blogger and Facebook user. At the launch of his newly published collection of essays and public talks, Yeo observed a “much healthier" level of civic participation in Singapore today, and put it down to two factors: the Government "letting go" to some extent, and new technology like social media "allowing sunlight to break through". He was referring to a speech he made in 1991, in which, as the Straits Times reports [23]:

> He used the analogy of the state as a banyan tree, calling for it to be pruned to let the sun through so that the undergrowth - or civic society - will not be stultified.

George Yeo, in his new book, writes [24]:

> As hierarchies give way to networks, it is younger members of society who adapt the most readily … If we fail to engage and involve the young, the transition from a hierarchical to a network society will be a troubled one.

The ‘motive force’ for this transition, Yeo claims, is the digital revolution:

> The new technologies unleashed by [it] enable better and higher forms of human organisation to emerge but not before old ones are brought down. The search for new pathways to that future is the story of today.

Heritage is at the heart of his understanding of that search. It was Yeo that persuaded the Chinese Chamber of Commerce to restore the Sun Yat Sen Villa, transforming it into a museum commemorating the founding father of the Republic of China, who visited Singapore nine times between 1900 and 1911 [25]. The restoration was not only a tribute to Singapore’s Chinese heritage, but a statement of Yeo’s intent to maintain strong economic ties with China in future. Yeo was also one of the founders of a project to revive Nalanda University, Bihar’s ancient seat of learning, and is now its new Chancellor, taking over from the Indian-American economist Amartya Sen, a fellow advocate of the power of debate and author of ‘The Argumentative Indian’ [26].

IV. GREECE: A CASE OF HERITAGE IN ACTION

Since 2008, after Wall Street imploded, Greece became the epicentre of Europe’s debt crisis. An article published by Suzanne Daley and Anemona Hartocollis in *The New York Times* in response to the Greek referendum of 5 July 2015 made the interesting claim that the ‘No’ was influenced by a “history of defiance”, founded in Greek history as illustrated here below in Figure 1.

The authors cite Nick Malkoutzis, the editor of *Macropolis. Greece in Perspective*, a political analysis website [27], in saying: “It is true that deep in the Greek psyche is the idea of glorious resistance against all odds. Moments of defiance, he added, are “written into the conscience of every Greek”. Instances used to justify this understanding of the Greek psyche are handpicked from both history and mythology, from legends of women throwing themselves off cliffs to escape slavery, to the iconic figure of the Spartan king Leonidas remembered for his death at the Thermopylae, to Greek fighters blowing themselves to avoid being captured by the Ottomans in the 1800s [28].

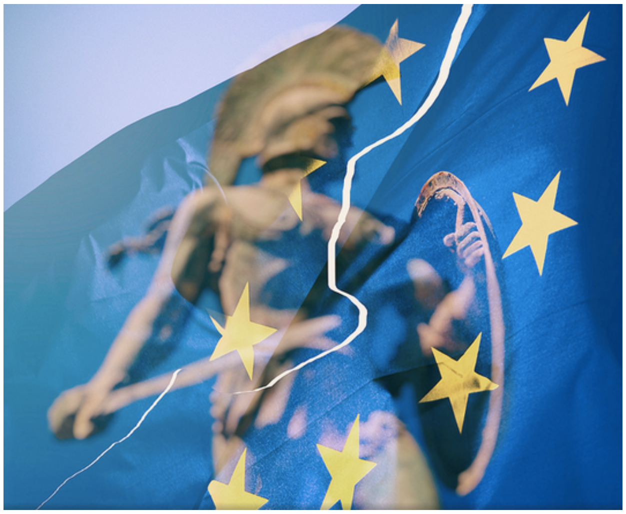


Figure 1. Greek "No" to the Troika and king Leonidas' legacy [29].

If Figure 1 shows a clear example of (ab)use of heritage to (emotionally) convey political statements, otherwise based on assumptions, Figure 2 shows the retired Greeks waiting to receive partial pension payments outside of a bank in Athens, who can hardly get relief by that “history of defiance” founded in Greek history, as described by Nick Malkoutzis. Our vision is that multi-user sharing platform can start to make the difference today in a strong collaboration between computer science, humanities, and news broadcasting tools (e.g., Apple News as firstly realised with iOS 9 in September 2015 [30], and Facebook Instant Articles [31]).

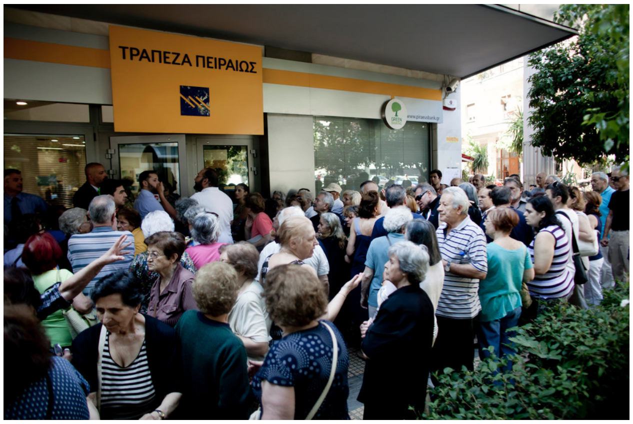


Figure 2. Photograph by Eirini Vourloumis for The New York Times.

The comparison between Figure 1 and Figure 2 provides a first example for what has been claimed here above in Section 3. Even if it is not a formal proof, yet, it shows the way towards it accomplishment: a comprehensive and systematic analysis and use of historical information would become here a key tool for sustainable and dynamic innovation in news broadcasting and appreciation. Let us think about a platform, which could automatically access the historical evidence and the many sub-sequential (politically and culturally informed/biased) interpretations related to the Spartan king Leonidas and visually match them with the key elements of Nick Malkoutzis' narrative and with the economic scenarios envisaged by the so called troika (European Commission, European Central Bank, and International Monetary Fund) using the methodology mentioned here below in Section 7.

The process of defining what is, and what has been, influences both our present day, our collective sense of the past, and our aspirations to the future. Our aspirations are formative, and therefore in shaping them, we also have a part in shaping the world to come.

The better we understand what's changing today, the greater chance we have of shaping the future we want. Futures practitioners (including the team at Forum for the Future) monitor the 'megatrends' shaping our lives—such as population growth, climate change and hyper connectivity—and collect signals of change: unprecedented instances of behaviours, technologies, designs and applications that could open new directions for current systems. These stories of ongoing and sudden change are used to create potential scenarios of different futures, so that organisations, sectors, industries or nations can make better decisions today.

Decisions are based on how we judge – both rationally and emotionally – various scenarios, and these judgements are based on assumptions of what is desirable and valuable. The stories we tell shape these assumptions. Today, some of the most influential storytellers—Disney, for instance [32]—condition young children to value riches over rags and conformity over difference: its happy endings are defined by wealth, marriage and conventional beauty.

Thus, if nations are founded upon narrative myths, and intergenerational repetition perpetuates the assumptions and values they contain, as demonstrated in 2011 by Caspar Hirschi [33], heritage can be understood as the system through which we—as communities, organisations and states—can choose these myths and share them.

Now we have the technology to implement a heritage-based better understanding of the future implications (on social, economic and environmental fronts) of the selections and (re)shaping of our mythical foundations. In the case of Greece, one of the most important outcomes of this approach—paraphrasing 90-year-old Jimmy Carter's interview with National Geographic—would be filling vacuums and things that governments do not do [34].

The vision is also to increase the participatory responsibility of citizens in the democratic processes of decision-making. The theoretical point has always been made both in philosophy and political science.

One could recall Plato's *Socrates' Apology*:

> The unexamined life is not worth living (Ὁ δε ἀνεξέταστος βίος οὐ βιωτὸς ἀνθρώπῳ, 38a).

Life is not worth living without ἔλεγχος/*elenchus*, i. e., without examination, argument of disproof or refutation, dialogue; cross-examining, testing, scrutiny especially for purposes of refutation. Such is the Socratic elenchus, often referred to also as *exetasis* or scrutiny and as *basanismus* or assay [35].

One could recall August Wilson, who signed both the Declaration of Independence and the Constitution, preached startlingly democratic theories - more democratic than the ideas of any other delegate to the Constitutional Convention. In his oration at Philadelphia on July 4, 1788, celebrating the adoption of the Constitution of the United States, he said [36]:

> You are responsible for the world that you live in. It is not the government's responsibility. It is not your school's or your social club's or your church's or your neighbour's or your fellow citizen's. It is yours, utterly and singularly yours.

One could recall the Greek composer Yorgos Tsangàris (1948-2008) [37]:

> There is no time for yourself if you want to become a Man (Δεν υπάρχει χρόνος για τον εαυτό σου αν θέλεις να γίνης Άνθρωπος).

But, it seems that only today we are having a new great opportunity: using heritage-based multi-user sharing platforms we could support collaborative visioning for the future and—as Marten Sheffer would say—"create safe operating space to avoid the loss of resilience that prepare the advent of crisis" [38].

V. SYRIA: FROM NEWS FEEDS TO MULTI-USER SHARING PLATFORMS

A key question today is about the support that technology and crowd-sourcing can offer to help us undertake the process of selecting and sharing our heritage collectively, as part of futures design.

One approach comes from News Deeply, an online media platform that aims to offer readers the cultural, political and historical context to current affairs, revalorising a systemic understanding of events – as opposed to the exclamatory hype of the tabloid press. It began with a news site focused exclusively on Syria [39], which features a timeline, conflict map and detailed sections covering 'the basics' since the March 2011 uprising, the regime, an understanding of the Islamic State of Iraq and the Levant/ash-Sham/Syria (better known as ISIS), a summary of key global players, and a reading room of leading thinkers.

Two elements, arguably, are missing: one is a section elaborating the heritage of Syria. This would enable the reader to conceptualise the scale of loss – both in terms of deepened empathy for the people and in imagining the damage to societal cohesion, as well as the loss of sites of cultural significance. The other missing element is the opportunity for users to contribute their own observations.

They can currently comment on shared stories, but not share their own.

Were these two desirable functions to be combined—creating a section in which users could contribute their understanding of Syria's pasts (Greek-Roman-Byzantine, Arab, Crusader, Ottoman, French, Independent) as well as the current civil war started in 2011—it would create a rich opportunity for discussion as to ways forward. Could the careful use of taxonomy in tagging the stories shared even allow users to study patterns of values, as represented in the heritage, and relate these to desired social outcomes? Our solution is proposed here in Section 7 and Section 8.

VI. Rwanda: A Shared Approach to narrative interpretation

Heritage as a means for collaborative design demands a shift away and a reorganization of traditional disciplines (history, anthropology, archaeology, etc.) towards a new awareness of collective sharing and interpretation.

An example of what this new awareness might draw upon in its development is offered by Dave Snowden, Founder and chief scientific officer of Cognitive Edge, with SenseMaker® [40], which allows people to act as their own ethnographers for complex analysis of their lived experiences.

Of course, anthropologists have always collected stories from the studied group; the difference in Snowden's approach is that the people not only share their experiences but also index them. His system means that each individual does not only have a voice but also the means of interpretation.

It is essentially a software tool that allows users to contribute a narrative and then asks them a set of questions to analyse what they have shared. This data is used to map their interpretation of their story against other users' interpretations of their own stories. The framework for this map might be a simple triangulation, or a grid. The SenseMaker® site explains:

> The output of SenseMaker is statistical data backed up by explanatory narrative. This means that advocacy is an integral part of the system. Numbers on their own appear objective but are not persuasive; anecdotes on their own may be persuasive but are not objective. […] "Instead of asking, "How do we create a culture of X?" we ask "How do we create more stories like this and fewer stories like that?"

Crucially, in sensitive contexts and for difficult subject matter, the interpretation can be made public and mapped, without individual's story being revealed. The system also allows people to share their experience in their own language and yet offer data that can be used widely, without costly translation. This enables the system to operate at significant scale – potentially even national or global.

To give an example, GirlHub—a Nike Foundation project—ran a pilot project with SenseMaker® to study the impact of its 12+ Young Empowerment Programme in Rwanda [41]. In Figure 4, the SenseMaker® matrix maps the descriptions caregivers gave of the intentions and actions of girls in the stories they shared. Caregivers were asked to share a true story about an experience in a girl's life. They were then asked whether, in their story, the girl wanted to do something, did not want to do something [mapped across the horizontal axis] did something, or did not do something [mapped on the vertical axis]. The researchers then focused their interest upon those stories where 'wanted to do something' met 'did something', to discern patterns that could lead to recommendations for future empowerment. For instance, they noted that, "The areas where girls and parents were viewed as having the most decision-making ability were education and health".

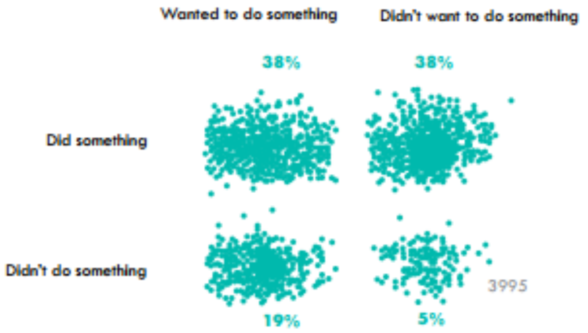


Figure 3. Pilot study conducted in 2015 in Rwanda by GirlHub.

A shared approach to narrative interpretation calls for the application of such a system for personal histories to develop a new collective heritage mapping as a foundation for possible sustainable futures as presented in the two following Sections 7 and 8.

VII. A complex-network approach to heritage data analysis

In the words of John H. Holland [42] and William Brian Arthur [43], complex systems are characterized by their dependence on contingencies; for example, event B happens because event A happened in the past, but not event C. These contingencies, also called path dependences, make the study of global histories highly bewildering, because of the concatenation of conditional probabilities [44].

Seeing history not as a linear progression of events, but as a complex, nonlinear network of contingencies gives us the correct frame of mind to respond to the issues raised by William A. Green in his article on world history periodization [45]:

> Periodization is rooted in historical theory. It reflects our priorities, our values, and our understanding of the forces of continuity and change. Yet periodization is also subject to practical constraints. For pedagogical reasons, world historians must seek reasonable symmetry between major historical eras despite huge discrepancies in the availability of historical data for separate time periods and for different areas of the world.

It also addresses to the issues raised by microhistory [46], which studies well-defined single historical units/events to ask—as defined by Charles Joyner—"large questions in small places" in contrast with large-scale structural views [47]. The most famous example being Carlo Ginzburg's *Il* *formaggio e i vermi* first published in Italian in 1976 [48]. In the book, which is considered to have initiated this research field in historical studies, the author wrote:

> The historians have long since learned that history is the history of men, not of the "great," and the closer you get up to everyday reality the better you decipher the past, and then grasp the sense of immediacy with the problems, the connections with today's present, i.e., history.

The solution proposed by Andrea Nanetti and Siew Ann Cheong to delimit time periods is to move away from focusing on main individual events, but to look instead at intensity in the flow of events in societies' natural nonlinear perception of time.

Figure 3 (left) shows the six W's of a narrative, and how they relates to the key actors, key events, key periods, key locations, key factors, and key actions in historical analysis. In Figure 3 (right) different link types (blue, green, and red, and possibly others as well) of socio-political relations (trade, diplomacy, conflicts, etc.) connect historical actors (a, b, c, d, e, f, g, h, i, j) in a complex network, from which, among other higher-level information, we can extract the power blocs shown here below as the clusters A and B.

Concerted effort and funding are now needed to enable such projects to proliferate.

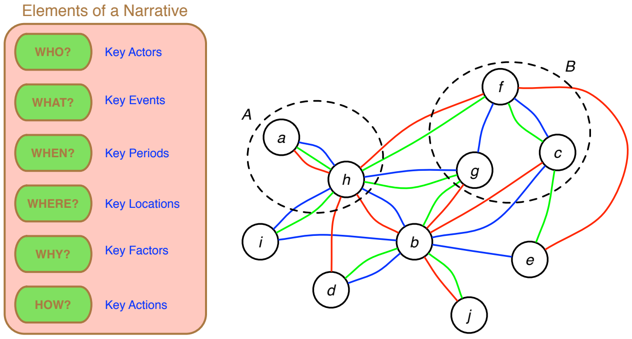


Figure 4. The 6 W's of a narrative as a complex network

Marten Scheffer's research team recently used eco-systems as a showcase to point out that complex systems theory associates regime shifts as critical transitions with higher intensities of events [49]. Building on his insights, Siew Ann Cheong and Andrea Nanetti decided to mine the complex network of intercontinental trade, diplomacy, conflicts and other interactions among cities, nations and continents during Late Middle Age and Early Renaissance (1205-1533 CE) to identify time and geography of such transitions. This time-resolved complex-network approach to historical analysis allows to automatically identifying the *key actors* driving a specific key event. At present, expert historians painstakingly piece together the key events, and thereafter examine the main actors in such events. With that method, this identification can be automated and different definitions of key actors adopted [50].

This methodological approach can easily move from the historical landscape to the social media historical memory, as intangible cultural heritage of humanity (see the UNESCO definition in Section 2).

VIII. Engineering Social Media Historical Memory as Intangible Cultural Heritage of Humanity

Engineering Historical Memory [51] is an experimental methodology and an on going research project for the organization of historical data in the digital age, that Andrea Nanetti theorized when he was Visiting Scholar at Princeton University in 2007, and further developed it when he was Visiting Full Professor at the University of Venice Ca' Foscari in 2012. Since 2013 he is carrying on the research at Nanyang Technological University Singapore at the intersections of humanities and data science/visualisation. The project was awarded best conference paper at 2013 Culture and Computing (Kyoto, Japan), and has been funded by Microsoft Research and Microsoft Azure (2014-2016).

Engineering Historical Memory is helping to develop heritage studies as a science in response to, and in anticipation of, the exponential growth of knowledge—encoded/embodied in complex interactions of written, pictorial, sculptural, and architectural records, oral memories, practices, and performed rituals—in our *glocal* society (i.e., reflecting or characterized by both local and global considerations). What sets it apart from other approaches is a focus on developing and applying computationally intensive techniques (e.g., pattern recognition, data mining, machine learning algorithms derived from other disciplines, and visualization solutions) to achieve this goal. It entails the creation and advancement of databases (relational, graph, and hybrid), algorithms, computational, statistical, and complexity techniques and theories to solve formal and practical problems arising from the study, interpretation, conservation, and management of cultural heritage data in the context already presented in Section 2.

The basic problem has been clearly framed by Larry Page in his TED's talk *Where's Google going next?* given on 21 March 2014 [52]:

> Google mission is to *organise* world's information and make it universally *accessible* and *useful*. People keep on asking: "Is it what you guys are still doing?" I think at it on myself and I am not quite sure about what to answer. Actually, when I think about *search*, it is such a deep thing for all of us: to really understand what you want, to understand the world's information... And we are still very much in the early stages of that. And it is totally crazy! We have been out for 15 years already, but it is not at all done.

In September 2015, Apple welcomed to its News App using the following advertisement [53].

> The best stories from sources you love, selected just for you. The more you read, the more personalised your News becomes.

To start the application the user is required to select a list of preferred sources. But there is no tool to cross the information and validate the single news.

On 23 September 2015, Wired published an article by Julia Greenberg referring to Facebook 360 videos in News Feed [54].

> "Over time the types of stories that people want to tell each other and the types of content they want to share with each other will get richer and more immersive" Facebook's VP of product Will Cathcart says. "So just as we have seen an evolution from text to photos, we are seeing a pretty big jump to video in the last couple of years. We think that's only going to continue.

From a media perspective, the challenge is to have a system that works on a visual base to be tested in two parallel experiences: one with the scholars (historians and art historians interested in data mapping and visualization) to investigate as deeper as possible at a global cultural scale the concepts of 'provenance' and 'validation' of the sources and their interpretations; and another in the social media to approach the actual shift from texts, to photographs, videos, and 360 immersive spaces in the community sharing processes.

IX. Conclusion and Future Work

> *It is true, for the first time in history all peoples on earth have a common present: no event of any importance in the history of one country can remain a marginal accident in the history of any other. Every country has become the almost immediate neighbour of every other country, and every man feels the shock of events, which take place at the other side of the globe. But this common factual present is not based on a common past and does not in the least guarantee a common future. Technology, having provided the unity of the world, can just as easily destroy it and the means of global communication were designed side by side with means of possible global destruction*
>
> (Johanna Arendt 1906-1975, Man in Dark Times, 1955/1968, p. 83)

As observed in Sections 3 to 6, the potential to share, select and index our heritage, and to use this as a tool in designing our future, has reached unprecedented scale. Heritage is considered as the *thesaurus* of human experiences (i.e., the comprehensive storage system of human knowledge and values) embedded in human artefacts and in nature as interactively experienced by different cultural communities, and biologically perceived by the human brain. In this way, heritage issues become the key-factor for innovation in the Anthropocene, the incoming era, during which human activity is becoming the dominant influence not only in climate and the environment but also in the human genetic and epigenetic heritages evolution.

In Section 7—following the solution proposed by Andrea Nanetti and Siew Ann Cheong in 2013—, has been proposed to use a complex-network approach to heritage data analysis as the core methodology, with which heritage science can support multi-user sharing platforms in developing effective and sustainable collaborative visioning for the future. We do not know what the next generation will need, value, and like, but we can display and discuss what humans needed, valued, and liked—and the reasons why they did it—using the results to make better decisions. The following Section 8 demonstrated how social media could become the repository of possible solutions, whether organized, treated, and investigated through the lens of a new upgrade of the methodologies developed by the century-old scholarly tradition of historical sciences.

In that way, this contribution can be used for theory and/or practice to develop ICT tools able to aggregate heritage data into news and social media platforms in order to increase the awareness of existing connections and possible future scenarios in a more scientific way.

All in all, heritage is not about the past but about a shared future: we believe that sharing and discussing our heritage via multi-user sharing platforms can support collaborative visioning for the future and towards a new art and science of living together on Earth.

Aknowledgment

The conception of this paper would not have been achieved without the continuous scholarly dialogue that since March 2013 Andrea Nanetti has been having with Siew Ann Cheong at Nanyang Technological University in the Complexity Program directed by Jan Vasbinder and later on in 2014 in the newly established Complexity Institute directed by Peter Sloot and Stephen Lansing. A key event was also the conference "Nurturing Leonardos" held at the ArtScience Musuem Singapore in May 2015, where the authors met and started to discuss about the topic of this paper.

References

[1] Available from http://www.bu.edu/pardee/9-27-05/ [accessed on 12 October 2015]. See: J. H. Miller, A Crude Look at the Whole: The Science of Complex Systems in Business, Life, and Society. New York: Basic Books, 2016.

[2] J. Donne, "Meditation XVII". [Online]. Available from: http://www.online-literature.com/donne/409/ [accessed on 12 July 2015].

[3] See the Introduction in A. Watts, Tao: The Watercourse Way. New York: Random House, 1975.

[4] J. Christman, "Liberalism and individual positive freedom," Ethics, vol. 101, no. 2, January 1991, pp. 343–359.

[5] R. A. Bentley, J. M. O'Brien, and A. W. Brock, "Mapping collective behavior in the big-data era," Behavioral and Brain Sciences, vol. 37, issue 1, February 2014, pp 63–76.

[6] Anna Simpson, Interview with C. Tonkinwise [conducted on 15 April 2015," unpublished.

[7] H. Haugh, "Community-led social venture creation," Entrepreneurship Theory and Practice, vol. 31, issue 2, March 2007, pp. 161–182, doi:10.1111/j.1540-6520.2007.00168.

[8] K. Alexiou, G. Alevizou, T. Zamenopoulos, S. de Sousa, and L. Dredge, "Learning from the use of media in community-led design projects," Cultural Science, vol. 8/1, Jan. 2015, pp. 30–40.

[9] G. Voss and N. Carolan, "User-led design in the urban domestic environment," Journal of Urban Technology, vol. 19, no. 2, 2012, pp. 69–87. Available from: http://www.unialliance.ac.uk/blog/2015/02/12/changing-healthcare-for-the-better-through-user-led-design-research [accessed on 12 July 2015].

[10] From Social Media to Social Product Development: The Impact of Social Media on Co-Creation of Innovation. [Online]. Available from: http://papers.ssrn.com/sol3/Papers.cfm?abstract_id=1975523 [accessed on 12 July 2015].

[11] Available from: https://visionmaker.us/nyc [accessed on 12 July 2015].

[12] M. Tulli Ciceronis, Epistularum ad Familiares, Liber IX, Ep. IV.

[13] Available from: http://www.thefuturescentre.org/topic-hubs/protein [accessed on 12 July 2015].

[14] Available from: http://www.iaria.org/conferences2015/ProgramCENTRIC15.html and http://www.iaria.org/conferences2015/SOTICS15.html [accessed on 26 July 2015].

[15] A. Nanetti, S. A. Cheong, and M. Filippov, "Interactive Global Histories. For a new information environment to increase the understanding of historical processes," in Proceedings of the International Conference on Culture and Computing 2013 (Kyoto, Ritsumeikan University, Sept. 16-18, 2013). Los Alamitos, CA: IEEE Computer Society, 2013, pp. 104–110, doi:10.1109/CultureComputing.2013.26.

[16] The complete videorecorded conference is avilable from: http://www.paralimes.ntu.edu.sg/Pages/Home.aspx.

[17] The complete videorecorded conference is avilable from: http://www.paralimes.ntu.edu.sg/NewsnEvents/HeritageandtheCreativeIndustry/Pages/Video%20Gallery.aspx.

[18] The complete videorecorded symposium is avilable from: http://oss.adm.ntu.edu.sg/symposium2015.

[19] UNESCO Charter for the Safeguarding of Intangible Cultural Heritage, 2003, Art. 2.

[20] A. Nanetti and S. A. Cheong, "Complexity Science and Man-Heritage-Landscape Systems. A Heritage Impact Factor Theory", in Proceedings of the 3nd International Conference on documentation, conservation and restoration of the architectural heritage and landscape protection, ReUSO 2015, organized by the School of Engineering Building at the Polytechnic University of Valencia, in association with the School of Architecture at the Polytechnic University of Madrid, the Architectural Department in the University of Florence, the Department of Civil Engineering and Architecture of University of Pavia, Forum Unesco-Universidad y Patrimonio, and the Instituto Universitario de Restauración del Patrimonio (Valencia, October 22-24, 2015), unpublished.

[21] T. C. Khoo, Chief Editor, Built by Singapore: from Slums to a Sustainable Built Environment. Singapore:m Centre for Liveable Cities, 2015 (Singapore Urban Systems Studies Series).

[22] Available from: http://www.thefuturescentre.org/signals-of-change/2951/crowd-sourcing-singapore-s-future-spaces [accessed on 12 July 2015].

[23] Available from: http://www.straitstimes.com/politics/singaporeans-more-vocal-today-bringing-new-tensions-and-possibilities-george-yeo [accessed on 12 July 2015]. Yeo calls himself a Taoist, referring to “the larger currents at play … to which we are subject” (citation in George Yeo on Bonsai, Banyan and the Tao, A. I. Latif and H. L. Lee, Eds.. Singapore: World Scientific, p. 3, 2015).

[24] A. I. Latif and H. L. Lee, Eds., George Yeo on Bonsai, Banyan and the Tao. Singapore: World Scientific, 2015, p. 5.

[25] Available from: https://en.wikipedia.org/wiki/Sun_Yat_Sen_Nanyang_Memorial_Hall [accessed on 12 July 2015].

[26] Available from: http://lkyspp.nus.edu.sg/wp-content/uploads/2013/04/pa_tk_ST_George-Yeo-A-Man-for-All-Seasons_090511.pdf [accessed on 12 July 2015].

[27] S. Daley and A. Hartocollis, “Greek ‘No’ May Have Its Roots in Heroic Myths and Real Resistance,” New York Times, 6 July 2015, p. 45.

[28] Available from: http://www.macropolis.gr.

[29] Available from: https://www.flickr.com/photos/helmet_13/19441895492 [accesed on 20 September 2015].

[30] Available from: http://www.apple.com/ios/?cid=wwa-us-kwg-features.

[31] Available from: https://instantarticles.fb.com.

[32] A. Simpson, The Brand Strategist’s Guide to Desire. London: Palgrave Macmillan, 2015, pp. 55–56.

[33] C. Hirschi, The Origins of Nationalism: An Alternative History from Ancient Rome to Early Modern Germany. Cambridge: Cambridge University Press, 2011.

[34] "My Work Since the White House and My Legacy [Interview with Jimmy Carter]," National Geographic, vol. 228, no. 4, October 2015, p. 11.

[35] G. Vlastos, "The Socratic Elenchus," Oxford Studies in Ancient Philosophy, vol. 1, 1983, pp. 27–58.

[36] Available from http://www.nccs.net/the-responsibility-of-citizens.php [accessed on 25 September 2015].

[37] Available from: http://www.antiwarsongs.org/do_search.php?idartista=12288&lang=it&stesso=1 [accessed on 25 September 2015].

[38] M. Sheffer et al., "Early-warning signals for critical transitions," Nature, vol. 461, 3 September 2009, pp. 53–59, doi:10.1038/nature08227.

[39] Available from: http://www.syriadeeply.org.

[40] Available from: http://cognitive-edge.com/sensemaker [accessed on 12 July 2015].

[41] Available from: http://old.cognitive-edge.com/wp-content/uploads/2015/04/GH-SenseMaker-brief.pdf [accessed on 12 July 2015].

[42] J. H. Holland, Emergence: From Chaos to Order. Oxford: Oxford University Press, 2000.

[43] W. Brian Arthur, Increasing Returns and Path Dependence in the Economy. Ann Arbor: University of Michigan Press, 1994.

[44] P. Bak and M. Paczuski, "Complexity, contingency, criticality," Proceedings of the National Academy of Sciences of the United States of America, vol. 92, 1995, pp. 6689–6696.

[45] W. A. Green, "Periodizing World History," History and Theory, vol. 34, issue 2, May 1995, pp. 99–111 (p. 99, for the quotation). See also the larger work by the same W. A. Green, "Periodization in European and World History," Journal of World History, vol. 3, 1992, pp. 13–53, and the paper by J. H. Bentley, "Cross-Cultural Interaction and Periodization in World History," The American Historical Review, vol. 101, no. 3, June 1996, pp. 749–770.

[46] C. Ginzburg, "Microhistory, two or three things that I know about it," in Idem, Threads and Traces. Berkeley: University of California Press, 2012, pp. 193–214.

[47] C. W. Joyner, Shared Traditions: Southern History and Folk Culture. Urbana: University of Illinois, 1999, p. 1.

[48] C. Ginzburg, Il formaggio e i vermi. Il cosmo di un mugnaio del '500. Einaudi: Torino, 1976. In English: The Cheese and the worms: the cosmos of a 16th-century miller. Transl. by J. Tedeschi and A. Tedeschi. Baltimore: Johns Hopkins University Press, 1980.

[49] M. Scheffer et al., "Early-warning signals for critical transitions," Nature, vol. 461, no. 7260, 2009, pp. 53–59.

[50] A. Nanetti, A. Cattaneo, S. A. Cheong, and C.-Y. Lin, "Maps as Knowledge Aggregators: from Renaissance Italy Fra Mauro to Web Search Engines," The Cartographic Journal, vol. 52, issue 3, August 2015, pp. 1-10.

[51] Available from: http://ehmazure.cloudapp.net.

[52] Available from: https://www.ted.com/talks/larry_page_where_s_google_going_next?language=en.

[53] Available from: http://www.apple.com/news.

[54] Available from: http://www.wired.com/2015/09/facebook-launches-360-video-immersive-star-wars-clip

The Effects of Travel Information Sources on Traveller's Resonance and The Travel Destination Decision-Making Process

Masayuki Kaneko*, Lourdes Morales-Villaverde†, Katsuko Nakahira T.*, Makiko Okamoto*
*Nagaoka University of Technology
Niigata, Japan,
Email:s123355@stn.nagaokaut.ac.jp
†University of California, Santa Cruz
USA

***Abstract*—In this paper, we propose a method to estimate the relationships between ′information acquisition′ and ′decision-making′. The method used is based on the ′vacation sequences′ described by van Raaij which consist of five stages: (1) general decision; (2) information acquisition; (3) decision-making; (4) participation; and (5) satisfaction or complaints. The method proposed in this paper focuses on steps (2) and (3), which involve active processes performed by a traveller. While performing information acquisition and decision-making processes, a traveller decides where to go according to his/her motivation for travel. These results should be guided by the degree of ′resonance′ between the available travel information and his/her motivation. The detailed structure of resonance might be different from traveller to traveller due to each traveller′s distinct personal characteristics. We assume that activities which involve destination determination are analogous to people's purchase behaviors as described by Howard's model. In this paper, we incorporate the steps from the purchase model into steps (2) and (3) by expanding them appropriately to propose a new model with two elements for information acquisition and three elements for travel destination recognition. The strengths of resonance are measured for steps (2) and (3) as the external information and their relationships are examined.**

Keywords–Travel-destination information; empathy; motivation; decision-making.

I. Introduction

In this paper, we focus on the relationships between information sources available to the traveller and how well they resonate with him/her in order to understand how these resources affect his/her destination decision-making processes.

Since the 1990s, the Japanese government has placed emphasis on tourism by initiating such programs as 'Welcome Plan 21 (1995)' and 'Visit Japan Campaign (2003–)' for foreign visitors, or 'Egg of Columbus (2002–)'[1] for domestic visitors. In response to these programs, many regions in Japan have begun to share information on regional tourist destinations via information media such as 'analogue' booklets and/or 'digital' web pages. As the amount of tourist information from a variety of sources increases, it becomes easier for tourists to obtain the necessary information to make decisions concerning their tours. On the other hand, these diverse sources with distinct characteristics in terms of the types of information each conveys may have different effects on decision-making processes as each tourist may have a different processing strategy or preference for information receipt. This should be supported by the following observation. With the growth of the Internet, it is reported that within the current generation (often called the 'smart phone native generation'), the number of cases in which travellers made decisions regarding travel destinations based on information they found on the Web (either from enterprise or past travellers) has increased. As a result, the information provided by travel information sources on the Web can have a tremendous impact on the travel destination decision-making process. We assume this diverse information resonates with their travel destinations, and this resonance offers a strong motivation to travel.

In previous study, the role of social media, such as twitter or blog discussed in the viewpoint of the relation between information flow and construction of human network.

In our research, we treat consumers' information in the Internet with words of mouth. In the Internet world, all Internet users can send information with their viewpoints. Recently, social media spread rapidly, and it makes virtual human connection via users' own murmur, which regards as words of mouth. In the sense, words of mouth is the information provided by consumer without seeking benefit. SNS is a service which promotes connection between people through comments or picture provided by them.

From these behavior, SNS provides words of mouth in the Internet, and it has impact on consumer who want to buy new products.

In this paper, we propose a method for estimating the impacts of the strengths of resonance between a traveller's personal characteristics and travel destination characteristics on the motivational aspect of the travel destination decision-making processes. This includes the motivation behind the decision on which destinations to visit, traveller's use of different information sources for leisure travel plans, and the information from enterprises or past travellers. We constructed the method by applying the following two processes: (1) measure travellers' personal characteristics to identify distinct decision-making categories, (2) measure the relationship between the type of information acquired and travel destination recognition in terms of the degree of 'resonance' between them. These results and the 'travel destination decision-making process' introduced by [2] are combined to estimate the effects of travel information sources on traveller's resonance and travel decision-making processes. In Section II we introduce the basic theory that this paper is based on and related works. In Section III we explain the method for measuring resonance derived from acquired information types and travel destination recognition. In Section VI we summarize the proposed method and describe the future plan.

II. Related works

In this section, we introduce related works with several topics.

A. Importance of the Internet study for tourism

The Internet is said to be important for marketing to attract people. Research on tourist destinations using social networking services (SNS) was studied by Milli [3]. The research target was the Split-Dalmatian County in the Republic of Croatia. This county has been using their own website for advertising since 2008 and they have had a presence on Facebook since 2010. The Facebook advertising campaign was recognized in other countries: Germany, Austria, the Czech Republic, Hungary, and Slovenia. This campaign triggered an increase in the total number of website fans from 5900 to over 10,000. This indicates the importance of the Internet and SNS for tourism. But the difference of influence on the words of mouth through websites and SNS is not clear.

B. Word of mouth

Word of mouth has been studied frequently in recent years. With growth of the Internet and SNS, word of mouth is an anticipated marketing method. Michael and Jeremy [4] studied word of mouth through the American buzz marketing site Yelp.com on which users evaluate restaurants. According to the results they obtained, restaurants increased their number of reservations by 19 percent when they received an extra half star rating on the site. This result indicates that information spread through word of mouth can possibly influence decision making. Judith and Dina [5] also studied word of mouth through Amazon.com and bn.com. They found that one-star reviews have a greater impact than five-star reviews on the same site. These two previous studies indicate that reviews have significant impacts on either increasing reservations or book sales. We already know that word of mouth has a tremendous impact on purchasing. However, we do not know how effective it is in the travel field. Also, the difference of influence from HP is not clear Therefore, we investigated this using the information provided by enterprises and consumers.

C. Resonance

Resonance is one of the important factors used in previous studies on decision making [6]. The Model Human Processor with Realtime Constraints (MHP/RT) reacts to input from external environments and internal generated input to make decisions and select action at any moment. MHP/RT processes four types of events: future events (consciously or unconsciously) or a past events (consciously or unconsciously). In the process of ‘past/conscious’, by reacting to activated pieces of knowledge through resonance processes, it reflects on and elaborates on a certain symbolic event. In our study, we positioned resonance as the positive or negative emotional result achieved when a person receives information or knowledge. Previous studies on the influence of SNS primarily focused on its characteristics such as how it can be used for global advertising or on its technical aspects. However, there are very few previous studies that have focused on personal characteristics in terms of SNS information. Therefore, our investigation centered on both personal characteristics and information source characteristics.

D. Travel destination decision-making process

In order to understand travellers ’ resonance and motivation(s), we described the travel destination decision-making process by incorporating Shimizu’s [7] model of the purchase decision-making process into van Raaij’s ‘vacation sequence’ [8].

Shimizu compared the information provided by consumer and the information provided by enterprise on observing purchasing behavior. Previous study indicated that the information provided by consumer does not play a big role on products recognition. But Shimizu suggested that the information provided by consumer was effective on the stage of products recognition. We considered travel action as one of purchasing behavior and we observed them. Our consideration is that information provided by consumer also has a big impact on the unknown place.

When describing the travel destination decision-making process, we assumed that the person already had the desire to travel in order to better describe the process between the information acquisition and decision-making stages. Selecting the travel destination is generally treated as a purchase behavior. This makes the travel destination decision-making process easier to understand based on the ideas suggested for information processing in the purchase decision-making process. Howard’s model is one of the most well-known purchase behavior models and it shows the decision-making process that leads to the purchase behavior after receiving external information as a source of stimulation [7]. Moreover, Howard’s model is a comprehensive model for predicting consumers’ action [7]. In [7] Shimizu expands Howard’s decision-making model in terms of the source of the information. When incorporating Shimizu’s model into van Raaij’s ‘ vacation sequence ’, we considered two sources of information as equally important: consumers (i.e. previous travellers) and enterprises (e.g. travel agencies, travel destination sites, government sites, etc.). The travel destination decision-making process as we described it consists of the following elements: ‘ certitude ’, which represents the traveller ’ s self-confidence in how much he/she will enjoy traveling based on the travel information he/she received; ‘attitude ’, which represents the traveller ’ s attitude towards the travel destination based on the travel information he/she received; and ‘ travel destination recognition ’, which represents the traveller’s recognition of a destination as a travel destination after receiving (travel) information. (These depend on description of [7] Figure 1 shows a part of the travel destination decision-making process. In the work presented here we assume that the travel destination decision is made based on the motivation that results from the strength of the traveller ’ s resonance. Traveller ’ s resonance refers to the traveller ’ s action for resonance when they obtain travel information. What is more, the traveller’s resonance can be influenced by personal characteristics (e.g. travellers ’ formative experiences, travel objectives, etc.) when he/she is provided with various kinds of information regarding travel destination candidates (e.g. their function, contents, etc.). The strength of the resonance is considered an important factor when deciding on the travel destination. Knowing the strength of the resonance and understanding the relationship between the personal characteristics of the traveller (e.g. formative experiences, travel objective), the environmental characteristics of the travel destination (e.g. its function, contents, etc.), and the acquired information (provided by consumers or enterprises) can help determine how to effectively attract individuals to travel destinations. (The table

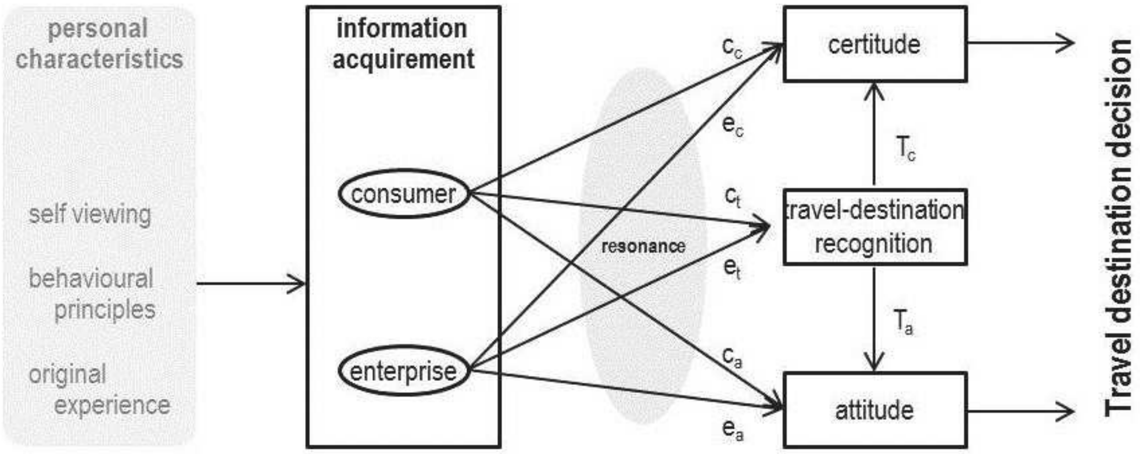


Figure 1. A part of travel destination decision-making process

of resonance presents this relationship and the resulting resonance.) Here we will focus on the travellers' resonance during the travel destination decision-making process, especially from the information acquisition stage to the travel decision stage. We will consider the strength of each element shown in Figure 1. First, let ' s discuss the meaning of letters. In here, c, e, and T indicate consumer, enterprise, and travel destination recognition. The suffix c, t, and a represent certitude, travel destination recognition, and attitude. (e. g. c_t indicates the resonance from consumer to travel destination recognition). The resonances are the strength of connection between each element shown in Figure 1. (c_c, e_c, c_t, e_t, c_a, e_a, T_c, T_a) These strength of resonances result from the combination of personal characteristics and environmental characteristics. Thus, we need to measure travellers' personal characteristics. They cannot be measured by numbers. Therefore, we need to conduct the experiment based on psychology.

E. Study 1 Ttravellers' Personal Characteristics

In order to better understand travellers' personal characteristics as they relate to the travel destination decision-making process, we conducted a questionnaire with 43 questions, in both open and closed format. This questionnaire was administered to 165 students in Nagaoka city (Niigata Prefecture).

The questionnaire is consisted with the following question items regarding to personal characteristics of travellers. 1) view of self (independent or interdependent)[9] 2)potential travel objective[10] 3)previous travel experience(s), (type : either family/friends or independent, objective : chosen from stimulation, cultural observation, communicate with local people, restore/improve health , have a new/unpredictable experience, experience nature, self-improvement, and develop personal relations, place, duration of travel 4)desires regarding future travel experiences(Except for duration, same items as 3)).

The objectives could be categorized into four factors by factor analysis. There were positive correlations between 2 of factors: 'expanding original experience' and 'have a new/unpredictable experience' and 'individual recognition' and 'the strength of assertion'.

We attained the following results. For past experiences, most of the participants travelled with their families or friends. Since the participants were young (15 to 20 years old), it is suggested that past travel experiences were mostly family or school trips. For the influence of past experience, the travel objective for past experience and future desired travel was matched. It was considered that the past experience became an anchor to make a decision. These results indicated that characteristics have some influence on decision making.[11]

In our study, we choose group of highly educated students as survey participants because they frequently use the Internet or SNS, so they have very small psychological and physical resistant for IT technology.

F. Study 2 Information acquisition and Resonance

In study 2 we investigate the strength of the connection between certitude, travel destination recognition, attitude; and travel intention when they receive travel information from different sources, enterprise or consumer(i.e. other travellers).

We considered travel action as one of purchasing behavior and we considered both of information provided by enterprise or consumer as travel information sources. From Shimizu's study, she showed that there is a big impact from the information provided by consumer on the products established as a bland. So we predicated that the information provided by consumer is also stronger for the unknown place and it may show similar results.

Procedure: To measure the effects of travel information sources on traveller's, we prepared two kinds of stimuli, which were related to cultural objective. Each set will consists of the following: Two pictures of the same travel destination(already known place and unknown place), one from the public website of the specific travel destination (enterprise) and the other from a traveller ' s Twitter account (consumer).

We consider that the combination of source of the travel information and travel destination as shown in Table I have a different effect on the strength of the traveller's resonance. We focused on two of travel destination(already known place and

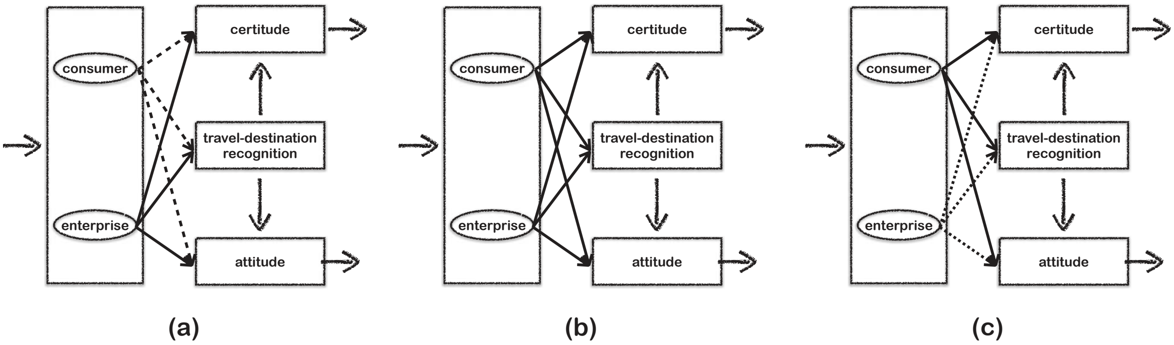


Figure 2. (a) visit historical sites, (b) experience nature, (c) stimulation

unknown place) because already known place and unknown place is on different situation that already known place is already constructed as a bland, but unknown place is ongoing to consist as a bland. So we considered that 2 of the travel destination also influence on traveller's resonance.

Question details: The items of the questionnaire was as follows. 1) Questions on participants' knowledge: This question asks whether or not the participants are familiar with the travel sites. 2) Questions on travel destination recognition: This question asks what information (provided by the enterprise or consumer) is a frequent source for new information for travel destination. 3) Questions on certitude: This question asks what information (provided by enterprise or consumer) invokes more confidence in the estimated value of the travel destination. 4) Questions on attitude: This question asks what information (provided by enterprise or consumer) suggests that the travel destination is a good place to visit. 5) Questions for intention: This question asks what information (provided by enterprise or consumer) invokes a desire to visit the travel destination. 6) Questions on satisfaction: This question asks what travel objectives will be satisfied from the travel information provided.

Through this method, we will estimate the relationship of resonance in terms of the source (enterprise vs. consumer) and the obscurity (already known place vs. unknown place). We conducted a questionnaire with 5 point scale of -2 to 2. We gave participants two pieces of information provided by enterprises and consumers. The objective was selected from the first survey. We focused on investigating of cultural observation. We analyzed travel intention by using a chi-square test. We attained the following results. For already known sites, participants attached more weight to the information provided by enterprises than that from consumers. For unknown sites, participants used both types of information. We considered that for already known sites, participants would like proper knowledge of the sites. For unknown place, participants would like to gather information including proper knowledge from the enterprises and comments from consumers who previously traveled there.[12]

TABLE I. THE PATTERNS FOR QUESTIONNAIRE. E REPRESENTS "ENTERPRISE" SOURCE, AND C REPRESENTS "CNSUMER" SOURCE

Travel objectives	Already known		Unknown	
	E	C	E	C
Cultural observation	A1	A2	A3	A4

III. THE RELATIONSHIP BETWEEN THEIR DESIRE AND INFORMATION ACQUIREMENT : ON THE VIEW POINT OF RESONANCE

Combining both results of study 1 – their objective for future desired travel – and study 2(resonance), we analyzed the relationships in terms of their characteristics(the objective for future desired travel) and travel information sources(enterprise/consumer).

We extrapolate about the connections for each of the objectives on which we focused. We believe that we will attain the following results as shown in Figure 2 (a) to (c). Each figures consists of two blocks: information resource (enterprise and customer) and travel destination recognition. These blocks are generated by the balance of certitude and attitude. Three lines are shown in Figure. 2: the dashed line, narrow line, and thick line. Hence, each line is defined as follows: the narrow line represents a normal strength of resonance, the dashed line represents a weak strength of resonance, and the thick line represents remarkable strength of resonance. This figure indicates the strength of resonance between the information (enterprise and consumer) and elements (certitude, attitude and travel destination recognition). If the travellers' objective is to 'cultural observation', then their resonance will be indicated as having remarkable strength based on the information provided by enterprise, as they would like to have reliable historical information. This suggests that their condition would be represented as shown (a) in Figure. 2. If the travellers' objective is to 'experience in nature' then their resonance will be indicated with almost equal strength for both information sources. They may want to know how good this site is from travellers who have previously visited the place, while they also want detailed information such as what kind of flowers can be observed. In this case, their condition would be represented as shown in (b) in Figure. 2. Finally, if they wish to travel without certain travel information provided by the enterprise in order to have a stimulating experience, their resonance will be indicated by remarkable strength for the information provided by consumers. In this case, their condition is represented as (c) in Figure. 2. In (c), the role of consumer's information is very important. Consumers'

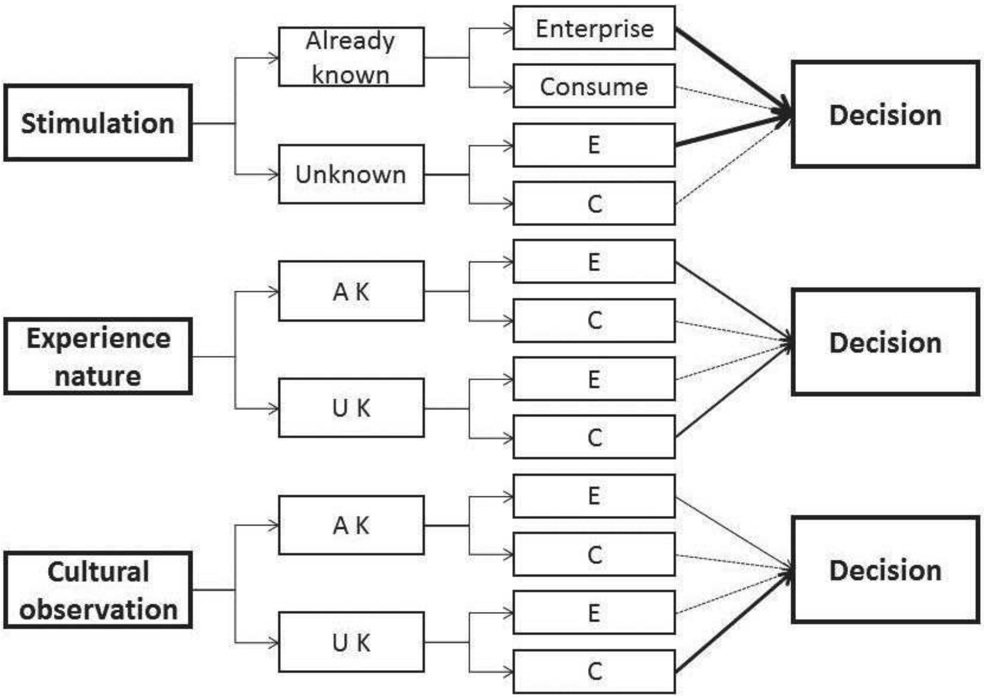


Figure 3. The result of the strength of resonance

TABLE II. THE RELATION BETWEEN SURVEY 1 AND 2

The objectives	Already known place	Unknown place
Stimulation	0.333	0.333
Experience nature	0.167	-0.333
Visit historical sites	0.100	-0.200

information is aggregated by many methods, but most popular way is social media. Currently, many people are able to update their comments, photos, videos, or audios on the Internet. They submit this content to blogs, Facebook, or many kinds of social media. Most social media contents can be searched with ease by Internet users. In a sense, social media is one of useful ways to aggregate consumers' tourism contents. This means the information acquirement layer in Figure. 1 is tremendously diverse and there are many resonance patterns found in the information. Hence, observing travellers ' resonance patterns is important to estimate the travel destination decision-making process.

For conducting surveys, we followed the research ethics guide line published in Nagaoka University of Technology.

IV. RESULTS

First we separated participants data(survey 1) with the objectives(future desired travel) of stimulation, cultural observation, and experience nature. We calculated the average points of survey 2 and we compared each of them. The results are shown in Table II and Figure 3.

In Figure 3, "A K", "U K", "E" and "C" represent Already known, Unknown, Enterprise and Consumer. In the figure, there are four levels of thickness for directional line; from thickest to thin, these represent more than 0.3 points of average score, from 0.2 to 0.3 points of average score, from 0.1 to 0.2 points of average score. And dashed directional line represents less than 0.1 points of average score. Although not observed significant difference for subjects was small, following trends were observed. 1) Participants whose objective was stimulation had more intention from the information provided by enterprise in either already known and unknown place. 2) Participants whose objective was experience nature had more intention from the information provided by enterprise for already known place and had more intention from the information provided by consumer for unknown place. 3) Participants whose objective was cultural observation had almost same results of the objective of experience nature. But for the participants whose objective was cultural observation got weaker result.

V. DISCUSSION

In this section, we discuss resonance reactions that were observed in our two participants groups: stimulation and cultural-observation-participants.

A. Resonance reactions by stimulation-participants

The degree of resonance shown by the participants who had the objective of "stimulation" was stronger toward enterprise-information than that of resonance toward Twitter. The result was true for both already-known-place and unknown-place.

Firstly, resonance would not occur when there is mismatch between the provided information content and the interest that a participant has. And therefore the stimulation participants showed little resonance toward the contents related to the objective of cultural observation.

Secondly, the amount of information provided by enterprise-information was richer than that provided by SNS. The former uses such media as bigger pictures, texts with a lot of words, and so on. Stimulation would be positively related with the amount of information. And therefore the stimulation participants showed stronger resonance towards such enterprise-information.

B. Resonance reactions by cultural-observation-participants

The degree of resonance shown by the participants who had the objective of “cultural observation” was stronger toward consumer-information such as Twitter than that of resonance toward enterprise-information about unknown-place.

As described before, the necessary condition for resonance reaction is the provision of matched information. A participant would examine the information he/she is interested in and he/she might induce a big resonance. However, even if the provided information matches the interests of the participant, it is not enough to induce resonance reaction: if the provided information does not have any useful and valuable “new” information to him/her, it should be estimated as worthless because it just provide the opportunity of just checking his/her experience pf ’having been the place’ and little resonance would be induced. We had the results that this is true for HP and SNS for already-known-place.

On the other hand, we found that cultural-observation-participants showed stronger resonances with the information provided by “consumer without seeking benefit” for unknown-places. This is because the participants had tendency to relay on the information provided by “consumer without seeking benefits.” In sum, the cultural-observation-participants resonated differently depending on the nature of the information, whether it is about already-known-place or unknown-place. This result is consistent with the one provided by Shimizu[7] who studied purchasing behavior.

VI. Summary and Future Works

In this paper, we explained the travel destination decision-making process in order to understand how travellers decide upon their travel destination. Travellers need to be motivated to travel. We assumed that the motivation is provided by the strength of resonance. It was suggested that personal characteristics need to be investigated in order to understand the travellers ’ travel destination making-process. In the process, we focused on the information acquirement process and travel destination making-process. Between these two steps, humans have many thoughts, but they are biased by human personal characteristics for information acquirement process, and many levels of resonance based on information type or contents itself. Considering this, we propose a method for measuring the effects of resonance strength on the relationships among destination decision factors. Our method is the means to acquire the diversity derived human and to estimate travel destination decisions. We administered the questionnaire to 165 students from Nagaoka city who are fluent in Japanese (not all were Japanese), under the supervision of their class advisors. We gave them two types of pictures based on information from the following combinations: enterprise, consumer, already known place, or unknown place. We focused on investigating travel intentions for cultural observation. We received the following results. For already known sites, participants attached more weight to the information provided by enterprises and for unknown sites, participants regarded both types of information equally. Finally, we conducted an analysis of the relationship between survey 1 (characteristics) and survey 2 (resonance). We attained the following results. The participants whose objective was stimulation attached more weight to the information provided by the enterprises. The participants whose objective was to experience nature and cultural observation attached more weight on to the information provided by consumer on the unknown sites.

Our future work will be the investigation of resonance of the people whose objective is stimulation, cultural observation or experience nature when the travel information which has a characteristics of stimulation or experience nature was given.

Also, the study group for this study was limited. Thus, we will conduct survey with considering age, gender, occupation and etc.

Acknowledgement

This work was supported by JSPS KAKENHI Grant Numbers 25350367, 15H02784.

References

[1] http://www.mlit.go.jp/common/000132607.pdf,2007, accessed 2015.07.12 in Japanese.

[2] Masayuki Kaneko, Katsuko Nakahira, and Makiko Okamoto, ”The description of the traveler’s decision making process induced by interaction of man - environment - information”, Proc.77th National Convention of Information Processing Society of Japan, Vol.1, 2015, pp.409-410.

[3] Mili Razović, ”Social Network and Promotion of Tourist Destination”, in Proceedings of the 22^{nd} Cromar Congress, Marketing Challenges in New Economy, 2011, pp746-766.

[4] Michael L. Anderson and Jeremy Magruder, ”Does Yelp Affect Restaurant Demand?”, Giannini foundation of Agricultural Economics, Vol.16, No.5, 2013, pp. 1-4.

[5] Judith A. Chevalier and Dina Mayzlin, ”The Effect of Word of Mouth on Sales”, Online Book Reviews. Journal of Marketing Research, August 2006, Vol. 43, No. 3, 2006, pp. 345-354.

[6] Muneo Kitajima and Makoto Toyota, ”Decision-making and action selection in Two Minds: An analysis based on Model Human Processor with Realtime Constraints”, Extended versions of selected papers from the Third Annual Meeting of the BICA Society, Vol.5, 2013 pp. 82-93.

[7] Mai Shimizu, ”The Effect of CGM on Consumer Information Processing: Comparison of Consumer Generated Media with Corporate”, Journal of Commerce, Vol.81, 2013, No3, pp. 93121, 2013.

[8] W. Fred van Raaij, ”Consumer research on tourism mental and behavioral constructs”, Annals of Tourism Research, Vol.13, No1, 1986, pp. 119.

[9] Toshitake Takata, ”On the scale for measuring independent and interdependent view of self”, Bulletin of Research Institute , No8, 2000, pp. 145-163.

[10] Yoshifumi Hayashi and Takehiro Fujihara, ”The Effects of the Evaluation of Travel Experiences on Tourist Satisfaction: From the Viewpoints of Tourism Motives and Past Travel Experiences”, SHAKAIGAKUBU-KIYO School of Sociology and Social Work Journal, No.114, 2012, pp. 199-212.

[11] Masayuki Kaneko, Katsuko Nakahira, and Makiko Okamoto, ”The effects of traveler’s empathy states for individual travel information sources on travel destinations decision processes”, Forum on Information Technology, vol.1, 2015, pp. 149-152.

[12] Masayuki Kaneko, Katsuko Nakahira, and Makiko Okamoto, ”Influences of Acquired Tourism Information on Resonance and Travel-Destination Decision-Making Process”, IEEE shin-etsu Session, 2015, p. 168.

Extracting Social Structure from DarkWeb Forums

Elizabeth Phillips, Jason R.C. Nurse, Michael Goldsmith, Sadie Creese
Cyber Security Centre,
Department of Computer Science,
University of Oxford, UK
Email: $\{firstname.lastname\}@cybersecurity.ox.ac.uk$

***Abstract*—This paper explores various Social Network Analysis (SNA) techniques in order to identify a range of potentially *'important'* members of Islamic Networks within Dark Web Forums. For this experiment, we conducted our investigation on five forums collected in previous work as part of the Dark Web Forum portal and built upon the tool support created in our previous research in order to visualise and analyse the network. Whilst existing work attempts to identify these structures through state-of-the-art Computational Linguistic techniques, our work relies on the communication metadata alone. Our analysis involved first calculating a range of SNA metrics to better understand the group members, and then apply unsupervised learning in order to create clusters that would help classify the Dark Web Forums users into hierarchical clusters. In order to create our social networks, we investigated the effect of repeated author resolution and various weighting schemes on the ranking of forum members by creating four social networks per forum and evaluating the correlation of the top n users (for $n = 10, 20, 30, 40, 50$ and 100). Our results identified that varying the weighting schemes created more consistent ranking schemes than varying the repeated author resolution.**

Keywords–Social Network Analysis; Dark Web Forum; Jihadist forums; Social Structure

I. Introduction

With the dawn of the digital revolution, the Internet has transformed the way in which people communicate with each other. One area in which this has been seen is within terrorist networks[1], [2]. The near-instantaneous responses from users now achievable with these developing technologies and the increased audience has meant that previous face-to-face interactions and radical discussions have migrated to online mediums [3], [4]. This migration has led to an increase in the popularity of Dark Web Forums as a means of sharing text based content as well as links to other sites, videos and rich Web 2.0 features [5].

Even before the increased adoption of dark web forums, researchers have used social network analysis to identify key individuals within these groups. Numerous researchers have investigated various techniques in order to retrospectively understand the inner-workings of the 19 terrorists involved in the September 11th attacks in 2001[6]. Early research involved sourcing news articles related to the known suspects after the event and analysing the structure of the group after the event and with known targets [7]. Others have focused on researching how different terrorists organisations work together [8] by performing social network analysis [9].

The widespread adoption of the Dark Web Forums have enabled researchers to analyse the rich source of data [10] and has enabled researchers to establish a greater insight into the inner workings of some of these groups. Whilst existing research has been undertaken into creating a social network topology for Dark web Forums [11], these existing research focus on small communities and networks and rely on sophisticated sentiment analysis techniques, which are difficult to scale when evaluating millions of messages[12].

Since the revelations of metadata collection exposed by Edward Snowden in 2013 [13] , the importance of metadata from emails has gained awareness. In light of these revelations, many individuals have been unsure of their own risk exposure using these metadata techniques alone. Existing work has been effective at establishing hierarchy from the metadata of email communications alone [14], [15]. Such metadata include the post's author and timestamp. In this paper, we set out to investigate whether similar techniques could be applied within Dark Web Forums and assess to what extent we can identify the social structure of the networks.

Our paper is focused on the research question "Using communication metadata alone, can we make reasonable inferences about the structure of terrorist groups". In particular, we set out to trial various Social Network Analysis (SNA) techniques in order to identify a range of potentially *'important'* members of the forums.

A. Social Network Analysis (SNA)

For decades, complex interactivity between entities has been modelled as networks. These include the internet [16], food webs and biochemical networks [17]. For each network, entities (such as computers, routers, animals, etc.) are considered as nodes or vertices that are connected together by links or edges (such as communications between computers, the flow of energy within a food network, etc.)

Link Analysis (LA) is the analysis of information flow within the networks above and has been a topic of study for several decades [18], [19]. A *Social Network* (SN) is defined as the representation of communication networks where the nodes are people and the edges correspond to the relationships between and *Social Network Analysis* (SNA) is defined as the application of Link Analysis to a social network. We can perform SNA on our newly created social network, where the flow of information corresponds to the flow of information on the forums which allows us to perform SNA on our network to determine hidden network structures.

II. Methodology

In order to assess the extent of the network discovery of our forums, we begin by collecting the forum posts. This in turn allowed us to extract the metadata from each forum from which to build our network. Once we had extracted the metadata mentioned above, the next step was to create a social network representation of each forum. We then set out to experiment ways in which we create a social network representation of the communication data [see section II-A]. Figure 1 shows an overview of our process.

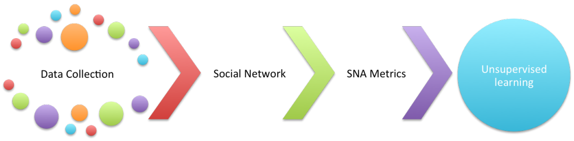


Figure 1. An outline of our approach to extract social structure from online communications

From this social network, we calculated the top-ranked SNA metrics from previous work [14]. After calculating the metrics for each node within the network, we then used unsupervised machine learning to identify clusters of interest within our network. Unsupervised machine learning was selected due to the unavailability of ground truth of the seniority of the forum members (other than identifying the top posters). Given the nature of our metrics chosen, the clusters will be grouped based on the similarity of their metrics, which in turn we hypothesise reflects the importance of the individual within the network.

A. Creating the Social Network

Once the metadata for each forum was collected, we then set out to experiment the ways in which to convert the metadata into a Social Network. We convert the metadata into a social network using algorithm 1.

Algorithm 1 Algorithm to convert a forum into a social network where w(a, s, m) is the weighted score of the message m from sender a to sender s and is determined based on our design decisions below and SNW(s_1, s_2) is the overall weight of the directed edge between s_1 and s_2 in our Social Network for a given forum f.

```
for each thread $t$ in forum $f$ do
    for each message $m$ in $t$ do
        $a$ = author($m$)
        for each previous sender $s$ in thread $t$ do
            SNW($a, s$) = SNW($a, s$) + w($a, s, m$)
        end for
    end for
end for
```

For this evaluation we experimented with two distinct methods of handling when an author posted multiple times in the same thread (Repeated Author Resolution) and two distinct ways of increasing the weight between two members based on the Time difference between responses.

In order to illustrate the application of our methods, Figure 2 shows an example of a forum with two members and four posts.

1) **Repeated Author Resolution:** In Figure 2, our forum posts shows two posts for $sender1$ and two posts for $sender2$. Our experiment tested 2 ways of handling repeated authors within a thread.

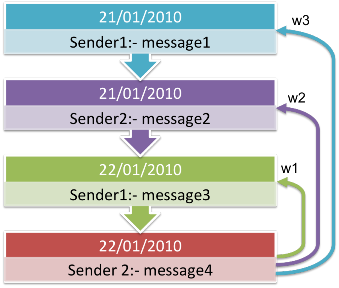


Figure 2. An example forum post with two members and four posts

Unique Senders: For this method, we only consider those senders that are unique within each thread. In this case, if we see an author reply to a thread they have previously commented on, we only add the weight for the most recent comment on the thread. In our example in Figure 2 when updating the scores after message4, only w_1 and w_2 would be added to SNA$(sender2, sender1)$ and SNA$(sender1, sender1)$ respectively and w_3 would not be calculated or added.

All Senders: For this method, we consider every sender within each thread and include duplicate senders. In this case, if we see an author reply to a thread they have previously commented on, we add the weight for each comment on the thread. In our example in figure 2 when updating the scores after $message4$, SNW$(sender2, sender1)$ is increased by $w_1 + w_3$ and SNW$(sender1, sender1)$ would be increased by w_2.

2) **Weighted Date Resolution:** When a new author a contributes a message m_a to a thread t where senders $s_1, s_2, \ldots s_k$ have previously commented, let d_{s_i} be the date that sender s_i sends a message m_{s_i} in the thread t. We then need to calculate w(a, s_i, m_a) for all i. In addition to evaluating the effectiveness of including all or unique senders, we also evaluated two distinct ways to calculate the weightings based on the time difference (d_t) between the two messages. The two methods we used to calculate the weightings were *'uniform weighting'* and *'inverse proportionality weighting'*.

Uniform weighting: Our first model assumed that the strength of the connection between two messages within a thread is independent on the amount of time taken to respond to a message. In this case, $\forall i$, w(a, s_i, m_a)= 1. Whilst this allows for a simple view of our forum, we set out to explore whether this model may not be representative as the same weight is given to a response to a message on the same day as a message with a response delay of 4 weeks.

Inverse Proportionality weighting: In order to overcome the issue outlined above, our second model was created on the assumption that the closer a response is in time, the stronger the weight of the connection between two users. In this case $\forall i$, w$(a, s_i, m_a) = \frac{1}{1+(d_a - d_{s_i})}$. This model ensures that $\forall i, 0 <$ w$(a, s_i, m_a) < 1$.

B. SNA Metrics

For the purpose of this paper, in order to measure the graph properties of each node that reflect importance, we used the

TABLE I. A TABLE OUTLINING THE METRICS USED TO EVALUATE OUR SOCIAL NETWORK

Attribute Name	**Description**
Sent Messages (SM)	The number of emails sent by an employee.
Received Messages (RM)	The number of emails received by an employee.
Degree Centrality (DCS)	The number of distinct employees within the network that an employee has sent emails to.
Betweenness Centrality Score (BCS)	The betweenness centrality measure for an employee.[20]
Pagerank Score (PRS)	The PageRank score an employee. [21]
Markov Ranking (MR)	The markov ranking of an employee. [22]
HITS Authority Score (HAS)	The authority score for an employee (if several users with high hub weights send an email t the user then they will have a higher authority score). [23]
HITS Hub Score (HHS)	The hub score for an employee (if the user sends emails to users with high authority scores then they will have a higher hub score). [23]
Clique Score (CS)	The number of cliques (maximal subgraphs) an employee is in using the Bron and Kerbosch algorithm.[24]
Weighted Clique Score (WCS)	The weighted clique score for each user, weighted by the number of users within each clique.
Average Distance Score (ADS)	The average distance between the user and all other users in the graph.
Clustering Coefficient (CC)	The extent to which vertices in a graph tend to cluster together. [25]

SNA metrics outlined in [14] and are shown in Table I. These metrics can be split into five main categories that can be used to identify relevant properties of our network. These categories are highlighted in Figure 3. By selecting a cross-section of metrics that cover all five main categories of metrics, our machine learning model is able to capture as much information from our network as possible.

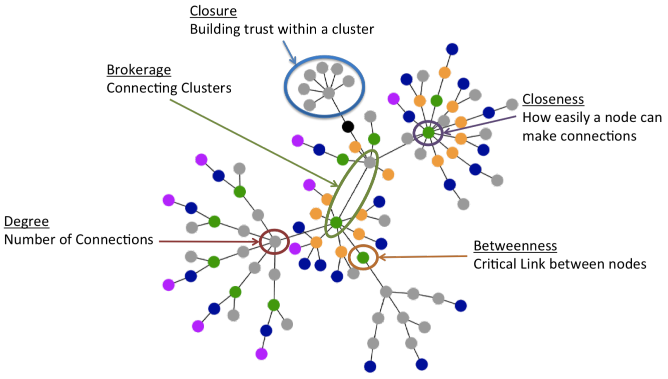

Closure
Building trust within a cluster
Brokerage
Connecting Clusters
Closeness
How easily a node can make connections
Degree
Number of Connections
Betweenness
Critical Link between nodes

Figure 3. An outline of the five main categories of SNA metrics, namely brokerage, degree, closeness, closure and betweenness.

III. EXPERIMENTAL SETUP

A. Datasets

For this experiment we evaluated 10 of the forums collected as part of the Dark Web Forum Portal dataset[26]. This dataset was collected by crawling a variety of radical websites [27] and allows us to compare the results and its applicability over multiple forums. Table II shows a breakdown of each forum used as part of the experiment.

In order to ensure that the connections in our graph reflect meaningful connections within the network, members whose edge weights are less than 10 are pruned from our graph. This allows the graph to reflect the strong connections whilst removing those connections that do not play a central role within the network. As such, given the different weighting approaches for each scheme, after pruning edges with edge weight less than 10, each social network contained varying numbers of vertices with 1 or more edges.

For each of our seven terrorist network datasets, we created four social networks using the combination of metrics below. Once the network was created, we then calculated our SNA metrics and performed unsupervised machine learning.

- Uniform weighting and unique senders
- Uniform weighting and all senders
- Inverse-Proportionality weighting and unique senders
- Inverse-Proportionality and all senders

In order to evaluate the appropriateness of each of our methods, we performed Unsupervised Learning using Expectation Maximisation (EM) [28] in order to create clusters in order to classify forum members. EM was chosen as it is able to handle unobserved data and missing datapoints, which can reflect the fact that we may only have a partial view of the network by observing the online forums alone.

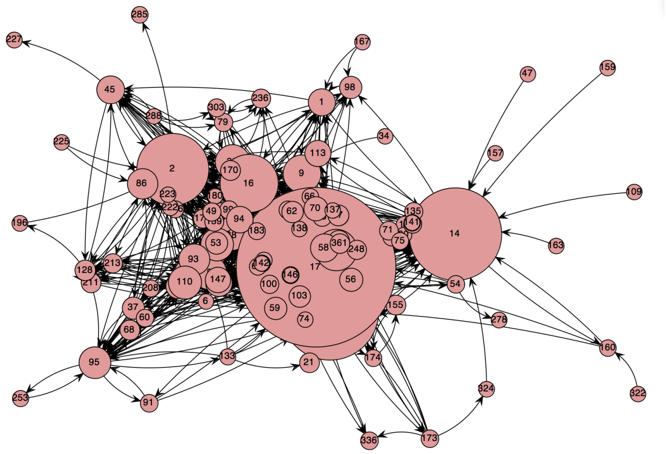


Figure 4. An example screenshot of our toolsuppport where each node is sized by the node's HITS hub score.

B. Tool Support

To allow us to visualise the results of our social network analysis and easily switch between the different weighting schemes and repeated author resolution, we built upon our existing tool [14] by adding additional features to allow users to provide their own weighting scheme to create their own Social Network Representation of the underlying forum data.

Figure 4 shows an example screenshot of our tool support displaying the social Network of the Ansar1 Network using repeated authors and inverse proportionality weighting.

TABLE II. AN OVERVIEW OF THE ENGLISH LANGUAGE FORUMS USED AS PART OF OUR EXPERIMENTS

Forum	#Messages:	# Threads:	# Members:	Start Date:	End Date:	Forum URL:
Ansar AlJihad Network (Ansar1)	29492	11244	382	12/08/08	01/20/2010	http://wwww.ansar1.info/
Gawaher (Gawaher)	372499	53235	9269	10/24/2004	06/07/12	http://www.gawaher.com
Islamic Awakening (IslamicAwakening)	201287	32879	3964	04/28/2004	05/22/2012	http://forums.islamicawakening.com
Islamic Network (IslamicNetwork)	91874	13995	2082	06/09/04	11/10/10	http://talk.islamicnetwork.com
Islamic Web-Community (Myiwc)	25016	6310	756	11/05/00	02/19/2010	http://www.myiwc.com/forums/index.php
Turn To Islam (TurnToIslam)	335338	41654	10858	06/02/06	05/20/2012	http://www.turnintonislam.com/forum/
Ummah	1491957	91527	21013	04/01/02	05/18/2012	http://www.ummah.com/forum/

IV. RESULTS AND ANALYSIS

In this section we outline some of the main highlights of our research findings. For our Ansar1 Dataset, we identified five clusters. $Cluster0$ contained 26 members, $cluster1$ contained 37 members, $cluster2$ contained 1 member, $cluster3$ contained 5 members and $cluster4$ contained 42 members. Table III provides further detail on the distribution of each cluster identified by the EM model.

TABLE III. OUTLINE OF OUR CLUSTERING ALGORITHMS FOR OUR ANSAR1 DATASET USING A UNIFORMED WEIGHTING SCHEME AND INCLUDING ALL SENDERS.

Attribute	**Cluster**				
	0	1	2	3	4
degree					
mean	0.0953	0.0095	0.8559	0.3568	0.0293
std.dev	0.0419	0.0021	0.1104	0.0944	0.0124
betweenness					
mean	30.5428	0.0535	5549.8267	551.7925	0.0002
std.dev	42.8349	0.3254	541.1433	337.2082	0.013
PageRank					
mean	0.0105	0.0031	0.1488	0.0464	0.0053
std.dev	0.005	0.0007	0.017	0.0223	0.002
Markov					
mean	0.0113	0.002	0.1513	0.0538	0.0048
std.dev	0.0064	0.0009	0.0185	0.0258	0.0027
HITS_authority					
mean	0.0977	0.0168	0.4142	0.2655	0.0477
std.dev	0.0451	0.0118	0.0715	0.0601	0.0273
HITS_hub					
mean	0.1085	0.0149	0.4118	0.2502	0.0445
std.dev	0.0423	0.0126	0.0712	0.0506	0.0291
weighted_cliques					
mean	538.4462	2.107	5108	3875.3263	17.6769
std.dev	666.9511	0.458	997.1299	669.2559	28.3311
cliques					
mean	5.968	1.0532	110	42.0026	1
std.dev	4.5005	0.2258	13.7192	14.3096	13.7192
avgDistance					
mean	1.5016	1.4613	1.8868	1.5895	1.474
std.dev	0.027	0.0414	0.0586	0.0342	0.0306
clusteringCoeff					
mean	0.6634	0	0.0589	0.257	1
std.dev	0.1631	0.0029	0.4387	0.0935	0.4387

Our evaluator identified 5 main clusters ranging in size from 1 to 42 users. From our results, we can immediately identify that cluster 2 (with 1 user) has a significantly greater score on almost all metrics compared to the other clusters followed by cluster 3. This leads us to believe that user 17 (the sole user in cluster 2) has significantly higher influence within the network, compared followed by those members in cluster 3 and as such, these users are potential candidates for being "important" users within the Ansar1 network. The significantly lower standard deviation for clusters 1 and 4 indicate that these users are likely to be less important in the network with their low scores in the ranking metrics (e.g. PageRank, HITS Hub, etc.).

TABLE IV. THE TOP TEN USERS USING ALL FOUR SOCIAL NETWORK METRICS SCHEME USING OUR ANSAR1 DATASET RANKED USING THE PAGERANK ALGORITHM

Inverse-1Date	Inverse-AllDates	Uniform-AllDates	Uniform-1Date
17	17	17	17
14	14	0	0
0	0	14	14
16	2	2	2
2	16	16	16
11	39	39	39
9	11	11	11
39	9	12	9
33	33	33	33

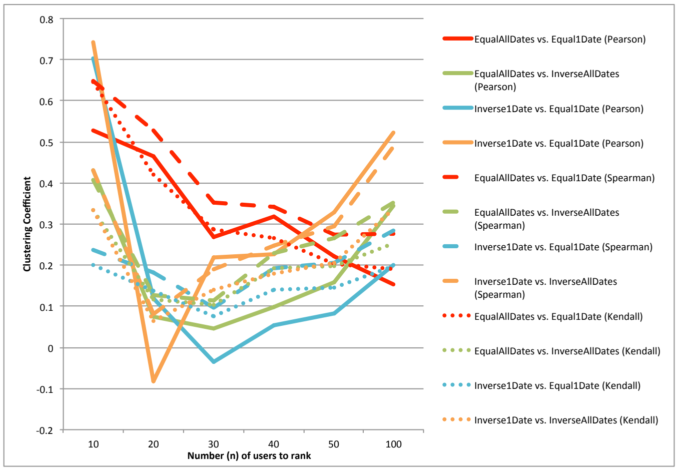


Figure 5. An outline of the correlation coefficients for each of our four networks using Spearman's, Kendall's and Pearson's correlation coefficients.

We also evaluated the difference in ranking of each created social network and Table IV shows the top ten ranked users (using the PageRank algorithm) for each of the four networks using the Ansar1 dataset. The table shows that all four networks identified user 17 as the most influential user and agreed on the top three users (but had different orderings of users 0 and 14). Similarly, they all agree on the top 10 users (with the exception of the UniformAllDates network, which substitutes user 9 for user 12.

In order to evaluate the effect that weighting and repeated author resolution has on our model, we evaluated the consistency of our rankings with schemes by comparing the top n ranked users (for $n = 10, 20, 30, 40, 50$ and 100) using Spearman's Coefficient Ranking [29], Kendall's Tau Measure [30] and Pearson's Correlation Coefficient [31]. Table V and Figure 5 shows the ranking scores of each comparison with only one degree of freedom (either identical weighting and different repeated author resolution or vice versa). Our results showed that on average, varying the repeated author resolution caused less variation in the ranking of the metrics in the final social network when applied for $n < 50$. However, the difference between rankings based on their repeated author

TABLE V. COMPARISON OF CORRELATION BETWEEN RANKING ALGORITHMS USING OUR ANSAR1 DATASET

Attribute	Number of users (n) to compare ranking using PageRank																	
	10			20			30			40			50			100		
	P	S	K	P	S	K	P	S	K	P	S	K	P	S	K	P	S	K
UniformAllDates vs.Uniform1Date	0.529	0.648	0.644	0.466	0.528	0.421	0.270	0.354	0.287	0.318	0.343	0.267	0.221	0.274	0.203	0.152	0.276	0.191
UniformAllDates vs. InverseAllDates	0.431	0.406	0.333	0.074	0.128	0.116	0.045	0.113	0.103	0.099	0.231	0.192	0.157	0.265	0.198	0.346	0.352	0.253
Inverse1Date vs. Uniform1Date	0.704	0.236	0.200	0.119	0.182	0.137	-0.035	0.097	0.076	0.054	0.192	0.141	0.084	0.205	0.146	0.202	0.285	0.197
Inverse1Date vs. InverseAllDates	0.741	0.430	0.333	-0.083	0.081	0.063	0.218	0.191	0.140	0.227	0.247	0.179	0.328	0.294	0.207	0.521	0.489	0.338

resolution diminishes as n increases. This leads us to believe that performing two social networks, one using repeated authors and one using unique authors only may provide subtly different views of the graph, which in turn will allow us to gain more insight from the social network.

V. CONCLUSIONS AND FUTURE WORK

In our paper, we set out to investigate the effectiveness of using metadata alone to identify influential and important users on Dark Web Forums. We set out to investigate the effect of repeated author resolution and various weighting schemes on our rankings by creating four social networks per forum and evaluating the consistency of the top n users (for $n = 10, 20, 30, 40, 50, 100$). We also performed unsupervised machine learning for each of network in order to identify clusters by user's importance. Our results showed us that difference in rankings from different weighting schemes were more consistent on average than those using different repeated author resolution techniques.

In order to allow us to further evaluate the validity of our work, we hope to establish an authoritative ground truth in order to assess the relative performance of our work. Another direction we hope to take the research is to perform dynamic analysis on our network and assess how the network changes over time. This technique could then be applied to identify potential insider threats within the Dark Web Forums by observing abnormal dynamic behaviour. In order to allow us to further evaluate the validity of our work, we hope to establish an authoritative ground truth in order to assess the relative performance of our work. Another direction we hope to take the research is to perform dynamic analysis on our network and assess how the network changes over time. This technique could then be applied to identify potential insider threats within the Dark Web Forums by observing abnormal dynamic behaviour.

ACKNOWLEDGMENT

This work has been made possible through the EPSRC Scholarship as part of the Oxford University Centre for Doctoral Training in Cyber Security.

REFERENCES

[1] K. Crilley, "Information warfare: new battle fields terrorists, propaganda and the internet," in Aslib Proceedings, vol. 53, no. 7. MCB UP Ltd, 2001, pp. 250–264.

[2] Y. Zhou, E. Reid, J. Qin, H. Chen, and G. Lai, "Us domestic extremist groups on the web: link and content analysis," Intelligent Systems, IEEE, vol. 20, no. 5, 2005, pp. 44–51.

[3] P. B. Gerstenfeld, D. R. Grant, and C.-P. Chiang, "Hate online: A content analysis of extremist internet sites," Analyses of social issues and public policy, vol. 3, no. 1, 2003, pp. 29–44.

[4] S. Mehrotra, Intelligence and Security Informatics: IEEE International Conference on Intelligence and Security Informatics, ISI 2006, San Diego, CA, USA, May 23-24, 2006. Springer Science & Business Media, 2006, vol. 3975.

[5] H. Chen, S. Thoms, and T. Fu, "Cyber extremism in web 2.0: An exploratory study of international jihadist groups," in Intelligence and Security Informatics, 2008. ISI 2008. IEEE International Conference on. IEEE, 2008, pp. 98–103.

[6] V. E. Krebs, "Mapping networks of terrorist cells," Connections, vol. 24, no. 3, 2002, pp. 43–52.

[7] V. Krebs, "Uncloaking terrorist networks," First Monday, vol. 7, no. 4, 2002.

[8] A. Basu, "Social network analysis of terrorist organizations in india," in North American Association for Computational Social and Organizational Science (NAACSOS) Conference, 2005, pp. 26–28.

[9] C. C. Yang, N. Liu, and M. Sageman, "Analyzing the terrorist social networks with visualization tools," in Intelligence and security informatics. Springer, 2006, pp. 331–342.

[10] T. Stevens, "Regulating the ?dark web?: How a two-fold approach can tackle peer-to-peer radicalisation," The RUSI Journal, vol. 154, no. 2, 2009, pp. 28–33.

[11] J. Xu and H. Chen, "The topology of dark networks," Communications of the ACM, vol. 51, no. 10, 2008, pp. 58–65.

[12] G. L'Huillier, S. A. Ríos, H. Alvarez, and F. Aguilera, "Topic-based social network analysis for virtual communities of interests in the dark web," in ACM SIGKDD Workshop on Intelligence and Security Informatics, ser. ISI-KDD '10. New York, NY, USA: ACM, 2010, pp. 9:1–9:9. [Online]. Available: http://doi.acm.org/10.1145/1938606.1938615 [Accessed: 2015-10-19]

[13] T. Guardian, "Edward snowden | world news | the guardian," 06 2015. [Online]. Available: http://www.theguardian.com/world/edward-snowden [Accessed: 2015-10-10]

[14] E. Phillips, J. R. C. Nurse, M. Goldsmith, and S. Creese, "Applying social network analysis to security," in International Conference on Cyber Security for Sustainable Society, 2015, pp. 11–27.

[15] M. Hay, G. Miklau, D. Jensen, D. Towsley, and P. Weis, "Resisting structural re-identification in anonymized social networks," Proceedings of the VLDB Endowment, vol. 1, no. 1, 2008, pp. 102–114.

[16] M. Faloutsos, P. Faloutsos, and C. Faloutsos, "On power-law relationships of the internet topology," ACM SIGCOMM Computer Communication Review, vol. 29, no. 4, 1999, pp. 251–262.

[17] A. Golightly and D. J. Wilkinson, "Bayesian parameter inference for stochastic biochemical network models using particle markov chain monte carlo," Interface Focus, 2011, p. rsfs20110047.

[18] L. Getoor and C. P. Diehl, "Link mining: A survey," SIGKDD Explor. Newsl., vol. 7, no. 2, Dec. 2005, pp. 3–12. [Online]. Available: http://doi.acm.org/10.1145/1117454.1117456 [Accessed: 2015-10-18]

[19] S. Wasserman, Social Network Analysis: Methods and Applications. Cambridge University Press, Nov. 1994.

[20] L. C. Freeman, "Centrality in social networks conceptual clarification," Social networks, vol. 1, no. 3, 1979, pp. 215–239.

[21] I. Rogers, The Google Pagerank algorithm and how it works, 2002. [Online]. Available: http://mira.sai.msu.ru/~megera/docs/IR/search/pagerank/pagerank_explained.pdf [Accessed: 2015-09-15]

[22] D. Koschtzki, K. A. Lehmann, L. Peeters, S. Richter, D. Tenfelde-Podehl, and O. Zlotowski, "Centrality Indices," in Network Analysis, ser. Lecture Notes in Computer Science, U. Brandes and T. Erlebach, Eds. Springer Berlin Heidelberg, 2005, no. 3418, pp. 16–61. [Online]. Available: http://link.springer.com/chapter/10.1007/978-3-540-31955-9_3 [Accessed: 2015-05-19]

[23] J. M. Kleinberg, "Authoritative sources in a hyperlinked environment," J. ACM, vol. 46, no. 5, Sep. 1999, pp. 604–632. [Online]. Available: http://doi.acm.org/10.1145/324133.324140 [Accessed: 2015-10-15]

[24] C. Bron and J. Kerbosch, "Algorithm 457: finding all cliques of an

undirected graph," Communications of the ACM, vol. 16, no. 9, 1973, pp. 575 – 577. [Online]. Available: http://dl.acm.org/citation.cfm?id=362367 [Accessed: 2015-06-15]

[25] S. N. Soffer and A. Vazquez, "Network clustering coefficient without degree-correlation biases," Physical Review Series E-, vol. 71, no. 5, 2005, p. 057101.

[26] "Dark web forum portal." [Online]. Available: http://cri-portal.dyndns.org/portal/Home.action [Accessed: 2015-10-10]

[27] H. Chen, "Dark web forum portal," in Dark Web. Springer, 2012, pp. 257–270.

[28] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum likelihood from incomplete data via the em algorithm," Journal of the royal statistical society. Series B (methodological), 1977, pp. 1–38.

[29] S. B. Lyerly, "The average spearman rank correlation coefficient," Psychometrika, vol. 17, no. 4, 1952, pp. 421–428.

[30] G. S. Shieh, "A weighted kendall's tau statistic," Statistics & Probability Letters, vol. 39, no. 1, 1998, pp. 17–24.

[31] P. Sedgwick et al., "Pearsons correlation coefficient," BMJ, vol. 345, 2012.

Towards an Index of Mental Wellbeing in Language

The relationship between time orientation, self-focus and mood during prolonged bed-rest.

Schmer-Galunder, Sonja
Smart Information Flow Technologies
Minneapolis, U.S.A.
sgalunder@sift.net

Wu, Peggy
Smart Information Flow Technologies
Minneapolis, U.S.A.
pwu@sift.net

Ott, Tammy
Smart Information Flow Technologies
Minneapolis, U.S.A.
tott@sift.net

Miller, Chris
Smart Information Flow Technologies
Minneapolis, U.S.A.
cmiller@sift.net

Abstract— **Monitoring the mental wellbeing and psychosocial states of human operators in safety-critical domains is key to improving crew performance, safety and security. However, acquiring objective data and insights to mental wellbeing and psychosocial states among human operators is difficult and has relied on the subjective reports of operators, which are prone to biases and distortions. Team communications in a broad range of contexts from business organizations to high criticality workplaces such as emergency response, airplane pilots and spaceflight could be significantly improved with quick access to reliable and objective data about the psychosocial health of a team. Data on the psycho-social dimensions of collaborative teams, such as social distance, power dynamics, affect, and a team's comfort working together are typically highly subjective, not readily computationally tractable, and are collected using self-reports such as think-aloud protocols or surveys that can confound the behaviors being studied. By developing a method to collect data using non- or minimally intrusive methods requiring low participant effort coupled with automated data processing, we unshackle researchers from the burdens of hand-coding raw data and enable them to make empirically based discoveries more rapidly. This paper presents a validated, cost-effective and fast alternative to the shortcomings of current assessment methods. We present the results from an automatic text analysis tool applied to large amounts of written text (i.e., journals kept by participants in a bed rest study) in order to identify topics of interest, the emotional valence (positivity or negativity) of topics, as well as changes in these metrics over time. These topics and aspects of the text were identified computationally and automatically. This research was performed on different groups of subjects participating in NASA analog studies, where the primary goal of our investigation was the identification of changes to psychosocial states. Our results show that it is possible to predict mood based on journal entries alone using Latent Semantic Analysis** ***and that we are able to identify non-conscious variables impacting well-being over time.***

Keywords—Latent Semantic Analysis; Sociolinguistics; Wellness; Psychosocial State Detection; Sentiment Analysis; Data Mining;

I. Introduction

Relying on self-reports to make assessments about mental wellbeing and psychological health on a continous basis has many potential pitfalls. This is particularly true in the context of long duration experiments such as longitudinal studies that span weeks or months. Respondent fatigue, where participants' attention and motivation to respond drops, is a well documented problem that affects data quality. Survey data is also subject to a participant's memory, biases, vigilance, and personality (e.g., some participants are simply not inherently very self-aware or reflective). Further, self-reports can affect task performance if they impose additional workload, or if the act of reporting is disruptive to the behavior or environment being studied. For example, in an emergency response scenario. In such scenarios, data is often gathered using human observation, which is also prone to errors and omissions.

In the domain of space exploration, NASA has identified the need to monitor individual behavioral and mental health, which is crucial in ensuring high performance and mission success, especially in its vision for deep space, long duration exploration missions. Of particular importance are the detection of changes in mood and mental wellbeing. These changes can be particularly difficult for crew members to detect themselves as fatigue can impair self-reflection. Lack of sleep has been associated with negative mood such as depression and anxiety, however recent findings show that the relationship between sleep and mood is bi-directional, such that daytime mood impacts sleep quality [1]. Sleep deprivation may increase ruminating, defined as compulsively focusing attention on a source of distress. This tends to be associated with remembering negative events from the past rather than contemplating the future [2], and ruminating is considered a vulnerability factor for depression [3]. In contrast, regularly timed exercise impacts mood stability positively and shortened sleep decreases mood stability [4]. Interestingly, time orientation, whether one focuses on the past or the future, might be a mediating factor. For example, Stolarski et al. [5] has found that vivid recall of past negative events worsens reported wellbeing while recall of past positive events has an energizing effect.

Using Linguistic Inquiry Word Count (LIWC) analysis, researchers have consistently found that writing has beneficial effects on wellbeing. For example, changes in insight and causal words over the course of writing are indicative of better mental health, because these word groups represent understanding and organization of thoughts into a more coherent and cohesive personal narrative (Pennebaker and Francis). Other studies found that persons who use a moderate number of negative emotion words, had fewer doctor than those who uses very few or a lot of negative words.

However, while reflective writing can have therapeutic effects on mental wellbeing, reflection can also turn into rumination. Rumination involves focusing repeatedly and passively on what one is feeling and why one is feeling a certain way. It has been shown to increase feelings of anger and aggression [6] and leads to higher levels of depressive symptoms over time [7]. Typically, depressed people tend to think more negatively about the past and tend to recall negative memories of the past more often.

At the same time, self-focus is associated with negative affect, where rumination is a strong mediator [8]. Interestingly, Mor and Winquiest also found that private self-focus is more strongly associated with depression and generalized anxiety, whereas public self-focus is more strongly associate with social anxiety.

But self-focus can take on different perspectives with different outcomes on personal mood and wellbeing. Kross et al found that memories of the past can be described using a self-distanced or a self-immersed perspective. Typically, when people recall negative emotional events, they take on a more self-immersed perspective; re-experience the past event in the first person, through one's own eyes [9]. However, if people take on a more self-distanced perspective, people reconstruct the event in ways that promote insight and closure, leading to less negative affect [10]. This kind of perspective taking can be detected in writing through the use of personal pronouns, past/present/future tense and affect words. However, because these processes happen often in a non-conscious domain, automatic and objective analysis of language can be useful in identifying an index of mental wellbeing.

With the vast improvements in speed and processing power of modern computers, it is now possible to automatically analyze very large text corpora with sophisticated computational models that reveal hidden, or latent, structures in written language that are associated with psychological and emotional states. These models can then be used to accurately and reliably predict psychological and emotional states from writing. These states may not be detectable through human observation. Developing tools like this will be crucial for monitoring space crews' mental health on long haul missions.

The goal of this work is to demonstrate the feasibility of automatically, unobtrusively and objectively detecting changes in mood that may result from participating in prolonged bedrest. Related factors impacting mood can be disturbed sleep, lack of exercise and cognitive factors. Our hypothesis is that mood is expressed unconsciously in the journal writing of participants, represented through the valence, frequency and co-occurrence of certain words. This work furthers the search for reliable and unobtrusive ways to monitor the psychological health of crews in extreme environments.

In preparation for future Mars missions, methods to augment current practices for evaluating psychological health and factors impacting behavior and performance (i.e., mood, sleep, etc.) need to be explored. It is difficult to overstate the value of an accurate, objective, repeatable, efficient, accepted and non-intrusive means of assessing relevant psychosocial states. An automatic, non-intrusive acquisition and processing capability allows researchers to learn and validate virtually everything else necessary for the psychosocial health and performance of future space missions.

II. Method

A. Bedrest Study

This study was carried out at NASA's Flight Analog Research Unit (FARU) Bedrest Facility at the University of Texas Medical Branch (UTMB), see Figure 1. The FARU bedrest facility's primary focus was to study the effects of microgravity on human physiology, for example, the effects of exercises and testosterone treatments on muscle and bone mass loss in space. Participants underwent 14 days of intake protocol and were then confined to bed-rest for 70 days, followed by 14 days of recuperation and post-treatment protocols (for more information see [9]). Some subjects work with trainers and exercise up to 9 times per week while maintaining a head-down position. Participants are selected for age, physiological and psychological characteristics analogous to the astronaut population. Subjects maintained a 6-degree head down angle for all activities during the 70-day bed rest period. They were monitored by a human 24/7 to ensure compliance. There were

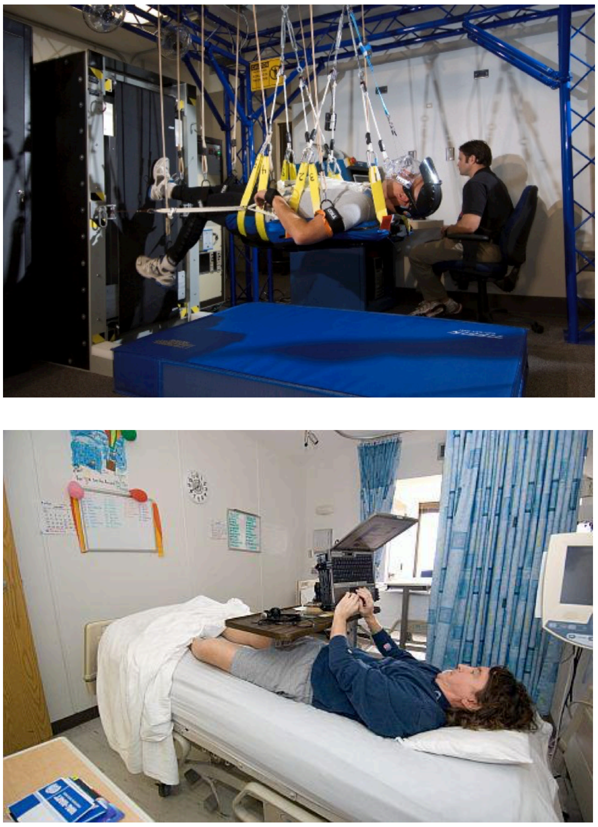


Figure 1. Bedrest Facilities UTMB FARU: (a) Bedrest Subjects on Vertical Treadmill. (b) After an inital 2 weeks of baseline data collection, subjects maintain a 6 degree headdown position for 70 days.

4 different conditions: exercise (E), control (C), exercise & testosterone (EX&T) and freefly (F). Subjects in the exercise condition trained in 9 sessions per week with a human trainer; in the testosterone condition exercisers were additionally injected with a very small amount of testosterone. The control group did not exercise but received a placebo injection. The freeflly condition did mild exercise and no injection/placebo. Physical health, as well as diet, were closely monitored and managed throughout the study. A clinical psychologist evaluated the psychological health of all subjects on a weekly basis.

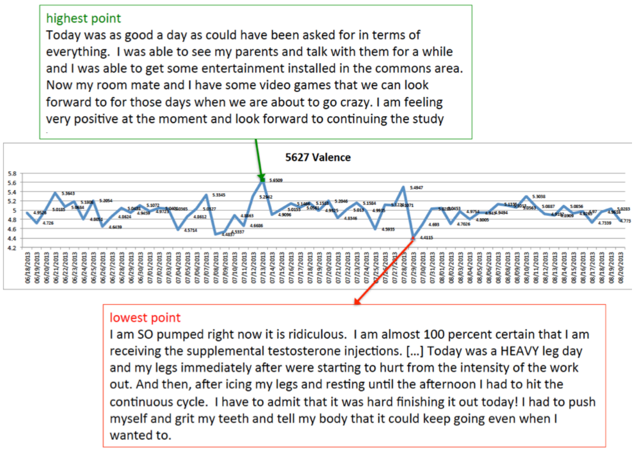
We collected journal data from 16 bedrest subjects with an average of 83.75 entries per subject. The total word count (WC) of the data corpus is 210,943, with the mean number of words per entry being 174.77. 59,371 (28%) words were provided during 2 weeks of pre-bedrest, with a mean of 242.33 per entry. 132,925 (64%) words were provided during bedrest, with a mean of 165.74 per entry. 18,647 (8%) words were provided during the 2 weeks of post bedrest, with a mean of 116.54 per entry. For this analysis, we included a total of 1174 journal entries.

B. *Measures and surveys*

Subjects were asked to complete the State Trait Anxiety Inventory (STAI) [11] once as part of intake, and complete the Positive and Negative Affect Schedule (PANAS) once daily. At the end of each day, they complete a 20-minute journal writing session, followed by a short survey asking general questions about their wellbeing, time orientation, self-vs. other focus, depth of thoughts and physical state. After they complete the 70-day bed rest, they are debriefed on their journaling results and general experience. Facility staff completed daily surveys about their general observations of a subject. The instructions for the written journal are as follows: “Reflect on your feelings throughout the day. Write in a quiet space without distractions for roughly 20 minutes. Make sure to write at least a half page of text (at least 250 words), though feel free to write more if you wish. Do not worry about spelling, grammar, or repetition.”

C. *Latent Semantic Analysis and Valence calculations*

Latent Semantic Analysis (LSA) is a computational algorithm that captures the occurrence and relative position of words and their semantic equivalents, and in our case also the valence of texts of interest. The term “valence” refers to general pleasantness vs. unpleasantness and it refers to a numeric value that we are able to deduce from the general tone of any text of interest, ranging from 0 (low valence) to 9 (high valence). To calculate valence, we used 1) LSA to generate a high-dimensional semantic space from the entire text corpus, resulting in a matrix where the columns correspond to text contexts (paragraphs, clauses or any other unit comprising a number of words), while every individual word is represented by a row. The matrix describes a high-dimensional space, which we reduced to 100 dimensions for computational efficiency. This is also the optimal number of dimensions that captures the semantic relationships in the text. The model depicts word incidences in different contexts. 2) We then used the Affective Norms of English Words (ANEW) [12] as a dependent measure for a multiple linear regression across the 100 dimensions in the semantic space. The result is the best-fitted valence of the words in the ANEW word list. This yields a predicted valence value for each expression of interest (words, sentences, documents), taking its surrounding context into consideration. Figure 2 shows two writing samples and calculated high/low valence points, based on the content.



highest point
Today was as good a day as could have been asked for in terms of everything. I was able to see my parents and talk with them for a while and I was able to get some entertainment installed in the commons area. Now my room mate and I have some video games that we can look forward to for those days when we are about to go crazy. I am feeling very positive at the moment and look forward to continuing the study

lowest point
I am SO pumped right now it is ridiculous. I am almost 100 percent certain that I am receiving the supplemental testosterone injections. [...] Today was a HEAVY leg day and my legs immediately after were starting to hurt from the intensity of the work out. And then, after icing my legs and resting until the afternoon I had to hit the continuous cycle. I have to admit that it was hard finishing it out today! I had to push myself and grit my teeth and tell my body that it could keep going even when I wanted to.

Figure 2. Samples of positive and negative valence of journal entries.

Using this method we calculated valence values for all journal entries and identified high and low points, captured changes over time in mood and we could characterize the contextual circumstances in which a topic is discussed. This essentially provides a method for quantitative analysis of highly qualitative data, enabling, for example, cross-validation of calculated valence with subject survey ratings for affect.

D. *Linquisitc Inquiry and Word Count (LIWC)*

Linguistic Inquiry and Word Count (LIWC) is a text analysis program based on work by Pennebaker, Booth and Francis [12]. LIWC counts the frequency of an occurrence of a word (or word root) in a given body of text and computes the percentage of total words in defined linguistic categories. The categories include negative emotion words (sad, angry), positive emotion words (happy, laugh), standard function word categories (first, second, and third person pronouns, articles, prepositions), and various content categories (e.g., religion, death, occupation). Words were assigned to specific categories by having groups of judges evaluate the degree to which about 2000 words or word stems were related to each of several dozen categories. LIWC uses the developed "dictionaries" to provide the percent of a given text that can be found in each category.

The computed LIWC score has been shown to correlate with specific attitudes and attributes in the writing of many populations (see [13], [14] for summaries). For example, a category of words associated with cognition and cognitive mechanisms is said to be associated with female speakers/writers, with negative emotions, and with complex reappraisals of situations such as following trauma [13], as well as increased mental health [14].

III. Results

A. *Valence predictions of mood over time*

Correlational analysis showed that general LSA valence for daily journal entries correlates with daily PANAS survey responses over all bed-rest subjects. Positive Affect (PA) and Negative Affect (NA) scores show significant correlations in the expected direction, where PA correlates positively with valence (r*s*[1174]= 0.159, p<0.001) and NA correlates negatively with valence (r*s*[1174]= -0.105, p<0.001). Thus, using automated tools it is possible to identify general mood without having to ask subjects to fill our surveys.

Moreover, using LSA, it is also possible to pick a random journal entry and *predict* with 63.6 % accuracy the experimental condition (E, C, T or F) of single subjects. The same analysis can be done with survey responses as the predictor variable. This analysis yields an accuracy rate of 59% accuracy for PANAS scores of single journal entries. While the accuracy may seem low, it is important to note that emotional variety of journal entries is relatively high, while PANAS scores represent a single score for affect. Moreover, text-based predictions of PANAS scores can be complimented by the capability to recognize changes such as upward or downward emotional trends, as well as the ability to examine the co-occurring predictions of other factors such as social connectedness, physical wellbeing or sleep. However, using this method, we can reliably predict features like experimental condition or at which day during the 90-day BR a particular journal entry has been written (Figure 3).

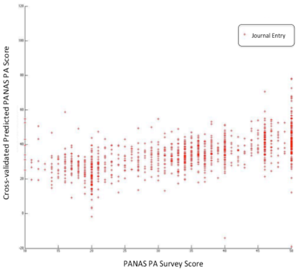


Figure 3. Predicted PANAS score of

B. *Positive and Negative topic identification and time orientation in corpora*

Apart from simply tracking general mood over time, we investigated larger text corpora (all journal entries of all subjects in the bed-rest study) in order to identify words (topics) with significantly higher (lower) valence in comparison to the rest of the corpus.

Figure 4(a) and 3(b) shows the result of this analysis - visualizations of arising topics in word clouds, where the size of the word indicates higher frequency/importance. During this study most positive topics contexts were "home", "friends", "games", "family", while words like "muscle", "pain", "sleeping". It is also interesting to note that time orientation in negative contexts is oriented towards the past as represented by the significantly more negative valence of words like "was", "were", "told", "didn't", "then" or "done" in comparison to words representative of the present, i.e. "today", "be", "play", "do", "make" or "here", which can be found in journal entries having more positive valence. We can also find that focus is more placed on "others" in positive valence contexts (i.e., "you", "friends", "family"), are more on the self in negative contexts (in particular own body parts i.e., "stomach", "leg", "thigh" or "shoulder"). These topics are derived from the sum of all journal entries and represent most frequent, or salient topics.

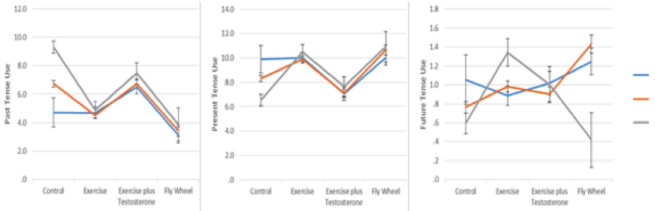
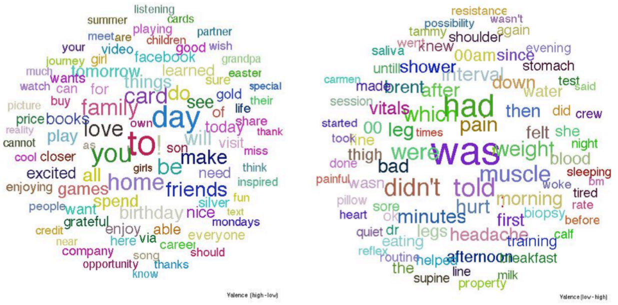


Figure 5. (a) Frequency of past tense by condition (Control, Exercise, Exercise&Testosterone), FlyWheel. (b) Frequency of present tense by condition (c) Frequency of future tense by condition.



4(a).

4b).

Figure 4. Positive and negative valenced word clouds in LSA space. (a) Substraction of low valenced words (word usage in positive contexts). (b) Substraction of high valenced (negative contexts).

C. *Self-focus, Time Orientation and Mood*

Because our analysis focused preliminary on the relationship between automated analysis of linguistic content and subjective survey responses, we focused on identifying correlations between subjectively reported experience and journal content. What we found in our data is that persons who responded on the self-focus survey being more "other-focused" also reported more Positive Affect (PA) on the PANAS scales, and less Negative Affect (NA), while persons who are more self-focused also feel more negative, and less positive affect. Correlations are significant at p=0.000 for PA with r=0.483 and NA r= -0.395. More self-focus is related to higher LIWC past tense word frequencies, while focus on others is related to less focus on the past. However, we found a negative correlation between past tense words and the word "I" (r=-0.151, p<0.000), and more "I" when writing about the present (r=0.286, p<0.000) and the future (r=0.182, p<0.000), meaning that subjects use less "I" when writing about the past.

One would expect that self-focus as expressed in the use of "I" would yield higher frequencies of "I" in journal entries with a past-focus. However, a more detailed valence analysis revealed that the valence of the word "I" is low in journal entries, which have a focus on the past, while valence of "I" is high when journal entries are more present and future oriented. Further correlational analysis showed that the use of personal pronouns is high when NA is high (r= 0.138, p<0.000) and low when PA is high (r=-0.159, p<0.000). Interestingly, both PA and NA affect responses are low when frequencies of past tense use is high (PA: r= -.263, p<0.000 and NA: r= -.148, p<0.000) while journal entries with a present-focus correlate positively with both PA (r= 0.125, p<0.000) and NA (r=0.221, p<0.000). We found further support for theories that link self-focus with anxiety and negative mood in correlations between time orientation and the STAI. State as well as trait anxiety correlated positively with frequency of past tense use (r= 0.224, p<0.000, r= 0.145, p<0.000), and negatively with future tense use (r = -0.202, p<0.000 for state anxiety, no sig. correlation for state anxiety). As expected, PANAS PA scores correlate negatively with STAI responses (state anxiety: r=-0.226, p<0.000, trait anxiety: r=-0.433, p<0.000) and positively with PANAS NA responses (state anxiety: r= .304, p<0.000, trait anxiety: r=0.196, p<0.000).

In sum, these results confirm current theories that a focus on self correlates negative mood and the recalling of an event in the past, while a focus on others is related to a more positive mood.

D. *Difference between Bedrest Conditions*

Next we looked at differences between conditions. Using a 4 Condition (Control, Exercise, Exercise&Testosterone, Freefly) x 3 Headdown (Head up, Head down, Post-Bedrest) ANOVA analysis, we found a main effect for condition for general valence F(3,1171)=6.697, p=0.000. Controls have lowest overall valence (M=4.95), Exercisers have higher overall valence (M=5.02), followed by the EX&T group (M=5.11) and the freefly group (M=5.09). We also found a main effect for PA F(3,1162)=5.293, p=0.001 and NA F(3,1162)=5.233, p=0.000. Thus, differences between the conditions are significant for both automatically retrieved as well as reported mood. However, there are some interesting differences. While generally speaking, Controls have lowest affect (low valence, high NA and low PA), subjects who exercised have higher reported PA *and* NA as well as higher valence in their journal entries. Subject who exercised, but also received testosterone report both lower PA and NA relative to the exercisers, while their overall valence is highest among all conditions. The freefly condition reports moderate PA, high NA and moderate/high valence. Clearly, there are differences in mood and wellbeing among the different experimental conditions, but in order to understand driving variables, we wanted to understand if self-focus and time orientation play a role. we looked for differences in past, present and future tense usage. We also found an interaction for past tense use, F(6,1172)=2.476, p<.05, present tense use, F(6,1172)=2.275, p<.05, and future tense use, F(6,1172)=3.683, p<.01, see Figure 5. Only the control group differed in past and present tense usage during the different phases of the study. While all groups differed in future tense use. The control group used the most future tense during head-up, the EX group used the most during post bedrest, while the Flywheel used the most during head-down. There was no difference for the EX+T group.

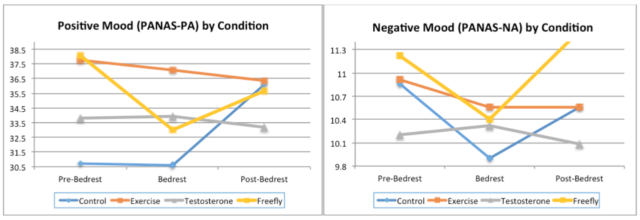


Figure 6. Scores by Group and and Headdown Conditions. (a) Positive Affect (PANAS PA), (b) Negative Affect (PANAS NA).

What we found is that Controls report less negative affect and less positive affect when talking about the past, but negative affect when talking about the future. This pattern is reversed for the EX&T condition, which report high PA when talking about the future and very low PA when talking about the past. Maybe the most interesting finding is that the EX&T condition, which has highest overall valence, uses more 3rd person pronouns when talking about the past, but no 3rd person pronouns when talking about the present, but instead more "I". Together with a high correlation of positive affect and present term word usage, this group may have highest overall valence because they are taking on a more self-distanced perspective, as reflected in 3rd person pronoun usage, more positive affect related to the "now" and an absence of negative affect. What is also interesting is that for the Controls, insight and cause words are less used in journal entries focusing on the past in comparison to the other condition, while insight worlds are used by all conditions, but the freefly condition in present-focused journal entries.

IV. Conclusions

Our results show that automated linguistic analyses on journaling data can provide insights into the mental wellbeing of single persons as well as a number of people of interest. Using LSA derived valence values, we found that is it possible to predict characteristics about journal entries, i.e. time when it was written or experimental condition, by simply looking at the semantic word contexts. Computational methods that are able to predict self-reported survey responses are an attractive alternative to time-consuming survey responses. These methods may also provide the means to send automatic messages back to earth that a team might be in difficulty early enough to take countermeasures. We were also able to identify variables impacting general mood, i.e. time orientation and self/other focus. What we found is that generally, people who are more self-focused also show more negative affect. However, taking on a distanced (vs. immersed) self-perspective, as reflected in the use of 3rd person pronouns as well as insight words, an indicator of better mental wellbeing can be found. For our purposes, PA reflects the activation of positive emotions, feeling enthusiastic, active and alert. High PA is a state of high energy and pleasurable engagement, whereas low PA is characterized by sadness and lethargy. In contrast, NA describes a general dimension of subjective distress and engagement in negative moods such as anger, fear, disgust or nervousness. Low NA is therefore a reflection of not being engaged in negative moods, and may include feeling calm and relaxed. Thus, what we found is that automated textual analysis of valence is able to identify activation of PA and NA. We believe that there are large individual differences in outward expression of emotion either in written or verbal communications, thus correlations may be stronger when averaged over either longer periods of time, or using higher numbers of observations. Longer periods with greater numbers of observations provide individuals with the opportunity to express their emotions through behaviors. This is consistent with our preliminary analysis of verbal communications data, which contains much more textual volume per sample, as well as stronger correlations between individual LSA valance components and PANAS.Within the context of long duration missions due to the lack of real-time communication, it is crucial to be able to detect and predict relevant behavioral and mental health metrics of the crew over extended periods of time. In other domains such as healthcare and interpersonal skills training where there is an abundance of narrative data, a process for automatic analysis is sorely needed to reduce the labor costs of manually coding observations.

While the use of journal writing may impose workload on the participant, the writing itself may in fact have positive psychological benefits [15]. In addition to positive and negative affect, the journal text can provide a rich dataset for analyzing multiple factors and sentiment on environmental factors and other contexts. Furthermore, once rich models are developed the system could guide crewmembers in their writing by suggesting the use of, for example, more 3rd person pronouns. This simple change in writing style might improve mood. Unlike surveys, free-form text and communications transcripts may also be mined for constructs that were not conceived a-priori to data collection. Minimally intrusive method of detecting attitudes towards environmental or situational factors and emotional state can help individuals make the connection between precipitating events to changes in moods and attitudes. For example, studies have shown that for 'healthy' persons, narratives of highly stressful events change over the course of time, including expect of their environment and new experience. But there is little change over time to the linguistic content of the narrative of less resilient persons [16]. Moreover, this method can be used for self-monitoring and can help increase individual self-awareness. It can also inform the author about how a message might be perceived by others before the message is sent. Thus, our results are able to provide an objective solution to identifying psycho-social dimensions wellbeing completely unobtrusively.

ACKNOWLEDGMENT

This work was sponsored by NASA's Human Research Program under contract #NNX12AAB40G. We would like to thank our NASA sponsors Lauren Leveton, Laura Bollweg, Brandon Vessey, Holly Patterson, the BHP element, and the subjects and staff at the various spaceflight analog facilities for their oversight, direction, and support.

REFERENCES

[1] M. Vandekerckhove and R. Cluydts, "The emotional brain and sleep: an intimate relationship," Sleep Med Rev, vol. 14, pp. 219-226, 2010.

[2] J. Smallwood and R. O'Connor, R. C., "Imprisoned by the past: Unhappy moods lead to a retrospective bias to mind wandering," Cognition and Emotion, vol. 25(8), pp. 1481–1490, 2011.

[3] S. Nolen-Hoeksema, B. E. Wisco, and S. Lyubomirsky, "Rethinking rumination," Perspectives on Psychological Science, vol. 3(5), pp. 400–424, 2008.

[4] J. R. Bowen, L. Balbuena, M. Baetz, and L. Schwartz, "Maintaining sleep and physical activity alleviate mood instability," Preventive Medicine, vol. 57(5), pp 461–465, 2013.

[5] M. Stolarski, G. Matthews, S. Postek, P. Zimbardo, and J. Bitner, "How We Feel is a Matter of Time: Relationships Between Time Perspectives and Mood, Journal of Happiness Studies, " vol. 15 , pp. 809-827, 2013.

[6] B.J. Bushman, "Does venting anger feed or extinguish the flame? Catharsis, rumination, distraction, anger, and aggressive responding," Personality and Social Psychology Bulletin, vol. 28, pp. 724-731, 2002.

[7] S. Nolen-Hoeksema and J. Morrow, "The effects of rumination and distraction on naturally occurring depressed mood," Journal of Abnormal Psychology, vol. 102, pp. 20-28, 1993.

[8] N. Mor and J. Winquist, J., "Self-focused attention and negative affect: A meta-analysis," Psychological Bulletin, vol. 128, pp. 638-662, 2002.

[9] For more information on the NASA's bedrest study, please see https://bedreststudy.jsc.nasa.gov/

[10] E. Kross and O. Ayduk, "Facilitating adaptive emotional analysis: Distinguishing distanced-analysis of depressive experiences from immersed-analysis and distraction," Pers. Soc. Psychol. Bulletin, vol 34, pp. 924–938, 2008.

[11] E. Kross, Ayduk, O., and W. Mischel, "When asking 'why' does not hurt: Distinguishing rumination from reflective processing of negative emotions," Psychological Science, vol. 16, pp. 09–715, 2005.

[12] C. D. Spielberger, R. L. Gorssuch, P. R. Lushene, P.R. Vagg, and G. A. Jacobs. "Manual for the State-Trait Anxiety Inventory. Consulting Psychologists Press, Inc., 1983.

[13] J. Pennebaker, R. Booth, and M. Francis, M, "Operator's Manual: Linguistic Inquiry and Word Count: LIWC2007" [Online]. Available from:http://homepage.psy.utexas.edu/homepage/faculty/pennebaker/reprints/LIWC2007_OperatorManual.pdf. 2013.08.28

[14] J. Pennebaker, "The Secret Life of Pronouns," New York: Bloomsbury, 2011.

[15] M. Purcell, M, "The Health Benefits of Journaling," [Online]. Available from: http://psychcentral.com/lib/the-health-benefits-of-journaling/000721, 2012.11.04.

[16] Y. Tausczik and J. Pennebaker, "The Psychological Meaning of Words: LIWC and Computerized Text Analysis Methods," Journal of Language and Social Psychology, vol. 29(1), pp 24–54, 2010.

[17] B. A. van der Kolk and W. Kadish, "Amnesia, dissociation, and the return of the repressed," In B.A. van der Kolk (Ed.), Psychological Trauma. American Psychiatric Press, Inc., Washington, D.C., 1987.