WEB 2020

The Eighth International Conference on Building and Exploring Web Based Environments

ISBN: 978-1-61208-789-4

September 27th – October 1st, 2020

WEB 2020 Editors

Daniela Marghitu, Auburn University, USA

WEB 2020

Forward

The Eighth International Conference on Building and Exploring Web Based Environments (WEB 2020) continued a series of events on Web-related theoretical and practical aspects, focusing on identifying challenges for building Web-based useful services and applications, and for effectively extracting and integrating knowledge from the Web, enterprise data, and social media.

The Web has changed the way we share knowledge, the way we design distributed services and applications, the way we access large volumes of data, and the way we position ourselves with our peers. Successful exploitation of Web-based concepts by Web communities lies on the integration of traditional data management techniques and semantic information into Web-based frameworks and systems.

We take here the opportunity to warmly thank all the members of the WEB 2020 technical program committee, as well as all the reviewers. The creation of such a high quality conference program would not have been possible without their involvement. We also kindly thank all the authors who dedicated much of their time and effort to contribute to WEB 2020. We truly believe that, thanks to all these efforts, the final conference program consisted of top quality contributions. We also thank the members of the WEB 2020 organizing committee for their help in handling the logistics of this event.

WEB 2020 Chairs

WEB 2020 Steering Committee

Michel Jourlin, Jean Monnet University, Saint-Etienne, France
Daniela Marghitu, Auburn University, USA
Mariusz Trzaska, Polish-Japanese Academy of Information Technology, Poland
Erich Schweighofer, University of Vienna - Centre for Computers and Law, Austria
Taketoshi Ushiama, Kyushu University, Japan

WEB 2020 Publicity Chair

Joseyda Jaqueline More, Universitat Politecnica de Valencia, Spain
Marta Botella-Campos, Universitat Politecnica de Valencia, Spain

WEB 2020 Industry/Research Advisory Committee

Krzysztof Walczak, Poznan University of Economics, Poland
Toyohide Watanabe, Nagoya Industrial Science Research Institute, Japan
Demetrios Sampson, Curtin University, Australia
Alexiei Dingli, University of Malta, Malta
Hossein Sarrafzadeh, Unitec Institute of Technology, New Zealand

WEB 2020
Committee

WEB 2020 Steering Committee

Michel Jourlin, Jean Monnet University, Saint-Etienne, France
Daniela Marghitu, Auburn University, USA
Mariusz Trzaska, Polish-Japanese Academy of Information Technology, Poland
Erich Schweighofer, University of Vienna - Centre for Computers and Law, Austria
Taketoshi Ushiama, Kyushu University, Japan

WEB 2020 Publicity Chair

Joseyda Jaqueline More, Universitat Politecnica de Valencia, Spain
Marta Botella-Campos, Universitat Politecnica de Valencia, Spain

WEB 2020 Industry/Research Advisory Committee

Krzysztof Walczak, Poznan University of Economics, Poland
Toyohide Watanabe, Nagoya Industrial Science Research Institute, Japan
Demetrios Sampson, Curtin University, Australia
Alexiei Dingli, University of Malta, Malta
Hossein Sarrafzadeh, Unitec Institute of Technology, New Zealand

WEB 2020 Technical Program Committee

Céline Alec, Université de Caen-Normandie, France
Leandro Antonelli, Lifia | Universidad Nacional de La Plata (UNLP), Argentina
Sofia Athenikos, Twitter, USA
Ismail Badache, LIS INSPE Aix-Marseille University, France
Maxim Bakaev, Novosibirsk State Technical University, Russia
Efthimios Bothos, Institute of Communications and Computer Systems (ICCS), Athens, Greece
Christos J. Bouras, University of Patras, Greece
Tharrenos Bratitsis, University of Western Macedonia, Greece
Rodrigo Capobianco Guido, São Paulo State University (UNESP), Brazil
Nadezda Chalupova, Mendel University in Brno, Czech Republic
Dickson Chiu, The University of Hong Kong, Hong Kong
Stefano Cresci, Institute of Informatics and Telematics (IIT) - National Research Council (CNR), Italy
Toon De Pessemier, Ghent University, Belgium
Annamaria Ficara, University of Palermo, Italy
Giacomo Fiumara, Università degli Studi di Messina, Italy
Raffaella Folgieri, Università degli Studi di Milano, Italy
Piero Fraternali, Politecnico di Milano, Italy
Marco Furini, University of Modena and Reggio Emilia, Italy
Jose Garcia-Alonso, University of Extremadura, Spain
Abigail Goldsteen, IBM Research - Haifa, Israel
Denis Gracanin, Virginia Tech, USA
Tor-Morten Grønli, Kristiania University College, Norway
Allel Hadjali, LIAS/ENSMA, Poitiers, France
Sebastian Heil, Technische Universität Chemnitz, Germany

Tzung-Pei Hong, National University of Kaohsiung, Taiwan
Muhammad Nihal Hussain, University of Arkansas at Little Rock, USA
Ivan Izonin, Lviv PolytechnicNational University, Ukraine
Yuji Iwahori, Chubu University, Japan
Girish Nath Jha, School for Sanskrit and Indic Studies - JNU, New Delhi, India
Roula Karam, Antares Vision, Italy
Sotirios Karetsos, Agricultural University of Athens, Greece
Hassan A. Karimi, University of Pittsburgh, USA
Kimmo Kettunen, The National Library of Finland, Finland
Pinar Kirci, Istanbul University-Cerrahpasa, Turkey
Fotis Kokkoras, University of Thessaly, Greece
Samad Kolahi, UNITEC, New Zealand
Anirban Kundu, Netaji Subhash Engineering College (under MAKAUT) / Computer Innovative Research Society, West Bengal, India
Claudio Lucchiari, Università degli studi di Milano, Italy
Avinash Malik, UniversityofAuckland, New Zealand
Abdul-Rahman Mawlood-Yunis, Wilfrid Laurier University, Waterloo, Canada
Daniela Marghitu, Auburn University, USA
Michele Melchiori, Università degli Studi di Brescia, Italy
Hector Menendez Benito, Middlesex University London, UK
Amin Mesmoudi, LIAS/ University of Poitiers, France
Manfred Meyer, Westfälische Hochschule - University of Applied Sciences, Bocholt, Germany
Héctor Migallón, Miguel Hernandez University, Spain
Andrea Millimaggi, Politecnico di Milano, Italy
Debajyoti Mukhopadhyay, Mumbai University, India
Prashant R. Nair, Amrita Vishwa Vidyapeetham University, India
Alexey Noskov, Philipps University of Marburg, Germany
Mahda Noura, Technische Universität Chemnitz, Germany
Guadalupe Ortiz, University of Cadiz, Spain
Wieslaw Paja, University of Rzeszów, Poland
Roberto Panerai Velloso, Universidade Federal de Santa Catarina, Brazil
Giuseppe Patane’, CNR-IMATI, Italy
Vladia Pinheiro, University of Fortaleza, Brazil
Agostino Poggi, DII - University of Parma, Italy
Tarmo Robal, Tallinn University of Technology, Estonia
Christophe Roche, Université Savoie Mont-Blanc, France
Marek Rychly, Brno University of Technology, Czech Republic
Fayçal Rédha Saidani, University of Mouloud Mammeri, Tizi-Ouzou, Algeria
Demetrios Sampson, University of Piraeus, Greece / Curtin University, Australia
Suzanna Schmeelk, St. John's University, USA
Mirco Schoenfeld, Technical University of Munich, Germany
Valentin Siegert, Technische Universität Chemnitz, Germany
Günther Specht, University of Innsbruck, Austria
George Tambouratzis, Institute for Language and Speech Processing - Athena R.C., Greece
Mariusz Trzaska, Polish-Japanese Academy of Information Technology, Poland
Taketoshi Ushiama, Kyushu University, Japan
Krzysztof Walczak, Poznan University of Economics, Poland
Liang Wu, Airbnb, USA

Jasy Liew Suet Yan, Universiti Sains Malaysia, Malaysia

Copyright Information

For your reference, this is the text governing the copyright release for material published by IARIA.

The copyright release is a transfer of publication rights, which allows IARIA and its partners to drive the dissemination of the published material. This allows IARIA to give articles increased visibility via distribution, inclusion in libraries, and arrangements for submission to indexes.

I, the undersigned, declare that the article is original, and that I represent the authors of this article in the copyright release matters. If this work has been done as work-for-hire, I have obtained all necessary clearances to execute a copyright release. I hereby irrevocably transfer exclusive copyright for this material to IARIA. I give IARIA permission or reproduce the work in any media format such as, but not limited to, print, digital, or electronic. I give IARIA permission to distribute the materials without restriction to any institutions or individuals. I give IARIA permission to submit the work for inclusion in article repositories as IARIA sees fit.

I, the undersigned, declare that to the best of my knowledge, the article is does not contain libelous or otherwise unlawful contents or invading the right of privacy or infringing on a proprietary right.

Following the copyright release, any circulated version of the article must bear the copyright notice and any header and footer information that IARIA applies to the published article.

IARIA grants royalty-free permission to the authors to disseminate the work, under the above provisions, for any academic, commercial, or industrial use. IARIA grants royalty-free permission to any individuals or institutions to make the article available electronically, online, or in print.

IARIA acknowledges that rights to any algorithm, process, procedure, apparatus, or articles of manufacture remain with the authors and their employers.

I, the undersigned, understand that IARIA will not be liable, in contract, tort (including, without limitation, negligence), pre-contract or other representations (other than fraudulent misrepresentations) or otherwise in connection with the publication of my work.

Exception to the above is made for work-for-hire performed while employed by the government. In that case, copyright to the material remains with the said government. The rightful owners (authors and government entity) grant unlimited and unrestricted permission to IARIA, IARIA's contractors, and IARIA's partners to further distribute the work.

Table of Contents

A Combined Approach to Dynamic Web Page Classification: Merging Structure and Content 1
Maria Niarou and Sofia Stamou

A Web-Based Platform to Teach Music Online 10
Fatemeh Jamshidi and Daniela Marghitu

Building Web-based Environment to Support Sponsored Research and University-wide Collaborations 16
Fatemeh Jamshidi, Abhishek Jariwala, Bibhav Bhattarai, Katherine Abbate, Daniela Marghitu, and Mallory Lucier-Greer

A Combined Approach to Dynamic Web Page Classification: Merging Structure and Content

Maria Niarou, Sofia Stamou
Department of Archives, Library and Museum Studies
Ionian University
Corfu, Greece
E-mails: {l14niar, stamou}@ionio.gr

Abstract — **Web data is constantly increasing at a very high pace. So does the need to come up with methods and tools that are able to process, organize and store this data effectively. To meet this need, several approaches have been proposed in the literature over the last decades, a critical amount of which focus on methods for classifying Web content in order to be able to retrieve relevant information in a cost-effective yet effortless manner. Motivated by the observation that the Web is changing not only with respect to content but also with respect to structure, we designed a combined classification method that encounters both textual and structural elements in the Web pages under examination. Our classification approach, presented here, investigates a number of parameters before assigning a Web page to a suitable category(-ies). A preliminary experimental evaluation of our method indicates that it accurately classifies web content both thematically and structurally.**

Keywords - Web data; dynamic data; similarities; semantics; classification; Web data structure.

I. Introduction

Many researchers have studied Web pages' classification. Existing research falls in two main categories: content-based and structure-based classification. Content-based classification tries to assign every Web page to a suitable thematic category [1]. To enable that, many approaches have been proposed for processing the textual content in Web pages, extracting thematic keywords, mapping them to existing ontologies and, therefore, identifying the most appropriate theme of the page. On the other hand, structure-based classification [2] relies on the pages' structural properties such as links, images etc. for grouping together pages of similar structure.

Despite the effectiveness of many of the proposed approaches, Web data classification still remains an open research challenge, basically because Web data is: (a) voluminous, (b) heterogeneous and (c) dynamic. In particular, the voluminous amount of online data makes it practically impossible to categorize every Web page regardless of the resources' power and availability. Web content is largely heterogeneous as it is represented via text, audio-visual material, multimedia, etc. Heterogeneity suggests that different elements should be encountered when trying to classify Web data and that different tools should be employed for elements' processing. This variance entails considerable communication overhead and increased complexity in any classification technique. Web content is dynamic and it constantly changes structurally and textually. Therefore, any classification attempt should account for the volatile nature of the material at hand and operate in a way that minimizes the need to re-examine already classified data unless it has significantly changed over time. The frequency of Web content change is also critical in Web data classification approaches because some pages might exhibit significant changes very often (e.g. news sites) while others might not change at all in their lifespan.

Driven by the idea that textual and structural classification are complementary, we designed a combined Web page classification approach that we present here. Our approach examines both content and structure for organizing Web pages and operates from an information retrieval point of view in the sense that it tries to group together pages that can serve similar information needs, thus lowering thus the cost and effort associated with user Web searches. Our method builds upon existing works and combines in a novel manner, elements in the Web pages' content and structure before concluding on the most appropriate category to assign every Web page. On top of that, our method accounts for the Web pages' changes (structural and textual) over time and, depending on the amount and frequency of changes, it reclassifies pages accordingly.

The experimental evaluation of our approach shows that our method manages to accurately classify Web pages when considering both their structure and contents, therefore implying that the combined investigation of structural and textual elements is successful in grouping together Web pages of similar themes or purposes.

The rest of the paper is organized as follows. In Section II, we review the related work. In Section III, we present the details of our classification approach and we justify the decisions we made with respect to the considered elements. Our approach combines three distinct yet complementary algorithms, which we discuss in detail. Section IV presents a preliminary evaluation we carried out in order to assess the effectiveness of our approach and reports the obtained results. We conclude the paper in Section V, where we also sketch our plans for future work.

II. Related Work

Web data classification has been a research challenge over the last decades [3]. To this end, several approaches have been proposed in the literature aiming at the effective and automated classification of Web data. As already we mentioned in the previous section, the proposed methods fall into two main categories: those that rely on the analysis of Web pages' content in order to organize them into thematically related groups, and those that exploit the pages' structural properties for enabling their classification.

The first approach focuses on the processing of text present in Web pages in order to learn the thematic categories it relates to. In this respect, existing works rely on machine learning for building classifiers, the most known of which are Naïve Bayes [4], Support Vector Machines [5], and decision trees [6]. Moreover, content-based classification techniques utilize semantic networks, ontologies, and hierarchies to create object clusters and, exploiting the relationships between the object categories, they organize the Web pages thematically [7]. The commonality across content-based classification methods is that they apply text pre-processing techniques to extract thematic keywords from text and, based on the frequency [8] and the (semantic) proximity of those keywords in the vector space representation, they are able to deduce their thematic category. Content-based classification can be greatly successful when classification algorithms have undergone a thorough training phase [9].

On the other hand, structure-based classification tries to organize Web pages based on their link properties [10], user clicks [11], sentiment analysis [12], etc. Despite the reduced time and effort associated with structural Web page classification, URL-based classifiers have to deal with a very small amount of information present in URLs, which is also noisy and may contain irrelevant features. Therefore, feature selection methods need to be applied, a process that increases the complexity of classification systems [13].

Although many of the above-mentioned techniques have proven successful, Web pages' classification still remains a challenging research quest for several reasons, but foremost because of the Web's volatile nature, which requires the re-examination of already classified Web pages. In this paper, we built upon existing research and we propose a classification approach that accounts for both the pages' content and structure in order to assign them to a suitable category. In addition, our method regularly re-examines classified data and, upon the detection of accountable changes, it re-classifies them accordingly.

Our approach differs from existing techniques as it is multi-dimensional and, at the same time, confronts the dynamic nature of the Web. Specifically, we propose a holistic approach for the Web pages' classification, exploiting a variety of features, some of which have been less used in existing work, and which, to the best of our knowledge, have not been combined into a single technique. At the same time, the methodology proposed integrates solutions that have previously been studied separately. Lastly, our classification method tries not only to identify a suitable category for assigning every Web page, but it also accounts for the changes a page might undergo, which in turn might require the re-classification of the page. The advantages of our proposed method are that it incorporates into a single technique the structural and content organization of Web pages. Moreover, our method regularly re-examines classified pages in order to detect changes and, upon doing so, to deduce if re-classification is needed. The details of our approach are presented next.

III. Methodology

Our classification approach is established on the ground that the content (mainly textual) of a Web page is informative of the subject(-s)/theme(-s) the page deals with, while the structure of the Web page hides information about its type, i.e., the intentions associated with user visits to its content. Motivated by the work of Jansen et al. [14], who showed that Web searches can be classified as either Navigational, Informational or Transactional and the significant impact this work had on Search Engine Optimization (SEO) approaches, we reverse the argument and we claim that Web pages can be classified in the same manner, i.e., as being either Informational, Transactional or Navigational. For instance, a Web page with a disproportionally small amount of text compared to non-textual elements - such as, links- is less likely to be an Informative page. Conversely, a page that requests the user to take some action on it (e.g., buy, pay, book) is more probably a Transactional page. Therefore, any attempt to classify Web content should account for both content and structure in the sense that content gives the theme and structure gives the type.

Taking all the above as our baseline assumption, we designed a classification algorithm that considers two complementary sets of Web page features: structural and textual. Our algorithm incorporates several components and operates on two complementary phases, namely structure-based and content-based classification, as described next.

A. MultiDimensional Page Classification

1) Structure-Based Classification

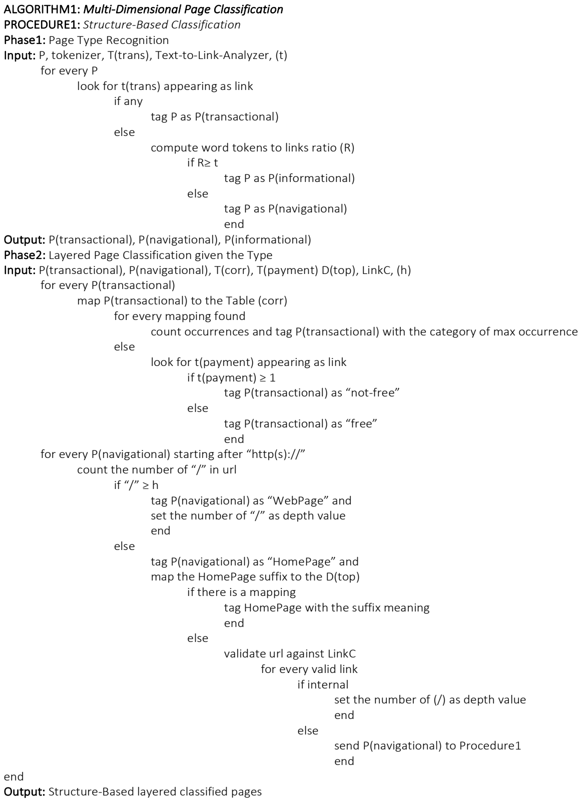
In Phase 1 (Page Type Recognition), the algorithm identifies the type of every page as follows. Given a page, it performs basic data processing in order to discriminate the page's textual and structural elements and then it utilizes a Text-to-Link-Analyzer [15] for extracting the page hyperlinks. Thereafter, it extracts the anchor text from every hyperlink and maps it against a list of transaction terms that are fed as input to the algorithm. The list of transaction terms has been manually constructed based on both empirical evidence and the findings of previous works [16]. Upon the detection of Transactional terms, it assigns a temporary tag to the page (i.e., type = *Transactional*) and further process it in Phase 2, as we will discuss next. In case no transactional terms are detected, the algorithm estimates the ratio between the page word-tokens to links and, if the ratio exceeds a given threshold t (experimentally set), it temporarily annotates the page as *Informational* and further process it in later steps. Finally, if the given page exhibits significantly more links than textual elements and lacks the presence of transactional terms, the algorithm annotates its type as *Navigational* and further process it in Phase 2. At the end of Phase 1, our classification algorithm enables a quick grouping of the pages under examination into one of the three categories, i.e., Informational, Navigational, Transactional, depending on their structural properties. To fine-tune this preliminary classification, we proceed to Phase 2 (Layered Page Classification) as follows.

The algorithm's input in Phase 2 consists of only the pages that have so far been characterized as either Transactional or Navigational along with a table of transaction correspondences, T(corr), a table of terms signifying payment actions, T(payment), and a list of Web top-level domains, D(top) [17]. The Transaction correspondences table, T(corr), lists the Transactional terms under the compatible transactional category and it was built based on the findings of an earlier unpublished work, where we asked human experts to identify within a set of Web pages the terminology (mainly verbs) that indicates actions that need to be taken on Web pages on behalf of users. The results were cross-evaluated with Google AdWords [18] before ending them up to the final list. Alternatively, one could apply hidden Web crawling techniques for determining the most common transactional terms within Web sites, but this puts an extra burden on the classification process, which is out of scope for our work. Similarly, the table of terms signifying payment actions has been manually defined based on manual linguistic analysis by experts of available domain-specific vocabularies. Lastly, the list of top-level domains was determined based on [19]. In Appendix B, indicative examples of the aforementioned table contents are given.

Taking the above input, during Phase 2, the algorithm begins with the likely Transactional pages and maps their transactional terms against the T(corr) table. It then estimates the occurrence frequency of every mapping found and tags the page with the most frequently occurring term. This term implies the type of transaction (e.g., pay, book, register, play) performed on the page. As an additional step, the algorithm maps the transactional terms of the page against the T(payment) table and, if there is a matching found, it characterizes the page as "not-free", otherwise it characterizes it as "free". This step figuratively validates the transaction. At the end of this step, every page that has been preliminary identified as Transactional is verified against terminological resources and, upon such verification, it is annotated with the type of transaction it entails and it is then classified as Transactional.

Afterwards, the algorithm examines the pages that have been preliminary characterized as Navigational and, based on their URL properties (e.g., number of '/', URL suffix, number of contained URLs), it annotates them either as Navigational_Homepage, if their URLs map against the list of top-level Web domains (D(top)), or as Navigational_Web page if they exhibit '/' above a threshold h (experimentally determined) and/or if they contain valid internal links. At the end of this step, every page that has been preliminary identified as Navigational, is either annotated as Navigational_Homepage/ Navigational_Web page or sent back to Phase1 for reprocessing. Figure 1 shows the pseudo-code of Procedure 1 of the algorithm, which classifies the pages structurally.

Having completed these two phases, our algorithm classifies Web pages into the most appropriate type (Transactional, Informational, Navigational) depending predominantly on their structural properties. The next procedure is to further examine the pages that have been classified by type and detect the main theme discussed in their contents so as to enable thematic classification.



```
ALGORITHM1: Multi-Dimensional Page Classification
PROCEDURE1: Structure-Based Classification
Phase1: Page Type Recognition
Input: P, tokenizer, T(trans), Text-to-Link-Analyzer, (t)
    for every P
        look for t(trans) appearing as link
            if any
                tag P as P(transactional)
            else
                compute word tokens to links ratio (R)
                    if R≥ t
                        tag P as P(informational)
                    else
                        tag P as P(navigational)
                        end
Output: P(transactional), P(navigational), P(informational)
Phase2: Layered Page Classification given the Type
Input: P(transactional), P(navigational), T(corr), T(payment) D(top), LinkC, (h)
    for every P(transactional)
        map P(transactional) to the Table (corr)
            for every mapping found
                count occurrences and tag P(transactional) with the category of max occurrence
            else
                look for t(payment) appearing as link
                    if t(payment) ≥ 1
                        tag P(transactional) as “not-free”
                    else
                        tag P(transactional) as “free”
                        end
    for every P(navigational) starting after “http(s)://”
        count the number of “/” in url
            if “/” ≥ h
                tag P(navigational) as “WebPage” and
                set the number of “/” as depth value
                end
            else
                tag P(navigational) as “HomePage” and
                map the HomePage suffix to the D(top)
                    if there is a mapping
                        tag HomePage with the suffix meaning
                        end
                    else
                        validate url against LinkC
                            for every valid link
                                if internal
                                    set the number of (/) as depth value
                                    end
                                else
                                    send P(navigational) to Procedure1
                                    end
end
Output: Structure-Based layered classified pages
```

Figure 1. MultiDimensional Page Classification_Procedure1 (Structure-Based Classification)

2) *Textual Based Classification*

To enable thematic classification, the algorithm begins by extracting textual elements from the page, anchor title and title. Having experimented with several textual features, we ended up with anchor title and title as the most informative of the theme of a page. Extracted anchor title and title terms are cross-matched and, upon the detection of exactly matching terms among them, we use those terms to annotate the theme of the page. Unless exact matchings are found, we apply traditional keyword extraction techniques to the pages' body and based on the top n-appearing keywords, we map them to WordNet lexical hierarchy [20]. Upon the detection of keyword mapping synsets, we extract the glosses of the latter and we look for overlapping terms within their definitions, i.e., glosses. The definition terms that are frequently overlapping across keyword matching glosses are utilized for verbalizing the theme of the page. Unless there are matchings between page keywords and WordNet synsets or unless there are no overlapping definition terms in the glosses of matching keywords, the theme of the page is deemed 'unknown' and the page is left unclassified. In Figure 2, we illustrate the pseudo-code of Procedure 2 of the algorithm.

```
ALGORITHM1: Multi-Dimensional Page Classification
PROCEDURE2: Content-Based Classification
Phase1: Textual Elements Extraction
Input: P
    for each P
        search for anchor title in Url
            if any
                tag as "P's anchorTitle"
                end
            else
                search for title in text body
                    if any
                        tag as "P's textTitle"
                        end
end
Output: P tagged with Textual elements
Phase2: Theme Detection
Input: (P's anchorTitle), (P's textTitle), WebPage Word Counter, PoS-Tagger,
Parser, WordNet, lemmatizer, (TF*IDF), (n)
For every page P look for common terms between P's anchorTitle and P's
textTitle
    if found
        use common terms as the theme(-s) to tag P
        end
    else
        PoS-tag and lemmatize P's text and extract the first n-appearing
    keywords
        check for overlapping terms between P's keywords and (P's anchor
        title and P's text title)
            if found
                use overlapping terms as the theme(-s) to tag P
                end
            else
                map P's first n-appearing keywords to WordNet and
                look for common senses between P's keywords and
                (P's anchor title and P's text title)
                    if found
                        use terms of common senses as the
                    theme(-s) to tag P
                        end
                    else
                        tag P as of unknown category
                    (Punknown)
```

Figure 2. MultiDimensional Page Classification_Procedure 2 (Content-Based Classification)

Based on the above phases and procedures, our algorithm classifies Web pages both by content and type. The output of the algorithm is the input Web pages annotated with a label indicating the exact type of the page as well as the theme of the contained information. What should be stressed is that the algorithm operates on the assumption that every page on the Web has a predominant intention, i.e., type, and, as such, the algorithm tries to detect this type and proceed accordingly. If there are pages of mixed types (e.g., Transactional and Informational) and those types are equally pronounced in their content, the algorithm deems the page type 'unknown' and proceeds with the textual classification of the page. To enable Web page classification in multiple types, we should adjust the threshold values accordingly. We defer this study for future work.

B. ReClassification based on Change Detection

A common challenge in Web data classification methods is how to account for the Web pages' dynamic nature, i.e., how to ensure that the classification outcome is up-to-date. To account for that, researchers have proposed various techniques for detecting Web changes [21] [22]. Such changes might be encountered to the content and/or structure of existing pages or they might concern the death or the birth of new pages.

Given that the Web is constantly evolving, we incorporated a re-classification component to our algorithm. The goal here is to detect, measure and identify the changes that possibly occur in the Web pages already classified both structurally and thematically. This procedure highlights the Web pages that need to be re-classified, in order to maintain data indexes always updated.

```
ALGORITHM2: Re-Classification based on Change Detection
Input: P(class, T), P'(unclass, T')
Procedure1: Re-Classification Decision based on Textual Changes
Input: (E(t) ∈ P), (E(t) ∈ P'), smlrtMetric, (m), (z)
    for each pair of (Pi ∈ P(class,T), (P'i ∈ P'(unclass, T'))
        compute sim(Pi, P'i)
            if sim(Pi, P'i) ≥ m
                tag P'i as thematically unchanged and classify P'i to the category of Pi
                end
            else
                tag P'i as thematically changed and
                compare (E(t) ∈ P'i) with (E(t) ∈ Pi)
                    count ((E(t) ∈ P'i) ≠ (E(t) ∈ Pi))
                        if ((E(t) ∈ P'i) ≠ (E(t) ∈ Pi)) ≤ z
                        go to Algorithm2Procedure2
                        end
                    else
                        send P'i to Algorithm1Procedure2
                        end
end
Output: thematically unchanged pages P' over time T'
Procedure2: Re-Classification Decision based on Structural Changes
Input: (E(s) ∈ P), (E(s) ∈ P'), smlrtMetric, (z)
    for each pair of ((Pi ∈ P(class, T), (P'i ∈ P'(unclass, T'))
        compare (E(s) ∈ Pi) with (E(s) ∈ P'i) and
        count ((E(s) ∈ Pi) ≠ (E(s) ∈ P'i))
            if ((E(s) ∈ Pi) ≠ (E(s) ∈ P'i)) ≤ z
                tag P'i as structurally unchanged and classify P'i to the category of Pi
                end
            else
                send P'i to Algorithm1Procedure1
                end
Output: structurally unchanged pages P' over time T'
Output: P'(ReClass, T'), thematically unchanged pages P' over time T', structurally unchanged pages P' over
time T'
```

Figure 3. Re-Classification based on Change Detection

According to Algorithm 2 (pseudocode shown in Figure 3), after its initialization, it compares, via similarity metrics, the structural and textual elements of any given page with their counterparts previously identified during the page's initial classification. The similarity metrics used in our approach are the Tree Edit Distance measures and Jaccard coefficient [23]. Based on the above, if the similarity between the pages' elements falls behind a predefined threshold value, the algorithm considers the page as changed and sends it back to Algorithm 1 for textual and/or structural (re-)classification.

Conversely, if both the structural and the textual elements of the page remain the same over time, the algorithm considers the page as unchanged and retains it to the category(-ies) it has been initially assigned by the classification algorithm. Note here that the value of thresholds can be experimentally fixed depending on the available data and the sought classification precision. Moreover, one could adjust thresholds dynamically, but, in the course of our experiment (as we discuss next), threshold values have been pre-determined.

C. *Optimized Re-Classification based on Change's Frequency Detection*

Having addressed the issue of changing Web content, we take a step further and we account for the changes' frequency. Change's frequency detection of a page, helps us determine our re-classification policy in order to save time and resources. In this framework, we capture the frequency with which Web pages change in order to optimize the runs of our Re-Classification algorithm. The idea was inspired by the work of Meegahapola et al. [24] and driven by the fact that Web pages change at different frequency rates.

According to the pseudocode of the algorithm, illustrated by Figure 4, we adjusted a timer, which is activated upon the initialization of the Re-Classification algorithm. The time intervals between the initialization of the Re-Classification runs are also predefined based on experimental set. Every time a change is detected between two chronologically different snapshots of a page, the timer records it. After several iterations of the Re-Classification algorithm on a single page at different time intervals, all the changes the page has undergone are recorded by the timer along with the timestamp of the change detection. This timeline of Web page changes helps us determine the best time period for a page to be revisited for change detection and if needed for re-classification.

```
Algorithm3: Optimized Re-Classification based on Change's Frequency Detection
Input: (P(class, T)), ((E(t) U E(s)) ∈ P(class, T)), P' ⊆ (P'(re-class, T'), ((E(t) U E(s)) ∈ P'(reClass,T')),
MaxFreqChange, MinFreqChange, Timer
    when Algorithm2 initializes, record Ts
    for every pair of ((Pi ∈ P(class, T), (P'i ∈ P'(re-class, T'))
        set Timer
            while ((E(t) U E(s)) ∈ Pi(class, T)) ≠ ((E(t) U E(s)) ∈ P'i(re-class,T')), record Ts
                if Ts ≥ MaxFreqChange
                    tag P'i as HighlyChanging Page and keep it in a secondary Index
                    end
                else
                    if Ts ≤ MinFreqChange
                        tag P'i as RarelyChanging Page and keep it in a secondary Index
                        end
                    else
                        tag P'i as RegularlyChanging Page and send it to Algorithm2
                        end
end
Output: Selection of Pages that need periodical Re-Classification
```

Figure 4. Optimized ReClassification based on Change's Frequency Detection.

Based on the above process, our method ensures that classified pages which undergo regular changes, are reconsidered by our classification algorithm when needed, in order to maintain their organization up-to-date.

Next, we present the experimental evaluation of our method and we report the obtained results.

IV. Experimental Implementation and Evaluation

To evaluate the effectiveness of our classification algorithm, we carried out a small-scale experiment in which we validated: (a) the classification performance of our method, and (b) the potential weaknesses of our approach. To collect our experimental dataset, we asked 10 experienced Web users to provide us with their bookmarked pages.

We informed our volunteers about our study objectives and we asked them to indicate for each of their shared bookmark the type of the page (by selecting between Informational, Transactional and Navigational) and the theme of the page. To familiarize our participants with the page types, we gave them brief instructions with respect to the definition of every type and we trained them by giving several examples. Moreover, we instructed them to indicate a single structural type for every page and in case of uncertainty to remove the page from their selection list. On the other hand, the theme of every page was self-determined by our volunteers and verbalized based on their understanding of the page's theme. We instructed our subjects to use as many keywords as they wished for verbalizing the underlying theme of a page, but upon the indication of several thematic keywords we asked them to point out the one that was in their opinion the predominant. The volunteers who supplied us with data were not further involved in the experimental process.

Based on the above dataset, we ended up with a gold-standard test-data of 2,330 pages, each of which was manually labeled by our participants with both structural type and thematic information. In TABLE I. we summarize the statistics of our experimental dataset.

TABLE I. EXPERIMENTAL TEST SET STATISTICS

Total set of experimental pages	***2,330***
Percentage of Informational pages	63%
Percentage of Transactional pages	17.99%
Percentage of Navigational pages	19.01%

This test set (cleaned up from any labelling) was given as input to our classification algorithm. Before proceeding with the details of our experiment, we should stress that the algorithm did not undergo any training phase, but rather it run in several iterations in order to fix the values of the thresholds it incorporates. During the first iteration, each threshold value was uniformly set to 0.5 suggesting equal probabilities so as to avoid any bias. Thereafter, in every subsequent iteration the values of each threshold were fine-tuned and based on the median values of all iterations, they were fixed as follows: the word tokens/links ratio (R) for discriminating between Informational and Navigational pages was set to 60/40 the number (h) of "/" in every page URL was set to 2, for the thematic classification we exploited the top 5 appearing keywords, whereas we retained the threshold values associated with the change detection algorithm to 0.5. Lastly, *MaxFreqChange* and *MinFreqChange* values were set to 2 and 1 respectively. Of course, one could modify the applied thresholds depending on the experimental set, or she/he could fix their values uniformly or finally determine thresholds dynamically depending on the application domain of the algorithm. Further experimenting with alternative threshold values falls beyond the scope of this study.

Having collected and pre-processed our experimental test pages and having determined threshold values, we run our algorithm and we evaluated its classification accuracy as follows. We compared both the structural and thematic categories our algorithm identified for each of the experimental pages against the respective structural and thematic categories our participants had manually indicated for the corresponding pages. That is, we evaluated the algorithm's structure-based classification accuracy (i.e. the ability to discriminate Informational, Navigational or Transactional pages) by comparing the structural type tags our participants had manually indicated to the tags our algorithm identified for labelling the respective pages. The matching tags across pages were deemed as correct structural classifications (true positives), whereas mismatching tags flagged shortcomings in the algorithm.

Similarly, we evaluated the algorithm's content-based classification accuracy (i.e. the ability to identify a suitable theme for every page) by comparing the list of thematic keywords our participants had indicated for every page to the thematic terms our algorithm had automatically identified for each page. Matching thematic keywords across pages were deemed as correct content-based classifications (true positives), whereas mismatches were interpreted as the algorithm's failure to identify a suitable thematic category for a page. At this point, we should stress that the terminology used for naming a thematic category was not always identical between that supplied by our volunteers and the terminology identified by our algorithm. Thus, to enable the comparison between the two we relied on WordNet against which we estimated the semantic similarity between the two. For measuring similarity, we utilized the Wu and Palmer similarity metric [25]. If the similarity values between terms (automatically detected by the algorithm and manually defined by our subjects) exceeded the value of 0.8 (values range between 0=no similarity and 1= exact match) we deemed the theme the algorithm identified as correct (i.e. true positive) else we deemed the theme as wrong (i.e. false positive).

The metrics we used for quantifying the algorithm's accuracy are classification recall and precision. ***Classification recall*** estimates the proportion of pages that the algorithm classified correctly (*TP: true positives*) out of all the pages examined in the test-set (*TP+FN: true positives+false negatives*), whereas ***classification precision*** indicates the proportion of pages that the algorithm classified correctly (*TP: true positives*) out of all the pages the algorithm managed to classify (*TP+FP: true positives+false positives*). Classification recall shows the algorithm's capacity in identifying a category for every page it examines, whereas classification precision shows the algorithm's capacity in identifying the correct category of a Web page. The formulas of the two metrics are given below:

$$recall = \frac{true\ postives}{true\ posititves + false\ negatives}$$

$$precision = \frac{true\ positives}{true\ positives + false\ positives}$$

Out of the 2,330 experimental pages, our algorithm managed to classify 1,966 (84.3%) and the remaining 364 (15.7%) pages were assigned to the class '*unknown*', meaning that no category could be identified by the algorithm for those pages. Results suggest that the algorithm is successful in detecting a category (structural and thematic) for the majority of the examined pages and a manual inspection to the pages it left unclassified revealed that this was due to multiple structural nature of the no tagged pages and lack of text in the thematically untagged pages.

To evaluate the algorithm's success in correctly classifying Web pages we applied the precision and recall metrics, as previously explained, and we report obtained results in TABLE II. and Figure 5, respectively.

TABLE II. CLASSIFICATION RECALL AND PRECISION VALUES

Recall on Navigational pages	0.75
Precision on Navigational pages	1
Recall on Informational pages	0.89
Precision on Informational pages	0.98
Recall on Transactional pages	0.78
Precision on Transactional pages	1
Recall on thematic classification	0.83
Precision on thematic classification	0.87
Overall classification recall	**0.8**
Overall classification precision	**0.99**

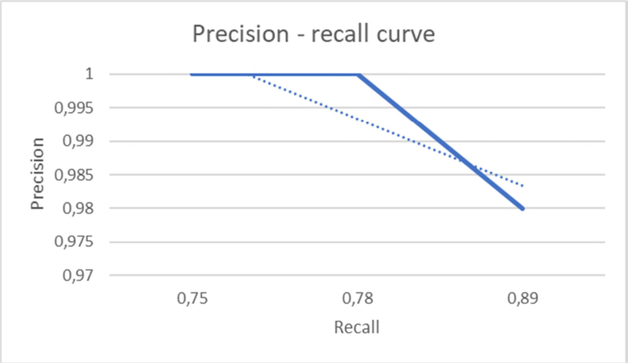


Figure 5. Classification precision-recall curve.

Results show that, out of all the Navigational pages in our test set (443 pages), the algorithm correctly identified 75% (332 pages) of them as Navigational and out of all the Informational pages in our test set (1,468 pages) the algorithm correctly identified 89% of them (1,307) as such. In addition, the algorithm correctly tagged as Transactional

78% (327 pages) of the examined Transactional pages (419 pages). Results show that the algorithm has a good overall accuracy in classifying pages based on their structural elements. This is further supported by the classification precision scores, which show that 100% of the pages the algorithm classified as Navigational were actually Navigational, 98% of the pages the algorithm classified as Informational were actually Informational and 100% of the pages the algorithm classified as Transactional were actually Transactional. Based on the classification precision and recall scores (overall and type-based), we may conclude that the algorithm has quite strict criteria for assigning Web pages to suitable structural types. This is attested by the very high precision scores and the lightly lower recall, which suggest that unless the algorithm has very strong indications of a page's type, it leaves it unclassified. A manual inspection of the cases the algorithm failed to identify the correct type of a page reveals that such pages had either mixed structural properties (e.g. their intention was both Informational and Navigational) or they contained very little information about their underlying type. With respect to the former, we already mentioned that in its current version the algorithm does not allow multiple structural classifications for a page, but this is an issue we are currently working on. To overcome the second shortcoming, we are testing additional elements that could signify the structural type of a page.

Regarding the algorithm's accuracy in identifying the theme of a page, our results show that the algorithm managed to identify a thematic category for 1,948 (i.e. 83.6%) of the 2,330 test pages and for those that it did find a category 87.4% (1,704 pages) were classified to the correct category (true positive classifications). Again, the manual inspection of our results showed that the algorithm failed to identify a theme for pages with very little content or with content verbalized with terms missing from WordNet. To overcome this terminological limitation, we are currently experimenting with a set of pre-defined categories to which we seek to organize the contextual elements of the Informational pages.

Thereafter, we evaluated the need for Web pages' re-classification as follows. We performed two additional downloads of the same set of pages after a period of one and three months respectively after the algorithm's first run. For every re-download, we run our re-classification algorithm as previously described and we computed the amount of changes between the content and structure elements of the examined pages, after one and three months since their first classification. The obtain results are reported in TABLE III. As the table shows, 67.9% of the pages had changed over a period of one month with the majority of changes being structural. In detail, 62% of the re-examined pages had undergone structural changes, 4.6% had undergone textual changes and only 1.3% of them had undergone both structural and textual charges. A close inspection to the amount of changes reveals that, the striking majority of them are minor (e.g. date changes) and thus there is no need to re-classify those pages. However, we found that in 6.6% of the changed pages, the structural and/or textual alterations were significant, therefore triggering the need to re-classify them.

Similarly, when re-examining the same set of pages after a period of three months, we found that 58.7% of them had changed, with the majority of changes pronounced mainly in structure (49.3%) and less in text (7.3%). Again, we observed that only a small portion of those pages (4.6%) exhibited significant alterations and, thus, should be re-classified.

As a last evaluation step, we computed for the pages that do need re-classification the frequency of changes they undergo over a period of three months since their first classification and we found that 46% of them change frequently, meaning that they had changed at least twice within a time slot of three months. This implies that for those changes regular re-examination is needed in order to keep classification up-to-date and that the algorithm should be periodically revisiting them. Based on the manual examination of a sample data out of those frequently changing pages, reveals that these concern among others news sites, online applications and so forth. Conversely, inspecting the results of the Re-Classification algorithm, when a page changes only with respect to the word tokens/links ratio (R), it is not being sent to Algorithm 1 for re-classification, as it is appearing with a small change percentage. However, this element is determined for page's type, so sometimes this change may be more significant than it seems. This downside could be overcome by transforming any structural and textual element to weighted element, according to its role in the classification decision. However, we defer this last issue for a follow up study of the current work.

TABLE III. CHANGING PAGES THAT NEED RECLASSIFICATION

Pages changed after 1 month	**67.9%**
Pages structurally changed	62%
Pages textually changed	4.6%
Pages structurally and textually changed	1.3%
Pages re-classified after 1 month	**6.6%**
Pages changed after 3 months	**58.7%**
Pages structurally changed	49.3%
Pages textually changed	7.3%
Pages structurally and textually changed	2.1%
Pages re-classified after 3 months	**4.6%**
Highly changing pages after 3 months	46%

V. CONCLUSION AND FUTURE WORK

In this paper, we present a novel Web pages classification approach that combines the structural properties and textual elements of Web pages in order to classify them in two dimensions, namely type and theme. Moreover, our method accounts for the changing nature of the Web and foresees a number of actions in order to determine whether a page needs to be re-classified after a time period as well as what is the best time period for re-classification. A preliminary experimental evaluation to our method against a manually annotated set of pages reveals that it manages to effectively capture both the type and the general theme of the examined Web pages.

We are currently improving our method to tackle some minor shortcomings already detected during experiments. More specifically, we are improving some parts of the Algorithm 1, so as to enable multiple type classifications of a given Web page if necessary. We are also experimenting with alternative/additional lexical resources for the detection of the pages' theme and we work on testing alternative similarity metrics. The prospective flexibility that characterizes our approach, respecting the resources mentioned above, would extend our methodology's significance. Close to the above, for the Transactional pages' detection, taking into consideration the images and/or symbols that represent a Transactional option for the user, and not only the verbal T(terms) appearing as links in the page's text body, can broaden algorithm's performance.

Observing the experimental results of the Re-Classification Algorithm, we grasped that the change of any textual and structural element is not equally important. In other words, each element can be less or more determinative for the page's re-classification decision. Based on this notice, we examine the elements assignment with weights. Additionally, in Algorithm 3, we consider as a challenge the dynamic definition of time intervals between the initialization of the Re-Classification runs. Additionally, we think about checking the algorithms' complexity and response time. Lastly, we are in the process of a large-scale experiment (larger dataset checked for longer time and after more time intervals), the results of which will be resented in a future work.

References

[1] L. Safae, B. E. Habib, and T. Abderrahim, "A Review of Machine Learning Algorithms for Web Page Classification," in *2018 IEEE 5th International Congress on Information Science and Technology (CiSt)*, Oct. 2018, pp. 220–226, doi: 10.1109/CIST.2018.8596420.

[2] A. Kumar and R. K. Singh, "A Study on Web Structure Mining," vol. 04, no. 1, p. 6.

[3] C.-W. Tsai, C.-F. Lai, H.-C. Chao, and A. V. Vasilakos, "Big data analytics: a survey," *J. Big Data*, vol. 2, no. 1, p. 21, Oct. 2015, doi: 10.1186/s40537-015-0030-3.

[4] R. Rajalakshmi and C. Aravindan, "Naive Bayes Approach for Website Classification," in *Information Technology and Mobile Communication*, vol. 147, V. V. Das, G. Thomas, and F. Lumban Gaol, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2011, pp. 323–326.

[5] S. Shinde, P. Joeg, and S. Vanjale, "Web Document Classification using Support Vector Machine," in *2017 International Conference on Current Trends in Computer, Electrical, Electronics and Communication (CTCEEC)*, Sep. 2017, pp. 688–691, doi: 10.1109/CTCEEC.2017.8455102.

[6] P. Kenekayoro, K. Buckley, and M. Thelwall, "Automatic classification of academic web page types," *Scientometrics*, vol. 101, no. 2, pp. 1015–1026, Nov. 2014, doi: 10.1007/s11192-014-1292-9.

[7] H. Li, Z. Xu, T. Li, G. Sun, and K.-K. Raymond Choo, "An optimized approach for massive web page classification using entity similarity based on semantic network," *Future Gener. Comput. Syst.*, vol. 76, pp. 510–518, Nov. 2017, doi: 10.1016/j.future.2017.03.003.

[8] A. Onan, S. Korukoğlu, and H. Bulut, "Ensemble of keyword extraction methods and classifiers in text classification," *Expert Syst. Appl.*, vol. 57, pp. 232–247, Sep. 2016, doi: 10.1016/j.eswa.2016.03.045.

[9] J. Qiu, Q. Wu, G. Ding, Y. Xu, and S. Feng, "A survey of machine learning for big data processing," *EURASIP J. Adv. Signal Process.*, vol. 2016, no. 1, p. 67, Dec. 2016, doi: 10.1186/s13634-016-0355-x.

[10] M. J. H. Mughal, "Data Mining: Web Data Mining Techniques, Tools and Algorithms: An Overview," *Int. J. Adv. Comput. Sci. Appl.*, vol. 9, no. 6, 2018, doi: 10.14569/IJACSA.2018.090630.

[11] Y. Liu, J.-Y. Nie, and Y. Chang, "Constructing click models for search users," *Inf. Retr. J.*, vol. 20, no. 1, pp. 1–3, Feb. 2017, doi: 10.1007/s10791-017-9294-x.

[12] F. Hemmatian and M. K. Sohrabi, "A survey on classification-techniques for opinion mining and sentiment analysis," *Artif. Intell. Rev.*, vol. 52, no. 3, pp. 1495–1545, Oct. 2019, doi: 10.1007/s10462-017-9599-6.

[13] R. Rajalakshmi and C. Aravindan, "A Naive Bayes approach for URL classification with supervised feature selection and rejection framework," *Comput. Intell.*, vol. 34, no. 1, pp. 363–396, 2018, doi: 10.1111/coin.12158.

[14] B. J. Jansen, D. L. Booth, and A. Spink, "Determining the informational, navigational, and transactional intent of Web queries," *Inf. Process. Manag.*, vol. 44, no. 3, pp. 1251–1266, May 2008, doi: 10.1016/j.ipm.2007.07.015.

[15] *https://sonovabitc.win/analyze.php* [retrieved: May, 2020].

[16] B. J. Jansen, D. L. Booth, and A. Spink, "Determining the user intent of web search engine queries," in *Proceedings of the 16th international conference on World Wide Web - WWW '07*, Banff, Alberta, Canada, 2007, p. 1149, doi: 10.1145/1242572.1242739.

[17] *https://en.wikipedia.org/wiki/List_of_Internet_top-level_domains [retrieved: May, 2020].*

[18] M. Raffinot and R. Rivière, "Optimizing Google Shopping Campaigns Structures with Query-Level Matching," *ArXiv170804586 Cs*, Aug. 2017, Accessed: Jul. 17, 2020. [Online]. Available: http://arxiv.org/abs/1708.04586.

[19] S. C. Jeeva and E. B. Rajsingh, "Intelligent phishing url detection using association rule mining," *Hum. -Centric Comput. Inf. Sci.*, vol. 6, no. 1, p. 10, Jul. 2016, doi: 10.1186/s13673-016-0064-3.

[20] G. A. Miller, R. Beckwith, C. Fellbaum, D. Gross, and K. J. Miller, "Introduction to WordNet: An On-line Lexical Database," *Int. J. Lexicogr.*, vol. 3, no. 4, pp. 235–244, 1990, doi: 10.1093/ijl/3.4.235.

[21] S. Flesca and E. Masciari, "Efficient and effective Web change detection," *Data Knowl. Eng.*, vol. 46, no. 2, pp. 203–224, Aug. 2003, doi: 10.1016/S0169-023X (02)00210-0.

[22] D. Yadav, A. K. Sharma, and J. P. Gupta, "Change Detection in Web Pages," in *10th International Conference on Information Technology (ICIT 2007)*, Rourkela, Orissa, India, Dec. 2007, pp. 265–270, doi: 10.1109/ICIT.2007.37.

[23] P.-N. Tan, M. Steinbach, and V. Kumar, *Introduction to data mining*, First edition, Pearson new international edition. Harlow: Pearson, 2014.

[24] L. Meegahapola, R. Alwis, E. Nimalarathna, V. Malawaarachchi, D. Meedeniya, and S. Jayarathna, “Detection of change frequency in web pages to optimize server-based scheduling,” in *2017 Seventeenth International Conference on Advances in ICT for Emerging Regions (ICTer)*, Sep. 2017, pp. 1–7, doi: 10.1109/ICTER.2017.8257791.

[25] Z. Wu, M. Palmer, “Verb Semantics and Lexical Selection,” pp. 133-138, 1994.

APPENDIX A

(E(s) ∈ P): list of P structural elements

(E(s) ∈ P'): list of P' structural elements

(E(t) ∈ P): list of P textual elements

(E(t) ∈ P'): list of P' textual elements

(E(t) U E(s)) ∈ P(class, T): list of P(class, T) textual and structural elements

(E(t) U E(s)) ∈ P'(re-class,T'): textual and structural elements of P'(re-class,T')

(h): threshold for homepages’ detection

(z): threshold for structurally unchanged pages’ detection

(n): threshold for keywords

(m): threshold for the thematically unchanged pages’ detection

(t): threshold for the informational pages’ detection

(TF*IDF): short for “term frequency–inverse document frequency”, is a numerical statistic that is intended to reflect how important a word is to a document in a collection or corpus.

(Ts): time stamp

D(top): list of web top-level domains

Lemmatizer: tool that groups inflected forms together as a single base form

LinkC: link counter (tool)

MaxFreqChange: maximum frequency “allowed” by the algorithm for changes to webpages

MinFreqChange: minimum frequency “allowed” by the algorithm for changes to webpages

P' ⊆ (P'(re-class, T'): subset of P’ that have been ReClassified based on Algorithm2 at time T’

P(class, T): pages classified from Algorithm1 at time T

P’(class, T’): pages classified from Algorithm1 at time T’

P(navigational): navigational pages

P(transactional): transactional pages

P(informational): informational pages

P: pages for classification

P’(ReClass, T’): textually or/and structurally changed pages P’ over time that need to be ReClassified

(P’s anchorTitle): page’s anchor title

(P’s textTitle): page’s text title

Parser: compiler or interpreter component that breaks data into smaller elements for easy translation into another language.

PoS-Tagger: software that reads text in some language and assigns parts of speech to each word (and other token), such as noun, verb, adjective, etc.

smlrtMetric: similarity metric

T(corr): table with transactions’ correspondences

T(payment): table with Payment terms [t(payment)]

t(payment): payment terms

T(trans): table with transactional terms

t(trans): transactional terms

Text-to-Link-Analyzer: tool for the calculation of WordTokens2Links Ratio

Timer: timer that calculates the time based on defined formula (1). If t is the time when a webpage is first examined, the timer will calculate every t_i time instance that we want to re-examine the same page, according to the formula (1): Each time instance of re-examination is derived from the multiplication product of its preceding time instance with the constant defined.

$$t_i=(t_i-1) *constant \qquad (1)$$

Tokenizer: tool for the tokenization of the text

WebPage Word Counter: tool for keywords’ extraction

WordNet: hierarchically organized dictionary

APPENDIX B

TRANSACTION CORRESPODENCES TABLE T(CORR).

Booking	**Download**	**E-commerce**	**Entertainment**	**Software**
Book a table	Download	Shop Now	Download	Download
Book Now	Free trial	(Add to) Bag	Find a table	Free trial
Find a table	Games	(Add to) Basket	Play	Games

TABLE WITH PAYMENT TERMS T(PAYMENT).

Payment Terms
Book Now
Buy Online
Buy
Pruduct+Price
Shop Now
Wish List

TABLE WITH SOME WEB TOP-LEVEL DOMAINS, D(TOP).

Name	**Entity**
.com	commercial
.org	organization
.net	network
.int	international organizations
.edu	education
.pr	Puerto Rico (United States)
.ps	Palestine
.pt	Portugal
.py	Paraguay

Copyright (c) IARIA, 2020.
ISBN: 978-1-61208-789-4

A Web-Based Platform to Teach Music Online

Fatemeh Jamshidi

Computer Science and Software Engineering
Auburn University
Auburn, Alabama, USA
Email: fzj0007@auburn.edu

Daniela Marghitu

Computer Science and Software Engineering
Auburn University
Auburn, Alabama, USA
Email: marghda@auburn.edu

Abstract—**The development of educational technology has encouraged music educators to consider different ways to teach music online. However, this change may need an online platform to support teachers when the size of the class is large. This paper aims to address two research questions. The first question is to determine the learning experiences of students and to understand their needs to support Web-based learning. The second question is to study the key technologies required for a Web-based music teaching system. In this paper, we try to determine the best ways to motivate students to study music and to make music teaching experiences more accessible, engaging, and fun for students.**

***Keywords*–Web based learning; Music education; Self-directed learning; Self-assessment.**

I. Introduction

With advancements in Computer Science (CS) technologies, the music discipline needs to make full use of new innovations to provide high-quality music learning resources to students. Applications of multimedia and Web-based learning technologies in teaching music can help teachers to keep students motivated and engaged. Some researchers have investigated students' Web-based music learning experiences [1]. The results have shown that multimedia teaching has many advantages:

1) A rich audio-visual experience can effectively enhance student's engagement and accelerate music learning through online activities.
2) Friendly interactive environments can increase the enthusiasm of the students in learning and practicing music.
3) The new way of teaching is in accordance with cognitive patterns.

With the use of blended strategies [2], this research tries to enhance student engagement and music learning through online activities in music courses. It also improves the effectiveness and efficiency of music classes by reducing lecture time and increasing the practice time. In the future, the traditional distinction between class time and non-class time will disappear.

The current generation is using technologies that connect single learners and collaborative varieties of learning to socialize, work, and learn [3]. Technology-based platforms to teach music should be able to cover resources for collaborative learning as well as individual capabilities. A Web-based learning environment offers a novel informal platform where individuals with different music backgrounds can learn and practice music. The aim of this research is to see the essential elements of current online learning approaches utilized in academic music courses and to contemplate their application within the development of an Internet pedagogical framework to show music online. Our primary prototype is designed to teach people with little to no music background, the basics of music theory and how to play the piano.

The rest of the paper is organized as following: In Section II, we cover some background about the project. Section III introduces the methodology of the website and covers the design and implementation of the website. Section IV presents the evaluation methods. Section V concludes the project along with suggestions for future enhancements.

II. Related Work

A. Score Following

Score following [4][5] is a technique in music technology to track the performance of the player in a given music score and is one of the most important components in automatic music accompaniment. The two most important components in score following are to express [6]:

1) the similarity between the current observation and the expected observation in each position in the music score that the player is playing.
2) the allowed temporal evolution of the score position.

Score following systems mostly do not recognize visual cues, that human musicians use for coordinating parts of the music without any musician playing [7]. For example, nodding gestures can be used to synchronize the introduction of a song. Some studies use visual information, such as the periodic hand motion of a player [8]. Visual cue is an important component when the audio signal is not enough for tracking the human musicians.

B. Pitch Tracking

A reliable estimate of the pitch of a monophonic sound recording (pitch tracking) is crucial to audio processing with multiple applications in music information retrieval. Pitch tracking is an essential component in music signal processing, where monophonic pitch tracking is used for generating pitch annotations for multi-track datasets.

Estimation of the pitch of a monophonic signal has been a longstanding topic for more than a half-century, and many well-founded methods have been proposed since [9]. Earlier methods mostly utilize a certain candidate-generating function, using pre and post-processing stages to produce the pitch curve. Those functions include the ”cepstrum [10], the autocorrelation function (ACF) [11], the average magnitude difference function (AMDF) [12], the normalized cross-correlation function (NCCF) as proposed by RAPT [13] and PRAAT [14], and the cumulative mean normalized difference function as proposed by YIN” [9].

A common method in previous approaches is that the derivation of a better pitch-tracking system depends on a robust candidate-generating function and/or sophisticated post-processing steps, i.e. heuristics. Furthermore, none of them are directly learned from data, except for manual hyperparameter tuning, which contrasts with other problems in music information retrieval such as chord ID [15], where data-driven methods have been shown to outperform heuristic approaches.

C. *Flowkey*

Flowkey [16] is an educational music app that teaches how to play your favorite songs on your digital or acoustic piano.

The app works on multiple devices and comes off useful and practical for beginners to learn piano or even for advanced musicians. It aims to help the beginner players who are not comfortable with reading sheet music notes yet.

The different modes and features of Flowkey are:

- Slow Mode - This feature allows the player to play along with the song at a slow speed to make the user feel comfortable with the virtual sheet music notes. In this section, the video also slows down without disturbing the audio.
- Fast Mode - This mode allows the player to play along with the song in the original tempo for that specific song.
- Loop Mode - the player will be able to choose a portion of the video tutorial and keep it on the loop until he/she gets it right.
- Hand Selection - This feature is beneficial for more complicated songs. It is designed for beginners in case the player initially gets confused by multiple keys being played at a time and would like to master on one hand first and then switch to another hand. Figure 1 shows the design of hand selection in Flowkey.
- Wait Mode - The virtual sheet music notes and the video wait for the player to play the notes after learning from the tutorial. Through the built-in microphone in the app, the wait mode detects the player’s movement without making him/her connect the actual digital or acoustic piano to the app. Therefore, the Flowkey app provides feedback on each notes the player plays.

D. *Playground Sessions*

Playground Sessions is a Web-based music learning software which helps users to subscribe to music theory lessons and provides a fun and effective experience for people to learn the piano online.

Playground Sessions has several different elements that contribute to the learning experience.

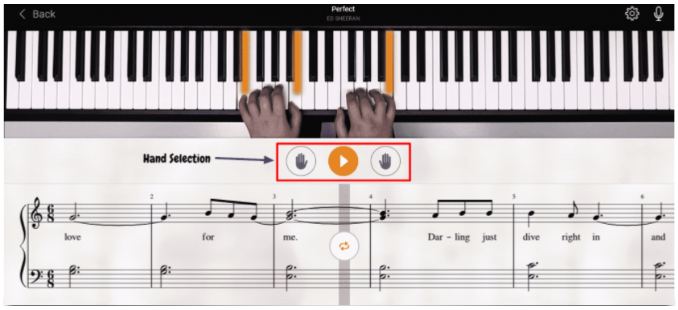


Figure 1. Design of Hand Selection in Flowkey.

- Interactive Lessons: This section is under the “Bootcamp” tab in the app where excerpts from well-known songs are designated to teach the player specific music concepts related to the song, with written instruction and game-like practice.
- Video Lessons: The video lessons are followed by interactive lessons. They cover more details about the interaction lessons and allow the player to practice what was taught.
- Forums: This is a place for users to share tips, stay up to date on Playground Sessions news, ask questions, and lodge complaints.

Unlike Flowkey and Playground Sessions, our Web-based music learning application consists of a curriculum that not only teaches music concepts to students, but also teaches them how to compose to a music piece. The other extra feature our application has compared to the other existing ones is that the player can review the other players’ performances and rate them. The user also may want to join the other users to play a duet, which is the ongoing feature of our application.

III. Methods

We developed a Web-based prototype to teach people with little to no music background the music theory basics and how to play the piano. To address this problem, we developed our prototype in two phases. The first one is for students to learn about the basics of music theory. Secondly, students will be able to practice what they learned by playing their favorite songs on the application with the app’s guidance, which is connected to their own digital or acoustic piano.

A. *Phase One: Curriculum and Teaching Strategies*

1) Design and development : We designed an online teaching platform, where teachers can have a one-on-one or a group session with their students. The software helps teachers to review the progress of their students throughout the week’s practice. Besides, teachers will be able to upload their own recorded courses for the students. Thirdly, we developed the music theory teaching scriptwriting, a work classification, courseware development, and other links. Figure 2 presents the primary teaching module architecture.

2) Development of the curriculum: In the non-technology-based teaching research, developing the different teaching models had to be according to students’ perceptions and music learning ability. This model could be practical when educators share the same goal, and students have the same understanding level. Under the current teaching models of College Music teaching, a teaching object’s attribute is not

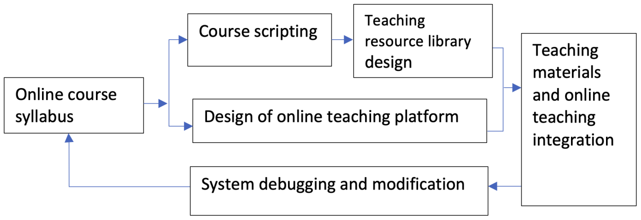

Course scripting
Teaching resource library design
Teaching materials and online teaching integration
Online course syllabus
Design of online teaching platform
System debugging and modification

Figure 2. Design and development process of online music theory courses (based on [17]).

consistent. Therefore, to conduct the trusted quality of teaching objectives, it is essential to develop students' multi-attribute professional features and multi-objective future careers.

In the research and practice of ”diversity”, teaching, content, and target design is based on the students' knowledge base, understanding ability, and so on. Therefore, in this project, we aim to implement an Artificially Intelligent application that can train itself based on the students' different expertise and performance skills. The students' attribute difference is not limited to the learning factor in the trusted teaching mode, but to individual behavior attribute, knowledge attribute, and personal attribute under corresponding training objective.

To determine students' needs to support Web-based learning, we designed a music-technology-based curriculum to evaluate students' interaction with the existing music learning applications and better address the missing components in our prototype.

We conducted this study in two phases, each with separate surveys and strategies. The first phase's purpose was to develop and evaluate an online pilot program's possibilities and requirements integrating computer technologies and music composition concepts for middle-school students. Opportunities to view and critique pilot online instructional units developed for the Knowledge Works Learning Academy (KWLA) were included as a regular part of class activities for two graduate music education courses at Auburn University: CTMU 7520-26 (Curriculum and Teaching in Music Education) and CTMU 7540-46 (Evaluation of Programs in Music Education).

The specific research questions for the study are as follows:

First Survey: Administrators

1) To what extent do comprehensive public high schools in the United States offer technology-based music classes?
2) To what extent does the district's socioeconomic status affect the likelihood of offering a technology-based music class?
3) To what extent does the district's geographic location affect the likelihood of offering a technology-based music class?
4) To what extent do/would school administrators value technology-based music classes?

Second Survey: Teachers

1) What is the curricular nature of these classes?
2) To what extent do these classes address nontraditional music students?
3) What is the professional background of teachers of technology-based music classes?
4) What types of software and hardware are being utilized in technology-based music classes?
5) How long have these classes been offered, and how were they initiated?
6) What level of support do school districts provide for these classes?

After analyzing the first phase, where music graduate students evaluated the curriculum's feasibility and potential student's needs, we made the curriculum changes accordingly. The results of our first phase study will be presented in future work. In phase two, we conduct a pilot study to monitor students' interaction with our music learning Web-based platform. This pilot study is planned for October-November 2020, and the results will be presented in future work.

B. Phase Two: Practicing Strategies

This phase is developed to help the students practice what they learned in the curriculum by playing their favorite songs on the application, connected to their own digital or acoustic piano. In other words, students will be able to play a duet with the application, play famous songs with the help of the application, and connect with other people using the application to play in a team (band).

The following are the key technologies required for a Web-based music teaching system:

1) Data Collection:

- Musicians: 17 music teachers (middle school, high school, or college teachers) are invited from the music education department at Auburn University to perform duet pieces and review the first phase of the application, including the designed curriculum.
- Music pieces: To collect the music pieces, we developed a survey to collect data from music teachers all over the USA. Each musician will perform every detail executed from the surveys for ten times in the different music expression. Therefore, our machine learning model will be trained based on theses music pieces rehearsals.
- Recording settings: Electronic pianos with Musical Instrument Digital Interface (MIDI) output will record the music pieces; therefore, all the parameters (dynamic, starting time, ending time, pedal) of every note can be recorded in real-time [18].
- Recording procedures: Musicians will practice the pieces for 30 minutes together (other than solo practices) and then start recording. Each recording session records approximately ten performances and lasts more than an hour.

2) Data Representation and Models: We are using various function approximations based on [18] to model the differences between one pianist's expression and another's. In this project, we start from music representations and models that only apply to specific music notes and gradually step to a high-dimensional phrase, which implies a whole piece of music.

Our artificial performer will generate (decode) its music expression by communicating with a player according to trained models. Piano notes can be represented by notes, beats, timing, dynamic, and pedal position.

3) Pitch Detection: Our Web-based music learning platform detects players' notes as they play along and provides feedback for the player to guide them toward the right path.

In this research, we used the CREPE pitch detection [9] algorithm to detect the player's notes in real-time. CREPE originally includes a deep convolutional neural network that operates on the time-domain audio signal for a pitch estimation. A diagram of the CREPE architecture is shown in Figure 3.

The CREPE input is a 1024-sample excerpt from the time-domain audio signal, with a 16 kHz sampling rate. There will be five convolutional layers that result in a 2048-dimensional latent representation connected to a 72-dimensional output vector y through sigmoid activation. Each of the 72 nodes in the output layer corresponds to a specific pitch value, defined in cents. Cent describes the relationship between musical intervals to a reference pitch f_{ref} in Hz, expressed as a function of frequency f in Hz:

$$\psi(f) = 1200 \times log_2 \frac{f}{f_{ref}} \qquad (1)$$

where $f_{ref} = 5$ Hz throughout the program. The 72 pitch values are noted as $\psi_1, \psi_2, \cdots, \psi_{72}$ and selected so that they cover six octaves with 20-cent intervals between C1 and B6, corresponding to 32.70 Hz and 1975.53 hz. $\hat{\psi}$ is the weighted average of the associated pitches ψ_i based output $\hat{y}$, which provides the frequency estimate in Hz:

$$\hat{\psi} = \frac{\sum_{i=1}^{72} \hat{y}\psi_i}{\sum_{i=1}^{72} \hat{y}}, \hat{f} = f_{ref}.2^{\frac{\hat{\psi}}{1200}} \qquad (2)$$

The target outputs we use to train the model are 72-dimensional vectors, where each dimension represents a frequency bin covering 20 cents.

The target is Gaussian-blurred in frequency to reduce the penalty for near-correct predictions, such that the energy surrounding a ground truth frequency decays with a standard deviation of 25 cents [9]:

$$y_i = \exp(-\frac{(\psi_i - \psi_{true})^2}{2.25^2}) \qquad (3)$$

The CNN is trained such that the binary cross-entropy between the target vector y and the predicted vector $\hat{y_i}$:

$$\Gamma(y, \hat{y}) = \sum_{i=1}^{72}(-y_i.log\hat{y}_i - (1 - y_i)log(1 - y_i)) \qquad (4)$$

Where both y_i and $\hat{y}_i$ are real numbers between 0 and 1. We are using Adam optimizer as our loss function, and the learning rate is 0.05. The best performing model is selected after training until the validation accuracy no longer improves for 12 epochs. One epoch consists of 300 batches of 12 examples randomly chosen from the training set.

A more general and comprehensive representation is designed to improve the model's generality further and predict the expressive timing more than the rhythm context. In particular, features are developed from four aspects of expressive collaborative performance, as shown in Figure 4.

4) Beat Tracking: Timing and dynamics are the two most fundamentals aspects of musical expression. In this project, we are planning to model different musicians' presentation as a co-evolving time series. "Based on this representation, we use a set of algorithms, including a sophisticated spectral learning method, to discover regularities of expressive musical interaction from rehearsals" [18].

Our Web-based application will be one of the first applications of spectral learning in the field of music. We consider adding some basic improvisation techniques where musicians have the freedom to interpret pitches and rhythms other than expressive timing and dynamics. We aim to implement a model that trains a different set of parameters for each measure and focuses on predicting the number of chords and the number of notes per chord. Given the model prediction, an improvised score is decoded using the nearest-neighbor search, which selects the training example whose parameters are closest to the determination [18]. We expect the model to generate more musical, interactive, and natural collaborative improvisation than a reasonable baseline based on mean estimation.

C. Web-based Platform Development

Our music learning platform is a Java-based Web application developed using HTML5 and CSS3 for the forum and JSP and Servlets as back-end. Initially, all the requirements were collected and analyzed based on Evolutionary Prototyping (EP).

1) Functional Requirements:

- Home Page (index.jsp): This is a Web page where music teachers and students can log in. The machine will recognize if the user is a student or professor based on the MySQL database's username. Students will be able to look through the curriculum and start their online lessons. Music teachers will set up a session, upload a new course, and embed the procedures provided by our application.
- Video Lessons: The video lessons include a score follower, pitch tracking, and beat tracking to provide real-time feedback for students as they play along. The design of the curriculum is provided in Figure 5.
- Forums: This is a place for students to share their performances with others and find other players to play a duet.
- Student Dashboard (dashboard student.jsp): This is a dashboard for students. Students can use their digital or acoustic piano to practice the lessons.

IV. EVALUATION

The evaluation part of the project is in the proposal phase.

A. Evaluation methods

This paper proposes to use both objective and subjective evaluations. The difference between the predicted results and the ground truth performances could be measured in both simulations and real-time accurate assessment arrangements. We propose to let subjects evaluate both the anticipated results and the ground truth performances for subjective evaluation. In particular, topics should be from two groups, which are non-music major and music major.

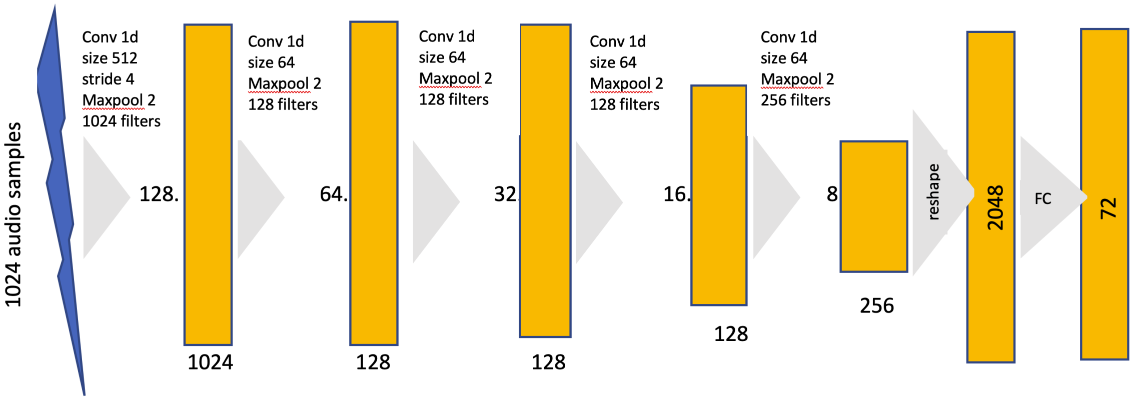


Figure 3. CNN Pitch Detection.

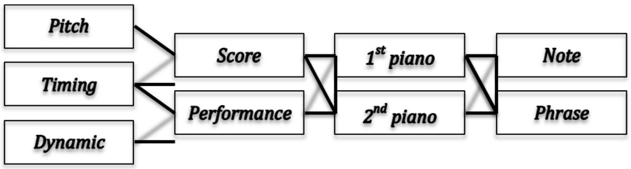
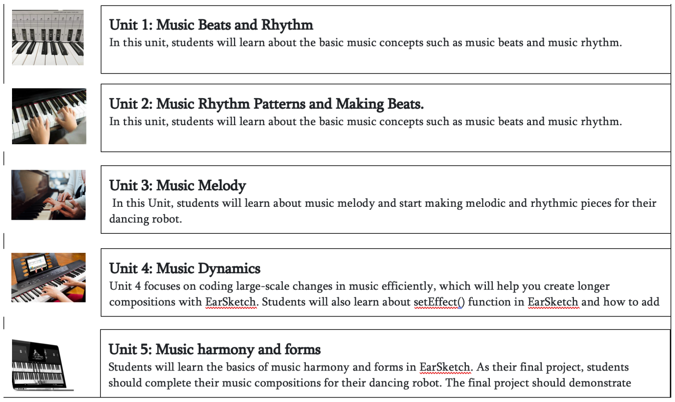


Figure 4. Proposed Architecture.



Unit 1: Music Beats and Rhythm
In this unit, students will learn about the basic music concepts such as music beats and music rhythm.

Unit 2: Music Rhythm Patterns and Making Beats.
In this unit, students will learn about the basic music concepts such as music beats and music rhythm.

Unit 3: Music Melody
In this Unit, students will learn about music melody and start making melodic and rhythmic pieces for their dancing robot.

Unit 4: Music Dynamics
Unit 4 focuses on coding large-scale changes in music efficiently, which will help you create longer compositions with EarSketch. Students will also learn about setEffect() function in EarSketch and how to add

Unit 5: Music harmony and forms
Students will learn the basics of music harmony and forms in EarSketch. As their final project, students should complete their music compositions for their dancing robot. The final project should demonstrate

Figure 5. Units Design.

B. *Criteria for Successful Completion*

In terms of scientific discoveries, successful completion means the introduction questions have been answered and implemented. In terms of system completion, successful completion means the tasks in Section 3 are mainly completed. In particular, pitch detection and instrument recognition are essential successful criteria, and beat tracking is advanced successful criteria. In terms of performance, successful completion means the artificial performer can learn how to sense and coordinate with human performers' music expression from a reasonable amount of rehearsal data. In other words, the artificial performer's synthetic behavior is the same as the ground truth performance and highly rated by subjective evaluations.

V. Conclusion and Future work

We released an initial version of the curriculum in May 2020 for music students at Auburn University. Since then, it has been viewed in academic courses. We have received informal, mostly positive feedback from music teachers (music graduate students) about the platform's user experience and efficiency and the curriculum. We have also received numerous suggestions and feature requests, especially from teachers, which we incorporate into the current version of the curriculum.

We plan to embed our curriculum, teacher training materials, and social media features directly into the interface instead of maintaining them on different websites in the coming year. Another addition to this music learning platform is to make the application more accessible so that students with visual impairment can also learn music concepts and play an instrument.

References

[1] S. W. Conkling, "Envisioning a scholarship of teaching and learning for the music discipline," College Music Symposium, vol. 43, 2003, pp. 55–64. [Online]. Available: http://www.jstor.org/stable/40374470

[2] I. Ruokonen and H. Ruismäki, "E-learning in music: A case study of learning group composing in a blended learning environment," Procedia-Social and Behavioral Sciences, vol. 217, 2016, pp. 109–115.

[3] I. Ruokonen, A. Sepp, A. Ojala, L. Hietanen, V. Tuisku, and H. Ruismäki, "A web-based music learning environment: A case study of users' experiences," The European Journal of Social & Behavioural Sciences, vol. 26, no. 3, 2019, pp. 2983–2993.

[4] R. B. Dannenberg and C. Raphael, "Music score alignment and computer accompaniment," Communications of the ACM, vol. 49, no. 8, 2006, pp. 38–43.

[5] M. Puckette and C. Lippe, "Score following in practice," in Proceedings of the International Computer Music Conference. INTERNATIONAL COMPUTER MUSIC ACCOCIATION, 1992, pp. 182–182.

[6] A. Maezawa and K. Yamamoto, "Muens: A multimodal human-machine music ensemble for live concert performance," in Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems, 2017, pp. 4290–4301.

[7] B. Vera, E. Chew, and P. G. Healey, "A study of ensemble synchronisation under restricted line of sight." in ISMIR, 2013, pp. 293–298.

[8] T. Itohara, K. Nakadai, T. Ogata, and H. G. Okuno, "Improvement of audio-visual score following in robot ensemble with human guitarist," in 2012 12th IEEE-RAS International Conference on Humanoid Robots (Humanoids 2012). IEEE, 2012, pp. 574–579.

[9] J. W. Kim, J. Salamon, P. Li, and J. P. Bello, “Crepe: A convolutional representation for pitch estimation,” in 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2018, pp. 161–165.

[10] A. M. Noll, “Clipstrum pitch determination,” The journal of the acoustical society of America, vol. 44, no. 6, 1968, pp. 1585–1591.

[11] J. Dubnowski, R. Schafer, and L. Rabiner, “Real-time digital hardware pitch detector,” IEEE Transactions on Acoustics, Speech, and Signal Processing, vol. 24, no. 1, 1976, pp. 2–8.

[12] M. Ross, H. Shaffer, A. Cohen, R. Freudberg, and H. Manley, “Average magnitude difference function pitch extractor,” IEEE Transactions on Acoustics, Speech, and Signal Processing, vol. 22, no. 5, 1974, pp. 353–362.

[13] D. Talkin and W. B. Kleijn, “A robust algorithm for pitch tracking (rapt),” Speech coding and synthesis, vol. 495, 1995, p. 518.

[14] P. Boersma, “Accurate short-term analysis of the fundamental frequency and the harmonics-to-noise ratio of a sampled sound,” in Proceedings of the institute of phonetic sciences, vol. 17, no. 1193. Amsterdam, 1993, pp. 97–110.

[15] E. J. Humphrey and J. P. Bello, “Rethinking automatic chord recognition with convolutional neural networks,” in 2012 11th International Conference on Machine Learning and Applications, vol. 2. IEEE, 2012, pp. 357–362.

[16] “Flowkey - learn piano with the songs you love,” https://www.flowkey.com/en [Last accessed: September 2020].

[17] Y. He, “Research on online teaching of music performance based on diversification and intelligence–take the online music teaching during the covid-19 as an example,” in 2020 International Conference on E-Commerce and Internet Technology (ECIT). IEEE, 2020, pp. 193–196.

[18] G. Xia and R. B. Dannenberg, “Duet interaction: learning musicianship for automatic accompaniment.” in NIME, 2015, pp. 259–264.

Building a Web-based Environment to Support Sponsored Research and University-wide Collaborations

Fatemeh Jamshidi
Department of Computer Science and Software Engineering
Auburn University
Auburn, Alabama, USA
Email: fzj0007@auburn.edu

Abhishek Jariwala
Department of Computer Science and Software Engineering
Auburn University
Auburn, Alabama, USA
Email: avj0007@auburn.edu

Bibhav Bhattarai
Department of Computer Science and Software Engineering
Auburn University
Auburn, Alabama, USA
Email: bzb0079@auburn.edu

Katherine Abbate
Project Manager, Military REACH
Auburn University
Auburn, Alabama, USA
Email: kma0057@auburn.edu

Daniela Marghitu
Department of Computer Science and Software Engineering
Auburn University
Auburn, Alabama, USA
Email: marghda@auburn.edu

Mallory Lucier-Greer
College of Human Sciences
Human Development and Family Science
Auburn University
Auburn, Alabama, USA
Email: mluciergreer@auburn.edu

***Abstract*—At the federal level, an organized partnership exists to serve military families, an association composed of the Department of Defense (DoD), the Department of Agriculture (USDA), and colleges and universities throughout the United States of America. Through this partnership, cooperative agreements are executed to support the needs of service members and their families. One such cooperative agreement between DoD, USDA, and Auburn University is Military REACH. This project aims to bridge the gap between military family research and practice by mobilizing peer-reviewed family science research into practical applications for military families and those who work on behalf of military families. At Auburn University, this project is an interdisciplinary collaboration between the academic libraries, the Department of Computer Science, and the Department of Human Development and Family Science. This paper aims to represent the Military REACH website and the new searching functionalities added to the project to increase users' numbers using Google Analytics results.**

Keywords*–*Military Families; Applications; Resources.

I. Introduction

For the past three years, the Auburn University Libraries and Computer Science Department have supported the University's research enterprise in a new way: by adopting a new collaborative model and serving as a high-level Information Technology (IT) and data-management consultants to faculty researchers who are pursuing external funding. A practical example of this model in action is the Military REACH project at Auburn University funded by the Departments of Agriculture and Defense (USDA/NIFA Award No. 2017-48710-27339; PI, Dr. Mallory Lucier-Greer). The purpose of Military REACH is to make research accessible to policy makers, helping professionals, and military families in a manner that is inviting, easily understood, and meaningful for their everyday context [1]. Our team, housed at Auburn University, works to critically evaluate empirical research related to military families and translate it into useful tools. These tools are actively disseminated to policy makers and military helping professionals to inform their decisions and practices as they work to support and enhance the lives of service members and their families. Specifically, the objective of this project is to provide high- quality resources to the Department of Defense (DoD) in the form of research and professional development tools across the spectrum of family support, resilience, and readiness. This work is primarily supported by the DoD's Office of Military Community and Family Policy. The purpose of this project is achieved through three primary deliverables, including:

- Provide timely, high-quality research reports at the request of DoD.
- Re-engineer, grow, and promote an online library of current research and its implications related to the well-being of military families.
- Design and market professional development opportunities, tools, and resources for youth development professionals.

The Military REACH Project is now in its fourth year and seems likely to continue; indeed, it has highlighted the library's value as an IT partner and led to research partnerships and collaborative funding proposals with other units on campus. This paper describes the related functions that are designed and implemented for each operator. The rest of the paper is organized as follows. In Section II, we cover some background about the project. Section III introduces our efforts to serve military families and covers the design and implementation of the website. Section IV demonstrates evaluation methods using Google Analytics. Section V provides evaluation results of the website. Section VI concludes the project along with suggestions for future enhancements.

II. Related Work

Military REACH started by evaluating existing research in the context of Research Infrastructures (RI) and Digital Libraries (DL). Recent reviews of digital preservation [2] and projects that promote research and awareness in the areas of digital preservation include CEDARS [3].

Two decades of research have worked to improve awareness of the digital preservation challenge and encouraged some organizations to improve the longevity of their digital resources. One of the most significant streams of research has been within cultural institutions, sometimes in collaboration with industry partners, to develop solutions to operational problems in these institutions [4]. National, regional, and University archives and libraries in Australia, Canada, Belgium, Denmark France, Germany, the Netherlands, New Zealand, Sweden, Switzerland, the U.K., the U.S., and elsewhere have investigated the implementation of institutional repositories, preservation, and strategies for Web archiving.

III. Collaborative Efforts to Serve Military Families

Working closely with the Military REACH team in the College of Human Sciences, the library's IT department contributed to the original funding proposal and has guided network architecture, Web development, IT tools and solutions, sustainability, data management, accessibility, usage statistics, and automated methods for identifying recently published research.

A. Design and Implementation

The REACH Web application has an architecture that can be implemented into three layers, as shown in Figure 1.

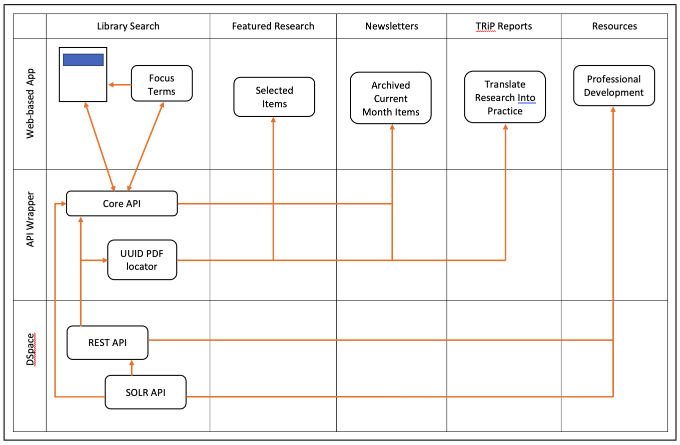


Figure 1. REACH System Architecture.

- Web-based app: This layer is the front-end of the application, where we mainly use Hypertext Markup Language (HTML), Cascading Style Sheets (CSS), and JavaScript in Java Server Pages (JSP). Also, the Cascade Content Management System (CMS) that we use to manage the JSP falls under this layer.
- Application Programming Interface (API) Wrapper: This layer is the back-end layer, where we use JAVA to write classes and methods that handle various functionalities of the website such as search, filter, sort, and so forth.
- DSpace: DSpace is an open-source repository software package mostly used to create open access repositories for the scholarly and published digital content. DSpace is the central database of the application. All Military REACH related research articles are stored in this layer. DSpace uses Apache SOLR based search for metadata and full-text contents, all of which are stored in a relational database and supports the use of PostgreSQL. Also, DSpace is used to manage and preserve all the formats of digital content (PDF, Word, JPEG, MPEG, TIFF files). Likewise, it also allows for a group based access to control the setting for level based permission to individual files.

1) Introduction to the Cascade Content Management Systems: To make the website easy to update, we are also making use of CMS. Cascade CMS is used in the Web application to manage site content, allowing multiple contributors to create, edit, and publish. Content created in a Cascade CMS is stored in Cascade as an XML file and displayed in a presentation layer based on a set of templates. Programming languages such as Extensible Stylesheet Language Transformations (XSLT), and Velocity [5] are used to transform the Extensible Markup Language (XML) file into HTML/JSP pages.

The typical features of a CMS can be as follows:

- Content creation (allows users to easily create and format content)
- Content storage (stores content in one place, in a consistent fashion)
- Workflow management (assigns privileges and responsibilities based on roles such as authors, editors and admins), and
- Publishing (organizes and pushes content live)

2) Cascade Content Management Systems Benefits: What makes Cascade extremely beneficial to a Web application, such as the Military REACH website, is the ease of updating it. Since it is an interface that can be easily used by non-programmers, once it is set up, people without a background in programming can use it to make substantial changes to the website. The What You See Is What You Get (WYSIWYG) editors included in the platform allow users to enter text and upload images even while they lack basic knowledge of HTML or CSS (languages that are vital for any Web application development).

The other advantage of having CMS in our website development process is its collaborative nature. Multiple users can log on and contribute, schedule or edit content to be published. Since the interface is browser-based, Cascade can, therefore, be accessed from anywhere by any number of users.

Similarly, Cascade CMS has an efficient, reliable way of sending frequent alerts to the users and site administrators of pages that have not been updated for a certain duration of time. The use of in-built features such as the daily content report, task manager, and content review dates helps an organization to stay fresh and take necessary steps to keep its users up-to-date.

Lastly, Cascade has a community of over 100,000 active users that are frequently using the platform and are readily available to voice their experiences with using features and capabilities of Cascade.

3) Use of Cascade Content Management System on Military REACH:

- The Teams page and Resources page of the website are entirely made in the Cascade CMS. These are the pages in the website that can be easily updated by even the non-technical team members in the organization.
- Other pages, such as the Home page, the Families page, the Contact Us page, and so on, are more hybrid, where all of the texts displayed in the pages can be edited from Cascade. Other major functionalities within the pages are, however, handled in the back-end by the developers.
- Therefore, having Cascade only pages and hybrid pages simultaneously gives a lot of flexibility for both the technical and non-technical team members involved in the organization.

IV. Evaluation Methods

A. Google Analytics

Military REACH has been using Google Analytics to access the user data since last year (March 1, 2019 - July 31, 2020). Google Analytics data do not include any personally identifiable information. They are presented in the aggregate data, making it a practical tool used in research settings without ethical concerns [6][7]. The Web development team installed Google Analytics by adding a tracking tag for Military REACH. The tracking tags are a combination of JavaScript and computer programming language used to develop the website. The tracking tag code allows contributing different forms of data related to the users' behavior on the website as soon as they visit the Military REACH website. The data can proceed from diverse avenues. For example, the URL of the page and the device used to access the site. Tracking code collects more data on the nature of the visit, such as the contents viewed, length of the session, average time on each page, location, etc. This information is in a real-time, interactive dashboard format that can be viewed by logging in to Google Analytics.

B. User Engagement

This project focuses on several indicators from Google Analytics to evaluate the level of engagement. These indicators contain the number of returning users (n), bounce rate, number of pages accessed per session (n), mean session, and time spent on each page (minutes, seconds).

The number of returning users mentions the number of sessions visited through the same client IP. A high number of returning users indicates a strong level of engagement with the Web-based platform [6][8].

The bounce rate is a percentage of single-page sessions in which there was no interaction with the page. A high bounce rate means minimal interaction with the page; however, it could also mean that users exit the page after finding what they were looking for right away. A low bounce rate can refer to a high overall engagement, especially for a multi-component platform like Military REACH. For example, there are not many available resources that would provide mental health support on the platform's home page. Therefore, users will often need to interact with various searching tools and Web pages to access the required information.

The number of pages per session indicates the number of Web pages that the user viewed in a single session. The mean session duration (minutes, seconds) means the mean duration of the time users spend on the website. Using these indicators to increase user engagement can be challenging. There are different interpretations. For example, many pages per session could occur from a high level of engagement, while it could also cause a superficial exploration of several pages. Besides, a long session duration can result from increased attention, but it could also be because the user keeps the Web page open while engaging in the other irrelevant activities.

C. Platform Improvement

Military REACH considers some other indicators from Google Analytics to inform the improvement of the platform. These indicators include page views, mean duration of visit, and bounce rate when accessing self-help tools (e.g., Family Focus page, TRIP reports page). Besides, the most visited pages were observed in terms of their overall average time spent on the page to understand which tools or pages were most beneficial or viewed.

The entrance rate illustrates a proportion of sessions starting from a given page. In comparison, the exit rate results from a ratio of sessions ending from a given page. The information regarding the entrance rate may explain which Web page serves as the first impression for the users. The exit rate may indicate when users felt disengaged or had adequate data needed for the session. Google Analytics provides information on the type of devices users are using to access the website. Such data can allow us to consider if implementing a mobile app for Military REACH would be practical or not. The three primary devices of interest to the current investigation are desktops, tablets, and mobile phones (counted here as mobile devices).

D. Marketing Strategy

Military REACH aims to reach as many users as possible. Therefore, we used Google Analytics to inform our marketing strategy. The research team has been reaching out to various military-connected organizations, especially around the United States. Twitter, Facebook, and LinkedIn accounts were also created to distribute awareness about the platform. To improve the marketing strategy, the ways used to access the website were analyzed. The methods include a direct link (i.e., typing the Web URL directly into a browser); organic search (i.e., entry through a search engine); and referrals via another website via social media via email. Understanding which ways are most accessible for users can help to improve the marketing strategy. Military REACH also uses the locations of users from different countries around the world.

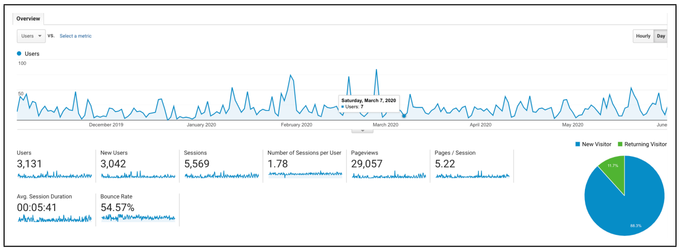
V. Evaluation Results

Military REACH started using Google Analytics since March 2019. The first version of the website was based on a single page application (March 2019 - November 2019). However, to better access our users' data, we switched to multiple page application using Java Server Pages (JSP) and Servlets (November 2019 - present). The following are the results from Google Analytics, which show the positive impact of this change in user engagement and platform functionalities to serve military families better.

A. User Engagement

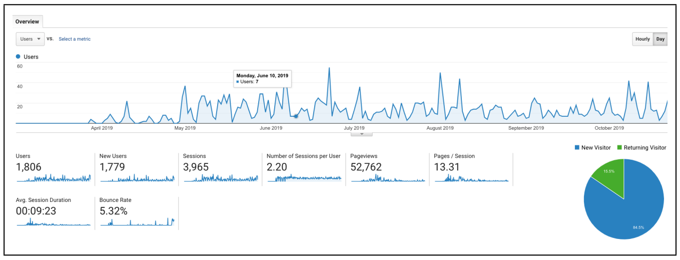
The last year of operation for the Military REACH platform saw 3,131 users from November 2, 2019 – June 11, 2020, and a total of 1806 users from March 1, 2019 - November 2, 2019 (Shown in Figures 2 and 3).

On average, users visited 5.22 pages per session from November 2, 2019 – June 11, 2020, and 13.31 pages from March 1, 2019 - November 2, 2019.



Users 3,131
New Users 3,042
Sessions 5,569
Number of Sessions per User 1.78
Pageviews 29,057
Pages / Session 5.22
Avg. Session Duration 00:05:41
Bounce Rate 54.57%

Figure 2. REACH overview presented in Google Analytics (Nov 2, 2019 - June 11, 2020).



Users 1,806
New Users 1,779
Sessions 3,965
Number of Sessions per User 2.20
Pageviews 52,762
Pages / Session 13.31
Avg. Session Duration 00:09:23
Bounce Rate 5.32%

Figure 3. REACH overview presented in Google Analytics (March 1, 2019 - November 2, 2019).

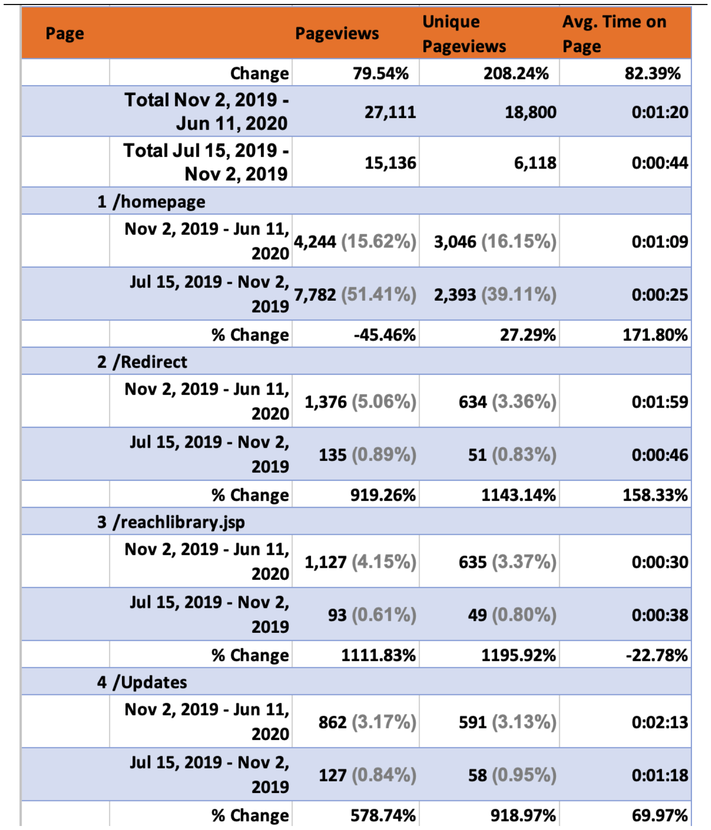
The results show that user engagement is increasing because of social media marketing, conferences, and overall better efficiency and effectiveness of the website.

B. Platform Improvement

Visits to the Military REACH Home Page comprised 15.62% (4244/27,111) of the total pageviews from November 2, 2020, to June 11, 2020, with an average duration spent of 1 minute and 20 seconds.

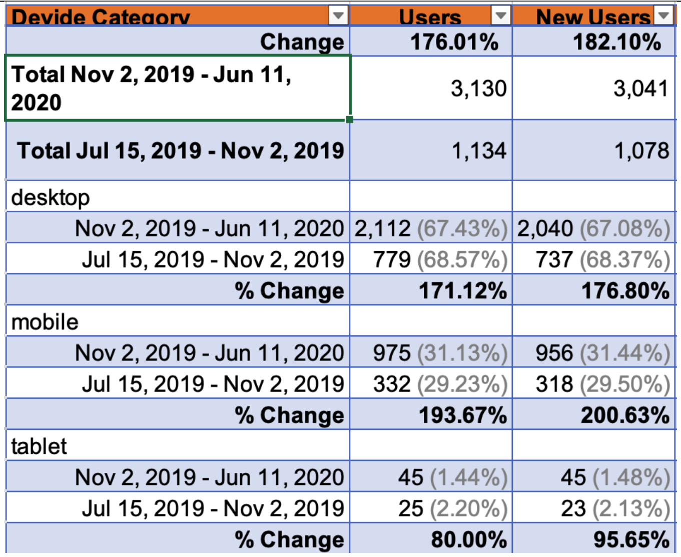
Table 1 presents details of the top ten most viewed pages. In March 2019 to Nov 2019, the Military REACH home page, which acts as the landing page, accounted for 51.41% (7782/15,136) of all entries when the website was still a single page application using Angular and Typescript. However, after transforming to multiple page applications, users can access the resources they are looking for, using shared links on our social media or email. A list of devices used by Military REACH users to access the site is presented in Table 2, indicating that the platform was accessed mostly via desktops (2112/3130, 67.43%). Furthermore, sessions completed via desktops had a higher average session duration than those completed via other devices.

TABLE I. REACH MOST VIEWED PAGES.



Page		Pageviews	Unique Pageviews	Avg. Time on Page
	Change	79.54%	208.24%	82.39%
	Total Nov 2, 2019 - Jun 11, 2020	27,111	18,800	0:01:20
	Total Jul 15, 2019 - Nov 2, 2019	15,136	6,118	0:00:44
1	/homepage			
	Nov 2, 2019 - Jun 11, 2020	4,244 (15.62%)	3,046 (16.15%)	0:01:09
	Jul 15, 2019 - Nov 2, 2019	7,782 (51.41%)	2,393 (39.11%)	0:00:25
	% Change	-45.46%	27.29%	171.80%
2	/Redirect			
	Nov 2, 2019 - Jun 11, 2020	1,376 (5.06%)	634 (3.36%)	0:01:59
	Jul 15, 2019 - Nov 2, 2019	135 (0.89%)	51 (0.83%)	0:00:46
	% Change	919.26%	1143.14%	158.33%
3	/reachlibrary.jsp			
	Nov 2, 2019 - Jun 11, 2020	1,127 (4.15%)	635 (3.37%)	0:00:30
	Jul 15, 2019 - Nov 2, 2019	93 (0.61%)	49 (0.80%)	0:00:38
	% Change	1111.83%	1195.92%	-22.78%
4	/Updates			
	Nov 2, 2019 - Jun 11, 2020	862 (3.17%)	591 (3.13%)	0:02:13
	Jul 15, 2019 - Nov 2, 2019	127 (0.84%)	58 (0.95%)	0:01:18
	% Change	578.74%	918.97%	69.97%

TABLE II. DEVICES USED TO ACCESS MILITARY REACH



Devide Category	Users	New Users
Change	176.01%	182.10%
Total Nov 2, 2019 - Jun 11, 2020	3,130	3,041
Total Jul 15, 2019 - Nov 2, 2019	1,134	1,078
desktop		
Nov 2, 2019 - Jun 11, 2020	2,112 (67.43%)	2,040 (67.08%)
Jul 15, 2019 - Nov 2, 2019	779 (68.57%)	737 (68.37%)
% Change	171.12%	176.80%
mobile		
Nov 2, 2019 - Jun 11, 2020	975 (31.13%)	956 (31.44%)
Jul 15, 2019 - Nov 2, 2019	332 (29.23%)	318 (29.50%)
% Change	193.67%	200.63%
tablet		
Nov 2, 2019 - Jun 11, 2020	45 (1.44%)	45 (1.48%)
Jul 15, 2019 - Nov 2, 2019	25 (2.20%)	23 (2.13%)
% Change	80.00%	95.65%

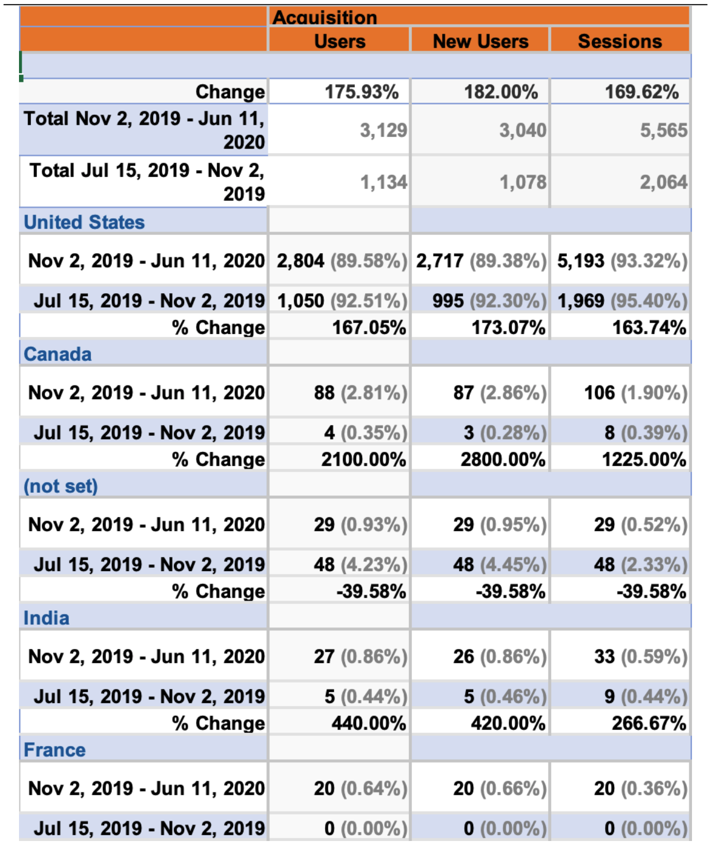
C. Marketing Strategy

Approximately 89.58% (2804/3129) of the users accessed the website from the United States. Table 3 shows that the users accessed the platform from around the world (Figure 4).

Google Analytics was a helpful tool to process the evaluation of the open-access, Web-based Military REACH platform.

The process evaluation provided information about the ways to keep users engaged, marketing strategies, and the aspects of the platform that required improvement.

TABLE III. LOCATIONS OF USERS FROM GOOGLE ANALYTICS.



	Acquisition		
	Users	New Users	Sessions
Change	175.93%	182.00%	169.62%
Total Nov 2, 2019 - Jun 11, 2020	3,129	3,040	5,565
Total Jul 15, 2019 - Nov 2, 2019	1,134	1,078	2,064
United States			
Nov 2, 2019 - Jun 11, 2020	2,804 (89.58%)	2,717 (89.38%)	5,193 (93.32%)
Jul 15, 2019 - Nov 2, 2019	1,050 (92.51%)	995 (92.30%)	1,969 (95.40%)
% Change	167.05%	173.07%	163.74%
Canada			
Nov 2, 2019 - Jun 11, 2020	88 (2.81%)	87 (2.86%)	106 (1.90%)
Jul 15, 2019 - Nov 2, 2019	4 (0.35%)	3 (0.28%)	8 (0.39%)
% Change	2100.00%	2800.00%	1225.00%
(not set)			
Nov 2, 2019 - Jun 11, 2020	29 (0.93%)	29 (0.95%)	29 (0.52%)
Jul 15, 2019 - Nov 2, 2019	48 (4.23%)	48 (4.45%)	48 (2.33%)
% Change	-39.58%	-39.58%	-39.58%
India			
Nov 2, 2019 - Jun 11, 2020	27 (0.86%)	26 (0.86%)	33 (0.59%)
Jul 15, 2019 - Nov 2, 2019	5 (0.44%)	5 (0.46%)	9 (0.44%)
% Change	440.00%	420.00%	266.67%
France			
Nov 2, 2019 - Jun 11, 2020	20 (0.64%)	20 (0.66%)	20 (0.36%)
Jul 15, 2019 - Nov 2, 2019	0 (0.00%)	0 (0.00%)	0 (0.00%)

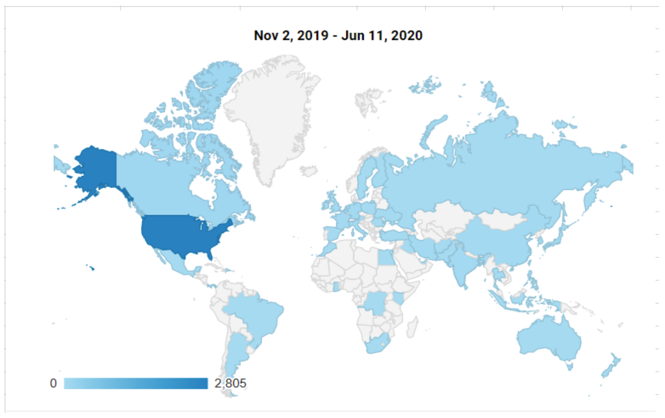


Figure 4. Map overlay about locations of users from Google Analytics.

Google Analytics was a helpful tool to process the evaluation of the open-access, Web-based Military REACH platform.

The process evaluation provided information about the ways to keep users engaged, marketing strategies, and the aspects of the platform that required improvement.

VI. CONCLUSION AND FUTURE WORK

Google Analytics results helped the Military REACH team analyze the website's usage to serve military families better. It shows that after adding more features to the search, function users are using the website in a practical way and spending more time on the website. Compared to the first two years, website usage almost tripled last year.

According to the Google Analytics results, 31% of users have access to the website through their phone. Therefore, to facilitate the accessibility of Military REACH resources, the team is investigating a mobile application (app).

In the future, Military REACH plans to conduct a pilot testing of a newly developed mobile app that will be used for the dissemination of REACH reports, mainly Translating Research Into Practice (TRIP) reports. We will conduct an efficacy study to examine the impact of our mobile app and TRIP reports specifically for helping professionals who directly serve military families. Survey data will be collected from participants (i.e., primary data collection) using Qualtrics (a survey software used at Auburn University), a secure online data collection tool. This data will help us understand the users' military family knowledge better, their confidence in serving military families, their satisfaction and reaction to the app, and make the military family research accessible to everyone.

REFERENCES

[1] L. Nichols, K. Abbate, C. W. O'Neal, and M. Lucier-Greer, "Mobilizing family research: Evaluating current research and disseminating practical implications to families, helping professionals, and policy makers," Southeastern Council on Family Relations Conference, Jul. 2019.

[2] H. R. Tibbo, "On the nature and importance of archiving in the digital age." Adv. Comput., vol. 57, Jan. 2003, pp. 1–67.

[3] K. Russell, "Digital preservation and the cedars project experience," New review of academic librarianship, vol. 6, no. 1, Apr. 2000, pp. 139–154.

[4] S. Ross and M. Hedstrom, "Preservation research and sustainable digital libraries," International journal on digital libraries, vol. 5, no. 4, Apr. 2005, pp. 317–324.

[5] A. Deshpande, A. Göllü, and L. Semenzato, "The shift programming language and run-time system for dynamic networks of hybrid automata," in Verification of Digital and Hybrid Systems. Springer, Jun. 2000, pp. 355–371.

[6] E. A. Song, "A process evaluation of a web-based mental health portal (walkalong) using google analytics," JMIR mental health, vol. 5, no. 3, Jul. 2018, p. e50.

[7] D. J. Clark, D. Nicholas, and H. R. Jamali, "Evaluating information seeking and use in the changing virtual world: the emerging role of google analytics," Learned publishing, vol. 27, no. 3, 2014, pp. 185–194.

[8] E. A. Vona, "A web-based platform to support an evidence-based mental health intervention: lessons from the cbits web site," Psychiatric Services, vol. 65, no. 11, Jan. 2014, pp. 1381–1384.