CONTENT 2019

The Eleventh International Conference on Creative Content Technologies

ISBN: 978-1-61208-707-8

May 5 - 9, 2019

Venice, Italy

CONTENT 2019 Editors

Hans-Werner Sehring, Namics, Germany

CONTENT 2019

Forward

The Eleventh International Conference on Creative Content Technologies (CONTENT 2019), held between May 5 - 9, 2019 - Venice, Italy, continued a series of events targeting advanced concepts, solutions and applications in producing, transmitting and managing various forms of content and their combination. Multi-cast and uni-cast content distribution, content localization, on-demand or following customer profiles are common challenges for content producers and distributors. Special processing challenges occur when dealing with social, graphic content, animation, speech, voice, image, audio, data, or image contents. Advanced producing and managing mechanisms and methodologies are now embedded in current and soon-to-be solutions.

The conference had the following tracks:

- Data Transmission and Management
- Web content
- Domains and approaches

Similar to the previous edition, this event attracted excellent contributions and active participation from all over the world. We were very pleased to receive top quality contributions.

We take here the opportunity to warmly thank all the members of the CONTENT 2019 technical program committee, as well as the numerous 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 that dedicated much of their time and effort to contribute to CONTENT 2019. We truly believe that, thanks to all these efforts, the final conference program consisted of top quality contributions.

Also, this event could not have been a reality without the support of many individuals, organizations and sponsors. We also gratefully thank the members of the CONTENT 2019 organizing committee for their help in handling the logistics and for their work that made this professional meeting a success.

We hope CONTENT 2019 was a successful international forum for the exchange of ideas and results between academia and industry and to promote further progress in the area of creative content technologies. We also hope that Venice provided a pleasant environment during the conference and everyone saved some time for exploring this beautiful city.

CONTENT 2019 Chairs

CONTENT 2019 Steering Committee

Raouf Hamzaoui, De Montfort University - Leicester, UK
Mu-Chun Su, National Central University, Taiwan
Nadia Magnenat-Thalmann, University of Geneva, Switzerland
Paulo Urbano, Universidade de Lisboa, Portugal
José Fornari, UNICAMP, Brazil

CONTENT 2019 Industry/Research Advisory Committee

Hans-Werner Sehring, Namics, Germany
René Berndt, Fraunhofer Austria Research GmbH, Austria
Daniel Thalmann, Institute for Media Innovation (IMI) - Nanyang Technological University, Singapore
Stefanos Vrochidis, Information Technologies Institute - Centre for Research and Technology Hellas, Greece

CONTENT 2019

Committee

CONTENT Steering Committee
Raouf Hamzaoui, De Montfort University - Leicester, UK
Mu-Chun Su, National Central University, Taiwan
Nadia Magnenat-Thalmann, University of Geneva, Switzerland
Paulo Urbano, Universidade de Lisboa, Portugal
José Fornari, UNICAMP, Brazil

CONTENT 2019 Industry/Research Advisory Committee
Hans-Werner Sehring, Namics, Germany
René Berndt, Fraunhofer Austria Research GmbH, Austria
Daniel Thalmann, Institute for Media Innovation (IMI) - Nanyang Technological University, Singapore
Stefanos Vrochidis, Information Technologies Institute - Centre for Research and Technology Hellas, Greece

CONTENT 2019 Technical Program Committee

Jose Alfredo F. Costa, Federal University - UFRN, Brazil
Hamed Alhoori, Northern Illinois University, USA
Mostafa Alli, Tsinghua University, China
Leonidas Anthopoulos, University of Applied Science (TEI) of Thessaly, Greece
Iana Atanassova, Université de Franche-Comté, France
Konstantinos Avgerinakis, CERTH-ITI, Greece
Kambiz Badie, Research Institute for ICT & University of Tehran, Iran
René Berndt, Fraunhofer Austria Research GmbH
Christos Bouras, University of Patras | Computer Technology Institute & Press <Diophantus>, Greece
Marcelo Caetano, INESC TEC, Porto, Portugal
Juan Manuel Corchado Rodríguez, Universidad de Salamanca, Spain
João Correia, University of Coimbra, Portugal
Raffaele de Amicis, Oregon State University, USA
Rafael del Vado Vírseda, Universidad Complutense de Madrid, Spain
Myriam Desainte-Catherine, LaBRI - Université de Bordeaux, France
Joël Dumoulin, HumanTech Institute | University of Applied Sciences of Western Switzerland, Switzerland
Miao Fan, Tsinghua University, China
José Fornari, UNICAMP, Brazil
Alexander Gelbukh, Instituto Politécnico Nacional, Mexico
Afzal Godil, National Institute of Standards and Technology, USA
Seiichi Gohshi, Kogakuin University, Japan
Raouf Hamzaoui, De Montfort University, Leicester, UK
Tzung-Pei Hong, National University of Kaohsiung, Taiwan

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

Real-Time Noise Level Detection for General Video

Chinatsu Mori, and Seiichi Gohshi

Kogakuin University
1-24-2 Nishi-Shinjuku, Shinjuku-ku, Tokyo, Japan
Email: gohshi@cc.kogakuin.ac.jp

***Abstract*—Currently, 4K TV is a standard TV and 8K broadcasting has began in December 2018. High-resolution in conjunction with low noise is an essential figure of merit in video systems. Unfortunately, any increase in resolution unavoidably increases noise levels. A signal processing method called noise reducer (NR) is often used to reduce noise. However, accurate noise level is needed when NR is employed. Noise level depends mostly on lighting conditions and is estimated by comparing adjacent frame difference. However, the frame difference is generated by moving objects as well as noise. Therefore, it is essential to determine whether the frame difference is caused by the moving objects or by the noise, which is a difficult task. Another difficulty arises from the fact that noise level detection must be achievable in real-time conditions since all video systems are required to work in real-time. This means that a complex method could not be used for noise level detection. In this paper, two noise level detection algorithms are presented. The combination of two of them is a concise algorithm able to accurately detect the noise level and work in real-time conditions.**

Keywords–Video noise reducer; 4KTV; 8KTV; Real-time; Nonlinear signal processing; Image quality.

I. INTRODUCTION

A dramatic change in imaging technologies has taken place in the 21st century. High Definition Television (HDTV) broadcasting started only 20 years ago and at that time HDTV sets were expensive. Today, HDTV is already a part of history. 4K TV broadcasting started a couple of years ago and 8K satellite broadcasting is started in December 2018. Although significant advances have been made in video resolution, imaging technologies are based on the same principle, i.e., the photoelectric effect. Imaging devices primarily comprise of photoelectric cells and the number of electrons generated by each cell is proportional to the number of photons received by the cell. As the resolution increases from HDTV to 4K and then to 8K, the size of the image cell decreases, i.e., the number of photons per image cell is inversely proportional to the resolution. Therefore, it is necessary to amplify the electric energy of a video signal at the output of a video camera.

The electrical energy generated by the image cell is amplified by a pre-amplifier for each pixel. An amplifying process always results in thermal noise called “Gaussian noise.” The level of noise is inversely proportional to the electric energy generated per cell. This is because fewer photons generate a lower voltage signal that requires amplification to achieve the appropriate voltage level. As HDTV, 4K, and 8K are high-resolution systems, the noise level increases because the size of the image cells becomes smaller due to the high-resolution. The best way to reduce noise in a high-resolution video is to increase the sensitivity of image cells’ photoelectric effect. However, in order to achieve this, there are technical limitations, which need to overcome. Even high-end mature HDTV cameras may have pulse noise called “Shot noise” under poor lighting conditions, such as night time shooting or shooting in a dark room.

Noise reducer (NR) is a technology able to reduce noise in video systems by using signal processing techniques. Although, a large number of NR algorithms have been reported most of them are complex and only compatible with still images. The use of such an algorithm in real-time video systems would cause a video to freeze. In other words, complex NR algorithms are not suitable for use in real-time video systems. Another issue is the ability to detect accurate noise levels in video/image systems before applying noise reducing techniques. In case of real-time video systems, noise levels should be detected in real-time as well. Adjacent frame difference is a basic method to detect noise levels. However, noise, as well as moving objects, is contained in the frame, which makes the detection of accurate noise levels in a real-time video a difficult task. In this paper, a real-time noise level detection method is proposed.

This paper is organized as follows. In Section II, related works of NR and noise level detection are explained. In Section III, two noise level detection algorithms are proposed. In Section IV, simulation results are presented. In Section V, the advantages and disadvantages of the algorithms are discussed and the combination of two of them is investigated. Finally, in section VI, conclusions of this work are presented.

II. RELATED WORKS

Conventional NR uses spatial or temporal digital filter to reduce noise [1]–[5]. Many NR methods are used for still images. They are spatial digital filters. Generally, the spatial digital filters cause image blurring. Although the common method is NR with wavelet transformation [6]–[9], the application of this method in videos is difficult: because real-time performance is required. Hence, an NR with a recursive temporal filter [10] is the only practical real-time method used for videos. However, it is necessary to know the accurate noise level for the NR to work. Generally, videos comprise a wide variety of content with different noise levels. The differences are also caused by lighting conditions. In the development of automatic, real-time NR hardware, the NR parameter must be set properly in accordance with the actual noise level of a video. Although the adjacent frame difference is the basis of noise level detection, the frame difference is the result of noise and the moving areas.

Only a few proposals for noise level detection methods in videos are available. The wavelet transformation is used

 ISBN: 978-1-61208-707-8

for the noise level detection [11], [12], but its real-time work application is difficult owing to its high processing cost.

The spatial and temporal digital filter is simple and is used for noise level estimation with low cost [13], [14]. Gaussian noise can be detected by applying high-pass filter, such as Sobel filter and Laplacian filter. However, these filters detect both noise and temporal moves of videos: the noise level is overestimated if the video includes fast and complex moves, such as camera works and object moves.

In the authors' previous works, a noise level detection method which uses a bilateral filter has been proposed [15]. However, the bilateral filter also comes with a high hardware cost. A noise level detection algorithm is essential not only for the real-time function but also the accurate determination of the actual noise level. The method that uses the bilateral filter fails to perform when the noise level is high. Therefore, some improvements are necessary to address these issues.

III. Proposed Methods

In this paper, two noise level detection algorithms are proposed and the combination of these methods is considered.

A. Noise Level Estimation

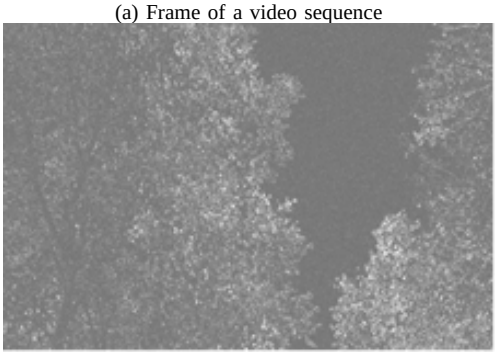
A video has three axes, namely, vertical, horizontal, and the frame. The plane that consists of the vertical and horizontal axes is called spatial, whereas the frame axis is called temporal. By comparing the correlations of spatial and temporal, the spatial correlation is stronger than the temporal. The conventional NR [10] uses the temporal characteristic, as does the noise level detection algorithm. However, the adjacent frame difference is the most effective method to detect the noise level, but it involves two types of signals: frame differences caused by noise and that by moving objects in a video.

Figure 1 illustrates some examples. Figure 1 (a) presents the frame of a video [16]. In the sequence, trees and leaves rustle in the wind. Figure 1 (b) shows the frame difference caused by the trees and leaves. The noise level can be obtained by the standard deviation of the frame difference values in the flat areas because the frame difference in the flat areas is created by noise. Thus, separating the flat areas with frame difference caused by noise from the areas with moving objects is necessary. There are two characteristics of the frame differences for separating the flat areas and moving areas. The frame difference caused by moving objects has shapes and areas, whereas that caused by noise is isolated. Moreover, moving objects have large frame difference values, whereas noise often generates small difference values. Based on these characteristics, we introduce two NR methodologies.

B. Frame Difference and Threshold Process

As discussed in Section III-A, the frame difference values caused by the moving objects are larger, thus, distinguishing these two using a threshold process is possible. Figure 2 shows the block diagram of the noise level detection with frame difference and threshold processing. The frame difference is detected using a frame memory and the input frame. In the threshold processing, only a small frame difference is selected, and its values and pixel numbers are sent to the noise level calculation block. In the noise level calculation block, the frame difference values and pixel numbers are accumulated. The average noise level can be measured using these two

(a) Frame of a video sequence



(b) Frame difference of (a)

Figure 1. Video frame and frame difference

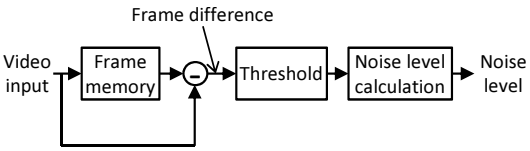


Figure 2. Frame difference and threshold process

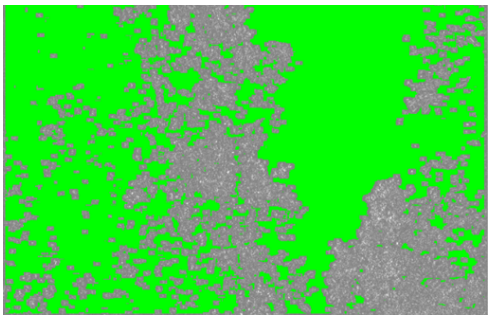


Figure 3. Areas detected using the frame difference and threshold process

values. Figure 3 shows the candidate of the flat areas using the frame difference and threshold process. However, this method also incorrectly identifies the frame difference caused by moving objects in the tree areas. The moving objects do not always produce large frame difference values. With the luminance-level difference between the moving objects and the background, the frame difference values are small and can sometimes generate similar values to those caused by noise. Although the frame difference between the blue sky and the trees is substantial in the video shown in Figure 1 (b), the frame difference among the tree leaves is minimal and similar to the values caused by noise. The incorrect identification due to similar magnitudes in change between moving objects and noise is the problem with this method.

 ISBN: 978-1-61208-707-8

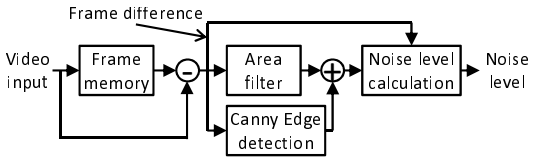
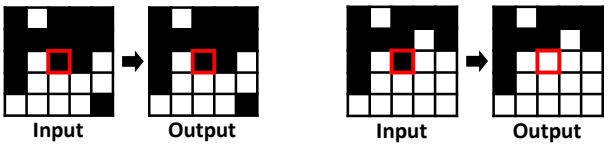


Figure 4. Proposed method 1



(a) Interest pixel as flat area (b) Interest pixel as moving area

Figure 5. Examples of area filter process.

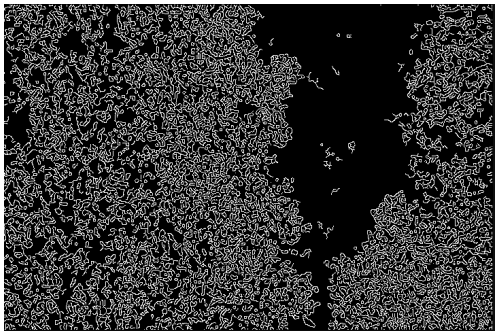


Figure 6. Output of the Canny edge detection block

C. Proposed Method 1: Area Filter and Edge Detection

As shown in Figure 3, the frame difference caused by the tree leaves is detected as the flat areas for determination of the noise level. Although the moving objects, that is, trees and leaves, result in large frame difference values, some can be quite similar to the nearby areas, such as white shining leaves and the blue sky. The shining leaves and the sky produce small frame difference, such as noise because they have similar luminance levels. To prevent this issue, we need to connect these areas and exclude the spaces from analysis. Thus, we introduce the area filter and Canny edge detection [17] illustrated in Figure 4, to improve the noise level accuracy.

Based on the input to the frame difference detection, Figures 4 and 2 are similarly presented. The frame difference is distributed into three blocks: the area filter, the Canny edge detection, and the noise level detection in Figure 4. The function of the area filter is illustrated in Figure 5, and is a symmetric nonlinear type of filter. The center pixel value is processed with the surrounding pixel values and has two parameters, the kernel size and the threshold level. The kernel size is 5×5, as shown in Figure 5. The input of the area filter is the frame difference and has positive and negative values.

In the area filter block, the frame difference is processed with an absolute function to render all values positive. The absolute values are identified using the algorithm presented in Figure 5. The white pixels indicate values exceeding the threshold level, whereas the black pixels are equal to or less than the threshold level. If the number of the surrounding pixels exceeding the threshold level is the majority, the area filter decides the interest pixels as the moving area, otherwise, it decides the interest pixels as the flat area. As shown in Figure 5 (a), the number of pixels exceeding the threshold is 11 (the white bocks), and the number equal to or less than the threshold is 14 (the black blocks). In this case, the output of the center pixel is as the flat area. As shown in Figure 5 (b), the number of pixels exceeding the threshold is 14 (the black blocks) and less than or equal to the threshold is 11 (the white blocks). Therefore, the output of the center pixel is as the moving area. By using the following method, we can detect most of the moving areas in Figure 3, but not quite all of them. Therefore, we also introduce the Canny edge detection. The Canny edge detection identifies the continuous edges in the frame difference. These edges are caused by the leaves. A couple of pixels around the Canny detected edges are obtained from the Canny edge detection block. The result of the Canny edge detection is shown in Figure 6. By using the logical OR on the area filter and edge detection blocks, the appropriate areas for the noise level detection are accurately detected.

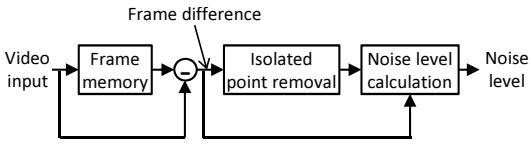


Figure 7. Proposed method 2-A

(a) Interest pixel in moving areas (b) Interest pixel in flat areas

Figure 8. Examples of isolated point removal process

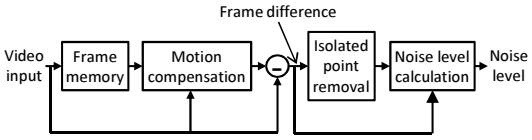


Figure 9. Proposed method 2-B

D. Proposed Method 2: Isolated Point Removal and Motion Compensation

The proposed method 1 shown in Figure 4 can accurately detect the noise level when standard deviation is less than 9. We will discuss the problem in the following section in detail. To address the problem that arises when standard deviation is higher than 9, we have proposed another method.

The signal flow of the proposed method 2-A for a high-level noise is shown in Figure 7. The frame difference detection process of the input and frame memory blocks in Figure 7 is the same as that in Figure 4. The frame difference is distributed into two blocks. The first one is the isolated point removal block and the other one is the noise level calculation block. As discussed in Section III-A, the frame differences are caused both by moving objects and noise. Given that noise level can be detected in flat areas, discriminating the flat areas with noise from the entire frame is necessary. Generally, the frame differences caused by noise in flat areas are isolated. When isolated point removal is used, the output of the isolated point removal block can be the same as the frame difference caused by the moving object, and the noise level can be estimated

 ISBN: 978-1-61208-707-8

using the areas excluding the detected moving areas.

The isolated point removal process is shown in Figure 8. The center pixel in Figure 8 is the interest pixel. Figure 8 (a) shows an example where the interest pixel is the moving area, and Figure 8 (b) illustrates the noise on the flat area. The input is the frame difference. Moreover, the absolute value of the frame difference is calculated and is binarized using the threshold level. The pixels shown in Figure 8 are the result of the binarization. The black areas are below or equal to the threshold level, indicating the flat area. Meanwhile, the white areas are higher than the threshold level, which are candidates similar to the moving areas or the noise on the flat areas. Using only the flat areas is necessary for the noise level estimation. Thus, in the isolated point removal process, the candidate pixels in the white areas are removed if the pixel is isolated and identified as the noise on the flat area. The parameter of the pixel size of the noise is used and the pixel size is set to 5 pixels, as shown in Figure 8. As presented in Figure 8 (a), the pixel size of the white area contains 7 pixels, which is larger than 5. The process identifies the area to be the moving area. As shown in Figure 8 (b), the pixel of the white area contains 4 pixels, which is less than 5. In this case, the pixel is determined to represent the noise, and it is removed.

Many frame differences are present in the frames. These differences have larger values when a video includes camera works, such as panning and tilting. However, the threshold process cannot detect the frame difference accurately for the noise level detection. Thus, we also introduce a block-based motion compensation to detect and reduce moving areas in the frame difference. The proposed method 2-A with motion compensation (method 2-B) is shown in Figure 9. The process of the motion compensation block; the frame is partitioned into blocks of pixels, and each pixel of a block is shifted to the position of the predicted block via the motion vector. This process is common in the discussions of video coding technologies, such as MPEG-2, MPEG-4, and HEVC. Furthermore, we verify and discuss the performance of the motion compensation in the following sections.

IV. EXPERIMENT

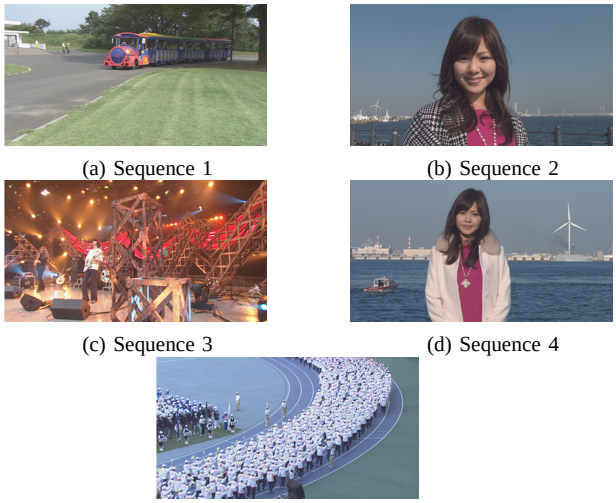
Simulation experiment was conducted to verify the performance of the proposed methods. Different levels of noise were added to video sequences, and the accuracy of the estimated noise level determined by each method was compared.

A. Test Sequences

Noise levels in general videos were estimated using the frame difference (Section 3.1), the proposed method 1 (Section 3.3), and the proposed methods 2-A and 2-B (Section 3.4). The five HDTV $(1,920 \times 1,080)$ video sequences [16] shown in Figure 10 were used in this experiment. All sequences included moving objects and various camera actions, such as panning and tilting. Gaussian noise with different standard deviations (1, 3, 5, 7, 9, 11, 13, and 15) was added to the videos.

B. Experimental Results

The experimental results are shown in Figures 11 and 12. Figures 11 (a)-(e) show the results for sequences 1-5 respectively. The figures show the estimated standard deviation for each level of added noise. The x-axis is the standard deviation of the noise added to the test sequence, and the y-axis



(a) Sequence 1 (b) Sequence 2

(c) Sequence 3 (d) Sequence 4

(e) Sequence 5

Figure 10. Test sequences

is the estimated standard deviation of the noise in the sequence. The marks show the median values of the estimated standard deviations. If the estimated noise level is correct, the result has the same value as the added noise standard deviation, i.e., $y = x$. The bars indicate the minimum to maximum range of the estimated noise standard deviation, which shows the variation of the results in the sequence.

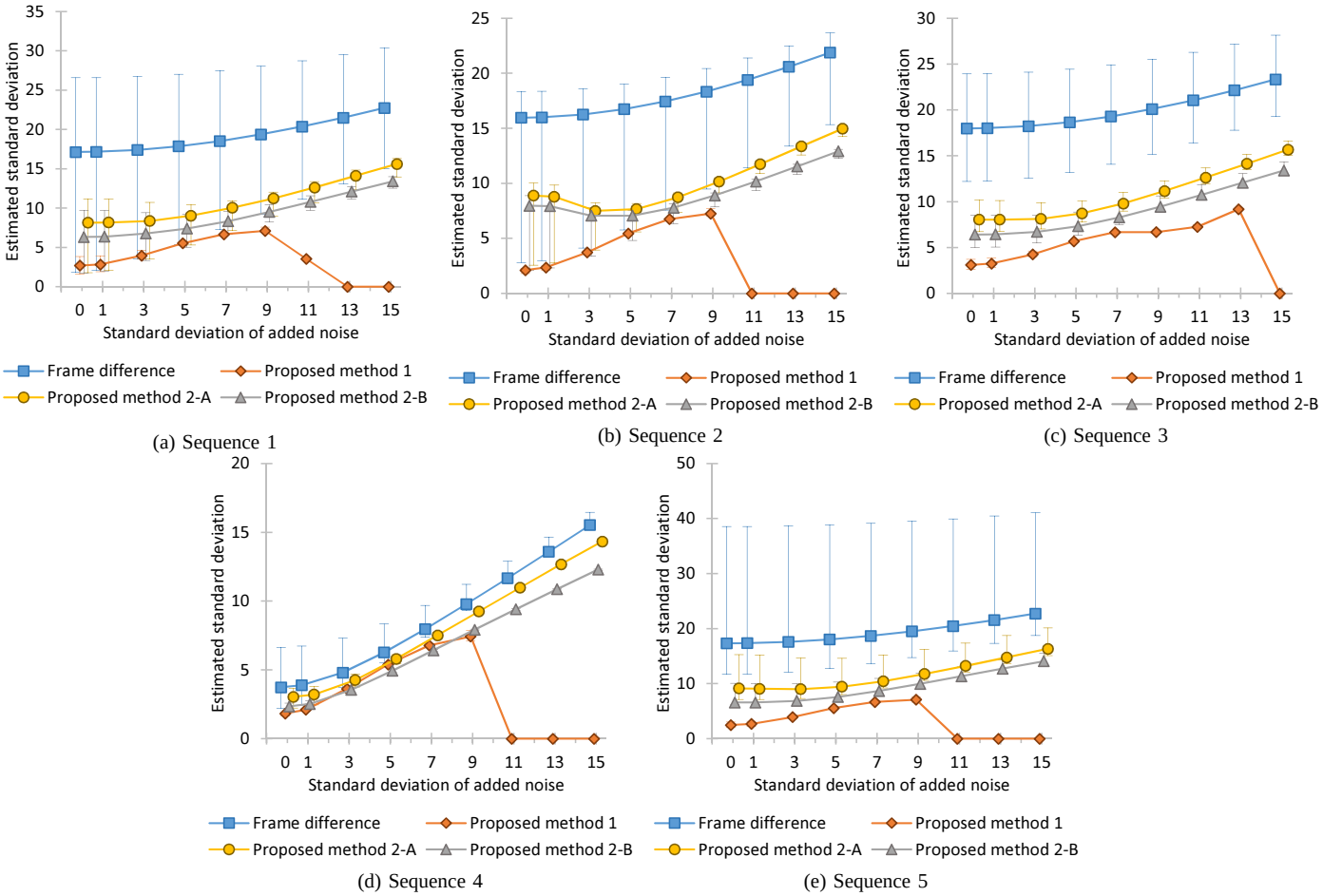
In Figures 11 (a)-(e), the results for the frame difference method are overestimated and demonstrate large variance. The estimated results for the proposed method 1 are the most accurate and have the smallest dispersion of results. However, the estimation is not possible with the noise standard deviation exceeding 9 because there are few or no appropriate areas for calculating noise standard deviation. The proposed methods 2-A and 2-B returned fewer errors and demonstrate more consistent estimated results than the frame difference. However, large errors tend to occur when the noise standard deviation is less than 3. A comparison of the results for the proposed methods 2-A and 2-B, with and without motion compensation, demonstrates that motion compensation is effective in certain cases. However, it increases the cost significantly because a real-time motion compensation requires large hardware.

Figure 12 shows the estimated noise standard deviation for all frames of sequence 1 (Figure 11 (a)). Figures 12 (a) and (b) show the estimation results for the proposed methods 2-A and 2-B when the noise level is larger than 9. Here the x-axis is the frame number, and the y-axis is the estimated standard deviation of the noise in the frame. The results become constant if the noise level estimation is correct.

In sequence 1, the train is moving with camera panning from 0 to 150 frames, then the panning stops. The train continues to move during frames 150 to 420. There is no motion in frames 420-450. As shown in Figures 12 (a) and (b), the effect of motion on the estimation result is negligible, and the results become constant.

Comparisons of the areas for noise estimation using the proposed method 1, and the proposed method 2-A are shown in Figure 13. The estimated noise areas for sequence 1 with added Gaussian noise are shown in Figures 13 (a)-(b) (standard deviation 3) and Figures 13 (c)-(d) (standard deviation 7). Here, the white areas are estimated moving areas; thus, only

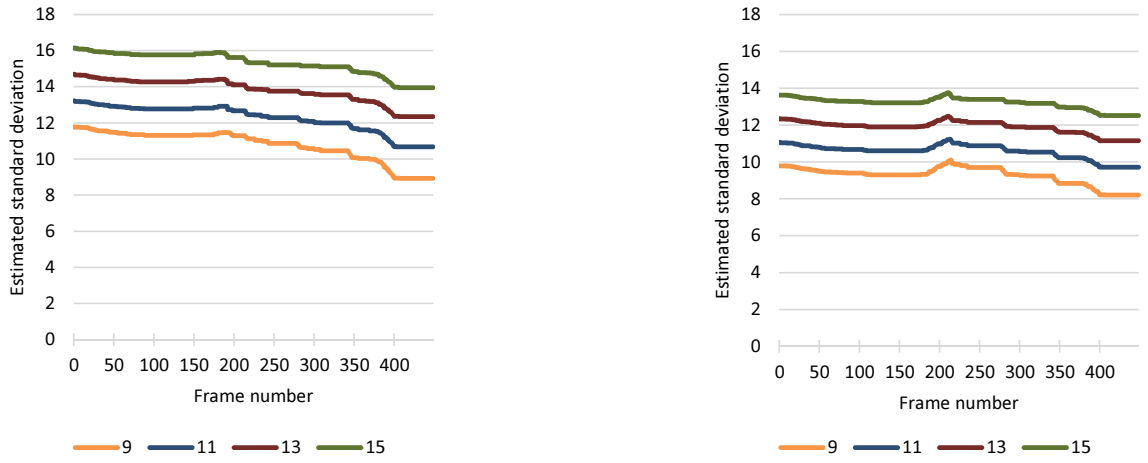
 ISBN: 978-1-61208-707-8



(a) Sequence 1 (b) Sequence 2 (c) Sequence 3

(d) Sequence 4 (e) Sequence 5

Figure 11. Results of estimated noise standard deviation. (a) - (e) show the results for sequences 1-5, respectively. The estimated results of all frames of the video sequence are accumulated. The marks show the median values of the estimated standard deviations. The bars indicate the maximum and minimum values.

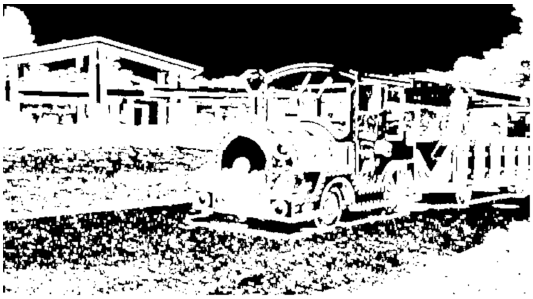


(a) Proposed method 2-A (b) Proposed method 2-B

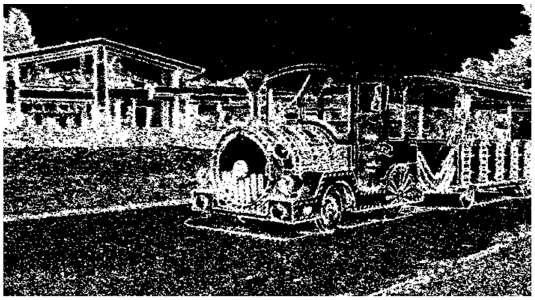
Figure 12. Results of estimated noise standard deviation in time axis for sequence 1 using the proposed methods (a) 2-A and (b) 2-B

the black areas are used for noise estimation. When comparing Figure 13 (a) with Figure 13 (b), and Figure 13 (c) with Figure 13 (d), the moving areas estimated using the proposed method 1 are thick; however, there are few areas for noise estimation when the noise level is high. Since the proposed method 1 fully eliminates moving areas, the noise level estimation becomes accurate. However, the noise estimation does not work with high level noise due to few or no available estimation areas. In contrast, the estimated moving areas using the proposed method 2-A are thin; therefore, the moving areas of the frame with high level noise are detectable.

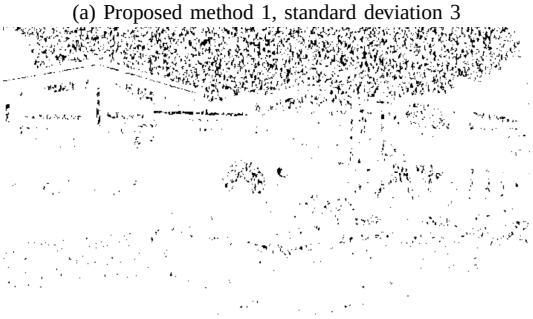
 ISBN: 978-1-61208-707-8



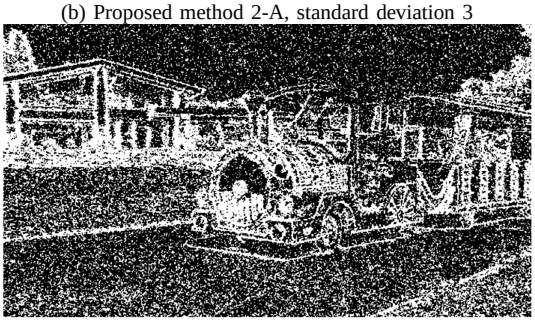
(a) Proposed method 1, standard deviation 3



(b) Proposed method 2-A, standard deviation 3



(c) Proposed method 1, standard deviation 7



(d) Proposed method 2-A, standard deviation 7

Figure 13. Areas of calculated noise standard deviations for one frame of sequence 1

V. DISCUSSION

As described in Section IV-B, the proposed method 1 can detect low level noise accurately when the standard deviation is less than 9. However, this method requires improvement to detect high level noise when the standard deviation is higher than 9. In contrast the proposed methods 2-A and 2-B can detect high level noise when the standard deviation is 5 or more. Therefore, we propose combining the proposed methods 1 and 2-A, i.e., when the detected noise level is less than 9, the proposed method 1 is appropriate and when the detected noise level is equal to or higher than 9, the proposed method 2-A is appropriate. Moving compensation can improve noise detection accurately; however, it requires significantly more expensive hardware.

VI. CONCLUSION

In this paper, real-time noise level detection algorithms for videos were proposed. The simulation results demonstrate that the best results can be realized by combining two methods. In future, we intend to develop a way to switch between methods automatically and to control NR using the proposed methods. Ultimately, we hope to develop real-time noise reduction hardware that controls noise level parameters automatically.

REFERENCES

[1] M. Kazubek, "Wavelet domain image denoising by thresholding and wiener filtering," *IEEE Signal Processing Letters*, vol. 10, no. 11, pp. 324–326, 2003.

[2] A. Pizurica, V. Zlokolica, and W. Philips, "Noise reduction in video sequences using wavelet-domain and temporal filtering," in *Wavelet Applications in Industrial Processing*, vol. 5266. International Society for Optics and Photonics, 2004, pp. 48–60.

[3] N.-X. Lian, V. Zagorodnov, and Y.-P. Tan, "Video denoising using vector estimation of wavelet coefficients," in *2006 IEEE International Symposium on Circuits and Systems*. IEEE, 2006, pp. 2673–2676.

[4] I. W. Selesnick and K. Y. Li, "Video denoising using 2d and 3d dual-tree complex wavelet transforms," in *Wavelets: Applications in Signal and Image Processing X*, vol. 5207. International Society for Optics and Photonics, 2003, pp. 607–619.

[5] A. Pizurica, V. Zlokolica, and W. Philips, "Combined wavelet domain and temporal video denoising," in *Proceedings of the IEEE Conference on Advanced Video and Signal Based Surveillance, 2003.* IEEE, 2003, pp. 334–341.

[6] N. Gupta, M. Swamy, and E. I. Plotkin, "Low-complexity video noise reduction in wavelet domain," in *IEEE 6th Workshop on Multimedia Signal Processing, 2004.* IEEE, 2004, pp. 239–242.

[7] R. O. Mahmoud, M. T. Faheem, and A. Sarhan, "Comparison between discrete wavelet transform and dual-tree complex wavelet transform in video sequences using wavelet-domain," *INFOS2008*, pp. 20–27, 2008.

[8] L. Jovanov, A. Pizurica, S. Schulte, P. Schelkens, A. Munteanu, E. Kerre, and W. Philips, "Combined wavelet-domain and motion-compensated video denoising based on video codec motion estimation methods," *IEEE transactions on circuits and systems for video technology*, vol. 19, no. 3, pp. 417–421, 2009.

[9] F. Luisier, T. Blu, and M. Unser, "Sure-let for orthonormal wavelet-domain video denoising," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 20, no. 6, pp. 913–919, 2010.

[10] J. C. Brailean, R. P. Kleihorst, S. Efstratiadis, A. K. Katsaggelos, and R. L. Lagendijk, "Noise reduction filters for dynamic image sequences: A review," *Proceedings of the IEEE*, vol. 83, no. 9, pp. 1272–1292, 1995.

[11] V. Zlokolica, A. Pizurica, and W. Philips, "Noise estimation for video processing based on spatio-temporal gradients," *IEEE Signal Processing Letters*, vol. 13, no. 6, pp. 337–340, 2006.

[12] V. Kamble and K. Bhurchandi, "Noise estimation and quality assessment of gaussian noise corrupted images," *IOP Conference Series: Materials Science and Engineering*, vol. 331, no. 1, pp. 1–10, 2018.

[13] M. Ghazal, A. Amer, and A. Ghrayeb, "A real-time technique for spatio–temporal video noise estimation," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 17, no. 12, pp. 1690–1699, 2007.

[14] S.-M. Yang and S.-C. Tai, "A fast and reliable algorithm for video noise estimation based on spatio-temporal sobel gradients," in *International Conference on Electrical, Control and Computer Engineering 2011 (InECCE)*. IEEE, 2011, pp. 191–195.

[15] K. Miyamae and S. Gohshi, "Noise level detection in general video," in *2018 International Workshop on Advanced Image Technology (IWAIT)*, Jan 2018, pp. 1–4.

[16] "Ite/arib hi-vision test sequence 2nd edition," 2009.

[17] M. J. B. Wilhelm Burger, *Principles of Digital Image Processing: Fundamental Techniques*. Springer, 2009, pp. 144–145.

 ISBN: 978-1-61208-707-8

Emotion-Based Color Transfer of Images Using Adjustable Color Combinations

Yuan-Yuan Su
Department of Computer Science
National Tsing Hua University
Hsinchu, Taiwan
email:steven.nthu@gmail.com

Hung-Min Sun
Department of Computer Science
National Tsing Hua University
Hsinchu, Taiwan
email:hmsun@cs.nthu.edu.tw

***Abstract*—This study developed a novel framework for the color-transfer between colored images, and that can further achieve emotion-transfer between colored images based on the human emotion (human feeling) and a predefined color-emotion model. In this study, a new methodology is proposed. The main colors in an image can be adjusted based on the complexity of the content of images. The others contributions in the study are the algorithms of the Touch Four Sides (TFS) and the Touch Up Sides (TUS), which can improve the identification of the background and foreground and the other main colors that are extracted from the images.**

Keywords-Emotion-Transfer; Color-Transfer; Color-Emotion; Image Content Analysis.

I. INTRODUCTION

Images are the important media for conveying human emotions. Colors are also the main component of an image. In recent years, researchers have intensively studied the usage of images to convey emotions, opinions [1]-[5], and be used for one method to take single main color or a fixed number of main colors combinations to implement color-transfer. Nevertheless, an unsolved problem is how to understand and describe the emotions caused by an image. Another problem is how to understand the inherent subjectivity of emotional responses by the user. This study develops a novel framework for the color-transfer focusing on color images and emotion-transfer implementation between color images based on human emotions and a predefined color-emotion model.

II. METHODS

The new emotion transfer method uses one scheme of dynamic and adjustable color combinations which is based on the complex content of a color image to determine which number of color combinations is used. Additionally, the proposed method can accurately identify the primary representative colors of the image, and also support both solutions, *i.e.*, using relative images and semantics which come from a predefined color-emotion model for emotion transfer. The method follows five steps as below.

1) *Color Emotion Model*
2) *Dynamic Extraction of the Main Colors*
   a) *Identification of the Main colors*
   b) *Determination of the amount of main colors*
3) *Identification of the Background and the Foreground*
4) *Emotion Transfer*
   a) *Matching for the Amount of Color Combinations*
   b) *Color Transfer*
   c) *Pixel Updates*
   d) *Gradient Preservation*
5) *Output Image Producing*

The flow chart of the proposed emotion-transfer framework is shown below.

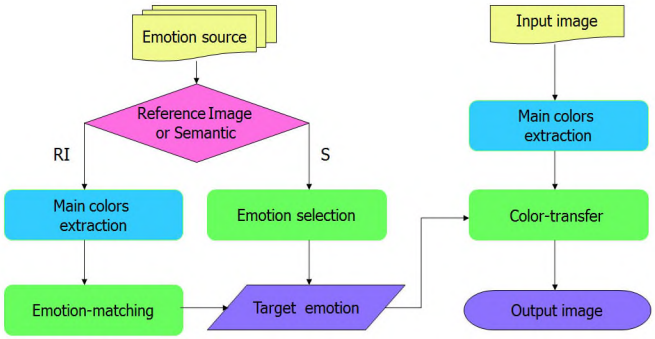


Figure 1. The flow chart of the proposed emotion-transfer framework

The target emotion from source can be separated in two ways. The first approach adopted the Reference Image (RI) to get a target emotion, which is acquired by two steps:

1) *Main colors extraction:* extract the main colors for the input image and reference image. These main colors can be categorized as three kinds of color combination: two-color, three-color, and five-color. Moreover, our method also supports other color-emotion model that may provide more than five-color combinations. Also, we combine two methods to identify the representative points from image-pixels, which are Independent Scalar Quantization (ISQ) [6] and Linde-Buzo-Gray (LBG) algorithm [7], which is similar to a K-means clustering algorithm. In order to determine the amount of main colors, the proposed method is then used to determine the background and other main colors.

2) *Emotion-matching:* analyze the input image to compare with other reference images based on the number of main colors. Here, there are two cases. If the number of main colors is the same, the main colors of reference images will be adopted as the color combinations of target emotion

 ISBN: 978-1-61208-707-8

directly. One example of this way is shown in Figure 5. When the number of main colors are different, the target emotion will be assigned with same emotion and search the closest color combination from a predefined color-emotion model. Another way is to allow users to choose a desired emotion directly from a predefined color-emotion model which is the model developed by Chou [8] and contains 24 emotions, each one includes 24 two-color combinations, 48 three-color combinations and 32 five-color combinations. All of the colors in the Chou model are mapped to the RGB color space and the CMYK color space. However, this study uses the CIELab [9] color space, which is converted from the RGB color space. In this study, the number of color combinations is chosen according to the complexity of the contents of the input image or reference image. The main colors extracted from the input image are mapped to the target emotion with the same amount of color combinations. Therefore, the number of main colors must be controlled within this range. After the color-transfer algorithm is used to transfer the target emotion to the input image, the output image is obtained, and the procedures of the color- emotion transfer is completed.

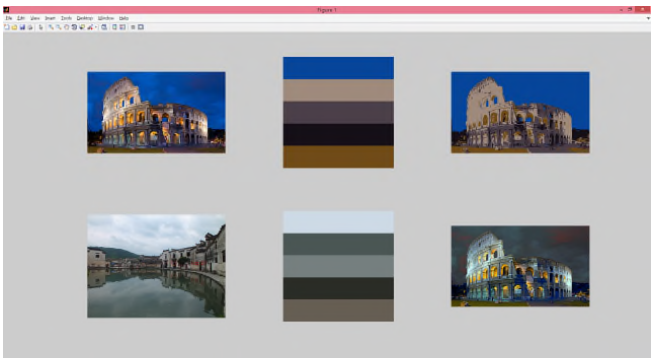


Fig. 5. When the input image and the reference image have the same number of main colors, the target emotion can adopt directly the main color from the reference image

III. Conclusion And Future Work

This study proposes a novel method and framework for performing emotion-transfer based on adjustable and dynamic color combinations for color images. The results for adjustable and dynamic color combinations have several advantages. They illustrate the emotions in images and enables color transfer to be performed separately in different regions of the images. The results show that the method improves the expression of color and emotion in images.

Non-professionals can also use the proposed method to describe objectively and efficiently for the communication of human emotions. The experimental results show that the proposed approach can naturally alter the emotions between photo and painting. It also allows emotionally rich images for art and design. Since previous color transfer algorithms use only one color or a fixed color combination to obtain new images, the images are usually not rich or natural in terms of colors and emotions, and the clustering result sometimes does not recognize the dominant main color correctly.

The proposed approach uses the methods of ISQ + LBG, which is more effective than LBG alone, and the quality of images are similar. Other new algorithms in the proposed approach, the Touch Four Slides (TFS) and Touch Up Slides (TUS) algorithms, can extract the background and dominant color correctly from the main colors for most images. Furthermore, the representative colors are obtained from the method of the adjustable and dynamic color combinations, which provides a closer link between human and emotion.

In the future, different color-emotion models could be used in this framework. For example, since the main colors extracted from color images are the most important representational colors, they can be used to identify specific images. In addition, it can also be used for the feature of the emotion. However, although the method always extracts the background color accurately, the dominant color may not be absolutely and correctly identified for all images. Therefore, the method may be still improved by additional elements to enhance the extraction of emotion from images, such as shape, texture, and so on.

Acknowledgment

This research was partially supported by the Ministry of Science and Technology of the Republic of China under the Grants MOST 107-2218-E-007-046 and MOST-106-2221-E-007 -026 -MY3.

References

[1] C.-K. Yang and L.-K Peng, “Automatic mood-transferring between color images,” IEEE Computer Graphics and Applications 28, 52-61,March 2008.

[2] W. Fuzhang, D. Weiming, K. Yan, M. Xing, C. P. Jean and Z. Xiaopeng, “Content-Based Color Transfer,” Computer Graphics Forum Volume 32, Issue1, pages 190-203, February 2013.

[3] T Pouli and E Reinhard, “Progressive histogram reshaping for creative color transfer and tone reproduction,” Computers & Graphics 35 (1), 67-80, 2011. 40, 2011.

[4] Y. Chang, S. Saito, K. Uchikawa and M. Nakajima, “Example based color stylization of images,” ACM Transactions on Applied Perception 2, 3 , 322-345, 2005.

[5] H. Li, Q. Hairong and Z. Russell, “Image color transfer to evoke different emotions based on color combinations”, Springer, SIViP, 9:1965-1973, 2015.

[6] J. D. Foley, A. V. Dam, S. k. Feiner, and J. F. Hughes, ” Comput. Graphics: Principles and Practice,” Ma Addison Wesley, 1990.

[7] Y. Linde, A. Buzo, and R. M. Gray, “An Algorithm for Vector Quantizer Design, ” IEEE Transactions on Communications, vol. 28, pp. 84-95, January 1980.

[8] C. Chien-Kuo, “Color Scheme Bible Compact Edition,” Grandtech information Co., Ltd, 2011.

[9] CIELAB, “CIELab - Color Models - Technical Guides,” internet:ba.med.sc.edu/price/irf/Adobe_tg/models/cielab.html, accessed Feb 2019.

 ISBN: 978-1-61208-707-8

Allowing Privacy-Preserving Smart City Open Data Linkage

Francesco Buccafurri
DIIES Dept.
University "Mediterranea" of Reggio Calabria
Reggio Calabria, Italy
Email: bucca@unirc.it

Celeste Romolo
DIIES Dept.
University "Mediterranea" of Reggio Calabria
Reggio Calabria, Italy
Email: celeste.romolo@gmailcom

***Abstract*—Open data are a crucial component of the smart-city ecosystem. In this scenario, many subsystems exist, like transport, shopping, cinema, theatre, utilities consumption, etc. Data coming from the interaction of citizens with these subsystems can be anonymized and published as open data, according to normative requirements and best practices. Therefore, any third-party (even government) entity can perform isolated data analytics, but it is not able to relate open data referring to the same citizen thus missing a lot of potential powerful information. In this position paper, we present a cryptographic approach aimed at allowing cross-correlation of smart-city open data only to authorized parties yet preserving citizens' privacy. The solution leverages the public digital identity system compliant with eIDAS (the European framework) by giving to the Identity Providers the role of Trusted Third Party.**

Keywords–Open Data; Privacy; Linked Data.

I. INTRODUCTION

The concept of *Smart City* is wide and involves in a truly integrated fashion all the components of a community like transport, shopping, cinemas, theaters, utilities consumption, etc. Among the other aspects, the capability of managing and exploiting the information flow underlying the working of the various components of a city is fundamental to make the community really *smart*. According to this paradigm, a very important task is to publish in an interoperable form data coming from the interaction between citizens and the various components of the city. Indeed, any third party can develop applications and perform powerful analysis by exploiting these data. This is basically the principle underlying the concept of *open data*, which both normative enforcement [1] and common best practices require to adopt in a smart community. Evidently, for privacy reasons, data can be published in an anonymous form, hopefully also by satisfying robust privacy protections, like *k-anonymity* [2] or *l-diversity* [3]. However, this as a negative side effect. Indeed, no correlation between data belonging to different subsystems can be done, thus missing a lot of potential powerful knowledge [3].

In this paper, we propose a new strategy, based on a multi-party cryptographic protocol, which allows us to keep anonymity of citizens, but enables data linkage for authorized parties, to recover the knowledge gap and increase the benefit of open data. The solution is practical, because it identifies how to map into real-life entities the different roles of the model, also by considering a public digital identity system compliant with eIdas (the European framework) [4] and by giving to the Identity Providers the role of Trusted Third Party.

The structure of the paper is the following. The next section contextualizes the proposal in the related literature. In Section III, background notions are provided. In Section IV, the problem is formulated and the proposed model is described. The detailed description of the solution is given in Section V. Finally, we draw our conclusions in Section VI.

II. RELATED WORK

A wide scientific literature exists highlighting the importance of open data in the context of Smart Cities. In [5] the role of big data and open data in smart cities is well explained and analyzed. In [6], the correlation between big data, smart cities and city planning is studied. The work [7], discusses how Mobile Application Clusters can be developed through competitions for innovative applications. The Smart City services that are developed in competitions benefit both the Mobile Application Cluster and the citizens. The function of the competition mechanism to encourage the development of new mobile applications utilizing Open Data is described with examples from the Helsinki Region. The authors of [8] sketch the rudiments of what constitutes a smart city, which we define as a city in which ICT is merged with traditional infrastructures, coordinated and integrated using new digital technologies. They highlight how to build models and methods for using urban data across spatial and temporal scales, and to apply them to many subsystems like transport and energy.

A considerable attention has been devoted to the problem of privacy in the context of Smart Cities mainly regarding the protection of information stored and managed by the City entities. In [9], the authors leverage some concepts of previously defined privacy models and define the concept of citizens privacy as a model with five dimensions: identity privacy, query privacy, location privacy, footprint privacy and owner privacy. By means of several examples of smart city services, we define each privacy dimension and show how existing privacy enhancing technologies could be used to preserve citizens privacy. The work [10] deals with problem of data over-collection. This problem arises from the fact that smartphones apps collect users' data more than its original function while within the permission scope. For the authors, this is rapidly becoming one of the most serious potential security hazards in smart city. In the above paper, the authors study the current state of data over-collection and study some most frequent data over-collected cases. The problem of security and privacy is deeply investigated in [11]. One of the main points of this paper is the observation that privacy can be achieved (i) by imposing high security requirements onto the used technology to avoid third party abuses; and (ii) by decoupling technical smart city data streams from the personal one to avoid abuse of data by insiders.

 ISBN: 978-1-61208-707-8

The problem of open data linkage has been considered in a number of papers in the past, especially in the field of health. The paper [12] is an evolution of the W3C SWEO community project, with the purpose of linking Open Data coming from various open datasets available on the Web as RDF, and to develop automated mechanisms to interlink them with RDF statements. In [13], the authors argue that Linked Data technology, created for Web scale information integration, can accommodate XBRL data and make it easier to combine it with open datasets. This can provide the foundations for a global data ecosystem of interlinked and interoperable financial and business information with the potential to leverage XBRL beyond its current regulatory and disclosure role.

Although privacy and linkability have been recognized, as shown above, as important problems in the field of Smart Cities, to the best of our knowledge, there is no paper trying to reach a compromise between the two features, which is, instead the contribution of this paper.

III. Background

Our solution leverages any public digital identity system compliant with the European framework eIdas [4]. Among these, we choose the Italian system SPID [14] to describe a concrete implementation of the general framework. SPID is based on the language Security Assertion Markup Language (SAML) [15], which is an XML-based, open-standard data format designed to exchange authentication and authorization messages between identity and service providers. It uses assertions (signed XML messages) to transfer information in such a way that federated authentication and authorization systems can be implemented. SAML messages included into the HTTP GET request, while longer messages exploits the mechanism of HTTP POST Binding. For security reasons, in SPID, HTTP must be used only on combination with TLS.

The SPID framework includes the following components:

1) **Users**. They are people using the system to authenticate for a service delivered by a Service Provider (see below). Besides an ID and all personal identifying information (such as social security number, name, surname, place of birth, date of birth and gender), other attributes can be associated with the users (like for example a professional status).
2) **Identity Providers**. They identify people in the registration phase (either frontally or remotely), create and manage IDs, and grant the assertion to the Service Providers to authenticate the users at the required level of assurance. The strength of the authentication of the user at the Identity Provider depends on the requested level of assurance. Identity Providers are private or public subjects certified by a Trusted Third Party.
3) **Service Providers**. They are public or private organizations adhering to the SPID system providing a service to authorized users and requiring a given level of assurance.
4) a **Trusted Third Party (TTP)**. It is a government entity (Agency for Digital Italy – AGID), which guarantees the standard levels of security required by the regulation and certifies the involved entities.
5) **Attribute Providers**. They are optional entities whose role is to certify attributes, such as possession of a degree, membership of a professional body, etc.

IV. The Smart-City Open Data Model and Problem Formulation

In this section, we introduce the Smart-City Open-data model which the solution proposed in this paper is applying to. The Smart City is composed of a number of subsystems, denoted as $\{S_1, S_2, \ldots S_n\}$. They are for example transport system, health facilities, schools, universities, shops, cinemas, theaters, utilities consumption, etc. Assume that each subsystem x has a pair $\langle I_x^i, D_x^i \rangle$ where I_x^i is the real identity of an individual i as known to the subsystem x and D_x^i is the set of data that every day (week, month, etc.) the subsystem x collects about that individual. Assume that each subsystem publishes as *open data* the pair $\langle P_x^i, \bar{D}_x^i \rangle$, where P_x^i, is a *pseudonym* of the real identity (as known to x) and $\bar{D}_x^i$ is a suitable transformation of the original data. Let us assume that:

1) $P_x^i = \alpha(I_x^i)$ where α is an anonymization function with the purpose of disguising the actual identity.
2) $\bar{D}_x^i = \delta(D_x^i)$ a transformation of the original data with the double purpose: (i) to hide useless details, and (ii) to make it difficult data de-anonymization (therefore, the function δ takes into account all threats contrasted by state-of-the-art approaches like *k-anonymity* [2] or *l-diversity* [3].

The current situation in real-life systems, and, to the best of our knowledge, in the scientific literature, is the following. The subsystems are independent each other and they use different functions α and δ. Therefore, there is no way to understand that the various data refer to the same individual, so one data are unlinkable. Observe that this, according to this model, this is an expected feature, because it is fundamental to really protect citizens' privacy. Indeed, each subsystem knows the real identity of the individuals, so linking its data with that of other subsystems may result in a potentially very dangerous information leakage. However, the side effect is that it is impossible, for any party, to reconstruct, even in anonymous form, the behavior of a single individual.

The aim of this paper, thus the problem faced by this work, is to recover the above gap of knowledge, without compromising citizens' privacy. The proposed solution is presented in the next section.

V. The Proposed Solution

In this section, we present a possible solution of the problem formulated in the previous section. The solution is both theoretical and practical, because is aware about how to map the entities of the model to concrete parties already playing a role in digital communities. The solution is tailored to a Country belonging the the European Union. Obviously, a more general case could be considered just by identifying different normative and infrastructural components.

We assume subsystems belong to the same Member State. Suppose this State adopts a Public Digital Identity System compliant with eIDAS regulation [4]. For example, in Italy

 ISBN: 978-1-61208-707-8

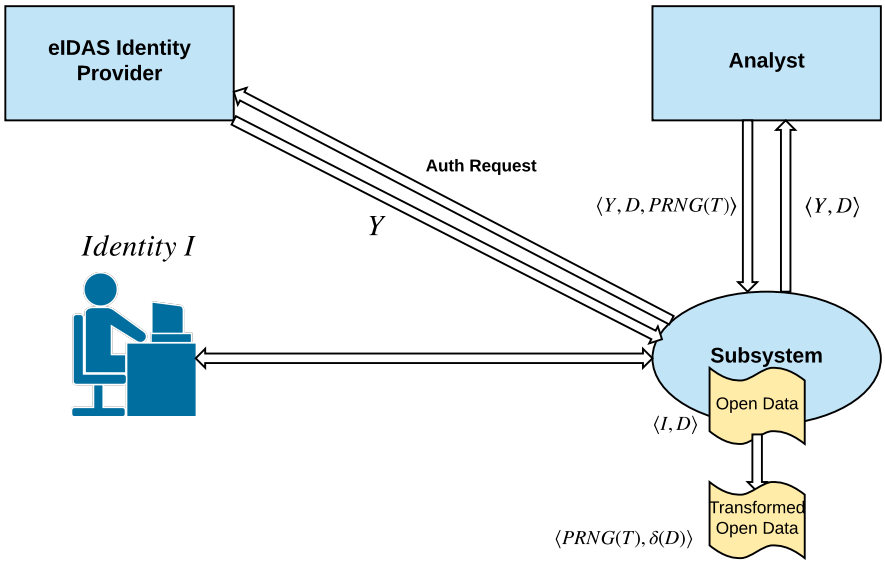


Figure 1. A simplified description of the protocol.

such a system exists and is named SPID (SAML-based authentication system) [15]. So, we assume that the users involved in the system own an eIDAS identity, so each is registered at an Identity Provider (for free). This is a realistic assumption because it is a scenario to which EU aims, according to the eIDAS regulations and others related Acts.

The actors of the systems are the following:

1) Users (i.e., citizens)
2) eIDAS (accredited) identity providers
3) subsystems, playing the role of (special) eIDAS service providers
4) Analysts $\{A_1, \ldots, A_t\}$, a dynamic group of parties empowered to data analysis.

Let U be a user whose identity is managed by the Identity Provider IP.

Let S be the subsystem that U is accessing. This transaction will result in a new (set of) open data. We focus on a single open data $\langle I, D\rangle$. Our proposal aims to modify the function α (see the previous section) in such a way that only for an Analysts Party it is possible to link open data published by different subsystems and referring to the same user. No change is required for the function δ.

The function is modified according to the following mechanism. According to the eIDAS system, the authentication request to a service provider S given by U is forwarded by S to IP. In our case, the request should be modified to enable the open-data mechanism (so, some modifications of the format of SAML messages of the eIDAS system are required). We could also require that the service provider (i.e., the subsystem) that wants to generate such open data must be previously registered and adhere to some common procedural rules.

When IP receives the authentication request, it activates the standard SAML mechanism, but the returned assertion (granting the authentication) will include also (here a modification of the SAML message is required):

1) An order number N, denoting the number of open-data authentications required so far by the user U
2) A value $Y = MAC(IDU, SIP)$, where MAC is a secure message authentication code (like for example HMAC [16]), IDU is the eIDAS identification number (and it can be considered as the identity value i above), and SIP is a secret owned by IP (this is done to avoid that S can invert Y and can find the identification number, which is not public). Moreover, as Y is the output of an hash function, no collision can be found, so Y is uniquely identifying the user.

The subsystem S, once the assertion is received, proceeds as follows:

1) Chooses an Analyst A_x (even at random);
2) Sends the triple $\langle Y, N, Id\rangle$ to A_x, where Id identifies the open data.

At this point, the Analyst A_x, computes:

1) $T = MAC(Y, X)$, where X is a secret shared with all the analysts (this can be obtained by using a dynamic group key agreement protocol – there are a number of efficient extensions of Diffie Hellman to do this [17])
2) T is the seed of a LEcuyer's PRNG [18]. So, A_x, computes $PRNGN(T)$.
3) A_x sends to S the message: $\langle Y, Id, PRNG(T)\rangle$. This, in other words, means that the new function α is defined as $\alpha(IDU) = PRNG(T)$.

At this point, the subsystem S matches the message $\langle Y, Id, PRNGN(T)\rangle$ to the corresponding open-data D and publishes it in the form $\langle PRNG(T), \delta(D)\rangle$.

Observe that when the same user U accesses another subsystem, say G, the protocol will associate the new open data with the pseudonym $PRNG(PRNG(T))$, so that the two open data are both anonymous but linkable. But they are linkable only for those that know the seed T (there is a seed for each user). So, full analytics can be performed only by any Analyst. No other party can do this.

Concerning the adoption of the public digital identity system, as observed earlier, it seems a realistic hypothesis, as the idea underlying this framework in EU member states is to use it as Single-Sign-On system for all the interactions between citizens and the public sector (eventually, with a unique interoperable system over the entire Europe). This is a practical solution because, thanks to the public identity, there is no need of a specific Registration Authority (Identity Providers play this role within their functions).

The proposed solution is summarized in Fig. 1, in which some messages are simplified for the sake of presentation.

VI. Conclusions

Open Data are a fundamental component of the Smart-City ecosystem. They allow transparency, e-participation, but also the development of application able to integrate different subsystems of the community, thus fulfilling the Smart-City paradigm. Typically, for privacy reasons the are anonymized in such a way that they are also unlinkable. However, a lot of potential powerful knowledge may derive from the correlation between data of the same user belonging to different subsystems. In this paper, we proposed a solution based on a multi-party cryptographic protocol also relying on the public digital identity system which appears as good compromise between privacy requirements and information power of data. This is a still work-in-progress paper. Therefore, a number of aspects need to be analyzed in more detail. Among these, the problem of de-anonymization of data, when linkable, assumes a different form than the case of unlinkable data. This is what we plan to do in the near future about this work, together with a *proof-of-concept* implementation of the system.

References

[1] "EU (2003): Directive 2003/4/EC of the European Parlia ment and of the Council of 28 January 2003 on Public Access to Environmental In formation." *In: Official Journal of the European Union, L 41/26,*, 2003.

[2] P. Samarati and L. Sweeney, "Protecting privacy when disclosing information: k-anonymity and its enforcement through generalization and suppression," Technical report, SRI International, Tech. Rep., 1998.

[3] A. Machanavajjhala, J. Gehrke, D. Kifer, and M. Venkitasubramaniam, "l-diversity: Privacy beyond k-anonymity," in *Data Engineering, 2006. ICDE'06. Proceedings of the 22nd International Conference on*. IEEE, 2006, pp. 24–24.

[4] C. Cuijpers and J. Schroers, "eIDAS as guideline for the development of a pan European eID framework in FutureID," *Open Identity Summit 2014*, vol. 237, pp. 23–38, 2014.

[5] I. A. T. Hashem, V. Chang, N. B. Anuar, K. Adewole, I. Yaqoob, A. Gani, E. Ahmed, and H. Chiroma, "The role of big data in smart city," *International Journal of Information Management*, vol. 36, no. 5, pp. 748–758, 2016.

[6] M. Batty, "Big data, smart cities and city planning," *Dialogues in Human Geography*, vol. 3, no. 3, pp. 274–279, 2013.

[7] H. Hielkema and P. Hongisto, "Developing the helsinki smart city: The role of competitions for open data applications," *Journal of the Knowledge Economy*, vol. 4, no. 2, pp. 190–204, 2013.

[8] M. Batty, K. W. Axhausen, F. Giannotti, A. Pozdnoukhov, A. Bazzani, M. Wachowicz, G. Ouzounis, and Y. Portugali, "Smart cities of the future," *The European Physical Journal Special Topics*, vol. 214, no. 1, pp. 481–518, 2012.

[9] A. Martinez-Balleste, P. A. Pérez-Martínez, and A. Solanas, "The pursuit of citizens' privacy: a privacy-aware smart city is possible," *IEEE Communications Magazine*, vol. 51, no. 6, pp. 136–141, 2013.

[10] Y. Li, W. Dai, Z. Ming, and M. Qiu, "Privacy protection for preventing data over-collection in smart city," *IEEE Transactions on Computers*, vol. 65, no. 5, pp. 1339–1350, 2016.

[11] A. Bartoli, J. Hernández-Serrano, M. Soriano, M. Dohler, A. Kountouris, and D. Barthel, "Security and privacy in your smart city," in *Proceedings of the Barcelona smart cities congress*, vol. 292, 2011.

[12] C. Bizer, T. Heath, D. Ayers, and Y. Raimond, "Interlinking open data on the web," in *Demonstrations track, 4th european semantic web conference, innsbruck, austria*, 2007.

[13] S. O'Riain, E. Curry, and A. Harth, "Xbrl and open data for global financial ecosystems: A linked data approach," *International Journal of Accounting Information Systems*, vol. 13, no. 2, pp. 141–162, 2012.

[14] "SPID-Agenzia per l'Italia Digitale," `http://www.agid.gov.it/sites/default/files/regole_tecniche/spid_regole_tecniche_v0_1.pdf`, 2015.

[15] "Security Assertion Markup Language (SAML)," `http://it.wikipedia.org/wiki/Security_Assertion_Markup_Language`, 2015.

[16] H. Krawczyk, R. Canetti, and M. Bellare, "Hmac: Keyed-hashing for message authentication," 1997.

[17] M. Steiner, G. Tsudik, and M. Waidner, "Key agreement in dynamic peer groups," *IEEE Transactions on Parallel and Distributed Systems*, vol. 11, no. 8, pp. 769–780, 2000.

[18] P. L'Ecuyer, "Uniform random number generation," *Annals of Operations Research*, vol. 53, no. 1, pp. 77–120, 1994.

 ISBN: 978-1-61208-707-8

Semantic Network Analysis for VR & Coding Education in Big Data

Su Jeong, Jeong
Creativity & Personality Laboratory
Tongmyong University
Busan, Korea
crystal06070@naver.com

Jeong Jin, Youn
Dept. Early Childhood Education
Tongmyong University
Busan, Korea
jjy@tu.ac.kr

***Abstract* —The purpose of this study is to investigate the informal Big Data to understand the discourse among the members of VR & Coding education. Therefore, we will understand the social meaning of VR & Coding education and discuss VR & Coding education required in society. For this research, we looked at Big Data about 'VR & Coding education' which has been collected through online channel for the last 5 years. We collected 4392 data sets from the online channel, and after refining, we constructed a semantic network with 200 important keywords. When we searched for 'VR + Coding Education', the most frequently appeared keywords were 'Experience', 'Virtual Reality', 'Fourth Industrial Revolution', 'Robot', 'Program'. Degree centrality values were 'experience', 'virtual reality', 'robot', 'fourth industrial revolution', 'infant'. Based on these results, we analyze the implications of VR & Coding education to society.**

Keywords-Semantic Network Analysis; VR & Coding Education; Big Data.

I. INTRODUCTION

We live in the age of the fourth industrial revolution. It is clear that the future will change rapidly into the era of artificial intelligence. Currently, 90% of the knowledge learned in elementary, middle, and high schools is said to be obsolete in ten years [1]. Like this, changes in the education that suits the changes of the 4th Industrial Revolution era are necessary. In particular, current society is transforming into a software-oriented society. Companies with software skills are entering a whole new market, dominating the market with new products and services, and introducing new software technologies into existing company systems to increase efficiency.

So how do we go about social change? We need preparation. Prepared individuals will survive in the future, but unprepared individuals can be predicted to decline or become more polarized. Therefore, in education, SW integration ability and creative problem solving ability should be more emphasized. To this end, it is necessary to introduce Software (SW) education for strengthening the capacity of new talent in the digital era in the educational field.

Virtual Reality (VR), which is a core area of the 4th industry, is considered as one of the technologies leading the fourth industrial revolution. Just as VR technology is applied in various industries such as games, movies, and tourism media, VR can be actively introduced in education. In addition, since coded education improves logical thinking ability and problem solving ability, it can nurture self-empowerment so that the child can actively solve the problem.

Since 2019, coding education has become mandatory in the 5th and 6th grades of elementary schools in Korea, and the importance of coding education is emerging. Therefore, this study aims to examine the social discourse on VR and coding education, which can be expected to improve children 's creativity and scientific knowledge through VR and coding education.

In this study, we collected and analyzed informal data such as Youtube and news to see how social discourse about VR and coding education is formed. In addition, this study suggested alternatives to activate VR and coding education. The 'research problem' of this study is 'What is social discourse about VR & coding education of Big Data?'

II. RESEARCH METHOD AND DATA ANALYSIS

We collected data on 'VR and coding education' through YouTube and news and data mining. In particular, we used Textom [5], a large-scale data analysis solution for data analysis and visualization. In addition, we used NetMiner to look at the degree centrality.

The data collection period is from the beginning of the discussion of coding education to the present (July 23, 2014 to April 7, 2019). The final collected data is 4392 cases. Text mining was performed based on data collected from Textom.

III. RESULTS

Keyword frequency analysis and Degree Centrality are summarized in the results.

A. Keyword Frequency Analysis

Based on the Big Data provided by Textom, the frequency analysis of the top 50 keywords through keyword analysis is shown in Figure 1. The result was 'Education', 'infant', 'progress', 'target', 'program', 'computer', 'start', 'software', 'student'. VR was analyzed by keywords such as 'game', 'virtual reality', 'content', 'experience' and so on. Figure 2

 ISBN: 978-1-61208-707-8

shows the result of visualizing with Word Cloud. Word Cloud is a technique for visualizing words that are related to key words [2].

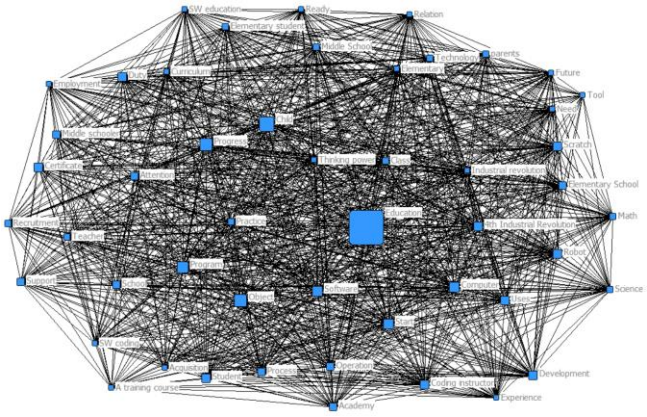


Figure 1. Coding training Frequency analysis of 50 keywords

Figure 2. Visualize VR as a keyword in Word Cloud

B. Degree Centrality.

We conducted a degree centrality analysis to identify the keywords related to 'VR and coding education'. Degree centrality is an indicator of direct connection between nodes, so it increases when there are many direct connections with other nodes [3]. In the analysis of degree of centrality, the keywords with the highest interaction were 'Experience', 'Virtual Reality', 'Robot', 'Fourth Industrial Revolution', 'Child'. We confirmed that common keywords have a significant influence on 'VR and coding education'.

IV. CONCLUSION

We investigated social discourse related to 'VR and coding education'. First, when we look through the keyword frequency analysis, we see many words of 'VR' and 'program'. This can be interpreted as a social atmosphere in which programs are presented based on VR in keeping with the changes in the era of the fourth industrial revolution. In addition, we can see that words such as 'elementary school', 'middle school', 'start' and 'obligation' appear. This means that the mandatory coding will be extended to elementary schools starting from middle and high schools in 2018 and 2019 [4]. In addition, it is understood that the coding education of younger children is also considered because of the mandatory coding education in grades 5 and 6 of elementary school. In particular, the need for content for VR and coding education applications is increasing [4]. Therefore, when one wants to apply the coding education to the younger age, you can consider providing the program using VR. In the course of coding education, we found through semantic network analysis that it is possible to provide contents through experiential education program as well as virtual reality such as VR as well as augmented reality. In other words, positive social perception about VR and coding education was confirmed overall, and it was recognized that application to younger children was considered considering that it is recognized as education suitable for the present age.

ACKNOWLEDGMENT

This work was supported by the Ministry of Education of the Republic of Korea and National Research Foundation of Korea (NRF-2017S1A5B6067046)

REFERENCES

[1] KBS Education Hope Project, "Learning is play," Aug 20, 2016. In press

[2] J. Su-Jeong, K. So-Eun and C. Ick-Joong, "Stepmother Images through Analyses of Twitter and News Articles," The Korea Contents Society, vol. 18(7). 2018, pp. 665-678.

[3] K. Yong-Hak, Social Network Analysis, Seoul: Park Young-sa, 2011.

[4] N. Hyun-Woo, "Study of VR/AR Coding Education Framework strategic for Elementary and Secondary School Students," Korea Science & Art Forum, vol. 36, 2018, pp. 85-97.

[5] Textom, Big Data Analysis Solution, 2018. http://www.textom.co.kr

 ISBN: 978-1-61208-707-8

Generative Content Co-creation:

Lessons from algorithmic music performance

Andrew R. Brown
Interactive Media Lab
Griffith University
Brisbane, Australia
email: andrew.r.brown@griffith.edu.au

Abstract—This article examines features of algorithmic music performance practices and considers how these might be applied to other generative content creation contexts. Based on the assumption that all generative processes are performative, the article draws from an analysis of live algorithmic music to outline lessons that may be more widely applicable to content co-creation with algorithmic systems. Methods discussed include algorithm selection and expression, the architecture of algorithmic system design, the effects of materiality on algorithmic performance, and how co-creative strategies openly embrace the influence of humans as agents in generative content systems. Having distilled and articulated these methods in this article it is anticipated that future research will apply them to generative media system beyond music performance for evaluation of their generalizability.

Keywords-Generative; Media; Creative; Content; Music; Algorithmic.

I. INTRODUCTION

Algorithmic content creation for fixed and dynamic works has a rich history, both in academic circles and in commercial contexts. For example, in the use of procedural content in computer games and interactive installations. Even if generative media outcomes are not dynamic the process of algorithmic content creation is, both in the system design and in the computational rendering. Algorithmic content creation always has a performative element. With this in mind, this article will reflect on the characteristics of performative algorithmic practices, in particular live coding and interactive computer music, and highlights aspects of these practices that are relevant for the design of generative content systems more broadly defined.

Generative media have been used in many contexts, from graphic design to architecture, and employ many techniques, from rule-based models to generative adversarial networks. Uses for generative media include customizing individual products based on templates or stylistic patterns, the design of complex artefacts with many dimensions of components, and the production of emergent or evolving experiences that adapt to changing contexts.

Within contemporary societies the production of media content is generally considered to be a creative act. Within this context a creative computational system can been defined as "a collection of processes, natural or automatic, which are capable of achieving simulating behavior which in humans would be deemed creative" [1]. Following Csikszentmihalyi [2], it is generally accepted that creative practice involves the production of novel and useful outcomes from work done within a conceptual space of acceptable outcomes. In the creative computation literature, there is often a distinction between (mere) generation for human selection from outcomes and generation for (self) evaluation by the machine [3]. In this article, the focus is on algorithmic systems used for generative co-creation [4] between humans and machines, such as those used in generative design or adaptive game music engines. Co-creation, in these contexts, is understood as "collaborative creativity where both the human and the computer take creative responsibility for the generation of a creative artefact" [5].

Section II will discuss approaches and considerations in algorithmic music practices that are relevant to generative media content creation. Section III will explore how performative interaction with generative systems is the basis for co-creation between people and machines.

II. LESSONS FROM LIVE ALGORITHMIC MUSIC

Generative algorithms are used in a range of music performance practices. These include those characterized as interactive music systems [6], live algorithms [7], networked music performance [8], or live coding [9]. Such practices will be collectively attributed here as live algorithmic music, and their practitioners as algorithmic musicians.

According to Lewis, live algorithmic music systems are interesting because they "produce a kind of virtual society that both draws from and challenges traditional notions of human interactivity and sociability, making common cause with a more general production of a hybrid, cyborg sociality that has forever altered both everyday sonic life and notions of subjectivity in high technological cultures" [10]. More specifically, it has been suggested that live algorithmic music practices convey three important attributes; algorithmic thinking, real-time creativity, and networked collaboration [11]. In the following sections each will be explored in turn.

A. Algorithm choice and design

Live algorithmic music is typically performed from memory and without time to consult reference materials, so it is sensible that algorithmic musicians seek to identify a small cohort of functions that are widely applicable to a variety of

 ISBN: 978-1-61208-707-8

musical circumstances. An advantage of focusing on a small set of coding patterns means these can be thoroughly understood and flexibly employed. This approach contrasts with many digital media software applications that boast the extensive array of functions available, most of which users will never use or, when they do, are 'one trick ponies' that soon become kitsch. Selection of such a set of foundational functions is likely to be media-specific and will have aesthetic implications. An example of such a selection is from the live coding duo *aa-cell* who suggest the following list of algorithms; probability, linear and higher order polynomials, periodic functions, modular arithmetic, set theory, and recursion [12]. They comment that "what has been a surprise to us, however, is just how much utility a small set of processes have provided" [12]. However, they also warn that correlations between mathematical functions and musical patterns are limited.

In addition to the choice of algorithm, algorithmic musicians often focus on "the way in which generative algorithms are represented in code to best afford interaction and modification during performance" [13]. They are aware that algorithm selection and design can significantly influence the flexibility and responsiveness of generative processes. So, they are concerned to choose the best parametric variations and constraints that maximize novelty whilst avoiding system misbehavior in the form of inappropriate output.

The balance of stability and novelty is at the heart of music making and is often stressed as key to other creative arts and to creativity itself. Algorithm design needs to enable the tuning of a system that walks a line between output that is neither boring nor inappropriate. Stability is often provided by incorporating domain knowledge into generative systems and novelty is often provided by processes with unpredictable outcomes.

One risk in overloading a system with domain specifics and constraints, is the restriction of output variety. Methods, such as genetic programming, can avoid such constraints by automating algorithm construction, not simply algorithm execution. However, the challenges of reliably automating meta-structural choices are many.

B. Compact code

It is important in live coding performances to manage the amount of code being typed on stage. Succinct expression of ideas and outcomes allows for efficient expression of ideas and responsiveness to contextual changes. Succinct expression relies on the language used, algorithm design, and interface for interaction.

Compactness of code is highlighted by Farnell as a desirable principle of procedural audio design, he argues that "compact code can be useful for purely software development reasons, being easy to understand, extend, and maintain" [14]. The advantage of code succinctness for creative expression is elsewhere emphasized: "The description length and complexity of an algorithm plays a large factor in its appropriateness for live coding... we consider algorithmic directness one of the most powerful aspects of live coding" [13].

However, compactness is not the algorithmic musician's only criteria. Many also seek 'descriptive transparency' defined as the "ability to further interact with the algorithm at a syntactic level" [13]. In practice this means that algorithms are designed to be modifiable through exposure of appropriate parameters. Descriptive transparency can be in tension with succinct expression as too concise a representation may obscure opportunities for interaction or variation by hiding key parameters or commitments.

C. Structure and abstractions

The choice of software abstractions makes a substantive difference to the ability to express ideas, in both a positive and a negative way [15]. The algorithmic musician constructs mental images of predicted outcomes and devises strategies to achieve them. In computing we use predefined structures, probabilistic decisions or situationally responsive processes, just like mental models—to guide prediction and planning. So computational abstractions need to be conditioned by limits or rules that guide effective outcomes.

Algorithmic musicians emphasize the importance of hierarchy in managing complexity and affording significant change with minimal high-level adjustments. This is necessary in live algorithmic music because "there is no possible way for us to deal with the complexity of the underlying operating system and hardware without levels of abstraction" [13]. In generative content systems, abstractions too are more than a structural convenience, they instil constraints and affordances and help represent a model of the content domain [16]. This push toward domain-specific abstraction has resulted in mini-languages being developed by algorithmic musicians to minimize cognitive load and enhance coding efficiency when performing.

Along with abstractions, the use of familiar metaphors in software design is also emphasized for reducing cognitive load for users. Brown and Sorensen [13] mention the deliberate use of familiar names for functions in their live coding practice—for example, bass, melody, and ambient—so that audience members unfamiliar with computer languages can make more sense of the projected code. When designing algorithmic music systems, Hoeberechts and co-authors [19] emphasize the importance of using real-world metaphors for system components, including 'instrument', 'performer', 'mood' and the like. Metaphors are also used to manage system organization. For example, in describing musical processes in terms of well understood practices such as describing software functions as 'players' that execute 'scores' on 'instruments'. The power of such analogy has also been emphasized by other studies of creativity and computation [18].

Metaphors can be useful for describing high-level parameters for algorithmic control. This should allow generative media systems to be guided by human interaction to produce a wide range of suitable outcomes. If abstractions are well chosen, then expressiveness and diversity of content output should be enhanced.

 ISBN: 978-1-61208-707-8

D. Networked architecture

Multiple dimensions can exist within one generative outcome. For example, a musical melody contains pitch, rhythm, timbre, volume and so on. However, relationships also exist between elements, like musical parts in a score, or visual components in a scene. Algorithmic musicians have found it useful to organize their code and their collaborations such that these elements correspond to a network of relationships. Networked musical performance practices include those with a focus on distributed interaction over the internet or, perhaps more pertinent here, distributed multi-agent systems whose architectures model the interdependence of sub components of the media being generated.

In many algorithmic music environments, concurrency is a key coding strategy to achieve interdependent modularity. Multiple concurrent operations are often conceived as 'loops' or 'processes' that act independently even if they share data. More formally, they have been implemented as 'temporal recursions'—code functions (closures) that call themselves periodically and maintain their own state [19].

Formalized protocols have been developed by algorithmic musicians for networked generative system architectures. A recent example is the Musebot (musical robot) framework designed to explore "the affordances of a multi-agent decision-making process" [20]. The Musebot protocol establishes "a parsimonious set of parameters for effective musical interaction between musebots" [20]. The open source specification includes a state-driven communication system for coordinating activities between agents. Messages are not themselves defined but are decided upon by cooperating developers. The objective of this approach is to compartmentalise the generative processes into components that manage complexity and enable flexible modular design and reuse.

E. The materiality of algorithms

Algorithms are made concrete using electromechanical means. When so constituted "an algorithm is a statement (in Foucault's meaning of the word), that operates, also indirectly, on the world around it" [21]. Algorithms that manifest as music machines have existed throughout history, as evident in the well-known player piano. The physicality of such machines conditions outcomes from them, for example through limits on speed of operation, resolution of output, and accuracy of calculation. In short, materiality matters.

According to Sorensen and Gardner, the "traditional view [in computer science] is to promote a strong separation between the *program*, *process* and *task* domains" [19]. Such separation may be counterproductive to effective media outcomes because it ignores the material implications of the world in which the computational processes are engaged. They suggest, instead, a model of 'cyber-physical programming' that acknowledges the temporal bounds of real-time computation and the interactions with physical media; such as sound playback systems, electronic circuitry, or 3D printing materials.

For algorithmic musicians, the affordances of computing machinery and software are particularly felt in relation to time. Music, as a temporal art form, relies on very precise timing for musical expression and sonic fidelity. Temporality is also pertinent for other interactive generative systems, such as computer games.

For performers or game players, feedback about the ongoing generative process is often expressed as real-time audio-visual output. The materiality of algorithmic media imposes limits on operations and provides feedback to human participants who can, in response to that feedback, become active agents in the generative process. Thus, materiality becomes the basis for interaction with generative process.

III. THE HUMAN IN THE LOOP

Almost by definition, algorithmic musicians are continually interacting with real-time generative processes. While such intimate co-creation may not always be true for all content generation systems, none are free from the influence of designers and programmers. Therefore, methods of co-creation with algorithms need to be taken into account.

A. Embodiment

In musical performance on acoustic instruments, sound strongly implies causality and agency [21]. In algorithmic music performance this connection can seem less direct, however, as Farnell suggests, "above all, it is important that we remain mindful of sound as a process of transforming energy" [14]. Even though Emmerson suggests that "electricity and electronic technology have allowed (even encouraged) the rupture of these relationships of body to object to sound" [22] the impact of gesture and (implied) action remain important in algorithmic media. When developing computational systems for generative media content, we should not lose sight of how human agency is implicated in the outcome.

Generative models of music often focus on emulating musical theories or musical cognition. Algorithms based on these theories need to take into account the performative aspects of music. When algorithmic musicians are producing music, they pay attention to sonic expression alongside compositional structure. Techniques that can be applied to both are discussed in [12] whilst techniques that focus on musical expression in particular are explored elsewhere [23].

Outside of music, the modelling of human creative gesture is well established. For example, in systems for digital drawing and character animation, or in the use of style transfer by machine learning systems for artistic practice. In music studies, the role of gesture is well explored [24], as is expressive gesture as a musical 'force' that guides expectations [25]. The implications of for generative content creation systems include consideration of a role for direct human motion in algorithm control, or for motion capture or physics simulation to animate parametric movements in an organic way.

 ISBN: 978-1-61208-707-8

B. *Interaction*

When addressing the role of visual feedback for audiences, live coders included in their TOPLAP manifesto (available online) a fundamental principle; "show us your screens". Projecting code during performances, it is hoped, will make visible the actions (typing) of the performer. For live coders themselves, visual feedback is also provided by the text editor which acts as their user interface to code acting as a musical score. In live coding, interaction is mediated by reading and writing code. Relatedly, recent explorations with the Musebot protocol have included human integration into multi-agent music systems using "a 'code-wrapper' around the human player—whimsically termed an *algoskin*" [26] that enables a human performer to appear to the network like another musebot. Other algorithmic musicians' employ various interfaces with algorithms, often via controllers employing combinations of buttons, dials and sliders that trigger functions and manipulate parameters. This field of interaction design for music is so active it has its own conference, i.e., New Interfaces for Musical Expression (NIME).

At issue here is the expression of 'liveness' [27], "a sense that the person playing is contributing to that emotive energy through the performance decisions being made" [28]. More generally our interest is in the contribution of performative interaction on the outcome of the generated digital media. This is particularly important for interactions between people and machines in co-creative algorithmic systems.

So, how can a person be an active co-creator? Dahlstedt suggests the following categorization; "You can play *on*, *in* or *with* an algorithm" [21]. Performing 'on' an algorithm means to control its parameters. Performing 'with' an algorithm means to undertake your own activities in parallel to the algorithm's without influencing it. Performing 'in' an algorithm means that actions of the algorithm and human are socially coupled [29] such that each interaction has an effect on other parts of the cybernetic system.

A fourth category of co-creation is the ability for the human to redefine the generative process as it executes. As is the case in virtuosic live coding performances. According to Magnusson this "seems to be a logical and necessary step in the evolution of human–machine communication" [11].

IV. CONCLUSION

The production of generative media requires creators to design the behaviors of algorithmic systems. In this way content outcomes are managed by the specification of creative behaviors rather than only by direct manipulation of materials. Behaviors are performative, and so we can learn from the performing arts how algorithmic behaviours lead to creative outcomes. Computational performing arts, such as live algorithmic music, have a special role to play in revealing pertinent practices applicable to generative content.

This article summarized live algorithmic music practices to assemble, for the first time, a consolidated set of lessons that may be helpful for co-creative content production. Methods that were identified include algorithm selection, algorithmic system architecture, the effects of materiality on algorithmic behaviour, and the influence of humans as co-creative agents. Future research will look at implementing these methods in prototype generative media systems for evaluation.

Lewis philosophically suggests that the impact of live algorithmic music may even reach beyond these lessons, that "perhaps our improvising computers can teach us how to live in a world marked by agency, indeterminacy, analysis of conditions, and the apparent ineffability of choice" [10].

REFERENCES

[1] G. Wiggins, "A preliminary framework for description, analysis and comparison of creative systems," *Journal of Knowledge Based Systems*, vol. 19, no. 7, pp. 449–458, 2006.

[2] M. Csikszentmihalyi, "The Domain of Creativity," in *Changing the World: A framework for the study of creativity*, London: Praeger, 1994.

[3] S. Colton, A. Pease, J. Corneli, M. Cook, and T. Llano, "Assessing Progress in Building Autonomously Creative Systems.," in *ICCC*, pp. 137–145, 2014.

[4] T. Lubart, "How can computers be partners in the creative process: classification and commentary on the special issue," *International Journal of Human-Computer Studies*, vol. 63, no. 4–5, pp. 365–369, 2005.

[5] A. Kantosalo, J. M. Toivanen, P. Xiao, and H. Toivonen, "From Isolation to Involvement: Adapting Machine Creativity Software to Support Human-Computer Co-Creation.," in *Proceedings of the International Conference on Computational Creativity*, Ljubljana, Slovenia, pp. 1–7, 2014.

[6] R. Rowe, *Interactive Music Systems: Machine listening and composing*. Cambridge, MA: The MIT Press, 1993.

[7] T. Blackwell and M. Young, "Live Algorithms," *Artificial Intelligence and Simulation of Behaviour Quarterly*, vol. 122, pp. 7–9, 2005.

[8] E. M. Schooler and J. Touch, "Distributed music: A foray into networked performance," in *Proceedings of the International Network Music Festival*, Santa Monica, CA, 1993.

[9] N. Collins, A. McLean, J. Rohrhuber, and A. Ward, "Live Coding in Laptop Performance," *Organised Sound*, vol. 8, no. 3, pp. 321–330, 2003.

[10] G. E. Lewis, "Why do we want our computers to improvise?," in *The Oxford Handbook of Algorithmic Music*, A. McLean and R. T. Dean, Eds. New York: Oxford University Press, pp. 123–130, 2018.

[11] T. Magnusson, "Herding cats: Observing live coding in the wild," *Comp. Music Journal*, vol. 38, no. 1, pp. 8–16, 2014.

[12] A. Sorensen and A. R. Brown, "aa-cell in practice: an approach to musical live coding," in *Proceedings of the International Computer Music Conference*, Copenhagen, pp. 292–299, 2007.

[13] A. R. Brown and A. Sorensen, "Interacting with Generative Music through Live Coding," *Contemporary Music Review*, vol. 28, no. 1, pp. 17–29, 2009.

[14] A. Farnell, "Procedural Audio Theory and Practice," in *The Oxford Handbook of Interactive Audio*, K. Collins, B. Kapralos, and H. Tessler, Eds. Oxford: Oxford University Press, pp. 531–540, 2014.

[15] H. Abelson and G. J. Sussman, *Structure and Interpretation of Computer Programs*, 2nd Edition. Cambridge, MA: The MIT Press, 1996.

[16] J. Rohrhuber, A. de Campo, and R. Wieser, "Algorithms Today: Notes on language design for just in time programming," in *Proceedings of the International Computer Music Conference*, Barcelona, 2005.

 ISBN: 978-1-61208-707-8

[17] N. Hoeberechts, J. Shamtz, and M. Katchabaw, "Delivering Interactive Experiences through the Emotional Adaptation of Automatically Composed Music," in *The Oxford Handbook of Interactive Audio*, K. Collins, B. Kapralos, and H. Tessler, Eds. Oxford: Oxford University Press, pp. 419–442, 2014.

[18] D. Hofstadter and E. Sanders, *Surfaces and Essences: Analogy as the fuel and fire of thinking*. New York: Basic Books, 2013.

[19] A. Sorensen and H. Gardner, "Cyber-physical programming with Impromptu," *ACM Sigplan Notices*, vol. 45, no. 10, pp. 822–834, 2010.

[20] A. Eigenfeldt, A. R. Brown, O. Bown, and T. Gifford, "Distributed Musical Decision-making in an Ensemble of Musebots: Dramatic Changes and Endings," in *Proceedings of the International Conference on Computational Creativity*, Atlanta, GA, pp. 88–95, 2017.

[21] P. Dahlstedt, "Action and Perception: Embodying Algorithms adn the Extended Mind," in *The Oxford Handbook of Algorithmic Music*, A. McLean and R. T. Dean, Eds. New York: Oxford Univerity Press, pp. 41–65, 2018.

[22] S. Emmerson, "'Losing Touch?': The human performer and electronics," in *Music, Electronic Media and Culture*, Hampshire, UK: Ashgate, pp. 194–216, 2000.

[23] P. Todd, "Simulating the Evolution of Musical Behaviour," in *The Origins of Music*, Cambridge, MA: The MIT Press, pp. 361–388, 2000.

[24] M. Leman, "Music, Gesture, and the Formation of Embodied Meaning," in *Musical Gestures: Sound, movement, and meaning*, R. I. Godøy and M. Leman, Eds. New York: Routledge, pp. 126–153, 2010.

[25] S. Larson, *Musical Forces: Motion, Metaphor and Meaning in Music*. Bloomington: Indiana University Press, 2012.

[26] A. R. Brown, et al., "Interacting with Musebots," in *Proceedings of New Interfaces for Musical Expression*, Blacksburg, VA, pp. 19–24, 2018.

[27] P. Auslander, *Liveness: Performance in a mediatized culture*. Oxon, UK: Routledge, 1999.

[28] M. Frengel, "Interactivity and Liveness in Electroacoustic Concert Music," in *The Oxford Handbook of Interactive Audio*, K. Collins, B. Kapralos, and H. Tessler, Eds. Oxford: Oxford University Press, 2014.

[29] H. R. Maturana and F. J. Varela, *The Tree of Knowledge: The biological roots of human understanding*. Boston: Shambhala, 1988.

 ISBN: 978-1-61208-707-8

Using Domain Taxonomy to Model Generalization of Thematic Fuzzy Clusters

Dmitry Frolov
National Research University "Higher School of Economics" Moscow, Russian Federation
Email: dfrolov@hse.ru

Susana Nascimento
Universidade Nova de Lisboa Caparica, Portugal
Email: snt@fct.unl.pt

Trevor Fenner
Birkbeck, University of London London, UK
Email: trevor@dcs.bbk.ac.uk

Boris Mirkin
National Research University "Higher School of Economics" Moscow, Russian Federation, and Birkbeck, University of London London, UK
Email: bmirkin@hse.ru

***Abstract*—We define a most specific generalization of a fuzzy set of topics assigned to leaves of the rooted tree of a domain taxonomy. This generalization lifts the set to its "head subject" in the higher ranks of the taxonomy tree. The head subject is supposed to "tightly" cover the query set, possibly bringing in some errors, both "gaps" and "offshoots". Our method globally minimizes a penalty function combining the numbers of head subjects and gaps and offshoots, differently weighted. We apply this to a collection of about 18000 research papers published in Springer journals on Data Science for the past 20 years. We extract a taxonomy of Data Science from the international Association for Computing Machinery Computing Classification System 2012 (ACM-CCS). We find fuzzy clusters of leaf topics over the text collection and use lifted head subjects of the thematic clusters to comment on the tendencies of current research in the corresponding aspects of the domain.**

Keywords–Generalization; gap-offshoot penalty; fuzzy cluster; spectral clustering; annotated suffix tree.

I. INTRODUCTION

The issue of automation of structurization and interpretation of digital text collections is of ever-growing importance because of both practical needs and theoretical necessity. This paper is concerned with an aspect of this, modeling generalization as a unique feature of human cognitive abilities.

The existing approaches to computational analysis of structure of text collections usually involve no generalization as a specific aim. The most popular tools for structuring text collections are cluster analysis and topic modelling. Both involve items of the same level of granularity as individual words or short phrases in the texts, thus no generalization as an explicitly stated goal.

Nevertheless, the hierarchical nature of the universe of meanings is reflected in the flow of publications on text analysis. We can distinguish between at least three directions at which the matter of generalization is addressed. First of all, there are activities related to developing taxonomies, especially those involving hyponymic/hypernymic relations (see, for example, [14] [17], and references therein). A recent paper [15] is devoted to supplementing a taxonomy with newly emerging research topics.

Another direction is part of conventional activities in text summarization. Usually, summaries are created using a rather mechanistic approach of sentence extraction. There is, however, also an approach for building summaries as abstractions of texts by combining some templates, such as Subject-Verb-Object (SVO) triplets (see, for example, [9]).

One more direction is what can be referred to as "operational" generalization. In this direction, the authors use generalized case descriptions involving taxonomic relations between generalized states and their parts to achieve a tangible goal, such as improving characteristics of text retrieval (see, for example, [12] [16].)

This paper begins a novel direction of research by using an existing taxonomy for straightforwardly implementing the idea of generalization. According to the Merriam-Webster dictionary, the term "generalization" refers to deriving a general conception from particulars. The "particulars", in our case, are represented by a fuzzy set of taxonomy leaves, whereas "the general conception" will be represented by a higher rank taxonomy node to embrace the fuzzy set as tight as possible. To the best of our knowledge, this approach has been never explored before. We experimentally show that our method leads to the type of conclusions which cannot be provided by other existing approaches to the analysis of text collections (see the end of Section III).

Our text collection is a set of about 18,000 research papers published by the Springer Publishers in 17 journals related to Data Science for the past 20 years. Our taxonomy of Data Science is a slightly modified part of the world-wide Association for Computing Machinery Computing Classification System (ACM-CCS), a 5-layer taxonomy published in 2012 [1].

The rest of the paper is organized accordingly. Section II presents a mathematical formalization of the generalization problem as of parsimoniously lifting of a given fuzzy leaf set to higher ranks of the taxonomy and provides a recursive algorithm leading to a globally optimal solution to the problem. Section III describes an application of this approach to deriving tendencies in development of the Data Science according to our Springer text collection mapped to the ACM-CCS. Its subsections describe stages of our approach to finding and generalizing fuzzy clusters of research topics. In the end, we point to tendencies in the development of the corresponding parts of Data Science, as drawn from the generalization results.

II. GENERALIZATION BY PARSIMONIOUSLY LIFTING A FUZZY THEMATIC SUBSET IN TAXONOMY: MODEL AND METHOD

Mathematically, a taxonomy is a rooted tree whose nodes are annotated by taxonomy topics.

We consider the following problem. Given a fuzzy set S of taxonomy leaves, find a node $t(S)$ of higher rank in the taxonomy, that covers the set S in a most specific way. Such a "lifting" problem is a mathematical explication of the human facility for generalization.

 ISBN: 978-1-61208-707-8

The problem is not as simple as it may seem to be. Consider, for the sake of simplicity, a hard set S shown with five black leaf boxes on a fragment of a tree in Figure 1. Figure 2 illustrates the situation at which the set of black boxes is lifted to the root, which is shown by blackening the root box, and its offspring, too. If we accept that set S may be generalized by the root, this would lead to a number, four, white boxes to be covered by the root and, thus, in this way, falling in the same concept as S even as they do not belong in S. Such a situation will be referred to as a gap. Lifting with gaps should be penalized. Altogether, the number of conceptual elements introduced to generalize S here is 1 head subject, that is, the root to which we have assigned S, and the 4 gaps occurred just because of the topology of the tree, which imposes this penalty. Another lifting decision is illustrated in Figure 3: here the set is lifted just to the root of the left branch of the tree. We can see that the number of gaps has drastically decreased, to just 1. However, another oddity emerged. A black box on the right belongs to S but is not covered by the head subject in the root of the left branch. This type of error will be referred to as an offshoot. At this lifting, three new items emerge: one head subject, one offshoot, and one gap. Which of the errors is greater?

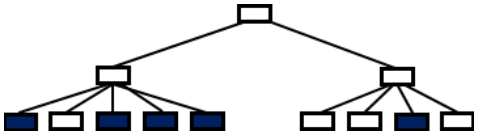


Figure 1. A crisp query set, shown by black boxes, to be conceptualized in the taxonomy.

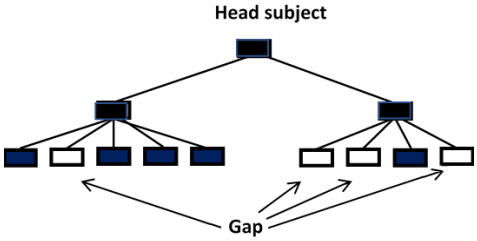


Figure 2. Generalization of the query set from Figure 1 by mapping it to the root, with the price of four gaps emerged at the lift.

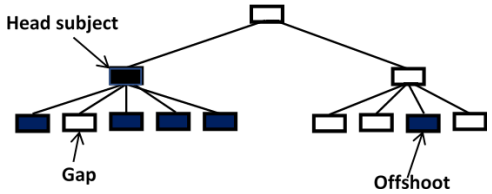


Figure 3. Generalization of the query set from Figure 1 by mapping it to the root of the left branch, with the price of one gap and one offshoot emerged at this lift.

We are interested to see whether a fuzzy set S can be generalized by a node t from higher ranks of the taxonomy, so that S can be thought of as falling within the framework covered by the node t. The goal of finding an interpretable pigeon-hole for S within the taxonomy can be formalized as that of finding one or more "head subjects" t to cover S with the minimum number of all the elements introduced at the generalization: head subjects, gaps, and offshoots. This goal realizes the principle of Maximum Parsimony (MP).

Consider a rooted tree T representing a hierarchical taxonomy so that its nodes are annotated with key phrases signifying various concepts. We denote the set of all its *leaves* by I. The relationship between nodes in the hierarchy is conventionally expressed using genealogical terms: each node $t \in T$ is said to be the *parent* of the nodes immediately descending from t in T, its *children*. We use $\chi(t)$ to denote the set of children of t. Each *interior* node $t \in T - I$ is assumed to correspond to a concept that generalizes the topics corresponding to the leaves $I(t)$ descending from t, viz. the leaves of the subtree $T(t)$ rooted at t, which is conventionally referred to as the *leaf cluster of* t.

A *fuzzy set* on I is a mapping u of I to the non-negative real numbers that assigns a membership value, or support, $u(i) \geq 0$ to each $i \in I$. We refer to the set $S_u \subset I$, where $S_u = \{i \in I : u(i) > 0\}$, as the *base* of u. In general, no other assumptions are made about the function u, other than, for convenience, commonly limiting it to not exceed unity. Conventional, or *crisp*, sets correspond to binary membership functions u such that $u(i) = 1$ if $i \in S_u$ and $u(i) = 0$ otherwise.

Given a fuzzy set u defined on the leaves I of the tree T, one can consider u to be a (possibly noisy) projection of a general concept, u's "head subject", onto the corresponding leaf cluster. Under this assumption, there should exist a head subject node h among the interior nodes of the tree T such that its leaf cluster $I(h)$ more or less coincides (up to small errors) with S_u. This head subject is the generalization of u to be found. The two types of possible errors associated with the head subject, if it does not cover the base precisely, are false positives and false negatives, referred to in this paper, as *gaps* and *offshoots*, respectively. They are illustrated in Figures 2 and 3. Given a head subject node h, a gap is a node t covered by h but not belonging to u, so that $u(t) = 0$. In contrast, an offshoot is a node t belonging to u so that $u(t) > 0$ but not covered by h. Altogether, the total number of head subjects, gaps, and offshoots has to be as small as possible. To this end, we introduce a penalty for each of these elements. Assuming for the sake of simplicity, that the black box leaves on Figure 1 have membership function values equal to unity, one can easily see that the total penalty at the head subject raised to the root (Figure 2) is equal to $1+4\lambda$ where 1 is the penalty for a head subject and λ, the penalty for a gap, since the lift on Figure 2 involves one head subject, the root, and four gaps, the blank box leaves. Similarly, the penalty for the lift on Figure 3 to the root of the left-side subtree is equal to $1 + \gamma + \lambda$ where γ is the penalty for an offshoot, as there is one copy of each, head subject, gap, and offshoot, in Figure 3. Therefore, depending on the relationship between γ and λ either lift on Figure 2 or lift on Figure 3 is to be chosen.

Consider a candidate node h in T and its meaning relative to fuzzy set u. An *h-gap* is a node g of $T(h)$, other than h, at which a *loss* of the meaning has occurred, that is, g is a maximal u-irrelevant node in the sense that its parent is not u-irrelevant. Conversely, establishing a node h as a head subject can be considered as a *gain* of the meaning of u at the node. The set of all h-gaps will be denoted by $G(h)$. A node $t \in T$ is referred to as *u-irrelevant* if its leaf-cluster $I(t)$ is disjoint from the base S_u. Obviously, if a node is u-irrelevant, all of

its descendants are also u-irrelevant.

An *h-offshoot* is a leaf $i \in S_u$ which is not covered by h, i.e., $i \notin I(h)$. The set of all h-offshoots is $S_u - I(h)$. Given a fuzzy topic set u over I, a set of nodes H will be referred to as a *u-cover* if: (a) H covers S_u, that is, $S_u \subseteq \bigcup_{h \in H} I(h)$, and (b) the nodes in H are unrelated, i.e., $I(h) \cap I(h') = \emptyset$ for all $h, h' \in H$ such that $h \neq h'$. The interior nodes of H will be referred to as *head subjects* and the leaf nodes as *offshoots*, so the set of offshoots in H is $H \cap I$. The set of *gaps* in H is the union of $G(h)$ over all head subjects $h \in H - I$.

We define the penalty function $p(H)$ for a u-cover H as:

$$p(H) = \sum_{h \in H - I} u(h) + \\ + \sum_{h \in H - I} \sum_{g \in G(h)} \lambda v(g) + \sum_{h \in H \cap I} \gamma u(h). \quad (1)$$

The problem we address is to find a u-cover H that globally minimizes the penalty $p(H)$. Such a u-cover is the parsimonious generalization of the set u.

Before applying an algorithm to minimize the total penalty, one needs to execute a preliminary transformation of the tree by pruning it from all the non-maximal u-irrelevant nodes, i.e., descendants of gaps. Simultaneously, the sets of gaps $G(t)$ and the internal summary gap importance $V(t) = \sum_{g \in G(t)} v(g)$ in (1) can be computed for each interior node t. We note that the elements of S_u are in the leaf set of the pruned tree, and the other leaves of the pruned tree are precisely the gaps. After this, our lifting algorithm ParGenFS applies. For each node t, the algorithm ParGenFS computes two sets, $H(t)$ and $L(t)$, containing those nodes in $T(t)$ at which respectively gains and losses of head subjects occur (including offshoots). The associated penalty $p(t)$ is computed too.

An assumption of the algorithm is that no gain can happen after a loss. Therefore, $H(t)$ and $L(t)$ are defined assuming that the head subject has not been gained (nor therefore lost) at any of t's ancestors. The algorithm ParGenFS recursively computes $H(t)$, $L(t)$ and $p(t)$ from the corresponding values for the child nodes in $\chi(t)$.

Specifically, for each leaf node that is not in S_u, we set both $L(\cdot)$ and $H(\cdot)$ to be empty and the penalty to be zero. For each leaf node that is in S_u, $L(\cdot)$ is set to be empty, whereas $H(\cdot)$, to contain just the leaf node, and the penalty is defined as its membership value multiplied by the offshoot penalty weight γ. To compute $L(t)$ and $H(t)$ for any interior node t, we analyze two possible cases: (a) when the head subject has been gained at t and (b) when the head subject has not been gained at t.

In case (a), the sets $H(\cdot)$ and $L(\cdot)$ at its children are not needed. In this case, $H(t)$, $L(t)$ and $p(t)$ are defined by:

$$H(t) = \{t\}, \quad L(t) = G(t), \quad p(t) = u(t) + \lambda V(t). \quad (2)$$

In case (b), the sets $H(t)$ and $L(t)$ are just the unions of those of its children, and $p(t)$ is the sum of their penalties:

$$H(t) = \bigcup_{w \in \chi(t)} H(w), \quad L(t) = \bigcup_{w \in \chi(t)} L(w), \\ p(t) = \sum_{w \in \chi(t)} p(w). \quad (3)$$

To obtain a parsimonious lift, whichever case gives the smaller value of $p(t)$ is chosen.

When both cases give the same values for $p(t)$, we may choose, say, (a). The output of the algorithm consists of the values at the root, namely, H – the set of head subjects and offshoots, L – the set of gaps, and p – the associated penalty.

It was mathematically proven that the algorithm ParGenFS leads to an optimal lifting indeed [5].

III. Highlighting tendencies in the current research by clustering and lifting a collection of research papers

Being confronted with the problem of structuring and interpreting a set of research publications in a domain, one can think of either of the following two pathways to take. One is so-to-speak empirical and the other theoretical. The first pathway tries to discern main categories from the texts, the other, from knowledge of the domain. The first approach is exemplified by the LDA-based topic modeling [2]; the second approach, by using an expert-driven taxonomy, such as ACM-CCS [1] (see, for example, [13]).

This paper follows the second pathway by moving, in sequence, through the following stages:

- preparing a scholarly text collection;
- preparing a taxonomy of the domain under consideration;
- developing a matrix of relevance values between taxonomy leaf topics and research publications from the collection;
- finding fuzzy clusters according to the structure of relevance values;
- lifting the clusters over the taxonomy to conceptualize them via generalization;
- making conclusions from the generalizations.

Each of the items is covered in a separate subsection further on.

A. Scholarly text collection

Because of a generous offer from the Springer Publisher, we were able to download a collection of 17685 research papers together with their abstracts published in 17 journals related to Data Science for 20 years from 1998-2017 [5]. We take the abstracts to these papers as a representative collection.

B. DST Taxonomy

Taxonomy building is a form of knowledge engineering which is getting more and more popular. Most known are taxonomies within the bioinformatics Genome Ontology (GO) project [6], Health and Medicine SNOMED CT project [8] and the like. Mathematically, a taxonomy is a rooted tree, a hierarchy, whose all nodes are labeled by main concepts of the domain the taxonomy relates to. The hierarchy corresponds to the inclusion relation: the fact that node A is the parent of B means that B is part, or a special case, of A.

The subdomain of our choice is Data Science, comprising such areas as machine learning, data mining, data analysis, etc. We take that part of the ACM-CCS 2012 taxonomy, which is related to Data Science, and add a few leaves related to more recent Data Science developments. The Taxonomy of Data Science, DST, with all its 317 leaves, is presented in [5].

C. Deriving fuzzy clusters of taxonomy topics

Clusters of topics should reflect co-occurrence of topics: the greater the number of texts to which both t and t' topics are relevant, the greater the interrelation between t and t',

the greater the chance for topics t and t' to fall in the same cluster. We have tried several popular clustering algorithms at our data. Unfortunately, no satisfactory results have been found. Therefore, we present here results obtained with the FADDIS algorithm developed in [11] specifically for finding thematic clusters. This algorithm implements assumptions that are relevant to the task:

- LN Laplacian Pseudo-Inverse Normalization (LaPIN): Similarity data transformation, modeling – to an extent – heat distribution and, in this way, making the cluster structure sharper.
- AA Additivity: Thematic clusters behind the texts are additive, so that co-relevance similarity values are sums of contributions by different hidden themes.
- AN Non-Completeness: Clusters do not necessarily cover all the key phrases available, as the text collection under consideration may be irrelevant to some of them.

1) Co-relevance topic-to-topic similarity score: Given a keyphrase-to-document matrix R of relevance scores is converted to a keyphrase-to-keyphrase similarity matrix A for scoring the "co-relevance" of keyphrases according to the text collection structure. The similarity score $a_{tt'}$ between topics t and t' is computed as the inner product of vectors of scores $r_t = (r_{tv})$ and $r_{t'} = (r_{t'v})$ where $v = 1, 2, \ldots, V = 17685$. The inner product is moderated by a natural weighting factor assigned to texts in the collection. The weight of text v is defined as the ratio of the number of topics n_v relevant to it and n_{max}, the maximum n_v over all v = 1,2,...,V. A topic is considered relevant to v if its relevance score is greater than 0.2 (a threshold found experimentally, see [4]).

2) Fuzzy thematic clusters: To obtain fuzzy clusters of topics we used a method FADDIS, that was developed in [10]. FADDIS finds clusters one-by-one. Paper [11] provides some theoretical and experimental computation results to demonstrate that FADDIS is competitive over popular fuzzy clustering approaches.

After computing the 317×317 topic-to-topic co-relevance matrix, converting it to a topic-to-topic LaPIN transformed similarity matrix, and applying FADDIS clustering, we sequentially obtained 6 clusters, of which three clusters appear to be obviously homogeneous. They relate to "Learning", "Retrieval", and "Clustering". These clusters, L, R, and C, are presented in Tables I, II, and III, respectively.

TABLE I. CLUSTER L "LEARNING": TOPICS WITH MEMBERSHIP VALUES GREATER THAN 0.15

$u(t)$	Code	Topic
0.300	5.2.3.8.	Rule Learning
0.282	5.2.2.1.	Batch Learning
0.276	5.2.1.1.2.	Learning to Rank
0.217	1.1.1.11.	Query Learning
0.216	5.2.1.3.3.	Apprenticeship Learning
0.213	1.1.1.10.	Models of Learning
0.203	5.2.1.3.5.	Adversarial Learning
0.202	1.1.1.14.	Active Learning
0.192	5.2.1.4.1.	Transfer Learning
0.192	5.2.1.4.2.	Lifelong Machine learning
0.189	1.1.1.8.	Online Learning Theory
0.166	5.2.2.2.	Online Learning Settings
0.159	1.1.1.3.	Unsupervised Learning and Clustering

TABLE II. CLUSTER R "RETRIEVAL": TOPICS WITH MEMBERSHIP VALUES GREATER THAN 0.15

$u(t)$	Code	Topic
0.211	3.4.2.1.	Query Representation
0.207	5.1.3.2.1.	Image Representations
0.194	5.1.3.2.2.	Shape Representations
0.194	5.2.3.6.2.1	Tensor Representation
0.191	5.2.3.3.3.2	Fuzzy Representation
0.187	3.1.1.5.3.	Data Provenance
0.173	2.1.1.5.	Equational Models
0.173	3.4.6.5.	Presentation of Retrieval Results
0.165	5.1.3.1.3.	Video Segmentation
0.155	5.1.3.1.2.	Image Segmentation
0.154	3.4.5.5.	Sentiment Analysis

TABLE III. CLUSTER C "CLUSTERING": TOPICS WITH MEMBERSHIP VALUES GREATER THAN 0.15

$u(t)$	Code	Topic
0.327	3.2.1.4.7	Biclustering
0.286	3.2.1.4.3	Fuzzy Clustering
0.248	3.2.1.4.2	Consensus Clustering
0.220	3.2.1.4.6	Conceptual Clustering
0.192	5.2.4.3.1	Spectral Clustering
0.187	3.2.1.4.1	Massive Data Clustering
0.159	3.2.1.7.3	Graph Based Conceptual Clustering
0.151	3.2.1.9.2.	Trajectory Clustering

D. Results of lifting clusters L, R, and C within DST

To apply ParGenFS algorithm, values of λ and γ should be defined first. This may highly affect the results. In the example above, lifting in Figure 2 is more parsimonious than lifting in Figure 3 if $\gamma > 3\lambda$, or the latter, if otherwise. We define off-shoot penalty $\gamma = 0.9$ to make it almost as costly as a head subject. In contrast, the gap penalty is defined as $\lambda = 0.1$ to take into account that every node in the taxonomy tree has about 10-15 children so that half-a-dozen gaps would be admissible. The clusters above are lifted in the DST taxonomy using ParGenFS algorithm with these parameter values.

The results of lifting of Cluster L are shown in Figure 4. There are three head subjects: machine learning, machine learning theory, and learning to rank. These represent the structure of the general concept "Learning" according to the text collection under consideration. The list of gaps obtained is less instructive, reflecting probably a relatively modest coverage of the domain by the publications in the collection (see in Table IV).

Similar comments can be made with respect to results of lifting of Cluster R: Retrieval. The obtained head subjects: Information Systems and Computer Vision show the structure of "Retrieval" in the set of publications under considerations.

For Cluster C 16 (!) head subjects were obtained: clustering, graph based conceptual clustering, trajectory clustering, clustering and classification, unsupervised learning and clustering, spectral methods, document filtering, language models, music retrieval, collaborative search, database views, stream management, database recovery, mapreduce languages, logic and databases, language resources. As one can see, the core clustering subjects are supplemented by methods and environments in the cluster – this shows that the ever increasing role of clustering activities perhaps should be better reflected in the taxonomy.

 ISBN: 978-1-61208-707-8

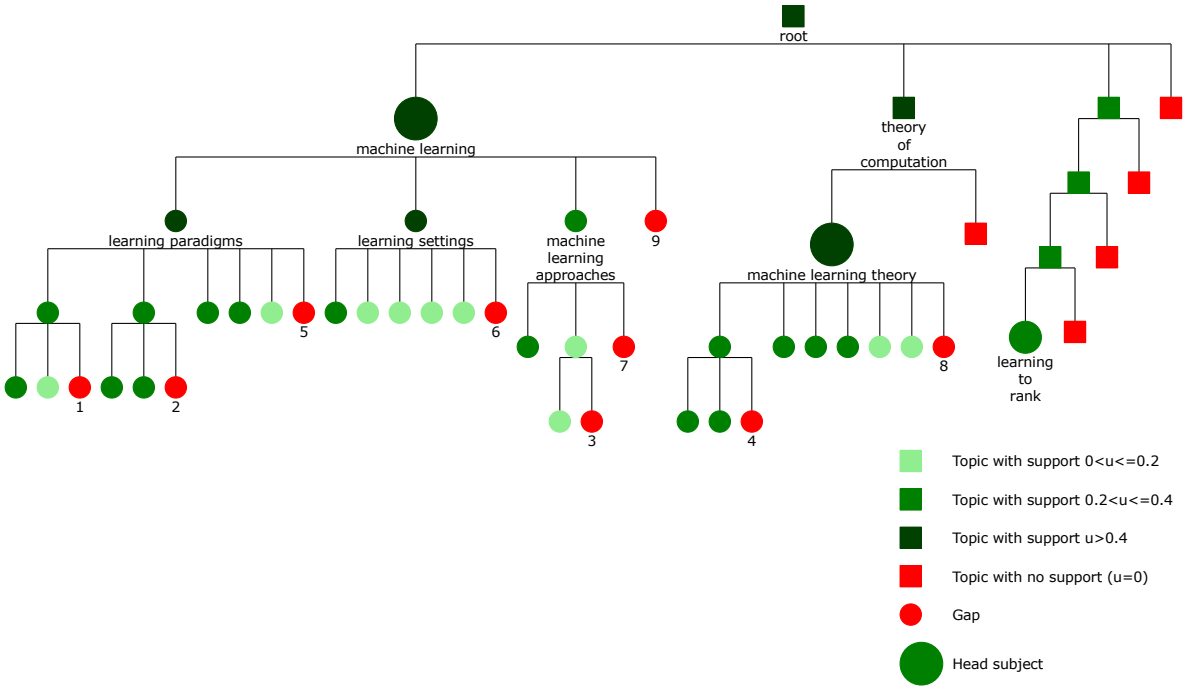


Figure 4. Lifting results for Cluster L: Learning. Gaps are numbered, see Table IV.

TABLE IV. GAPS AT THE LIFTING OF CLUSTER L

Number	Topics
1	ranking, supervised learning by classification, structured outputs
2	sequential decision making in practice, inverse reinforcement learning in practice
3	statistical relational learning
4	sequential decision making, inverse reinforcement learning
5	unsupervised learning
6	learning from demonstrations, kernel approach
7	classification and regression trees, kernel methods, neural networks, learning in probabilistic graphical models, learning linear models, factorization methods, markov decision processes, stochastic games, learning latent representations, multiresolution, support vector machines
8	sample complexity and generalization bounds, boolean function learning, kernel methods, boosting, bayesian analysis, inductive inference, structured prediction, markov decision processes, regret bounds
9	machine learning algorithms

E. Making conclusions

We can see that the topic clusters found with the text collection do highlight areas of soon-to-be developments. Three clusters under consideration closely relate, in respect, to the following processes:

- theoretical and methodical research in learning, as well as merging the subject of learning to rank within the mainstream;
- representation of various types of data for information retrieval, and merging that with visual data and their semantics; and
- various types of clustering in different branches of the taxonomy related to various applications and instruments.

In particular, one can see from the "Learning" head subjects (see Figure 4 and comments to it) that main work here still concentrates on theory and method rather than applications. A good news is that the field of learning, formerly focused mostly on tasks of learning subsets and partitions, is expanding currently towards learning of ranks and rankings. Of course, there remain many sub-areas to be covered: these can be seen in and around the list of gaps in Table IV.

Moving to the lifting results for the information retrieval cluster R, we can clearly see the tendencies of the contemporary stage of the process. Rather than relating the term "information" to texts only, as it was in the previous stages of the process of digitalization, visuals are becoming parts of the concept of information. There is a catch, however. Unlike the multilevel granularity of meanings in texts, developed during millennia of the process of communication via languages in the humankind, there is no comparable hierarchy of meanings for images. One may only guess that the elements of the R cluster related to segmentation of images and videos, as well as those related to data management systems, are those that are going to be put in the base of a future multilevel system of meanings for images and videos. This is a direction for future developments clearly seen from lifting results.

Regarding the "clustering" cluster C with its 16 (!) head subjects, one may conclude that, perhaps, a time moment has come or is to come real soon, when the subject of clustering must be raised to a higher level in the taxonomy to embrace all these "heads". At the beginning of the Data Science era, a few decades ago, clustering was usually considered a more-or-less

 ISBN: 978-1-61208-707-8

auxiliary part of machine learning, the unsupervised learning. Perhaps, soon we are going to see a new taxonomy of Data Science, in which clustering is not just an auxiliary instrument but rather a model of empirical classification, a big part of the knowledge engineering.

It should be pointed out that analysis of tendencies of research is carried out by several groups using co-citation data, especially in dynamics (see, for example, a review in [3]). This approach leads to conclusions involving "typical", rather than general, authors and/or papers, and, therefore, is complementary to our approach.

IV. Conclusion and Future Work

This paper presents a formalization of the concept of generalization, an important part of the human ability for conceptualization. According to Collins Dictionary, conceptualization is "formation (of a concept or concepts) out of observations, experience, data, etc." We assume that such an operation may require a coarser granularity of the domain structuring. This is captured by the idea of lifting a query set to higher ranks in a hierarchical taxonomy of the domain.

The hierarchical structure of taxonomy brings in possible inconsistencies between a query set and the taxonomy structure. These inconsistencies can be of either of two types, gaps or offshoots, potentially emerging at the coarser "head subject" to cover the query set. A gap is such a node of the taxonomy, that is covered by the head subject but does not belong in the query set. An offshoot is a node of the taxonomy, that does belong in the query set but is not covered by the head subject. The higher the rank of a candidate for the conceptual head subject, the larger the number of gaps. The lower is the rank of the head subject, the larger the number of offshoots. Our algorithm ParGenFS allows to find a globally optimal lifting to balance the numbers of head subjects, gaps, and offshoots depending on relative penalties for these types of inconsistencies.

The proposed approach to generalization can be used in a number of similar tasks, such as positioning of a research project, interpretation of a concept which is not present in the taxonomy, annotation of a set of research articles. These all are parts of the processes of long-term research analysis and planning at which our approach should be positioned.

Among major issues requiring further development in this direction, two of the most relevant are taxonomy developments and specifying penalty weights. The former needs more attention both from research communities and planning committees. Specifically, most urgent directions for development here are: developing better methods to automate the process of taxonomy making and open discussion of the taxonomies at conferences and meetings of research communities and committees. Our current approach could be used for automation of updating taxonomies at the situations at which there are too many head subjects, like in the case of "Clustering" cluster in this paper. As to the latter, a reasonable computational progress over penalty weights can be achieved, in our view, by replacing the criterion of maximum parsimony by the criterion of maximum likelihood if each node of the taxonomy can be assigned probabilities of "gain" and "loss" of topic events.

Acknowledgment

D.F. and B.M. acknowledge continuing support by the Academic Fund Program at the National Research University Higher School of Economics (grant 19-04-019 in 2018-2019) and by the International Decision Choice and Analysis Laboratory (DECAN) NRU HSE, in the framework of a subsidy granted to the HSE by the Government of the Russian Federation for the implementation of the the Russian Academic Excellence Project "5-100". S. N. acknowledges the support by FCT/MCTES, NOVA LINCS (UID/CEC/04516/2013)

References

[1] The 2012 ACM Computing Classification System. [Online]. Available:*http://www.acm.org* /about/class/2012 (Retrieved 17 March, 2019).

[2] D. Blei, "Probabilistic topic models,"*Communications of the ACM*, 55 (4), pp. 77–84, 2012.

[3] C. Chen, "Science mapping: A systematic review of the literature", *Journal of Data and Information Science*, vol. 2, no. 2, pp. 140, 2017.

[4] E. Chernyak, "An approach to the problem of annotation of research publications." *Proceedings of the 8th ACM international conference on web search and data mining*, ACM, pp. 429-434, 2015.

[5] D. Frolov, B. Mirkin, S. Nascimento, and T. Fenner, "Finding an appropriate generalization for a fuzzy thematic set in taxonomy", Working paper WP7/2018/04, Moscow, Higher School of Economics Publ. House, 60 p., 2018 (URL: *https://wp.hse.ru*/data/2019/01/13/1146987922/WP7_2018_04_____ .pdf, retrieved 17 March, 2019).

[6] Gene Ontology Consortium, "Gene ontology consortium: going forward", *Nucleic Acids Research*, vol. 43, pp. D1049-D1056, 2015.

[7] R. Klavans and K. W. Boyack, "Which type of citation analysis generates the most accurate taxonomy of scientific and technical knowledge?", *Journal of the Association for Information Science and Technology*, 68(4), pp. 984-998, 2017.

[8] D. Lee, R. Cornet, F. Lau, and N. De Keizer, "A survey of SNOMED CT implementations," *Journal of Biomedical Informatics*, vol. 46, no. 1, pp. 87-96, 2013.

[9] E. Lloret, E. Boldrini, T. Vodolazova, P. Martnez-Barco, R. Munoz, and M. Palomar, "A novel concept-level approach for ultra-concise opinion summarization", *Expert Systems with Applications*, 42(20), pp. 7148-7156, 2015.

[10] B. Mirkin and S. Nascimento, "Additive spectral method for fuzzy cluster analysis of similarity data including community structure and affinity matrices," *Information Sciences*, vol. 183, no. 1, pp. 16-34, 2012.

[11] B. Mirkin, *Clustering: A Data Recovery Approach*, Chapman and Hall/CRC Press, 2012.

[12] G. Mueller and R. Bergmann, "Generalization of Workflows in Process-Oriented Case-Based Reasoning", In FLAIRS Conference, pp. 391-396, 2015.

[13] S. Nascimento, T. Fenner, and B. Mirkin, "Representing research activities in a hierarchical ontology," in *Procs. of 3rd International Workshop on Combinations of Intelligent Methods and Applications* (CIMA 2012), Montpellier, France, August, pp. 23-29, 2012.

[14] Y. Song, S. Liu, X. Liu, and H. Wang, "Automatic taxonomy construction from keywords via scalable bayesian rose trees," In *IEEE Transactions on Knowledge and Data Engineering*, 27(7), pp. 1861-1874, 2015.

[15] N. Vedula, P.K. Nicholson, D. Ajwani, S. Dutta, A. Sala, and S. Parthasarathy, "Enriching Taxonomies With Functional Domain Knowledge," In *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval*, ACM, pp. 745-754, 2018.

[16] J. Waitelonis, C. Exeler, and H. Sack, "Linked data enabled generalized vector space model to improve document retrieval," In *Proceedings of NLP & DBpedia 2015 workshop in conjunction with 14th International Semantic Web Conference (ISWC)*, CEUR-WS, vol. 1486, 2015.

[17] C. Wang, X. He, and A. Zhou, "A Short Survey on Taxonomy Learning from Text Corpora: Issues, Resources and Recent Advances," In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pp. 1190-1203, 2017.

 ISBN: 978-1-61208-707-8

An Integrated Model for Content Management, Presentation, and Targeting

Hans-Werner Sehring
Namics – A Merkle Company
Hamburg, Germany
e-mail: hans-werner.sehring@namics.com

Abstract—**The World Wide Web is the basis for increasingly many information and interaction services. Personalization provides users with information and services that are adequately tailored to their current needs. Targeting, a form of implicit personalization for groups of users, comes to broader practical use for a growing number of commercial websites. The wider adoption results from the availability of platforms that incorporate targeting. Solutions are usually built on top of content management systems used for the production of websites. The definitions required for targeting are related to content, but they are superimposed in the sense that they are not an integral part of the content model or the content itself. This paper presents an initial model that is used to study the integration of models for content, content visualizations, and content targeting. Potential benefits from an integrated model are manifold. It allows expressing personalization rules along with the content they refer to in a consistent way. This way, personalization is applied by putting content in context rather than through superimposed targeting rules. By expressing personalization rules in the same context-dependent and evolvable way as content, they can also evolve over time and can be adapted to different user contexts. On top of that, they can be defined and maintained by content editors and other users of a content management system.**

Keywords-personalization; targeting; segmentation; context-aware content management; content management.

I. INTRODUCTION

The World Wide Web has undergone a tremendous development. For over two decades now, there is research on *personalization* of contents published on the web and of the presentations used for publication.

There is a wide range of personalization approaches for different purposes and goals [1]. These approaches differ in several aspects [2], e.g., in the way personalizations are derived: explicitly by users stating their preferences or implicitly by deriving them from users behavior and habit. An example for explicit personalization are websites that allow the user to name their interests or that allow to individually rearrange parts of the web site. Implicit personalization is achieved, e.g., by observing interactions of a user with a website [3] or by taking previously visited websites into consideration (customer journeys, at best).

Personalization approaches also differ in the subject of the individual adaptations, e.g., content or content representations (visualizations of content created for publication). Content personalization can be found, e.g., in online shops where users receive individual pricing. Content visualizations are personalized by, e.g., using lists of content entries where these lists are ordered in a user-specific way.

Personalization has already been adopted to a range of specific, innovative websites, in particular those that confront the user with large amounts of content. Such websites use it to filter and prioritize content based on assumed user preferences.

Currently, *targeting* is applied by an increasing number of commercial websites. We consider targeting as implicit personalization of content for user groups. The adaptation of content is limited to predefined points, though. Typically, part of the content is selected from building blocks that are prepared for the different user groups.

A set of tools that has emerged during the past years constitutes the basis that allows configuring websites for personalization. Examples are personalization engines built into content management systems and commerce platforms, as well as external personalization services that allow adjusting websites to specific user groups.

There is a lack of models that would cover multiple kinds of personalization approaches [4] and, therefore, allow different usage scenarios to be integrated in one solution.

Typically, commercial products use means of personalization that are superimposed to a (non individualized) base system. A content management system, e.g., allows defining a content model according to which content will be edited, managed, and published. This content model is defined in a uniform way for all users and application scenarios. On a different layer, personalization is added by other means, typically rules that define how to adjust content representations of specific user groups.

Therefore, there is no coherence between content models, content visualization layouts, and personalization rules in such systems. Instead, content has to be defined with all possible audiences and usage scenarios in mind, visualizations have to provide the variations to be offered as personalizations, and personalization rules may be defined on the basis of these definitions.

Contemporary products typically require fixed content models and visualizations (or at least ones that cannot be changed by content editors). This only leaves such personalization rules at the content editors' disposal that can be defined with respect to the possibilities and constraints raised by content models and content visualizations.

The aim of this paper is providing first studies towards a fully integrated model that combines many aspects of

 ISBN: 978-1-61208-707-8

content and its personalized utilization. For this study, we use the Minimalistic Meta Modeling Language (M3L) [5] as a testbed. This language is well-suited for content models since it covers variations and contexts of content in a direct way. Insights into a variety of personalization options originate from previous research on Concept-oriented Content Management [6]. These insights are transferred to M3L models.

Future research will investigate how to employ such integrated models to cover a wide range of personalization approaches and applications. With the help of such models it will be possible to use more than the set of predefined configuration options that contemporary systems exhibit. Instead, these models are expected to unveil personalization capabilities over all aspects of services, their content, and their appearance, as well as to give the possibility of utilizing the interconnections between these.

The rest of this paper is organized as follows. Section II describes targeting approaches typically found in commercial software products. Section III provides a short introduction into the M3L. Section IV presents a first modeling experiment to utilize the M3L for expressing and integrating the common targeting approach into content models for websites. Conclusions and acknowledgement close the paper.

II. TARGETING IN COMMERCIAL PRODUCTS

There is a wide range of approaches to personalization that can be found in the literature and in prototype implementations. In this paper, we constrain ourselves to targeting which is of particular importance for commercial websites. Targeting is a form of implicit personalization of content assembled for presentation with respect to a customer group. The personalization itself is directed by rules set up by content editors.

A. Segment-based Targeting Rules

For targeting, as it is found in many commercial products, users of a web site are assigned *segments*. Segments are categories describing a user's interest or preferences. These are predefined for a particular website (though there are scientific approaches that include deriving segments by, e.g., means of clustering [7]).

The assignment of segments to users is based on *tracking* (or *analytics*) used during web page delivery. By tracking, accesses to web pages are recorded. Depending on the granularity required, interactions on smaller parts than a whole page may be counted [8].

From the web pages visited by a user, her or his interests are derived by collecting the topics covered by those web pages. The web pages considered in this collection could be, e.g, those web pages that have been visited most often, or the web pages for which the visits exceed a given threshold.

The segments assigned to a user (by that time) are used as a parameter to content selection and production of documents from content. This way, content and its representations are personalized for user groups, namely groups consisting of users with the same segments assigned.

B. Related Work

Targeting is found in diverse systems and services, e.g., in Content Management Systems (CMSs), commerce systems, and marketing suites.

1) Personalization Engines in Content Management Systems. Some CMSs have means of segmentation built in. These systems allow equipping content with rules for the selection of content to be included in published web pages based on user segmentation. Like many others, the CMS products of CoreMedia [9] and Sitecore [10] work this way.

2) Superimposed Personalization. Instead of an integrated personalization engine inside a CMS, personalization can alternatively be applied on the basis of an external service. Adobe Target [11] is a prominent representative of this personalization approach.

3) Consideration of Additional Information on Users. Instead of just considering user behavior in the form of web page access profiles, increasingly many applications are also based on explicit customer data. Such data come from, e.g., a Customer Relationship Management (CRM) system, from the history of transactions in a commerce system, from the history of cases in a support system, or from feedback given by means of ratings. Personalization may additionally be based on context information, e.g., the time of day, the device the visitor uses, or some kind of work mode she or he is in [12]. Such context information is partially considered in commercial personalization engines.

III. THE MINIMALISTIC META MODELING LANGUAGE

The *Minimalistic Meta Modeling Language* (*M3L*, pronounced "mel") is a modeling language that is applicable to a range of modeling tasks. It proved particularly useful for context-aware content modeling [13].

For the purpose of this paper, we only introduce the static aspects of the M3L in this section. Dynamic evaluations that are defined by means of different rules are presented in the subsequent section.

The descriptive power of M3L lies in the fact that the formal semantics is rather abstract. There is no fixed domain semantics connected to M3L definitions. There is also no formal distinction between typical conceptual relationships (specialization, instantiation, entity-attribute, aggregation, materialization, contextualization, etc.).

A. Concept Definitions and References

A M3L definition consists of a series of definitions or references. Each definition starts with a previously unused identifier that is introduced by the definition and may end with a period, e.g.:

```
Person.
```

A reference has the same syntax, but it names an identifier that has already been introduced.

We call the entity named by such an identifier a *concept*.

The keyword is introduces an optional reference to a *base concept*, making the newly defined concept a *refinement* of it.

 ISBN: 978-1-61208-707-8

A specialization relationship as known from object-oriented modeling is established between the base concept and the newly defined derived concept. This relationship leads to the concepts defined in the context (see below) of the base concept to be visible in the derived concept.

The keyword is always has to be followed by either a, an, or the. The keywords a and an are synonyms for indicating that a classification allows multiple sub-concepts of the base concept:

```
Peter is a Person. John is a Person.
```

There may be more than one base concept. Base concepts can be enumerated in a comma-separated list:

```
PeterTheEmployee is a Person, an Employee.
```

The keyword the indicates a closed refinement: there may be only one refinement of the base concept (the currently defined one), e.g.:

```
Peter is the FatherOfJohn.
```

Any further refinement of the base concept(s) leads to the redefinition ("unbinding") of the existing refinements.

Statements about already existing concepts lead to their redefinition. For example, the following expressions define the concept Peter in a way equivalent to the above variant:

```
Peter is a Person.
Peter is an Employee.
```

B. Content and Context Definitions

Concept definitions as introduced in the preceding section are valid in a context. Definitions like the ones seen so far add concepts to the top of a tree of contexts. Curly brackets open a new context, e.g.:

```
Person { Name is a String. }
Peter is a Person{"Peter Smith" is the Name.}
Employee { Salary is a Number. }
Programmer is an Employee.
PeterTheEmployee is a Peter, a Programmer {
  30000 is the Salary.
}
```

We call the outer concepts the *context* of the inner, and we call the set of inner concepts the *content* of the outer.

In this example, we assume that concepts **String** and **Number** are already defined. The sub-concepts created in context are unique specializations in that context only.

As indicated above, concepts from the context of a concept are inherited by refinements. For example, Peter inherits the concept **Name** from **Person**.

M3L has visibility rules that correlate to both contexts and refinements. Each context defines a scope in which defined identifiers are valid. Concepts from outer contexts are visible in inner scopes. For example, in the above example the concept **String** is visible in **Person** because it is defined in the topmost scope. **Salary** is visible in **PeterTheEmployee** because it is defined in **Employee** and the context is inherited. **Salary** is not valid in the topmost context and in **Peter**.

C. Contextual Amendments

Concepts can be redefined in contexts. This happens by definitions as those shown above. For example, in the context of **Peter**, the concept **Name** receives a new refinement.

Different aspects of concepts can explicitly be redefined in a context, e.g.:

```
AlternateWorld {
  Peter is a Musician {
    "Peter Miller" is the Name.
  }
}
```

We call a redefinition performed in a context different from that of the original definition a *conceptual amendment*.

In the above example, the contextual variant of Peter in the context of **AlternateWorld** is both a **Person** (initial definition) and a **Musician** (additionally defined). The **Name** of the contextual **Peter** has a different refinement.

A redefinition is valid in the context it is defined in, in sub-contexts, and in the context of refinements of the context (since the redefinition is inherited as part of the content).

D. Concept Narrowing

There are three important relationships between concepts in M3L.

M3L concept definitions are passed along two axes: through visibility along the nested contexts, and through inheritance along the refinement relationships.

A third form of concept relationship, called *narrowing*, is established by dynamic analysis rather than by static definitions like content and refinement.

For a concept c_1 to be a narrowing of a concept c_2, c_1 and c_2 need to have a common ancestor, and they have to have equal content. Equality in this case means that for each content concept of c_2 there needs to be a concept in c_1's content that has an equal name and the same base classes.

For an example, assume definitions like:

```
Person { Sex. Status. }
MarriedFemalePerson is a Person {
  Female is the Sex.
  Married is the Status.
}
MarriedMalePerson is a Person {
  Male is the Sex.
  Married is the Status.
}
```

With these definitions, a concept

```
Mary is a Person {
  Female is the Sex.
  Married is the Status.
}
```

is a narrowing of MarriedFemalePerson, even though it is not a refinement of that concept, and though it introduces separate nested concepts Female and Married.

E. Semantic Rule Definitions

For each concept, one *semantic rule* may be defined.

The syntax for semantic rule definitions is a double turnstile ("|=") followed by a concept definition. A semantic rule follows the content part of a concept definition, if such exists.

A rule's concept definition is not made effective directly, but is used as a prototype for a concept to be created later.

 ISBN: 978-1-61208-707-8

The following example redefines concepts **MarriedFemalePerson** and **MarriedMalePerson**:

```
MarriedFemalePerson is a Person {
  Female is the Sex. Married is the Status.
} |= Wife.
MarriedMalePerson is a Person {
  Male is the Sex. Married is the Status.
} |= Husband.
```

The concepts **Wife** and **Husband** are not added directly, but at the time when the parent concept is evaluated. Evaluation is covered by the subsequent section.

Concepts from semantic rules are created and evaluated in different contexts. The concept is instantiated in the same context in which the concept carrying the rule is defined. The context for the evaluation of a rule (evaluation of the newly instantiated concept, that is) is that of the concept for which the rule was defined.

In the example above, the concept **Wife** is created in the root context and is then further evaluated in the context of **MarriedFemalePerson**.

Rules are passed from one concept to another by means of inheritance. They are passed to a concept from (1) concepts the concept is a narrowing of, and (2) from base classes. Inheritance happens in this order: Only if the concept is not a narrowing of a concept with a semantic rule then rules are passed from base concepts.

For example, **Mary** as defined above evaluates to **Wife**.

F. Syntactic Rule Definitions

Additionally, for each concept one *syntactic rule* may be defined.

Such a rule, like a grammar definition, can be used in two ways: to produce a textual representation from a concept, or to recognize a concept from a textual representation.

A semantic rule consists of a sequence of string literals, concept references, and **the name** expressions that evaluate to the current concept's name.

During evaluation of a syntactic rule, rules of referenced concepts are applied recursively. Concepts without a defined syntactic rule are evaluated to/recognized from their name.

For example, from definitions

```
WordList {
  Word. Remainder is a WordList.
} |- Word §" " Remainder.
OneWordWordList is a WordList |- Word.
Sentence { WordList. } |- WordList "."
HelloWorld is a Sentence {
  Words is the WordList {
    Hello is the Word.
    OneWordWordList is the Remainder {
      World is the Word.
} } }
```

the textual representation `Hello World.` is produced.

Syntactic rule evaluation is not covered in this article.

IV. A Model of Content Personalization

This section provides a first simple M3L model of content, its visualization on web pages, website users, web page accesses, and the targeting of the web pages to the users based on past web page accesses.

```
WebPage.
SegmentingWebPage is a WebPage {
  Topic is a Segment. }
User.
Visit {
  Visitor is a User.
  ViewedPage is a WebPage. }
Segment.
```

Figure 1. Base model for targetable websites.

A. A Web Page and User Behavior Model

Figure 1 shows the essence of a M3L model for a web page, its users, the web page accesses, and segments in which to classify users.

Actual web pages are defined as refinements of the **WebPage** concept. Such concepts contain content as needed and they evaluate syntactically to HTML code for the presentation of that page. Figure 2 shows an example.

A **SegmentingWebPage** has a **Topic** assigned. The topic is represented by a **Segment** (see below).

The **User** concept serves as the identity of a web page visitor. It may contain user data.

A **Visit** records the access of a user to a web page. In real-world applications, typically a tracking tool is used for this purpose.

Segments are used in a twofold manner: On web page accesses, they name the topic of a web page in order to derive the area of interest of a visitor. When delivering the web page in a personalized way, a user's segment is used to select and evaluate personalization rules.

Segments might be managed in a structure like shown in the example. Only the segments themselves are significant.

B. Tracking Web Page Visits

Targeting is based on the users' behavior. Behavior is analyzed by tracking web page accesses. In the example of the M3L model we do so by creating (or finding) a matching **Visit** instance for a web page and user.

If the user is unknown, we create a **User** concept instance at the time of the first request.

```
Teaser.
RessortPage is a SegmentingWebPage {
 Title is a String.
 MainContent is a String.
 NewsTeaser is a Teaser.
} |- §"<html>"...§"</html>"
SoccerOverviewPage is a RessortPage {
 Soccer is the Title.
 "On this page..." is the MainContent.
 Segments{Ressorts{Sports.}} is the Topic.
}
Segments {
 Ressorts {
  Politics is a Segment.
  Sports is a Segment. } }
```

Figure 2. Example of a targetable website.

```
Tracking {
 Score {
  SegmentedUser is a User.
  AssignedSegment is a Segment.
  Value. }
 Visit
 |= Score {
  Visitor is the SegmentedUser.
  ViewedPage { Topic. }
   is the AssignedSegment.
 }
 ScoreUpdate is a Score
 |= Score { 1 is the Value. }
 ScoreIncrement is a ScoreUpdate {
  Value is an Integer. }
 |= ScoreIncrement {
  Integer {
   Value is the Pred.
  } is the Value.
 }
}
```

Figure 3. Base model for tracking.

The assignment of segments to a user is based on the *score* a segment got for a user. This score is the number of visits of a user to web pages with a topic that equals that segment.

In order to measure scores, we introduce the base concept **Integer** with just enough conception in order to have the ability to count. To this end, **Integer**s have a reference **Pred** to their predecessor. Using this reference, the order of integers is defined. The numerical value of an **Integer** is thus the length of the chain of its predecessors. In M3L:

```
Integer { Pred is an Integer. }
0 is an Integer.
1 is an Integer { 0 is the Pred. }
```

The concepts defined in Figure 3 are used to manage scores. The **Value** of a **Score** that a segment has for a user is assigned an **Integer** concept as a refinement. **Visit**s have assigned the user and the visited page.

On every request of a user *u* for a web page *p*, the web server issues a

```
CulturePage17 is a WebPage {
 "Museums and Exhibitions" is the Title.
 ReportOnNewExhibition is the MainContent.
}
Segments { Ressorts {
 Politics {
  CulturePage17 {
  LatestPollResults is a NewsTeaser. } }
 Sports {
  CulturePage17 {
   SoccerExhibition is a NewsTeaser.
   RunningGameScore is a NewsTeaser. } }
} }
```

Figure 4. Example of targeting definitions.

```
SegmentDetermination {
  InitialThreshold is an Integer.
  SegmentsOfUser {
    UTS is a User.
  }
  |= Score { UTS is the SegmentedUser. }
  Score_rec is a Score {
    Value is an Integer.
  } |= Score {
    Value { Pred. } is the Value.
    Threshold { Pred. } is the Threshold.
  }
  IncludedScore is a Score_rec {
    0 is the Threshold.
  } |= AssignedSegment.
  ExcludedScore is a Score_rec {
    0 is the Value.
  } |= Segments.
}
```

Figure 5. Base model for segmentation.

```
Tracking {
  Visit {
    u is the Visitor. p is the ViewedPage.
  } |= Score {
    Visitor is the SegmentedUser.
    p { Topic. } is the AssignedSegment. } }
```

Visit is here extended by a semantic rule in order to represent a function that updates the score of a segment for a user. The concept **Tracking** provides a scope for individual function invocations.

If such a score already exists with the given user and the web page's topic assigned (recognized by **Value** being an **Integer**), then it will be narrowed to the matching **ScoreIncrement**. That concept in return will increase the value by one. This addition is done by setting value to the successor of the current value.

Else, the semantic rule will initialize the score by setting the **Value** to the **Integer 1**.

C. *Applying Targeting Rules*

When users are segmented, the segmentation can be used to create personalized web pages for users.

Figure 4 shows a simple example of a personalizable web page. The **CulturePage17** has a static title and static textual content. It also may contain a list of news teasers that is filled in the context of a user's segment(s). To target web pages to users, each request of a user *u* for a page *p* will lead to an evaluation of *p* in the context of *u*'s segment(s).

The segment(s) of a user typically are derived from the scores they have for that user. In the case of selecting the segments with a certain threshold, the definitions from Figure 5 are used in the selection process.

The highest ranked segments of a user are evaluated inside the concept **SegmentDetermination**, that serves as a scope for executions. The concept **SegmentsOfUser** acts as a function from **User**s to segments with scores above the threshold. That function is invoked in the scope.

 ISBN: 978-1-61208-707-8

The evaluation is based on an **InitialThreshold** that is set inside **SegmentDetermination**. Is is set to the value that has to be reached by scored segments.

The first "invocation" of **SegmentsOfUser** for a user collects all **Score**s of the given user. These scores are then narrowed down during function evaluation. Each iteration of the evaluation starts through the concept **Score_rec** that decreases both **Value** and **Threshold** by one.

If the **Threshold** reaches **0**, then the score is narrowed down to **IncludedScore**. In that case, the value was greater than the threshold. The score is replaced with the segment in this case, thus terminating the recursion.

If the **Value** reaches **0**, however, then the value was less than the threshold. In this case the recursion ends without a specific result by replacing it with **Segments**.

By using the results of the evaluation for the segment contexts used in Figure 4, requests to the sample page **CulturePage17** from a user *u* are targeted to this user:

```
SegmentDetermination {
  SegmentsOfUser { u is the UTS. }
    { CulturePage17. } }
```

At the same time as the targeted web page is derived, a request for a web page will also increment the matching score as defined in the previous subsection. This concludes the circle of segmenting and targeting.

This example just demonstrates the selection of content to display at a given position in a web page, as it is also possible with commercial products. With the approach demonstrated here, however, it will also be possible to personalize other aspects of a web page.

V. SUMMARY AND OUTLOOK

The paper concludes with a summary and an outlook.

A. Summary

Many forms of personalization are discussed in literature for quite some time now. Still, integrated models covering most or all aspects of personalization are missing in practice.

This paper presents a study on such an integrated model, that combines content modeling with personalization, and that allows expressing various forms of personalization.

The initial modeling approach achieves to integrate content, content representation, users, page visits, segments, user segmentation, and targeting "rules". This integration allows coherent definitions of targeted web sites.

B. Outlook.

This paper concentrates on implicit personalization of presentations for groups of users, in practice called targeting.

A next step would be to extend the model to other forms of personalization in order to investigate whether these fit in equally well and can be combined within one model.

Content delivery and consumption depends on the context of the user. The utilization of context information for personalization should fit the models well using the M3L. Still, this needs to be studied.

This paper covers an analysis based on a hypothetical model only. It now needs to be connected to a working web server in order to gain practical results.

To increase practical relevance, further information on users should be integrated into the segmentation process. Such information may come from a Customer Relationship Management (CRM) system, from transaction processing systems like shop solutions, and from customer journeys.

ACKNOWLEDGMENT

Targeting is one building block in many digitization projects. Valuable insights have, therefore, been gained thanks to colleagues, partners, and customers.

Though the model presented in this paper is not related to the work at Namics, the author is thankful to his employer for letting him follow his research ambitions based on experience made in customer projects.

REFERENCES

[1] K. Riemer and C. Totz, "The Many Faces of Personalization: An integrative economic overview of mass customization and personalization," in The Customer Centric Enterprise, M. M. Tseng and F. T. Piller, Eds. Berlin, Heidelberg: Springer, pp. 35-50, 2003.

[2] J. Blom, "Personalization: a Taxonomy," Proc. CHI EA '00 Extended Abstracts on Human Factors in Computing Systems, Apr. 2000, pp. 313-314.

[3] J. Zhang, "The Perils of Behavior-Based Personalization," Marketing Science, vol. 30, pp. 170-186, Dec. 2011.

[4] A. L. Montgomery and M. D. Smith, "Prospects for Personalization on the Internet," Journal of Interactive Marketing, vol. 23, pp. 130-137, Jul. 2008.

[5] H.-W. Sehring, "Content Modeling Based on Concepts in Contexts," Proc. Third International Conference on Creative Content Technologies, Sep. 2011, pp. 18-23.

[6] J. W. Schmidt and H.-W. Sehring, "Conceptual Content Modeling and Management," Perspectives of System Informatics, Lecture Notes in Computer Science, vol. 2890, pp. 469-493, 2003.

[7] T. Jiang and A. Tuzhilin, "Improving Personalization Solutions through Optimal Segmentation of Customer Bases," IEEE Transactions on Knowledge and Data Engineering, vol. 21, pp. 305-320, Mar. 2009.

[8] K. S. Kuppusamy and G. Aghila, "A Model for Web Page Usage Mining Based on Segmentation," International Journal of Computer Science and Information Technologies, vol. 2, pp. 1144-1148, 2011.

[9] Website of the CoreMedia AG. [Online]. Available from: https://www.coremedia.com/ [retrieved: March, 2019]

[10] Sitecore. *Behavioral Targeting and A/B Testing: What's Appropriate for Your Business?* [Online]. Available from: https://www.sitecore.com/de-de/company/blog/416/behavioral-targeting-and-a-b-testing-what-39-s-appropriate-for-your-business-4063 [retrieved: March, 2019]

[11] Adobe Target. *Behavioral targeting*. [Online]. Available from: https://www.adobe.com/marketing/target/behavioral-targeting.html [retrieved: March, 2019]

[12] M. Gorgoglione, C. Palmisano, and A. Tuzhilin, "Personalization in Context: Does Context Matter When Building Personalized Customer Models?" Proc. 6th IEEE International Conference on Data Mining (ICDM 2006), Dec. 2006, pp. 222-231.

[13] H.-W. Sehring, "Context-aware Storage and Retrieval of Digital Content: Database Model and Schema Considerations for Content Persistence," International Journal on Advances in Software, vol. 11, pp. 311-322, Dec. 2018.

 ISBN: 978-1-61208-707-8

Semantically-driven Competitive Intelligence Information Extraction: Linguistic Model and Applications

Iana Atanassova, Gan Jin, Ibrahim Soumana, Peter Greenfield and Sylviane Cardey

CRIT - Tesnière, Université de Bourgogne Franche-Comté
30 rue Mègevand, 25000 Besançon, France
Email: { surname.name } @univ-fcomte.fr

***Abstract*—In a competitive environment and in the current context of rapid technological advances, competitive intelligence is a key strategic need in the private sector and requires the development of Web content tools capable of robust and semantically-driven text classification. In this paper, we present a method for the information extraction and semantic classification of text segments. Our approach to text processing makes use of linguistic clues to populate an ontology of competitive intelligence. We have developed a method for the automatic identification and classification of sentences into predefined semantic classes by using linguistic models and a knowledge-based approach. This method has been tested on a dataset of journal articles in horology and aeronautics, and can be extended to other domains. The tool that we have developed is part of the WebSO+ platform for competitive intelligence. We present the overall methodology for annotation and information extraction, our experimental protocol and the results obtained from the evaluation.**

Keywords–Web content; Information Extraction; Competitive Intelligence; Sentence Classification; Semantic Annotation; Linguistic Model.

I. INTRODUCTION

Today, innovation in the private sector is conditioned by the context of a highly competitive international environment and rapid technological advances. For these reasons, competitive intelligence has become a key strategic need, i.e., companies have to monitor constantly the new technologies in their domains, but also market changes and emerging trends, the activities of competitors and partners, etc. [1]. In addition to this, the amount of information generated daily in each particular domain may lead to information overload. The traditional information sources have been enriched by new media such as social networks and customer opinions on the Web. For these reasons, competitive intelligence is becoming more and more costly in terms of the time and human effort to process the information. For example, [2] analyzes recent practices in European firms and arrives at the conclusion that competitive intelligence "has grown well beyond competitors to include customer related intelligence, technology, market, etc.".

Our work tackles the problem of the development of Web content tools capable of the automatic processing of news articles and user feedback in order to extract and classify specific types of information relevant to competitive intelligence. The major objective is to facilitate and accelerate the task of competitive intelligence in companies by providing tools for the efficient monitoring of both published information sources (e.g., news articles) and also customer feedback. We have developed a method for the automatic identification and classification of sentences into predefined semantic classes by using linguistic models and a knowledge-based approach.

As observed by [3], the majority of commercial applications in information extraction make use of rule-based approaches, while statistical and machine-learning (ML) approaches are widely used in the academic world. Table I gives the proportions of the use of various approaches employed in academic research and in the private sector (vendors of products for Information Extraction) for the implementations of entity extraction. These results were obtained by [3] in a study of conference papers over a 10 year period. Among the reasons for which the rule-based approaches dominate the commercial market is their advantage in terms of flexibility and traceability. The most recent efforts by the research community in this field are directed towards providing standardized rule languages and rule editing tools [4][5].

TABLE I. TYPES OF APPROACHES USED FOR INFORMATION EXTRACTION

	Rule-based (%)	*Hybid (%)*	ML based (%)
Scientific articles	3.5	21	75
Large vendors	67	17	17
All vendors	45	22	33

The problem of text classification in general has been the subject of many studies [6], most of which can be considered as document retrieval tasks in the sense that they work at the level of the document. In our approach, we focus on sentence extraction and sentence classification, which better correspond to the user need, i.e., to identify the exact information related to the competitive intelligence task, rather than retrieving the document that contains such information.

In this paper, we describe our approach to text processing which makes use of linguistic clues to populate an ontology of competitive intelligence. The module for Information extraction that results from this approach has been implemented as part of the WebSO+ platform for competitive intelligence that provides an environment for monitoring various types of information sources. The textual data is obtained by a separate module that scrapes lists of web sites and customer review databases on a daily basis. In this paper, we focus on the automatic classification module that is used for information extraction and allows filtering relevant text segments amongst

 ISBN: 978-1-61208-707-8

the large mass of data retrieved daily, and which are presented to the end user.

The experiment that we report here has been carried out on two specific industrial sectors which are horology and aeronautics. These sectors were chosen because they are well developed and strategically important in the Region of Franche-Comté in France and in Switzerland. The linguistic resources have been developed specifically for the processing of news articles and customer feedback in these two domains. However, our methodology, which relies on a general approach, can be applied to other domains after adapting the linguistic resources to take into consideration the new domains. These resources consist of linguistic rules that follow a grammar specifically designed for this purpose and which make use of lists of regular expressions.

The rest of the paper is organized as follows. In the next section, we present the linguistic model and the overall approach for automatic classification of text segments. In Section III, we give the details of the evaluation that has been carried out on the classification module, and the results. In Section IV, we discuss some of the difficulties in the processing and give examples of errors and possible solutions. Section V presents our conclusion and future work.

II. LINGUISTIC MODEL AND AUTOMATIC CLASSIFICATION METHOD

In this section, we first present the semantic subclasses that form our ontology of competitive intelligence. These subclasses are then used for the classification of sentences and information extraction based on our linguistic model.

A. Ontology of Competitive Intelligence

The objective of our method is firstly to identify sentences that are relevant to the task of competitive intelligence, i.e., that contain explicit information on competitors, market trends, innovations etc. At the same time, we classify these sentences into several subclasses, which are presented in Figure 1, where the names of the subclasses are given in English, and the French translation is in brackets.

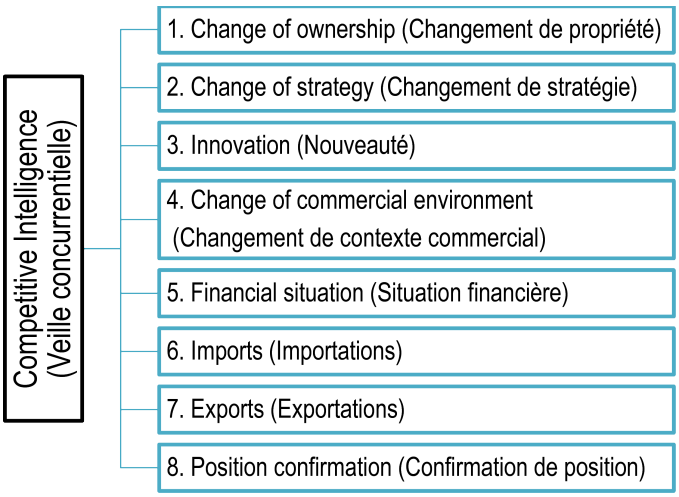


Figure 1. Ontology of Competitive Intelligence: subclasses used for the classification of sentences

Each of the subclasses can be expressed in texts using various expressions and linguistic structures. The table II presents examples of sentences that correspond to some of the subclasses. These examples have been extracted from articles that belong to the corpus described in the following section.

B. Linguistic Model

To develop the linguistic resources, we have analyzed the ways in which the subclasses are expressed in texts by studying a corpus of about 2,000 documents in French of different types: news articles, scientific publications, patents, customer feedback, etc. We then established sets of linguistic structures that are directly implementable and that allow identifying these types of information in new texts. This approach uses the SyGuLAC theory that stems from the microsystemic approach and discrete mathematics [7] in order to propose tools for linguistic analyses and their generalization.

The analysis is considered from the point of view of sense-mining in order to make possible the identification of relevant information in texts. Unlike data-mining systems that search for keywords in a sentence or in a text, in order to identify relevant information, we propose working at the level of sense which is present at all levels of analysis: lexical, syntactic and semantic and their intersections: morpho-syntax, lexico-syntactic-semantics, etc. [8]. However, our aim is not to construct a model that describes the language in its entirety with a global representation of its different levels separately: lexical, syntactic, mophological, semantic. Rather, we concentrate only on one specific objective at a time, which is, in the current study, the identification of information related to competitive intelligence. Thus, in our approach we consider only the elements that are necessary and constitutive of the problem at hand, which can be lexical, morphological, syntactic, etc. in nature. These elements are represented in linguistic structures that are directly implementable in terms of sets of regular expressions or other features that are identifiable in strings [9][10].

C. Implementation

The linguistic structures are represented in our systems as regular expressions following a grammar that was designed specifically for this task. As an example, the structure in Figure 2 represents a part of a subclass ”1. Change of ownership” of textual segments in French in the domain of horology. In this structure, several operators are used, e.g., *Verbe, opt*, that correspond to abstract representations in our model. The linguistic structure uses microsystems that are defined in order to tackle one specific problem. The specifications of the operators and constraints in the grammar are defined as in [7]. Several hundred such structures are associated with each of the above subclasses.

This architecture for information extraction has several advantages:

- the linguistic resources (structures) are independent of the processing model and the implementation of the information extraction engine;
- this methodology can be adapted and used in domains that correspond to precise needs in industry, where machine learning is impossible due to the lack of large scale corpora;

 ISBN: 978-1-61208-707-8

TABLE II. EXAMPLES OF SENTENCES THAT CORRESPOND TO THE SUBCLASSES OF COMPETITIVE INTELLIGENCE

Example	*Class*
Le groupe horloger hispano-suisse Festina a repris les actifs de la société neuchâteloise Technotime.	1.
Alors que les groupes de luxe n'ont eu de cesse ces dernières années de consolider - quand ce n'est pas posséder - leur réseau de distribution, ils comptent désormais sur des blogs ou autres sites spécialisés pour effectuer du e-commerce.	2.
Le suisse Longines présente une montre équipée d'un mouvement à pile nouvelle génération.	3.
Car aujourd'hui, c'est certain, l'amateur n'est plus un collectionneur affectionnant les mécanismes comme dans le passé, mais un adepte averti de la valeur des choses.	4.
L'année 2017 devrait bien se présenter pour Tag Heuer, a noté Jean-Claude Biver, se disant toutefois prudent face aux incertitudes économiques et géopolitiques du monde et tablant sur une croissance "à un chiffre" pour la marque.	5.
Les exportations horlogères suisses ont continué de reculer en février, accusant une baisse de 10% à 1,5 milliard de francs suisses (1,3 milliard d'euros), a annoncé mardi la Fédération de l'industrie horlogère suisse (FH).	7.
"Nous avons consolidé notre quatrième position dans l'horlogerie suisse en matière de chiffre d'affaires, derrière Rolex, Omega et Cartier", souligne celui qui entré chez Longines en 1969 et qui dirige la société depuis 1988.	8.

listEH/opt(Arg1) + Verbe(v-ach) + opt((listEH)/Arg2) + (...) + opt(Ctxt)

Figure 2. Example of a linguistic structure for the subclass "1. Change of ownership"

- the linguistic analysis is based essentially on the structures that are defined by linguists, unlike in other approaches that rely on machine learning of keyword distributions;
- there is a complete traceability of all the linguistic structures involved in some task resolution, which means that the sources of errors can be identified and corrected by modifying the erring structures;
- the modification or the improvement of one linguistic structure can be done independently of the other linguistic structures, which means that the performance of the system can be incremented fairly easily by correcting existing structures to improve the precision or adding new structures to achieve higher recall.

The linguistic analyses take into consideration large context spans in the textual segments. In fact, from a linguistic point of view, we know that the presence of words or particles very far away in the linear representation of a sentence can have a significant impact on the overall meaning. For this reason, the linguistic structures that we use take into consideration the entire sentences. This approach to language modeling presents a considerable advantage for the description of linguistic phenomena compared to other methods that are inspired by the 'bag of words' model.

III. EVALUATION

We have performed an evaluation in order to quantify the capacity of our system to correctly identify and classify sentences that contain information relevant to the task of competitive intelligence for French. In this section, we describe the method that we adopted for this evaluation and the results obtained.

A. Corpus

We have constructed a corpus of news articles in the two sectors of horology and aeronautics in French. The corpus was collected manually by exploring online journals and by using search engines to identify relevant sources and articles on the Web. The sources of the articles are of two types: general newspapers and sites (e.g., Le Monde, Figaro, Le Parisien, Le Point, Les Echos, Le Temps (CH), Tribune de Genève), and specialized magazines and sites in the domains of horology, aviation, new technologies, stock exchange (e.g., www.montres-de-luxe.com, fr.worldtempus.com, www.meltystyle.fr, www.journal-aviation.com, last access 1/3/2019) that are published both in France and in Switzerland at the beginning of 2017. The evaluation corpus contains a total of 45 documents and 1,027 sentences. Table III gives more details of the size of the corpus.

TABLE III. DESCRIPTION OF THE EVALUATION CORPUS

Domain	*Documents*	*Sentences*
Horology	30	745
Aeronautics	15	282
Total	45	1 027

The types of documents and their format in the evaluation corpus are similar to the real-case use for which we have designed the information extraction engine.

B. Evaluation Protocol

We compared the results of the automatic classification with a gold standard obtained by manually classifying the sentences in the corpus. The evaluation was done following the four stages described below.

Stage 1. Automatic segmentation of the corpus into sentences.

Stage 2. Manual classification of all sentences. This was done by students and researchers in the domain of Natural Language Processing. Each sentence in the evaluation corpus was examined and identified as relevant or irrelevant to the task of competitive intelligence. Then, all relevant sentences were assigned one of the 8 subclasses.

Stage 3. Automatic classification of the sentences in the corpus by running our classification module.

Stage 4. Calculation of the precision (P), recall (R) and F-measure (F) [11].

C. Results

Figure 3 shows the number of sentences in the evaluation corpus that were manually classified into each subclass considering the two domains of horology and aeronautics. We observe that the subclass "3. Innovation" contains most of the sentences related to horology, and the subclasses "3. Innovation" and "5. Financial situation" contain the largest sets of sentences in aeronautics.

 ISBN: 978-1-61208-707-8

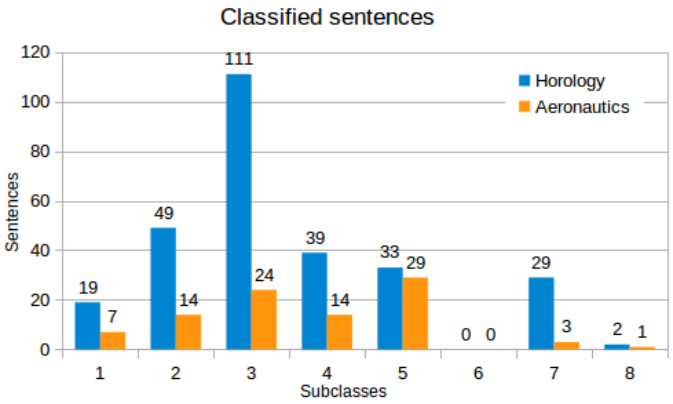


Figure 3. Distribution of the sentences according to the 8 subclasses in the manual classification

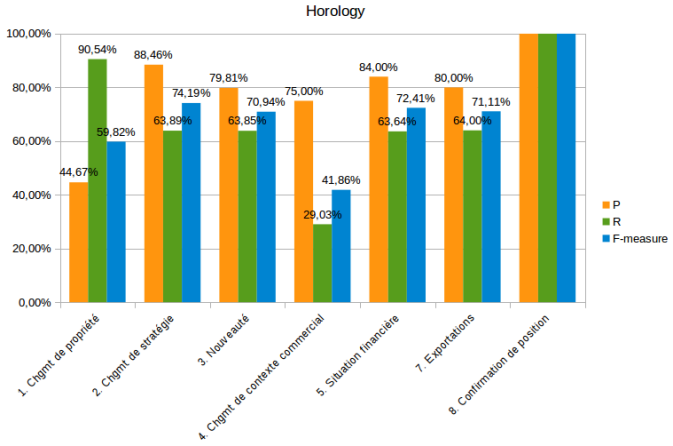


Figure 4. Results of the evaluation in the domain of horology

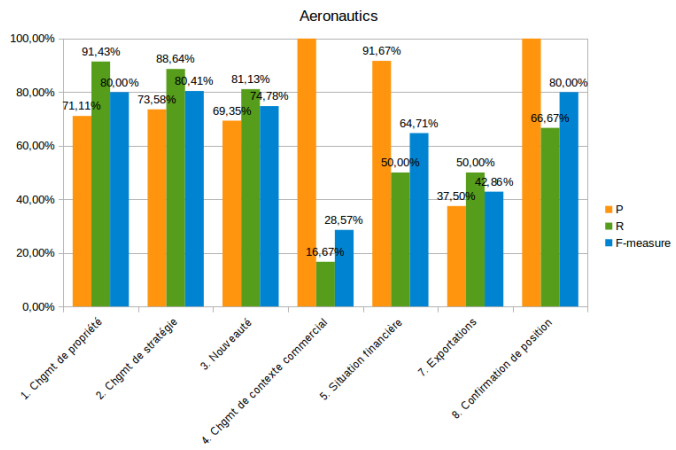


Figure 5. Results of the evaluation in the domains of aeronautics

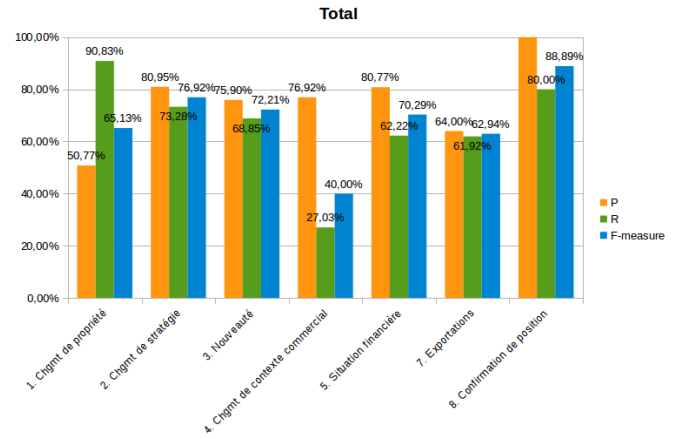


Figure 6. Results of the evaluation in both domains of horology and aeronautics

Figures 4 and 5 present the values of the precision, recall and F-measure calculated by comparing the manually and automatically classified sentences.

In Figure 4, we observe that in horology the values for the precision are high (above 75%) for all subclasses except for ”1. Change of ownership”. This score is due to the fact that this first subclass is complex as it contains different types of sentences that express various kinds of merger-acquisition transactions, as well as liquidations and joint ventures. To improve the system’s performance for this subclass, the corresponding linguistic structures can be identified and corrected. At the same time, the value of the recall for this subclass is above 90%. Considering the subclass ”4. Change of commercial environment”, the values of the recall are low (around 29%) because these kinds of changes can be expressed in many various ways. A larger number of linguistic structures should be considered to improve the coverage.

In Figure 5, the values of the precision are also quite high except for the subclass ”7. Exportations”. In fact, the manual classifications for this subclass show that in some cases a confusion can be made between the subclasses ”7. Exportations” and ”1. Change of ownership” for sentences that express large investments or purchases in aircraft companies. The value of the recall for the subclass ”4. Change of commercial environment” is low and, as in the domain of horology, this subclass needs a more comprehensive set of linguistic structures.

Figure 6 presents the results of the evaluation for both the domains of horology and aeronautics. The overall performance of the system is satisfactory as the majority of the precision and recall values are above 70%. As we have noted, the subclasses of ”1. Change of ownership” and ”4. Change of commercial environment” still need some improvement.

IV. Discussion

The results presented above show that the methodology that we used based on sets of linguistic structures is adequate for the identification and the classification for most of the subclasses. However, improvements can be made for 2 subclasses. In this section, we consider some of the typical errors and discuss the possible improvements that can be made using our methodology.

Table IV gives some typical examples of errors that were extracted from the evaluation corpus. We have analyzed all sources of errors and listed the causes and actions needed

 ISBN: 978-1-61208-707-8

TABLE IV. EXAMPLES OF SENTENCES THAT WERE NOT CORRECTLY IDENTIFIED OR CLASSIFIED BY THE SYSTEM

Example	*Automatic class*
Le siège est installé à La Chaux-de-Fonds, dans le canton de Neuchâtel, à quelques dizaines de kilomètres de la frontière du Doubs ; elle compte trois usines dans le Jura suisse et un centre de recherche à Palo-Alto, en Californie.	1.
Le rythme ultra-cadencé des renouvellements de produits, de caducité des composants et la chute des prix ont ainsi causé la faillite de Pebble, pionnier du secteur.	not identified
L'ascension rapide et le potentiel d' Akrone ont séduit Christophe Courtin, qui, avec son fonds Courtin Investment, vient de prendre 25 de son capital.	not identified
Cela explique, comme le souligne Alain Zimmermann, CEO de Baume & Mercier, pourquoi toutes les marques de luxe recentrent depuis peu leur intérêt sur des modèles à forte identité.	not identified
Et de rappeler que lorsque Tag Heuer a lancé sa montre connectée en 2015, 4 millions d'impressions Internet ont été enregistrées en deux jours.	4.
Ali Nouri – c'est son nom – mise beaucoup sur cette clientèle potentielle et il ne lancera la production de ses premiers modèles, en Chine, une fois seulement qu'il aura engrangé suffisamment de commandes.	3.
Doté d'un calibre Dior Inversé et d'une masse oscillante fonctionnelle visible à l'avant du cadran, le garde-temps évoque par sa structure mécanique les tournoiements d'une délicate robe de grand soir.	not identified
D'autant que les acheteurs, en particulier masculins, se projettent aisément et font de leur montre une sorte de talisman qui les transforme en héros d'une saga qu'ils écrivent dans leur tête.	not identified
Une relance constatée aussi par Swatch Group, dont un tiers de l'activité se fait en Chine.	not identified
L'ascension rapide et le potentiel d' Akrone ont séduit Christophe Courtin, qui, avec son fonds Courtin Investment, vient de prendre 25 de son capital.	not identified

to improve the performance of the system. These can be summarized in the following several points:

- Noise in the automatic classification due to some structures that are "too general" and identify a large number of sentences. These structures can be improved in order to identify only relevant sentences by adding new lists and constraints.
- Silence in the system due to the presence of rare words, expressions or neologisms in some sentences. This problem can be solved by adding new linguistic structures.
- Some sentences make use of figurative language (metaphors, comparisons, etc.) in the news articles. This problem is difficult to tackle in general, but the most frequent cases can be studied and the linguistic structures can be adjusted to take into consideration some of these phenomena.
- Errors related to the sentence segmentation: the limits of the identified segments are of crucial importance for the application of the linguistic structures as they take into consideration entire sentences. In some rare cases the segmentation is not correct and this can be improved.
- Errors due to the use of negation in the sentences: negation in French is expressed in most cases by two words that surround the verb form (*ne ... pas, ne ... aucun, ne ... jamais, ...*). Such cases need more constraints.

Our approach has been developed to respond to the specific needs of competitive intelligence in the private sector, and the choice of the classes addresses these needs. While other methods exist for sentence classification, such as machine learning or neural network approaches using pretrained word embeddings [12], such methods depend heavily on the availability of large annotated datasets that are necessary for the training of the model. Obtaining such datasets with good quality annotations is expensive. To our knowledge, no such datasets exist in the field of competitive intelligence, and therefore the use of the latest deep learning approaches in this contexts is not practically applicable. In the approach that we propose, the major effort is concentrated on the development of the linguistic model for the classifier rather than the manual annotation of a dataset. If the results are satisfactory, this method could be used as a bootstrap process to produce large annotated text corpora that could in turn be used as training datasets for neural network models.

V. Conclusion and Future Work

In this paper, we have presented an overall approach for the automatic classification of sentences based on sets of linguistic structures and its implementation for the task of competitive intelligence. We report on the results of the experimentation on news articles in two specific domains that are horology and aeronautics in French. The same methodology can be adapted to other domains if necessary.

Our classification module is part of a comprehensive platform for competitive intelligence, where the automatic classification helps users rapidly identify relevant information and thus deal with large volumes of data. In a real case scenario, hundreds of sources need to be scanned daily by the user. The capability of the system to highlight automatically classified sentences related to competitive intelligence plays an important role in diminishing the workload and enabling the user to digest ever larger amounts of information.

Our future efforts will be directed in two directions. Firstly, we aim to develop Semantic Web APIs, in order to render the data available through SPARQL requests and build new interfaces. Secondly, this methodology should be applied and evaluated on datasets of articles in other domains, such as the automobile industry and smart cities.

Acknowledgments

Part of this research has been funded by the FEDER (Fonds européen de développement régional) and selected by the French-Swiss programme Interreg V: WebSO+ project.

The authors thank Laurence Gaida and Philippe Payen de la Garanderie, members of the CRIT laboratory of the University of Bourgogne Franche-Comte, for their participation in the evaluation of the linguistic model.

References

[1] C. A. Bulley, K. F. Baku, and M. M. Allan, "Competitive intelligence information: A key business success factor," Journal of Management and Sustainability, vol. 4, no. 2, 2014, p. 82.

[2] J. Calof, R. Arcos, and N. Sewdass, "Competitive intelligence practices of european firms," Technology Analysis & Strategic Management, vol. 0, no. 0, 2017, pp. 1–14.

 ISBN: 978-1-61208-707-8

[3] L. Chiticariu, Y. Li, and F. R. Reiss, “Rule-based information extraction is dead! long live rule-based information extraction systems!” in Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing, 2013, pp. 827–832.

[4] P. Kluegl, M. Toepfer, P.-D. Beck, G. Fette, and F. Puppe, “UIMA Ruta: Rapid development of rule-based information extraction applications,” Natural Language Engineering, vol. 22, no. 1, 2016, pp. 1–40.

[5] W. Wang and K. Stewart, “Spatiotemporal and semantic information extraction from web news reports about natural hazards,” Computers, environment and urban systems, vol. 50, 2015, pp. 30–40.

[6] M. Allahyari et al., “A brief survey of text mining: Classification, clustering and extraction techniques,” arXiv preprint arXiv:1707.02919, 2017.

[7] S. Cardey, Modelling language, ser. Natural Language Processing Series. John Benjamins Publishing Company, 2013.

[8] S. Cardey et al., “A model for a reliable automatic translation, the TACT multilingual system, LISE project (Linguistics and Security),” in Proceedings of WISG’09, Workshop Interdisciplinaire sur la Sécurité Globale, Troyes, France, 2009.

[9] G. Jin, “A system for French-Chinese automatic translation in the domain of global security,” Ph.D. dissertation, University of Franche-Comté, Besançon, France, 2015.

[10] G. Jin, I. Atanassova, I. Souamana, and S. Cardey, “A model for multilingual opinion and sentiment mining,” in Conference TOTh 2017, Terminology & Ontology : Theories and applications, Chambery, France, 2017, pp. 283–287.

[11] C. J. Van Rijsbergen, “Foundation of evaluation,” Journal of Documentation, vol. 30, no. 4, 1974, pp. 365–373.

[12] Y. Zhang, S. Roller, and B. Wallace, “MGNC-CNN: A simple approach to exploiting multiple word embeddings for sentence classification,” arXiv preprint arXiv:1603.00968, 2016.

 ISBN: 978-1-61208-707-8