Profile-based Approach for Age and Gender Identification
Notebook for PAN at CLEF 2016

Ma. José Garciarena Ucelay[1], Ma. Paula Villegas[1], Dario G. Funez[1],
Leticia C. Cagnina[1,2], Marcelo L. Errecalde[1],
Gabriela Ramírez-de-la-Rosa[3], and Esaú Villatoro-Tello[3]

[1] LIDIC Research Group,
Universidad Nacional de San Luis, Argentina
[2] Consejo Nacional de Investigaciones Científicas y Técnicas (CONICET)
{mjgarciarenaucelay,villegasmariapaula74,funezdario}@gmail.com
{lcagnina,merrecalde}@gmail.com
[3] Language and Reasoning Research Group, Information Technologies Dept.,
Universidad Autónoma Metropolitana (UAM) Unidad Cuajimalpa, México
{gramirez,evillatoro}@correo.cua.uam.mx

Abstract. This paper describes the participation between the LIDIC research group of the UNSL from Argentina and the Language and Reasoning research group of the UAM Cuajimalpa from Mexico at the PAN's 2016 Author Profiling task. For the proposed method we adopted a profile-based approach, which has been successfully applied in the Authorship Attribution problem. Thus, we proposed a variation of this technique for tackling the Author Profiling task. Performed experiments showed that using about 8000 most frequent character n-grams for the construction of the different profiles, our proposed method obtains a better performance for both the same genre of documents as well as for the cross-genre scenario.

Keywords: Profile-based approach, Author Profiling, Natural Language Processing.

1 Introduction

Lately, the Author Profiling (AP) task is among the challenges that has been very attractive for the scientific community, specially for fields such as Natural Language Processing, Forensics, Marketing, and Internet Security. As known, the main goal of the AP is to distinguish, from a given text, among different authors' categories and not to identify the author itself; the latter is known as Authorship Attribution [1]. Thus, the AP task aims at modelling, through more general set of features, groups of authors. Ideally speaking, such features will represent, to some extent, how different categories of authors employ their language depending on its age, gender, native language, political preference, personality, etc. [2].

One of the very first works on facing the problem of AP are [2,3], where it was shown the pertinence of statistical techniques for distinguishing among authors' gender and age. Since then, many approaches have been proposed for facing the AP challenge [3,4,5,6,7]. A common approach among these research works is the use of textual representations, which have shown being effective enough when the revised documents represent formal texts, for instance, news reports, scientific papers, books, etc. Nonetheless, most of traditional approaches face several difficulties when provided documents are from a more informal source, such as blogs, chats, or social media texts (*e.g.*, tweets).

As part of the efforts in providing effective solutions to the AP challenge, the PAN@CLEF[4] proposes a competitive evaluation exercise for uncovering plagiarism, authorship, and social software misuse. For this year PAN campaign the focus of AP shared task is on cross-genre age and gender identification [8], meaning that, the training documents will be on one genre (*e.g.* Twitter, blogs, social media, etc.) and the evaluation will be on a different one.

The rest of this document is organized as follows, Section 2 describes some of the most relevant research works that have tried to solve the problem of AP with a profile-based paradigm. Section 3 describes the ideas that motivate this work. Next, Section 4 describes our proposed method for approaching the AP problem and, Section 5 shows the obtained results on the PAN 2016 dataset. Finally, Section 6 depicts our future work ideas and the obtained conclusions.

2 Related work

In the field of Author Analysis, there are several tasks that fall under the same type of stylistic analysis; these tasks are Author Attribution, Plagiarism Detection and Author Profiling. In the Author Attribution problem, there are two predominant paradigms: instance-based paradigm and profile-based paradigm. The former is the common one and also is the most used in the other related tasks of Author Analysis; this paradigm assumes each document of an author as independent. However, the profile-based paradigm, in which all the documents for the same author are treated as one, despite its simplicity is not very common.

The most recent research that uses the profile-based paradigm is the one proposed by Potha and Stamatatos [9]. They evaluated the profile-based paradigm for the author attribution task and tested the paradigm against methods that use an instance-based paradigm from the PAN-2013 participants. The authors established four parameters for their method, such as the length of the n-grams, the length of the unknown document, the length of the profile and the dissimilarity function. Results showed that their method, using a set of global and local settings, outperforms single methods from the participants of PAN 2013 for the author authorship track.

Another researches, also for Author Attribution, use hybrid approaches, that is, some characteristics are taken from both paradigms (i.e., instance-based and

[4] `http://pan.webis.de/clef16/pan16-web/`

profile-based) [10,11]. In these researches the authors use each document for each author as independent in the same way the instance-based paradigm does, but a profile is built for each author.

As the previous works show, profile-based approaches have been given competitive results for author attribution tasks. In this sense, we want to test this simple approach in another author analysis problem, *i.e.*, author profiling task. As in [9], we set some parameters such as the length of the profile and the length of the n-grams in an cross-domain scenario.

3 Profile based approaches

Profile-based methods have been successfully used for addressing problems related to the authorship attribution (AA) task [1]. In a typical AA problem, a text of unknown authorship is assigned to a candidate author, given a set of candidate authors for which we have available texts of undisputed authorship. In this context, for each class of author these methods build a profile containing information extracted from a collection of documents written by the author [12]. Figure 1 summarizes graphically the process of generating the profiles of each author.

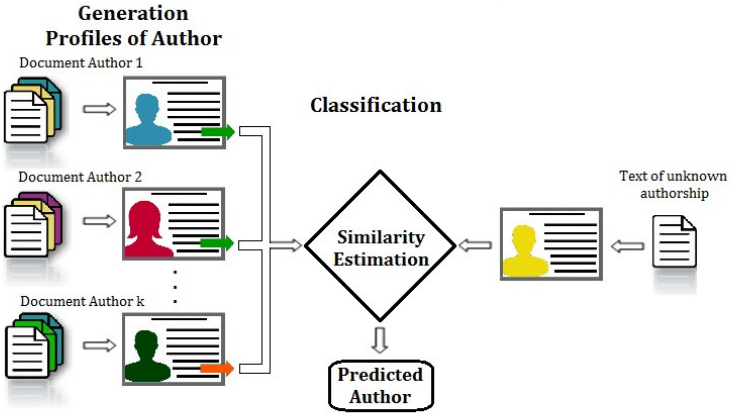

Fig. 1. Generation of profiles process (left) and classification of a test document (right) for AA task.

The information extracted from the documents for the construction of the profiles can be related to the writing style or the text content as we briefly describe below.

- **Style-based features**: such as frequency or number of pronouns, articles and prepositions, number of hyperlinks, words average, etc. [13]. One of the

most used is the frequencies of n-grams of characters. The n-grams are substrings of n consecutive characters [14]. In particular for English language, n-grams of characters with n=3 have demonstrated to be effective. These features capture interesting information depending on the gender and the age of the author. For example, women in blogs use more pronouns and affirmative-negative words.
- **Content-based features**: consider the words related to different topics [13]. For example, the women usually write words related to personal concerns such as shopping, mom, etc. Instead, the men usually write about politic and technology.

In order to obtain the author profiles, these methods consider a set of documents of each author and extract the set of features. As the set could be too large, the profile will consider only the L more frequent features from the whole set. Then, before classifying a target document, the method will construct a profile with that unique document and using a similarity measure with respect to all authors' profiles, it will determine the authorship [15].

Some similarity (or distance) measures used in the profile-based approaches are:

1. Kešelj's Relative Distance (KRD) [16]: calculates the distance K between two profiles P_1 and P_2 as:

$$K = \sum_{x \in X_{P_1} \cup X_{P_2}} \left(\frac{2 \times (P_1(x) - P_2(x))}{P_1(x) + P_2(x)} \right)^2 \tag{1}$$

where $P_i(x)$ is the frequency of the term x in profile P_i, and X_{Pi} is the set of all terms that occur in the profile P_i.

2. Simplified Profile Intersection (SPI) [17]: calculates the amount of features that belong to both profiles P_1 and P_2 as:

$$K = \sum_{x \in X_{P_1} \cap X_{P_2}} \left(\frac{2 \times (P_1(x) - P_2(x))}{P_1(x) + P_2(x)} \right)^2 \tag{2}$$

As profile-based approaches have been successfully used for the AA task, we propose to use these for the Author Profiling task.

4 *Sistema de Perfiles*: the proposed method

Our study focuses on predicting the age and gender of the author (female or male), for the languages English, Spanish and Dutch. For the age, the task considers the following ranges of ages: 18-24, 25-34, 35-49, 50-64 and 65-xx years old, only for the English and Spanish texts [18].

In order to use a profile-based approach, we represent a specific *class* of author with a profile. Then, for predicting the gender and the age, we made 10 different profiles which comprise information combined about the possible gender and age of the authors. Thus, we obtained profiles for the following categories: *female_18-24*, *male_18-24*, *female_25-34*, *male_25-34*, *female_35-49*, *male_35-49*, *female_50-64*, *male_50-64*, *female_65-XX* and *male_65-XX*.

Regarding the features for the construction of the profiles, preliminary experiments showed that the use of character n-grams were adequate. The complete system named *Sistema de Perfiles* (SP) was implemented in two stages. In the first one we constructed the profiles for each category for each language separately. We used the documents (*i.e.*, training set) provided by Author Profiling task at PAN-PC-2016 [8]. To getting the profiles of each category (each language separately) we applied the following steps considering all the training set:

- Unification of each separate xml files in a single txt file (concatenation). One for category.
- Preprocessing of the txt file obtained for each category: tags and images are removed.
- Generation of the n-grams using the txt file and calculate the frequencies of each one. Sort the n-grams considering those most frequent at first[5]. This step is performed for each category.
- Save the profile of the category considering only the L most frequent n-grams obtained in the previous step.

The second stage is the classification of a test document in a particular language (this information is provided). SP receives an input xml file then, the following steps are performed:

- Preprocessing of the input file: tags and images are removed of the file and it is saved as a txt file.
- Obtaining the n-grams and sorting those considering only the L most frequent (profile document).
- Check for similarity with the profiles of each category using the SPI function described above. It compares the profile document with the corresponding to each category returning the label of that which is closer. Take into account that the profiles considered in this step are those with similar language of the input file.

5 Experiments and results

5.1 Intra-Domain Study

We first studied the performance of SP in a intra-domain experiments. Regarding the parameter L of SP, we consider that choosing an appropriate value is important to achieve a correct balance between an acceptable execution time and a good percentage of instances correctly classified. Moreover, if the L value

[5] We used the library *Morphadorner* for this step, which is an open-access Java library for NLP supplied by the Northwestern University.

Table 1. Accuracy obtained in intra-domain classification by gender using SP for Dutch (**DU**), English (**EN**) and Spanish (**SPa**) languages.

L	**3**			**3+4**			**3-5**			**4**			**4+5**			**5**		
	DU	**EN**	**SPa**	**DU**	**EN**	**SPa**	**DU**	**EN**	**SPa**	**DU**	**EN**	**SPa**	**DU**	**EN**	**SPa**	**DU**	**EN**	**SPa**
2000	64	65	71	59	60	62	57	60	60	68	67	74	60	61	60	63	67	62
4000	77	74	74	65	72	68	61	61	60	65	69	80	61	70	62	65	72	77
6000	67	68	74	65	70	80	61	68	65	69	66	77	68	67	68	77	72	85
8000	61	54	77	72	70	82	67	66	71	65	65	85	65	68	77	69	68	85

Table 2. Accuracy obtained in intra-domain classification by age using SP for English (**EN**) and Spanish (**SPa**) languages.

L	**2**		**3+4**		**3-5**		**4**		**4+5**		**5**	
	EN	**SPa**	**EN**	**SPa**	**EN**	**SPa**	**EN**	**SPa**	**EN**	**SPa**	**EN**	**SPa**
2000	37	77	32	48	36	34	43	51	37	42	41	51
4000	38	82	41	60	38	57	44	82	38	48	46	77
6000	38	77	39	77	40	54	44	85	40	74	43	80
8000	38	85	44	77	41	74	46	80	46	80	41	82

is small it occurs an *underfitting*. On the contrary, if the L value is excessively large, SP can generate an *overfitting* of the classification. This is because the generated profiles would be adjusted too much over the corpus used for training.

Then, we carried out some preliminary intra-domain experiments, using only the training corpus provided by PAN 2016 competition. Although the competition stated that Author Profiling task would focus on cross-genre age and gender identification, we believed convenient to try different values of L using the same corpus for both, training and testing. PAN 2016 corpus consists of 436 documents written in English, 250 in Spanish and 384 in Dutch language. We splitted this collection taking the 80% to train, and leaving the remaining 20% to test.

Tables 1 and 2 show the results of experiments for gender in Dutch, English and Spanish languages, as well as for age in the case of the latter two. We consider the percentage obtained of correctly classified instances, in other words, the accuracy as a measure of performance. Rows of Tables 1 and 2 indicate the different values for L (from 2000 to 8000) and columns point out different models of representation, that is, only 3-grams of characters or the combination from 3-grams to 5-grams, and so on.

We can observe that, in general, the best values of accuracy were reached when L was 4000 and 3-grams were utilized. In some cases, 5-grams work similarly to the use of 3-grams, but the reason for choosing the latter was given by the time incurred in the execution. Building the profiles based on 5-grams took twice as long as the construction of the profiles based on 3-grams.

5.2 Cross-genre Study

As we mentioned before, this years PAN Author Profiling task was stated as cross-genre classification [8]. In this context, “genre” refers to the type of source

Table 3. Baseline (accuracy) obtained in cross-genre classification by age and gender using Naïve Bayes, tf-idf word representation, PAN2016-training corpus to train and for testing we used PAN-2014 (social-media (sm) and blogs (blg)) and PAN2015 twitter corpora.

	PAN2014				**PAN2015**		
	English		**Spanish**		**English**	**Spanish**	**Dutch**
	sm	blg	sm	blg	twitter	twitter	twitter
Gender	50	50	52	59	53	52	71
Age	25	35	25	24	33	30	-

Table 4. Accuracy of SP for cross-genre classifications by age and gender in English. PAN-2016 training corpus to make the profiles and PAN-2014 sub-corpora of social media (sm) and blogs (blg) to test.

L	**500**		**2000**		**4000**		**6000**		**8000**		**10000**	
	sm	blg	sm	blg	sm	blg	sm	blg	sm	blg	sm	blg
Gender	50	54	51	54	50	51	51	54	50	58	51	57
Age	26	15	28	15	24	6	24	38	28	43	24	42

from which the texts proceed, for example, Twitter, blogs and social media. For the experimentation we constructed the profiles for SP from the complete training corpus provided by the competition at PAN-2016.

In order to test our SP method in a cross-genre scenario, we used two different corpus: a representative subset of the collection supplied by the competition in PAN-2014 [19], and the complete corpus of PAN-2015 competition [20]. For the former collection we only considered the texts obtained from blogs and social media, both in Spanish and English languages. For the latter test collection we used all the available texts, which were obtained from Twitter; for Dutch we only evaluated the gender identification problem.

At first, we obtained a general baseline in order to have values to compare with. Thus, using the training and test sets mentioned in the previous paragraph, with the Naïve Bayes classifier and the tf-idf word representation, we reached the results shown in Table 3.

The results obtained with our SP method are shown in Table 4 and Table 5. We show the accuracy obtained for classification by gender and age, with different L values using 3-grams. The results in both tables correspond to the PAN-2014 collection to test for English and Spanish language. As we can see, SP with L=8000 achieves in the most of the cases, the highest percentage of classification (over the baseline).

Table 6 shows the accuracy obtained using PAN-2015 collection for testing with different L values. Although there is not a L value which is the best in all languages for both age and gender, we can conclude that L=8000 still per-

Table 5. Accuracy of SP for cross-genre classifications by age and gender in Spanish. PAN-2016 training corpus to make the profiles and PAN-2014 sub-corpora of social media (sm) and blogs (blg) to test.

L	**500**		**2000**		**4000**		**6000**		**8000**		**10000**	
	sm	blg	sm	blg	sm	blg	sm	blg	sm	blg	sm	blg
Gender	48	34	50	49	55	56	61	60	58	60	58	59
Age	12	13	12	30	29	27	28	38	30	45	27	43

Table 6. Accuracy of SP for cross-genre classifications by age and gender. PAN-2016 training corpus to make the profiles and PAN-2015 to test for English (**EN**), Spanish (**Spa**) and Dutch (**DU**) language.

L	**500**			**2000**			**4000**			**6000**			**8000**			**10000**		
	EN	**SPa**	**DU**	**EN**	**SPa**	**DU**	**EN**	**SPa**	**DU**	**EN**	**SPa**	**DU**	**EN**	**SPa**	**DU**	**EN**	**SPa**	**DU**
Gender	52	52	56	58	57	65	59	65	65	67	57	62	64	62	76	64	61	65
Age	42	35	-	45	43	-	37	40	-	32	44	-	32	44	-	34	43	-

forming well in the most of the cases. In fact for Dutch language (the only experimentation performed) with this value of L, SP obtained the best result.

Finally, for simplicity, we have set, for all categories and all languages, our SP system with L=8000 and as a similarity measure the SPI metric for the final submission in the PAN competition. This decision was determined based on the averages of the results obtained and shown in the tables above. All the experiments were run using the TIRA platform [21,22].

Figure 2 summarizes the obtained performance of our system when it is tested with different corpora using the PAN-2016 data set for building the profiles. It is worth noting that in all considered cases (PAN-2014 and PAN-2015) the accuracy values are good when L=8000.

6 Conclusions and future work

This paper described the joint participation of the LIDIC research group of the UNSL from Argentina and the LyR research group of the UAM Cuajimalpa from Mexico at the PAN-2016 Author Profiling task.

We presented a profile-based method for the Author Profiling task. Our proposal uses profiles of character 3-grams for representing information about the different categories of authors. We performed experiments in intra and cross genre scenarios and we showed that using the 8000 most frequent character 3-grams, our method obtains the best performance of classification for genre and age.

In future works we plan to test different features for the construction of the profiles and the use of different similarity measures for comparing the profiles.

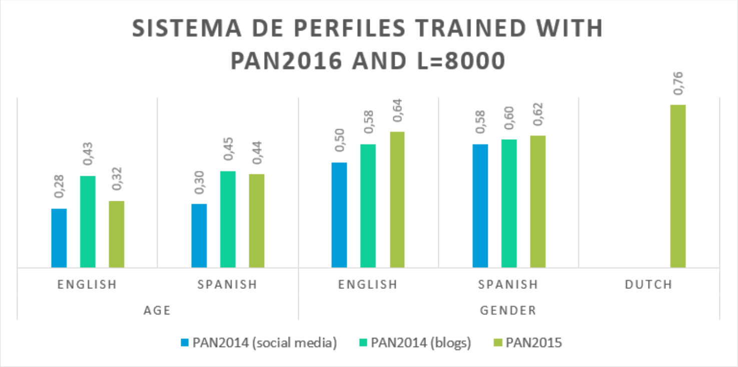

Fig. 2. Summary of the performance of the *Sistema de Perfiles* submitted to the PAN-2016 competition. For the submitted system L was set to 8000 and the SPI metric was used as similarity measure.

Acknowledgments. This work was partially funded by CONACyT under the Thematic Networks program (Language Technologies Thematic Network projects 260178, 271622). We also thank to UAM Cuajimalpa, CONACyT (Project grant number 258588) and SNI-CONACyT for their support.

References

1. E. Stamatatos, “A survey of modern authorship attribution methods,” *J. Am. Soc. Inf. Sci. Technol.*, vol. 60, pp. 538–556, Mar. 2009.
2. S. Argamon, M. Koppel, J. Fine, and A. R. Shimoni, “Gender, genre, and writing style in formal written texts,” *TEXT*, vol. 23, pp. 321–346, 2003.
3. M. Koppel, S. Argamon, and A. R. Shimoni, “Automatically categorizing written texts by author gender,” *Literary and Linguistic Computing*, vol. 17, no. 4, pp. 401–412, 2002.
4. J. D. Burger, J. Henderson, G. Kim, and G. Zarrella, “Discriminating gender on twitter,” in *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, EMNLP ’11, (Stroudsburg, PA, USA), pp. 1301–1309, Association for Computational Linguistics, 2011.
5. C. Peersman, W. Daelemans, and L. Van Vaerenbergh, “Predicting age and gender in online social networks,” in *Proceedings of the 3rd International Workshop on Search and Mining User-generated Contents*, SMUC ’11, (New York, NY, USA), pp. 37–44, ACM, 2011.
6. D. Nguyen, R. Gravel, D. Trieschnigg, and T. Meder, “How old do you think i am?; a study of language and age in twitter,” in *Proceedings of the Seventh International AAAI Conference on Weblogs and Social Media*, AAAI Press, 2013. Reporting year: 2013.
7. A. P. López-Monroy, M. M. y Gómez, H. J. Escalante, L. Villaseñor-Pineda, and E. Stamatatos, “Discriminative subprofile-specific representations for author profiling in social media,” *Knowledge-Based Systems*, vol. 89, pp. 134 – 147, 2015.

8. F. Rangel, P. Rosso, B. Verhoeven, W. Daelemans, M. Potthast, and B. Stein, "Overview of the 4th Author Profiling Task at PAN 2016: Cross-genre Evaluations," in *Working Notes Papers of the CLEF 2016 Evaluation Labs*, CEUR Workshop Proceedings, CLEF and CEUR-WS.org, Sept. 2016.
9. N. Potha and E. Stamatatos, *Artificial Intelligence: Methods and Applications: 8th Hellenic Conference on AI, SETN 2014, Ioannina, Greece, May 15-17, 2014. Proceedings*, ch. A Profile-Based Method for Authorship Verification, pp. 313–326. Cham: Springer International Publishing, 2014.
10. H. V. Halteren, "Author verification by linguistic profiling: An exploration of the parameter space," *ACM Trans. Speech Lang. Process.*, vol. 4, pp. 1:1–1:17, Feb. 2007.
11. J. Grieve, "Quantitative authorship attribution: An evaluation of techniques," *Literary and Linguistic Computing*, vol. 22, no. 3, pp. 251–270, 2007.
12. H. J. Escalante, M. M. y Gómez, and T. Solorio, "A weighted profile intersection measure for profile-based authorship attribution," in *Proceedings of MICAI 2011*, vol. 7094, pp. 232–243, 2011.
13. J. Schler, M. Koppel, S. Argamon, and J. W. Pennebaker, "Effects of age and gender on blogging," in *AAAI Spring Symposium: Computational Approaches to Analyzing Weblogs*, pp. 199–205, 2006.
14. W. B. Cavnar and J. M. Trenkle, "N-gram-based text categorization," in *Proceedings of SDAIR-94, 3rd Annual Symposium on Document Analysis and Information Retrieval*, pp. 161–175, 1994.
15. R. Layton, P. Watters, and R. Dazeley, "Recentred local profiles for authorship attribution," *Natural Language Engineering*, vol. 18, pp. 293–312, 2012.
16. V. Kešelj, F. Peng, N. Cercone, and C. Thomas, "N-gram-based author profiles for authorship attribution," *Proceedings of the conference pacific association for computational linguistics, PACLING*, vol. 3, pp. 255–264, 2003.
17. G. Frantzeskou, E. Stamatatos, S. Gritzalis, and S. Katsikas, "Source code author identification based on n-gram author profiles," in *Artificial Intelligence Applications and Innovations*, vol. 204 of *IFIP*, pp. 508–515, Springer US, 2006.
18. "9th evaluation lab on uncovering plagiarism, authorship, and social software misuse (PAN 2013)." `http://pan.webis.de/`, 2013.
19. F. Rangel, P. Rosso, I. Chugur, M. Potthast, M. Trenkmann, B. Stein, B. Verhoeven, and W. Daelemans, "Overview of the 2nd Author Profiling Task at PAN 2014," in *CLEF 2014 Evaluation Labs and Workshop*, pp. 15–18, CEUR-WS.org, 2014.
20. F. Rangel, F. Celli, P. Rosso, M. Potthast, B. Stein, and W. Daelemans, "Overview of the 3rd Author Profiling Task at PAN 2015," in *CLEF 2015 Evaluation Labs and Workshop*, pp. 8–11, CEUR-WS.org, 2015.
21. T. Gollub, B. Stein, S. Burrows, and D. Hoppe, "TIRA: Configuring, Executing, and Disseminating Information Retrieval Experiments," in *9th International Workshop on Text-based Information Retrieval (TIR 12) at DEXA* (A. Tjoa, S. Liddle, K.-D. Schewe, and X. Zhou, eds.), (Los Alamitos, California), pp. 151–155, IEEE, Sept. 2012.
22. M. Potthast, T. Gollub, F. Rangel, P. Rosso, E. Stamatatos, and B. Stein, "Improving the Reproducibility of PAN's Shared Tasks: Plagiarism Detection, Author Identification, and Author Profiling," in *Information Access Evaluation meets Multilinguality, Multimodality, and Visualization. 5th International Conference of the CLEF Initiative (CLEF 14)* (E. Kanoulas, M. Lupu, P. Clough, M. Sanderson, M. Hall, A. Hanbury, and E. Toms, eds.), (Berlin Heidelberg New York), pp. 268–299, Springer, Sept. 2014.