Analyzing Email Patterns with Timelines on Researcher Data

Jangwon Gim[1], Yunji Jang[1], Do-Heon Jeong[1,*], Hanmin Jung[1]

[1] Korea Institute of Science and Technology Information (KISTI)
245 Daehak-ro, Yuseong-gu, Daejeon (305-806), South Korea
{jangwon, yunji, heon, jhm}@kisti.re.kr

Abstract. This paper proposes a procedure that easily extracts a feature that helps differentiate between similar researcher names in articles. We examined email patterns and their timelines to identify researchers. Our statistical analysis results show multiple email address usage patterns are found in the case of approximately 43% researchers, and 5% of the patterns are overlapped. Base on the statistics, we conclude that the identification of researchers is still required to enhance performance of the researcher-centric analytics systems and applications.

Keywords: researcher name disambiguation, feature selection, researcher data set, timeline

1 Introduction

With ever-increasing amounts of research data and advancements in technology in big-data environments, a paradigm shift is required. Accordingly, studies on new business intelligence services are being conducted and forecasting and analysis methods are being developed. Prescriptive analytics first appeared in 2013 among several analytical methods and offers diverse strategies for achieving the objectives of and improving business competence. The 2014 Gartner Hype Cycle Special Report predicted that prescriptive analytics will advance rapidly and reach a technology maturity stage within the next ten years[1)]. InSciTe Advisory is a service developed in 2013 for strengthening researcher research skills by using the 5W1H method with prescriptive analytics [1,2]. The service analyzes a researcher's skill set and provides analytical results by means of the 5W1H method in order to assist a researcher in attaining a role model group. However, exact diagnosis and analysis of researchers is required to provide them with an optimum strategy for reaching their research goals. To achieve this objective, a researcher's basic information as well their research data must be collected completely in order to examine research results and identify fields of study. This ensures that the researcher is properly identified. For example, a

* Co-responding author

1) http://www.comworld.co.kr/news/articleView.html?idxno=48181

researcher's research information is often confused with that of other researchers and thus retrieved together. This happens because of similar full or abbreviated names of researchers. Accurately identifying a researcher is thus difficult. If research results are integrated without accurate identification of the researcher in question, analysis of this researcher and his or her studies can be either overestimated or underestimated. In this study, we propose an accurate data acquisition procedure to properly identify researchers. Our proposed researcher identification method extracts researchers' email usage and timeline patterns. The structure of this paper is as follows. We discuss related studies in Section 2 and describe the feature selection procedure in Section 3. In Section 4, we present and analyze test data and results. Section 5 concludes our study and states avenues for our future research.

2 Related Work

The amount of academic literature published on the World Wide Web is ever-growing [3]. In this environment, researchers spend much time analyzing the following: research fields that are growing rapidly, well-known academic literature in specific research fields, and authors and their work that are most pertinent to their own research. Accordingly, researcher competence strengthening services that provide researchers with the most relevant and desirable information are being widely developed. These services help to ensure accuracy of researcher data and thus a researcher's credibility. Proper identification of researcher data is critical and many studies are being conducted in this area. Such studies on the accurate identification of research data have been published in databases such as DBLP and PubMed, which are popular sites for reviewing and collecting high quantities of researcher data [4,5].

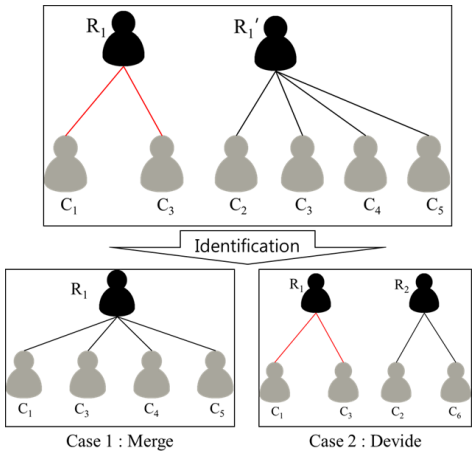

Fig. 1 An example of researchers who has the same name

Figure 1 shows an example of researchers who can have the same name. The authors might be divied two different researchers or might be merged as a researcher.

Ensuring the accuracy of classification is difficult when the network automatically classifies Researcher 1 as the same person associated with data collected on specific research papers. Therefore, methods for automatically identifying researchers are necessary when largescale literature data is considered. In addition, a correct answer set with high accuracy and an experimental data set are required when researchers conduct studies based on researcher data. To achieve this, accurate identification of a researcher is required for certain works which are part of researcher data. Therefore, currently operating authentication services such as Elsevier SciVal Expert and ORCID are designed so that researchers can provide relevant information directly and manage it by themselves [6,7]. The accuracy of researcher information is improved through these services. However, researcher identification remains a problem when we try to integrate the data provided by these services with previously published data.

Therefore, studies on researcher identification based on researcher meta information extracted from papers and scientific literature data have been proposed. Studies exist that examine the use of researcher email and affiliation information. The study in [8] examines the email content of specific researchers and extracts names for identification purposes. The similarity of names is identified by examining the extracted name and a sentence containing that name. The researchers in [8] performed identification based on email contents, but they did not consider characteristics of email addresses themselves such as character strings. The study in [9] tried to solve disambiguation problems related to author names on the basis of researcher affiliation information. To accomplish this, it proposed the pairwise factor graph (PFG) method. This method generates pairs by randomly combining two papers a researcher has published and attempts to identify the researcher based on similarity information. In addition, it examines the distribution of atomic clusters by using the pairs to compare co-authors with the researcher, affiliation names, and titles of papers. However, identifying the exact author is difficult when another or several researchers exist who have the same name. Labeling Oriented Author Disambiguation (LOAD) method using a machine learning algorithm [10]. In LOAD, data was clustered with Precision Clusters (HPCs) and High Recall Clusters (HRCs). It clustered meta information, which can be extracted from each paper, including email and affiliation, and distinguished a different person with the same name by clustering papers by each author based on HPC. Comparing it to the existing automatic homonymy algorithm, LOAD improved the accuracy of disambiguation issues and can save a time for a human to label a specific cluster. However, it does not consider the timeline information of the features. One of the most important factors for identifying a researcher in researcher data is timeline information [11]. Email address and affiliation information of a researcher can be changed, added, or deleted. Therefore, a researcher's activity history can be tracked if timeline information is used for identifying a researcher [12].

In this paper, we introduce the extracting procedure for certain features, such as email address and affiliation name, which can play an important role in identifying a researcher. Further, we propose an analysis method based on the timeline, and state our experimental results.

3 Feature Selection

3.1 A procedure for selecting email patterns

This chapter explains the feature extraction procedure from researcher data.

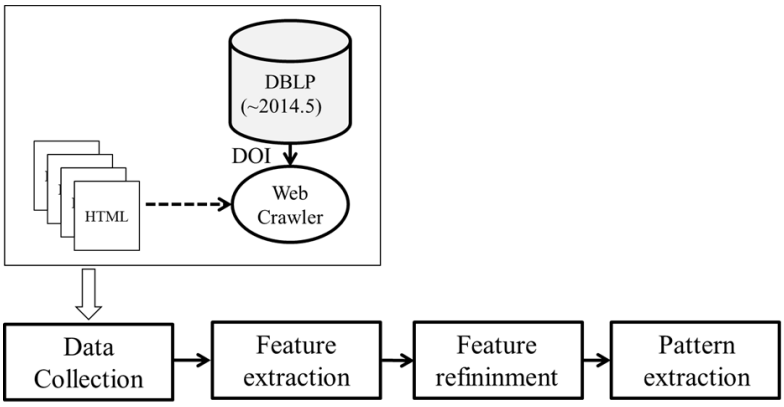

Fig. 2 A feature selection procedure

The feature extraction procedure involves four stages as shown Figure 2. The first stage involves the collection of researcher data; we collect the meta information of the published papers that are on the web to identify disambiguation of a researcher's name. To do so, we collect the Digital Object Identifiers (DOIs) of the papers; using these DOIs, we collect the published information on the pertinent sites. Since the websites where a paper is published are structured in different forms, we develop customized crawlers to collect data, taking into consideration the structure of each web page. The second stage is the feature extracting stage; email addresses of researchers are extracted. During this stage, the year when an email address was generated is extracted together with the email address to obtain the timeline information for the email addresses extracted. The third stage is the refining stage; during this stage, we remove the unnecessary data that may exist in the pertinent feature. For example, in addition to an affiliation name, the address and postal code of that organization are mentioned in some scientific papers; the same organization may be mistaken as different organizations owing to a different address or postal code mentioned in the papers, and therefore, such unnecessary information is removed. The last stage is the pattern extraction stage; during this stage, the unique pattern of a researcher is derived from the extracted pattern information, and this unique pattern can be used to accurately identify a researcher.

3.2 Email patterns of researchers

Figure 3 shows the general pattern of a researcher's email address based on a timeline. If a researcher's email address is considered with the timeline information,

the usage period of the email address can be defined. To estimate this period, the start and end time of a particular email address in use are defined as the time of first usage and the time of the latest usage (the last appearance of the pertinent email address), respectively. For example, more than 2 email addresses appeared for a particular researcher, and the periods during which each email address was used constitute a coprime relationship, i.e., the usage period for both the email addresses do not overlap like Case 1 as shown in Figure 3. The case 2 in Figure 3 shows that email addresses of researchers with the same name are different from each other but the appearance periods are overlapped; in the case, it is necessary to identify the researchers as same or different because it is possible that they are actually different researchers although their names are the same.

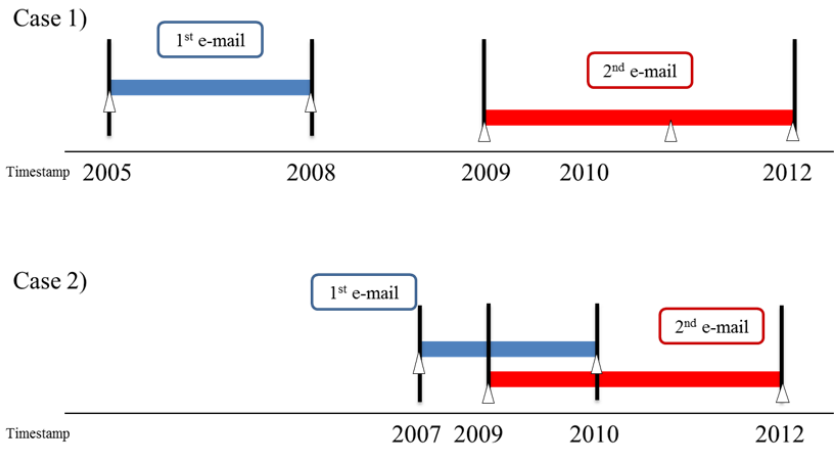

Fig. 3 Two cases of Email address patterns with timelines

4 Statistical Analysis

4.1 Data set

Metadata (title of the paper, the publishing year, coauthors, DOI, etc.) of a paper, as well as the researcher's name, are included in the researcher data extracted by DBLP. However, it does not include the email address and affiliation information of a researcher. Therefore, we implemented an experimental data set according to the procedure explained in Section 3.1. The number of researchers present on DBLP as of September, 2014 is about 1,465,700, and the number of papers is about 4,122,000. To obtain an experiment data set from the pertinent data, we collected website contents using the DOI of papers. The email information is collected automatically by a crawler, so the email addresses of all authors are collected. Therefore, it is not necessary that the n-th email address is the n-th author's email address. Thus, we considered only those email addresses for our experimental data, which had the number of authors equal to the number of email addresses. As a result, 64,802

researchers were extracted for the experimental data. To extract the first author, we compared the pertinent researcher's name with the co-author list, located it at the first instance, and deleted the overlapped name; finally an experimental data set including 18,867 researchers was implemented. We found that these researchers whose names were extracted using the aforementioned process, published 3,790 papers, which is 46.64% of the entire experimental data set.

4.2 Discussions

Through the statistical analysis of the results, we found that the number of overlapped email addresses whose appearance frequency are more than twice is 5.28%. In the data set, 3,162 researchers published at least two papers and a total of 8,126 papers were published by them. We set the minimum number of paper publications at two as a condition because the number of email addresses can be considered as one when the number of paper published is one. Further, 1,371 authors' emails appeared at least twice in the data set. Among the researchers, a total of 167 researchers showed the pattern depicted in the right side of Figure 4, and they published 574 papers (7.06% of the entire experimental data set). The result about the overlapped email address is lower than 2% of the result about the overlapped papers.

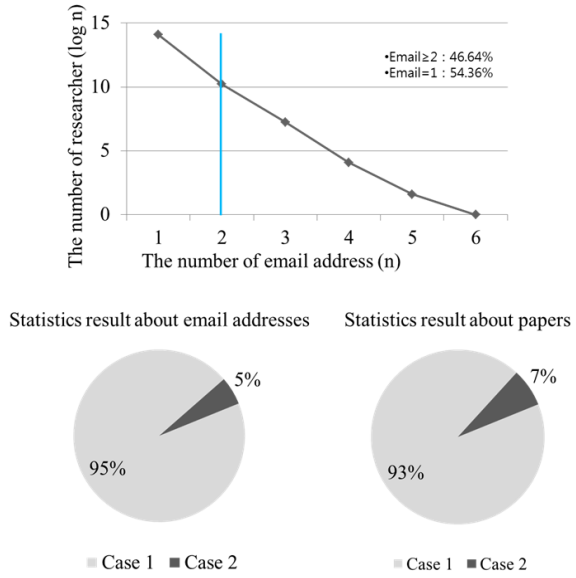

Fig. 4 Statistics results about two cases in terms of email and papers

It means that the productivity of researchers who have overlapped email patterns become high, and we awared that additional methods are needed to classify them because 43% of the researchers who have more than 2 email addresses.

5 Conclusion and future studies

In the big data environment, academic data are generated at a very fast rate. In this vein, researchers need to quickly obtain and accurately grasp the information presented in studies related to their research, the possible co-author network, and the trend in a specific research field to strengthen researchers' research competence. Accordingly, prescriptive analytics are required for accurate analysis and establishing strategies. However, these services have to be implemented based on accurate data to establish customized strategies for researchers. To this end, the accurate identification of a researcher's name and improvement in the information credibility with regard to researcher data becomes important.

This paper proposed an extraction procedure for important features from researcher data to identify ambiguous researcher's names. To improve researcher identification, we defined the email address usage pattern by considering the timeline characteristic of the researcher's email information and carried out experiments based on the DBLP data set; we verified that our identification method based on email addresses and that considers timeline characteristic is effective, and can be used as an important factor for identifying a researcher. As a future study, we will find a unique pattern representing a researcher by collecting and extracting the affiliation information considering the timeline characteristic from researcher data, and will research on an automated researcher identification system method by applying the obtained pattern to identify researchers; to verify its effectiveness, we will implement an accurately refined data set and compare its performance with the experimental data set.

6 Acknowledgments

This work was supported by the IT R&D program of MSIP/KEIT. [2014-044-024-002, Developing On-line Open Platform to Provide Local-business Strategy Analysis and User-targeting Visual Advertisement Materials for Micro-enterprise Managers].

7 References

1. Sa-kwang Song, Jinhyung Kim, Myunggwon Hwang, Jangwon Kim, Do-Heon Jeong, Seungwoo Lee, Hanmin Jung, Wonkyung Sung, "Prescriptive Analytics System for Improving Research Power," Proceedings of the 16th International Conference on Computational Science and Engineering (CSE), pp. 1144-1145, 2013.
2. Jinhyung Kim, Myunggwon Hwang, Jangwon Gim, Sa-Kwang Song, Do-Heon Jeong, Seongwoo Lee, Hanmin Jung, "Researcher Performance Analysis and Role Model Recommendation Model for Prescriptive Analytics," Proceedings of the Korea Computer Congress 2013 (KCC2013), Vol. 40, No. 2, pp. 241-243, 2013.
3. Madian Khabsa, Clyde Lee Giles, "The Number of Scholarly Documents on the Public Web," PLoS One, Vol.9, No.5, 2014 (DOI: DOI: 10.1371/journal.pone.0093949).

4. Ley, Michael. "The DBLP computer science bibliography: Evolution, research issues, perspectives," String Processing and Information Retrieval. Springer Berlin Heidelberg, pp. 1-10, 2002.
5. PUBMED, http://www.ncbi.nlm.nih.gov/pubmed
6. Emily Vardell, Tanya Feddern-Bekcan, Mary Moore, "SciVal Experts: a Collaborative Tool," Medical Reference Services Quarterly, Vol. 30, No. 3, pp. 283-294, 2011.
7. ORCID, http://orcid.org/
8. Einat Minkov, William Weston Cohen, Andrew Ng, "Contextual Search and Name Disambiguation in Email using Graphs," Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval,Vol. 29, pp. 27 - 34 , 2006.
9. Xuezhi Wang, Jie Tang, Hong Cheng, Philip S. Yu, "Adana: Active name disambiguation," Proceedings of the 11th International Conference on Data Mining (ICDM), pp. 794 - 803, 2011.
10. Yanan Qian, Yunhua Hu, Jianling Cui, Qinghua Zheng, Zaiqing Nie, "Combining machine learning and human judgment in author disambiguation," Proceedings of the 20th ACM international conference on Information and knowledge management, pp. 1241-1246, 2011.
11. Pei Li, Xin Luna Dong, Andrea Maurino, Divesh Srivastava, "Linking temporal records," Proceedings of the VLDB Endowment, Vol. 4, No. 11, 2011.
12. Jangwon Gim, Myunggwon Hwang, Sa-Kwang Song, Jinhyung Kim, Do-Heon Jeong, Hanmin Jung, "Researcher History Tracking Service for Prescriptive Analytics based on Researcher Activities," In Journal of KIISE : Computing Practices and Letters, Vol. 20, No. 6, pp. 0359-0363, 2014.