Estimating users' areas of research by publications and profiles on social networks

Petr Saloun
VSB-Technical University of
Ostrava, 17. listopadu 15
70833 Ostrava
Czech Republic
petr.saloun@vsb.cz

Adam Ondrejka
VSB-Technical University of
Ostrava, 17. listopadu 15
70833 Ostrava
Czech Republic
adam.ondrejka.st@vsb.cz

Ivan Zelinka
VSB-Technical University of
Ostrava, 17. listopadu 15
70833 Ostrava
Czech Republic
ivan.zelinka@vsb.cz

ABSTRACT

We focus on estimating a research area of a researcher/user by finding a unique identity in digital libraries and social networks and by analyse of public metadata of their publications and published information on social networks profiles. The lack of content of the metadata in some of the publications is solved by the information retrieval using techniques of NLP. We estimate the author's domain by extracting keywords from abstracts as well as by information published on social profiles. The result of this work is a design, an original algorithm and experimental verification of the algorithm.

Categories and Subject Descriptors

H.5.4 [**Information Interfaces and Presentation**]: Hypertext/Hypermedia—*User issues*; H.3.7 [**Information Interfaces and Presentation**]: Digital Libraries—*User issues*

General Terms

Design

Keywords

digital library, identify user, social media, information retrieval, natural language processing

1. INTRODUCTION

There are situations in life when we need to find works of a specific researcher, for example when we organize a conference. One of the most common way to solve this problem is to search for the information about this researcher, either by looking at the institutions and his publications or by examining the topics he had on various conferences, and then create a profile of the researcher manually. With the boom of social networking people began to publish more openly accessible data than before. Using the data may reveal an interesting complement to the true identity of a person. Unfortunately, the expansion and the emergence of various social networks caused a relatively large fragmentation and users publish specific information about themselves to a social network focusing on the specific topic. The fact that people can have the same name is another obstacle, therefore it is necessary to verify that it is actually a profile of the right person and not of his namesake. The main objective of this work is to identify researchers on social networks and digital libraries. Based on the public information on these sites, we estimate the area of a person's research. The results are keywords that serve both as a description of the person and as an input for further research in finding suitable reviewers of publications presented at conferences and for detecting the violations of a copyright.

2. ESTIMATING AREA OF AUTHOR'S RESEARCH

To find the right profiles we used a technique which compares specific attributes by different weights. Details are described in [2]. We used a modified version shown in the Equation 1 (similar work is mentioned in [3]).

$$sim_{u,p} = \begin{cases} \sum_{i=0}^{n} w_i \cdot sim(a_{i,u}, a_{i,p}) & \text{if } sim(a_{name}) > th_{name} \\ 0 & \text{otherwise} \end{cases} \tag{1}$$

where $sim(a_{name})$ is similarity between author and user profile names, th_{name} is threshold value to decide if names are the same or not, n is count of compared attributes, w_i is weight of compared attributes, a_p is set of user's profile attributes, a_u is set of user's attributes by his publications, $sim(a_{i,u}, a_{i,p})$ is similarity between attributes. The text comparison is done by fuzzy matching to include potential typing errors in attributes.

As shown in the Algorithm 1 the input is the name of the researcher. Then the search requests to all the digital libraries are executed and it downloads the publications. Each publication is then categorized by the defined criteria. Initially we eliminated all the articles that were similar or equal and were occurring in multiple libraries. Then we categorize the publications by affiliations using the text similarity algorithm and also by their co-authors. Now we have groups of possible unique authors. There is an issue now with the author publishing on his own or being active in multiple affiliations, because then the algorithm divides him into more groups. To handle the situation, we included a comparison with user's connections retrieved from social networks and additional information about skills, experiences and so on. After that we categorized the keywords by social profile similarity. We found all the profiles associated with the researcher name. Then we tried to find common connections and affiliations, and if there were at least one in each pair we would assign them together with the compared social profiles. The process was repeated for every found co-authors and referred publications with the input of the previously found authors, so the results would be more accurate. People with the same name are not merged into one identity, because of the classification by connections and affiliates. It is highly unlikely that these people will have the same co-authors, friends and jobs. More information about a unique user identity is described in [1].

The research domain is obtained by analysing the keywords of all found publications and by extracting the additional information from the social profiles. Because of lacking and incompletely chosen keywords in many publications we had to use our original technique to get additional keywords from abstract. We do not go into detail describing our original technique, because of the page limit of this poster.

```
Data: Author's first name and last name
Result: User's identities
firstName, lastName ← {user raw input};
for searcher in DigitalLibrariesSearchers do
    publications ←SearchAuthor(firstName, lastName);
end
GroupByPublication(publications);
GroupByAffiliates(publications);
for searcher in SocialNetworkSearchers do
    publications ←SearchAuthor(firstName, lastName);
end
for group in groups do
    for publication in publications do
        groupKeywords +=
        AnalyzePublication(publication);
    end
end
finalGroups = GroupBySocialProfiles(groups);
```

Algorithm 1: Finding unique author identity on digital libraries and social networks

3. EXPERIMENT

From the digital libraries we chose IEEExplorer[1], ACM Digital Library[2] and SpringerLink[3]. In this work, the researchers are found on LinkedIn[4] and Researchgate[5] social networks. In the experiment we check if we can find unique identities and research domains of 180 randomly selected researchers. The search of user identities in digital libraries has been tested by at least 180 researchers, by downloading and analysing about 3100 publications (Table 1). The researchers were chosen randomly and included people of different nationalities. Initially there were users grouped only by co-authors and affiliates. There were 118 authors grouped correctly ("R") with rate 65 %. 3 authors had assigned other author's publications ("POA", 2 % error rate) because of fact that searched author had publications with namesake co-authors and it was poorly evaluated as same person, error rate in this case was 59 authors were not merged correctly ("NA"), there were too many created identities of which should be same one author. This was caused by publications with no or one co-author and different affiliations, it was not possible to find connection between them. Error rate of this category was 33 %.

Table 1: Experiment of finding identities

	R	POA	NA	Precision	Recall
Co-authors	118	3	59	97 %	66 %
C-A + Social	132	3	45	98 %	74 %
C-A + S + K	166	14	0	92 %	100 %

In the next step we included comparisons of authors by data found on their social profiles. 132 users were identified correctly (73 %) and to 3 same authors were again assigned wrong publications due to the same reasons, error rate remained 2 %. The only improvements were made in the case when one author was in two different groups ("NA") and when there were connections found in social profiles between them, so error rate decreased to 25 %.Finally we added comparisons by keywords between publications with a single author and publications with multiple authors. 166 users were identified correctly, correct rate increased to 92 %. There was no situation with a one author in more groups ("NA", error rate of this category decreased to 0 %). Unfortunately 14 users had assigned wrong publications ("POA"), error rate increased to 8 %. It was caused by errors in extracting of keywords and the associated bad detection of a similarity between researchers and publications.

4. CONCLUSION

The goal of our work was to create algorithm to estimate research area of users by finding their identities in digital library and social networks and by analyse found data. As the results from our experiment show, the algorithm for identifying research identities on digital libraries and social networks was successful in 92 % of all the attempts in final. This work was the first step in the research of recommending publications to authors and finding violations of copyrights. We would want to try to add comparing authors' domains detected from publications and information on the Internet to classic full-text search approach. This work is input for further research in finding suitable reviewers of publications presented at conferences and for detecting the violations of a copyright.

5. ACKNOWLEDGMENT

The following grant is acknowledged for the financial support provided for this research: Grant of SGS No. SP2014/42, VSB - Technical University of Ostrava, Czech Republic.

6. REFERENCES

[1] K. Kostkova, M. Barla, and M. Bielikova. Social relationships as a means for identifying an individual in large information spaces. In M. Bramer, editor, *Artificial Intelligence in Theory and Practice III*, volume 331 of *IFIP Advances in Information and Communication Technology*, pages 35–44. Springer Berlin Heidelberg, 2010.

[2] E. Raad, R. Chbeir, and A. Dipanda. User profile matching in social networks. In *Network-Based Information Systems (NBiS), 2010 13th International Conference on*, pages 297–304, Sept 2010.

[3] J. Vosecky, D. Hong, and V. Y. Shen. User identification across social networks using the web profile and friend network. *IJWA*, 2(1):23–34, 2010.

[1] http://ieeexplorer.ieee.org
[2] http://dl.acm.org
[3] http://www.springerlink.com
[4] http://linkedin.com
[5] http://www.researchgate.com