Influence of Time on User Profiling and Recommending Researchers in Social Media

Chifumi Nishioka
ZBW – Leibniz Information Centre for Economics
Kiel, Germany
Kiel University, Kiel, Germany
chni@informatik.uni-kiel.de

Gregor Große-Bölting
Kiel University, Kiel, Germany
ggb@informatik.uni-kiel.de

Ansgar Scherp
ZBW – Leibniz Information Centre for Economics
Kiel, Germany
Kiel University, Kiel, Germany
a.scherp@zbw.eu

ABSTRACT

We conduct two experiments to compare different scoring functions for extracted user interests and measure the influence of using older data. We apply our experiments in the domains of computer science and medicine. The first experiment assesses similarity scores between a user's social media profile and a corresponding user's publication profile, in order to evaluate to which extend a user's social media profile reflects his or her professional interests. The second experiment recommends related researchers profiled by their publications based on a user's social media profile. The result revealed that while the functions using spreading activation produce large similarity scores between a user profile and publication profile, the scoring functions with statistical methods (e.g., an extension of BM25 with spreading activation) perform best for recommendation. In terms of the temporal influence, the older data have almost no influence on the performance in the medicine dataset. However, in the computer science dataset, while there is a positive influence in the first experiment, the second experiment demonstrated a negative influence when adding too old data.

CCS Concepts

•**Applied computing → Document analysis;**

Keywords

user profiling, social media, temporal analysis

1. INTRODUCTION

Social media platforms such as Twitter are used to share professional thoughts and to build professional networks [10]. Thus, we assume that it is possible to learn a user's professional profile from his or her social media activities. Understanding a user's professional interests is an important task, because it enables, e. g., better recommendations of scientific publications. However, it is difficult to construct a user's professional profile from social media activities as professional fields are closely connected to each other. For instance, how is it possible to distinguish between a computer scientist whose professional field is data mining and one whose professional field is machine learning from social media activities? Existing works like Abel et al. [3] compared the performance of extracting professional profiles from different social media platforms. They observed that the resulted professional profiles contained a lot of noise. Kapanipathi et al. [9] developed an approach for extracting cross-domain user interests from Twitter with a manually-created hierarchical knowledge base that is extracted from Wikipedia. They employed different scoring functions that reveal user interests which are not mentioned directly in Twitter and gave a score. Our previous work [8] provided approaches to extract professional interests from social media activities and publications, using a domain-specific knowledge base and spreading activation functions.

This paper extends our prior work [8] by conducting two experiments to assess different existing and new entity scoring functions and investigate the temporal influence of the used data. We employ a domain-specific knowledge base to reduce noise that Abel et al. [3] observed. Different from Kapanipathi et al. [9], we apply entity scoring functions that combine entity extraction, spreading activation, and statistical methods (e.g., TF-IDF [13] and CF-IDF [7]). In addition, we analyze the influence of older data. To the best of our knowledge, the temporal influence of data for profiling professional interests has not been assessed so far. Both experiments are applied on two datasets, one in the field of computer science and the other in medicine. While Kapanipathi et al. [9] conducted a user experiment, we do automatic assessments like Abel et al. [3]. We employ Twitter as a social media platform, because many scientists use it to disseminate their professional thoughts [10]. In the first experiment, we compute similarity scores between the users' social media profiles derived from social media activities and their publication profiles made by their own scientific publications. We assume that a user's social media profile reflects the content of publications. This experiment extends our prior work [8] by analyzing whether older publications boost the similarity scores and investigating the influence of the number of publications and number of tweets. The second experiment recommends related researchers profiled by their publications based on a user's social media profile. It

i-KNOW '15, October 21 - 23, 2015, Graz, Austria

ISBN 978-1-4503-3721-2/15/10…$15.00

DOI: http://dx.doi.org/10.1145/2809563.2809601

investigates how well the approaches enable to distinguish a user's publication profile from other users' publication profiles. Regarding time, we investigate whether publication profiles with the older publications enhance the performance of the recommendation. Additionally, we examine how the number of publications and social media items affect on the results of the experiments and whether the abstracts of the publications have a strong influence. In summary, the research questions are: (i) the effectiveness of different entity scoring functions, (ii) the temporal influence (i.e., do older publications enhance the performance?), (iii) the influence of the number of publications and social media items, and (iv) the influence of using abstracts for publication profiles (i. e., do the abstracts enhance the publication profiles?).

The results reveal that while the scoring functions based on spreading activation theory produce large similarity scores between a user profile and publication profile, the scoring functions with the statistical methods perform best for recommendation. In terms of the temporal influence, the older data has almost no influence on the performance in the medicine dataset. In the computer science dataset, while there is a positive influence in the first experiment, the second experiment demonstrated a negative influence when adding too old data. Regarding the number of publications, it correlates with the performance in both experiments. On the other hand, we observe no or weak correlation between the performance and the number of social media items. Abstracts make only a slight improvement in both experiments.

The subsequent section presents related work. Section 3 briefly describes our research methodology, the entity scoring functions, and datasets used. Sections 4 and 5 present the results of the two experiments. The results are discussed in Section 6, before we conclude the paper.

2. RELATED WORK

Social media have been seen in the past as a rich source for user profiling and extracted profiles have been applied to recommender systems [1, 2, 4, 7]. Abel et al. [3] attempted to extract professional interests from social media, including LinkedIn, Delicious, and Twitter. They compared the tag-based approach (for Delicious), bag-of-words approach (for LinkedIn and Twitter), and semantic entity-based approach (for Twitter) in the scenario of recommending publications. They observed that Twitter profiles seem to cover more professional interests than other social media platforms, but also include more noise. Regarding different profiling functions, the semantic entity-based profiles obtained from Twitter outperformed the others. In our study, we utilize a domain-specific knowledge base for entity detection, therefore user profiles might not contain noisy entities.

Although traditionally user profiles have been represented using the vector space model [13], recent studies have introduced graph-based user profiles and proved its effectiveness: Shen et al. [15] extracted named entities from Twitter and disambiguate them by utilizing a graph structure of an external knowledge base. Kapanipathi et al. [9] created the hierarchical-structured user profiles where each node denoted a semantic entity. They created their own, cross-domain knowledge base from Wikipedia. They proposed several propagation functions to give a score to each node based on spreading activation theory [5]. In contrast, we use an existing external knowledge base and combine these approaches with statistical methods (e.g., TF-IDF [13]).

In terms of the temporal influence of the data, De Pessemier et al. [6] investigated the effect of the older data for collaborative filtering. Their results indicate that the accuracy increases by extending user profiles with additional older consumption data. In contrast, there is the opposite effect for user-generated content, i.e., involving older consumption data has a negative influence on the recommender accuracy. Zheng et al. [17] also evaluated the effect of data generated over a different time period on recommendation precision using several popular model-based collaborative filtering algorithms. Their results show that while more recent data have larger impacts, the usefulness of older data cannot be ignored as long as there are sufficient old samples. However, the addition of insufficient amount of old data seems to have negative impacts. Thus, since the older data has both positive and negative influence for profiling depending on algorithms and contexts, we need to carefully investigate and evaluate the influence of the older data. In this paper, we look into the influence of the older publication data for profiling professional interests and recommending researchers based on a user's social media profile.

3. EVALUATION METHODOLOGY

First, we briefly present how to extract entities from texts (i.e., social media items and publications), following previous work [8]. Subsequently, we list the entity scoring functions applied in this paper in Section 3.2. In Section 3.3, we provide details about the two datasets used in this paper.

3.1 Entity Detection

We detect entities from social media items, using a domain-specific knowledge base, following the previous work [8]. We employ a domain-specific knowledge base for entity detection to avoid noise as Abel et al. [3] observed in their user profiling. Each entity stored in a knowledge base retains its relations to higher order (generalized) entities and lower order (specialized) entities and has labels, including synonyms, alternate diction, and abbreviations. Entities do not share labels. We extract entities, employing a naive string matching approach. Thus, labels in a knowledge base are used as a dictionary of entities. In order to reduce the number of false positives, we only take labels composed of at least four characters. Consequently, we obtain a set of entities for the social media profile, where each entity is given a score by a scoring function. Likewise, we applied the same procedure to extract entities from the publications of a user. Below, we describe the entity scoring functions used in the paper.

3.2 Scoring of Entities

We apply the entity scoring functions to give a score to each extracted entity. The functions that use spreading activation theory [5] and a hierarchial structure of a knowledge base (i.e., Basic, Bell, BellLog, HCF-IDF, BM25HC) give a score to entities that are not mentioned explicitly in the analyzed text. Below, $score(c, d)$ denotes a score of an entity c in a text d.

Frequency: This function gives a frequency (i.e., number of appearances) of an entity as a score like Abel et al. [3].

$$score_{freq}(c, d) = freq(c, d), \tag{1}$$

where $freq(c)$ returns the number of appearances of c in d.

Basic Spreading Activation (Basic): The basic spreading activation function described in [9] gives a score as below.

$$score_{basic}(c,d) = freq(c,d) + \lambda \cdot \sum_{c_j \in C_l(c)} score_{basic}(c_j,d), \tag{2}$$

where $C_l(c)$ returns the set of entities located in a lower order of the entity c. λ denotes the decay parameter. As mentioned in the previous work [8], we set $\lambda = 1.0$.

Bell Spreading Activation (Bell): The distribution of entities across the different levels of a hierarchical knowledge base may follow a bell curve. Based on this observation, Kapanipathi et al. [9] developed bell spreading activation defined in Equation 3.

$$score_{bell}(c,d) = freq(c,d) + F_c \sum_{c_j \in C_l(c)} score_{bell}(c_j,d), \tag{3}$$

where $F_c = \frac{1}{nodes(h(c)+1)}$. Here, $h(c)$ returns the level where an entity c is located in a knowledge base and *nodes* provides the number of entities at a given level in a knowledge base. Thus, F_c indicates the number of entities at the lower level of c.

Bell Logarithmic Spreading Activation (BellLog): To reduce the impact of the raw count, Kapanipathi et al. [9] employed log scale for Bell. BellLog is defined as Equation 3 replacing $score_{bell}$ with $score_{belllog}$ and F_c with FL_c defined as $FL_c = \frac{1}{log_{10}(nodes(h(c)+1))}$.

CF-IDF: Concept Frequency Inverse Document Frequency (CF-IDF) [7] is an extension of the traditional TF-IDF [13], that counts entities instead of terms.

$$score_{cfidf}(c,d) = freq(c,d) \cdot \log \frac{|D|}{|d \in D : c \in d|}, \tag{4}$$

where $|D|$ denotes the number of texts in an entire text corpus and $|d \in D : c \in d|$ means the number of documents d in D containing c.

HCF-IDF: Hierarchical Concept Frequency Inverse Document Frequency (HCF-IDF) is an extension of CF-IDF, using the hierarchical structure of a knowledge base. We compute HCF-IDF, using the scores of BellLog, since Kapanipathi et al. [9] reported that BellLog best performed (except the method *PriorityInterest* that is inapplicable here).

$$score_{hcfidf}(c,d) = score_{belllog}(c,d) \cdot \log \frac{|D|}{|d \in D : c \in d|}, \tag{5}$$

where $|d \in D : c \in d|$ denotes the number of texts containing c after applying BellLog.

BM25C: BM25C is an extension of Okapi BM25 [12]. Like CF-IDF, we count entities instead of terms.

$$score_{bm25c}(c,d) = IDF(c,D) \cdot \frac{freq(c,d) \cdot (k+1)}{freq(c,d) + k \cdot (1 - b + b \cdot \frac{|d|}{avgdl})}, \tag{6}$$

where both k and b are parameter and *avgdl* denotes the average length (i.e., the average number of entities) of the texts in a corpus. We set $k = 1.6$ and $b = 0.75$ following [11]. $IDF(c,D)$ is defined by Equation 7.

$$IDF(c,D) = \log \frac{|D| - |d \in D : c \in d| + 0.5}{|d \in D : c \in d| + 0.5} \tag{7}$$

BM25HC: In addition to BM25C, we develop BM25HC, an extension of BM25C, using a hierarchical structure of a knowledge base. The $score_{bm25hc}$ is computed as defined in Equation 6 but by replacing $freq(c,d)$ with $score_{belllog}$. Regarding $IDF(c,D)$, we use Equation 7 where $|d \in D : c \in d|$ denotes the number of texts containing c after applying BellLog.

In addition to the aforementioned entity-based scoring functions, we also experiment the traditional TF-IDF [13] and Okapi BM25 [12].

3.3 Used Datasets

We used two datasets from the scientific domains of computer science and medicine. Twitter was chosen as social media platform because of its predominance among social media platforms and its strong use among researchers to disseminate their thoughts [10].

3.3.1 Computer Science

For entity detection in the field of computer science, we used the ACM Computer Classification System (CCS)[1] as a domain-specific knowledge base. The CCS contains 2,299 entities in the field of computer science as well as 9,086 labels. Thus, on average an entity has 4.95 labels ($SD = 3.59$). According to Kapanipathi et al. [9], the number of entities over the different levels in a knowledge base should follow a normal distribution for applying Bell and BellLog. We verified this by visual inspection of the CCS taxonomy.

In terms of Twitter data, we collected 88 Twitter users working in the field of computer science (following the procedure of our previous data collection [8]). Specifically, we retrieved tweets that mentioned one of the A*-rated[2] 26 computer science conference hashtags via Twitter API. A*-rated conferences were chosen because of their importance. We used only the hashtags that were officially used on the conference web pages or official conference Twitter accounts. Subsequently, we filtered the obtained users and kept only users who also appeared in DBLP records. Although conference hashtags are not necessarily unique, we assume that users who have publication records on DBLP use the hashtags to indicate computer science conferences. Through this procedure, we identified 88 Twitter accounts with corresponding DBLP records of 88 computer scientists. Then, we retrieved their tweets, using Twitter API. Please note that we could obtain 3,200 tweets at most for each user due to the limitation of the Twitter API. A user published on average 697.58 tweets (SD: 443.17).

In order to obtain publications for each user, we used the extended DBLP dataset[3]. From the dataset, we obtained

[1] http://www.acm.org/about/class/class/2012, last access: May 17, 2015

[2] CORE ranking from 2014, see http://103.1.187.206/core/, last access: June 2, 2015

[3] AMiner Citation Network Dataset, http://arnetminer.org/lab-datasets/citation/DBLP_citation_Sep_2013.rar, last access: June 9, 2015

titles and abstracts of publications authored by one of the 88 users. In total, we got $1,059$ titles and 325 abstracts. On average, a user has 12.03 titles (SD: 13.45) and 3.69 abstracts (SD: 5.12). The publications of 29 of the 88 users have no abstract. The average year of publication is 2006.74 (SD: 4.94). The latest publication was published in 2013 and the oldest publication was in 1983.

3.3.2 *Medicine*

In addition to the domain of computer science, we conduct the experiments in the domain of medicine. For named entity detection, we used the Medical Subject Headings (MeSH)[4] as a domain-specific knowledge base[5]. The MeSH contains $27,300$ entities in the field of medicine as well as their relations and $224,368$ labels. Thus, on average an entity has 8.22 labels (SD: 9.19). In addition, a visual inspection confirmed that the number of entities over the different levels follows a normal distribution.

Regarding Twitter data, 64 Twitter users were obtained by searching Twitter for the top five journals[6]. We queried each of the five journal hashtags via Twitter API and extracted users who mentioned at least one of those hashtags. Subsequently, we filtered the obtained users and kept only users who also appeared on PubMed records. Through this procedure, we identified 64 Twitter users. A user published on average 1508.13 tweets (SD: 1282.62).

To obtain publications of the 64 users, we accessed the PubMed database[7] and obtained publications via URL calls E-utility functions[8]. On average, a user has 50.34 publications (SD: 65.95) with on average 43.27 abstracts (SD: 60.23). Four of the 64 users have no abstract. The average year of publication is 2010.40 (SD: 3.64). The latest publication was published in 2015 and the oldest publication was in 1976.

4. USER PROFILING

We start with the experiment to assess how well a user's social media profile reflects a user's professional interests that are retrieved from publication profiles. Different from our previous work [8], we investigate not only the effectiveness of the different entity scoring functions but also the temporal influence (i.e., do older publications reflect user interests?), the influence of the number of publications and tweets, and the influence of using abstracts for publication profiles (i.e., do abstracts enhance the publication profiles?)

4.1 Procedure and Metrics

We investigate how similar a user's social media profile and corresponding user's publication profile are. First, we extract entities from both social media items and publications as described in Section 3.1. Subsequently, we apply one of the entity scoring functions described in Section 3.2. The entity scoring functions are applied to both social media items and publications. In order to examine how well a user's social media profile reflects a user's professional interests retrieved from publication profiles, we compute similarity scores between a user's social media profile and the corresponding publication profile. We employ the cosine similarity to compute similarity scores.

$$sim_{cos} = \frac{\vec{S_u} \cdot \vec{E_u}}{\parallel \vec{S_u} \parallel \parallel \vec{E_u} \parallel}, \tag{8}$$

where $\vec{S_u}$ and $\vec{E_u}$ denote the profile vectors converted from a user u's social media profile and a publication profile. We compute sim_{cos} for all users and report the average and standard deviation.

4.2 Results

Entity Scoring Functions. Table 1 shows the similarity scores with respect to each entity scoring function described in Section 3.2. The columns labeled with "all" denotes that both titles and abstracts are used for publication profiles. As shown in our previous work [8], Basic results in the largest similarity scores between user profiles and publication profiles. We further look into the differences between entity scoring functions by running a one-way repeated-measure ANOVA with a Greenhouse-Geisser correction (for all analyses, it is $\epsilon = .23$). We applied Shaffer's modified sequentially rejective Bonferroni procedure [14] to assess significance of pair-wise differences between entity scoring functions. We use a standard significance level of $\alpha = 0.05$. Referring to the computer science dataset and its two variants (i.e., "titles" and "all"), there are pairwise significances between the scoring functions (for "titles" $t(87)$ is in $[2.64, 12.19]$, $p < .04$, for "all" $t(87)$ is in $[3.32, 14.40]$, $p < .01$) except CF-IDF and HCF-IDF and BM25C and BM25HC. Regarding CF-IDF and BM25HC, there is a significance for "all" (i.e., both titles and abstracts) but not for "titles". In terms of the medicine dataset, there are pairwise significant differences ("titles": $t(63)$ is in $[3.73, 12.38]$, $p < .01$, "all": $t(63)$ is in $[3.06, 17.09]$, $p < .01$) except between BM25C and TF-IDF and BM25HC and TF-IDF. Regarding the difference between Freq and BellLog, there is a significance for "all" but not for "titles". Details are omitted for reasons of brevity.

Table 1: Average cosine similarity of the different entity scoring functions (SD in parentheses). The column "all" denotes that both titles and abstracts are used. The best results are marked in bold font.

	Computer Science		Medicine	
	title	all	title	all
Freq	.17 (.20)	.20 (.21)	.21 (.22)	.31 (.24)
Basic	**.33** (.24)	**.38** (.24)	**.35** (.24)	**.46** (.24)
Bell	.23 (.22)	.28 (.22)	.25 (.23)	.35 (.24)
BellLog	.17 (.20)	.20 (.21)	.21 (.22)	.31 (.24)
TF-IDF	.05 (.05)	.05 (.05)	.08 (.11)	.07 (.10)
CF-IDF	.13 (.16)	.14 (.16)	.13 (.17)	.16 (.20)
HCF-IDF	.14 (.16)	.14 (.15)	.14 (.18)	.17 (.20)
BM25	.03 (.03)	.03 (.03)	.04 (.04)	.04 (.04)
BM25C	.09 (.10)	.10 (.09)	.08 (.09)	.08 (.10)
BM25HC	.10 (.10)	.10 (.10)	.08 (.09)	.09 (.10)

[4]2015 MeSH "Descriptor Records" retrieved May 16, 2015, http://www.nlm.nih.gov/mesh/filelist.html, last access: June 2, 2015

[5]We convert the original .xml file into the .nt file using the convertor HIVE https://code.google.com/p/hive-mrc/, last access: August 3, 2015

[6]http://impactfactor.weebly.com/medicine.html, last access: June 2, 2015

[7]http://www.ncbi.nlm.nih.gov/pubmed, last access: June 2, 2015

[8]http://www.ncbi.nlm.nih.gov/books/NBK25500/, last access: June 2, 2015

Temporal Influence. Publications in both datasets are released in various years. Thus, we investigate whether older publications are noise for profiling publications or boost the cosine similarity scores. Specifically, we start to measure similarity scores between a publication profile created from publications in the most recent year and a social media profile. Subsequently, we incrementally add publications published in the older years and measure similarity scores. Although the most recent years of publications are 2013 in the computer science dataset and 2015 in the medicine dataset, we start 2012 and 2014, respectively, because the number of publications in the most recent years are much fewer than the previous years. In addition, we ignore the publications published before 1999 (computer science) and 2001 (medicine), because publications published before those years are few, too. Figures 1 and 2 plot the similarity scores for the computer science dataset and medicine dataset. In the figures, all the publications released after a year shown in the x-axis are taken into account for publication profiles. In both datasets, similarity scores increase when more older publications are added. Especially, the similarity scores of the scoring functions without the statistical methods (i.e., Basic, Bell, BellLog) are boosted a lot, when adding publications published in 2011, 2010, and 2009 in the computer science dataset (see Figure 1). Compared to the computer science dataset, the similarity scores in the medicine datasets are static over the years.

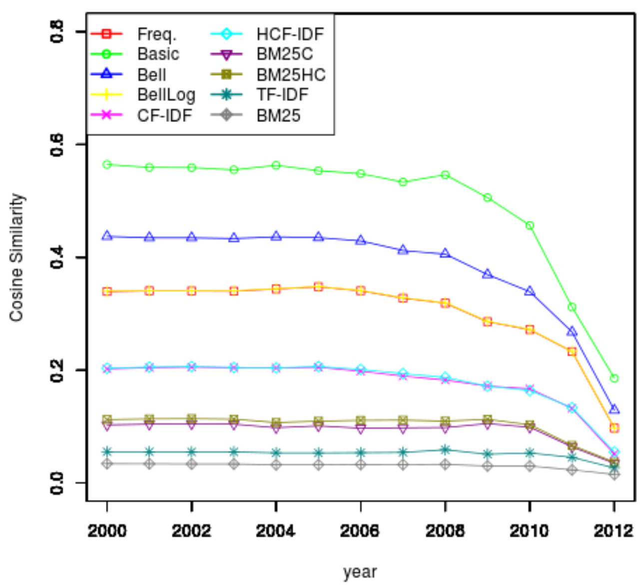

Figure 1: Effectiveness of the older publications on similarity scores for the computer science dataset. All the publications published after a year shown in the x-axis are used for publication profiling. The y-axis represents the average cosine similarity.

Influence of the Number of Publications and Tweets. We investigate whether the number of publications and number of tweets have an influence on user profiling. We compute Kendall's rank correlation coefficient τ to measure correlation between the similarity scores and the number of publications and between the similarity scores and the number of tweets. Tables 2 and 3 show the results. In both domains, we observe medium correlations. We observe that the similarity scores correlate with the number of publications more strongly than the number of tweets.

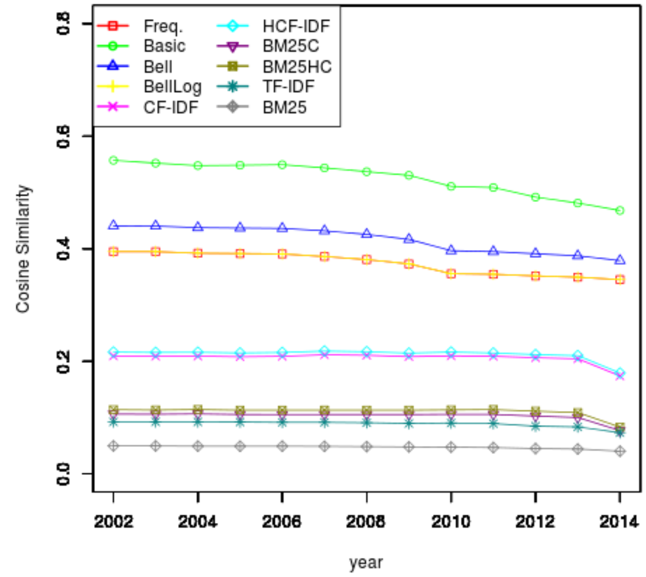

Figure 2: Effectiveness of the older publications on similarity scores for the medicine dataset. All the publications published after a year shown in the x-axis are used for publication profiling. The y-axis represents the average cosine similarity.

Table 2: Kendall's τ between cosine similarity and the number of publications. "all" denotes that titles and abstracts are used. The p-values in parentheses are marked in bold font if $\leq .05$.

	Computer Science		Medicine	
	title	all	title	all
Freq	.36 (**.00**)	.42 (**.00**)	.42 (**.00**)	.36 (**.00**)
Basic	.34 (**.00**)	.41 (**.00**)	.49 (**.00**)	.45 (**.00**)
Bell	.36 (**.00**)	.45 (**.00**)	.44 (**.00**)	.37 (**.00**)
BellLog	.35 (**.00**)	.42 (**.00**)	.42 (**.00**)	.36 (**.00**)
TF-IDF	.30 (**.00**)	.32 (**.00**)	.35 (**.00**)	.35 (**.00**)
CF-IDF	.30 (**.00**)	.33 (**.00**)	.37 (**.00**)	.36 (**.00**)
HCF-IDF	.30 (**.00**)	.34 (**.00**)	.39 (**.00**)	.38 (**.00**)
BM25	.30 (**.00**)	.31 (**.00**)	.43 (**.00**)	.40 (**.00**)
BM25C	.27 (**.00**)	.28 (**.00**)	.42 (**.00**)	.43 (**.00**)
BM25HC	.28 (**.00**)	.29 (**.00**)	.41 (**.00**)	.42 (**.00**)

Influence of using Abstracts. In Table 1, we can observe that adding abstracts to the titles (columns "all") slightly boosts the similarity scores. While the similarity scores are increased by adding abstracts when using the entity scoring functions without the statistical methods (i.e., Basic, Bell, BellLog), the difference of similarity scores are small when using the scoring functions using the statistical methods (i.e., TF-IDF, CF-IDF, HCF-IDF, BM25, BM25C, BM25HC).

Table 3: Kendall's τ between cosine similarity and the number of tweets. "all" denotes the titles and abstracts. The p-values in parentheses are marked in bold font if $\leq .05$.

	Computer Science		Medicine	
	title	all	title	all
Freq	.15 (**.00**)	.09 (.23)	.13 (.12)	.16 (.07)
Basic	.22 (**.00**)	.17 (**.02**)	.08 (.33)	.10 (.23)
Bell	.21 (**.00**)	.12 (.11)	.12 (.16)	.14 (.09)
BellLog	.20 (**.00**)	.10 (.16)	.13 (.12)	.16 (.07)
TF-IDF	.11 (.14)	.13 (.08)	.14 (.09)	.17 (.05)
CF-IDF	.18 (**.01**)	.11 (.15)	.16 (.06)	.16 (.06)
HCF-IDF	.22 (**.00**)	.14 (.06)	.14 (.11)	.15 (.09)
BM25	.09 (.26)	.14 (.07)	.20 (**.02**)	.24 (**.00**)
BM25C	.18 (**.02**)	.12 (.11)	.19 (**.03**)	.21 (**.01**)
BM25HC	.24 (**.00**)	.17 (**.02**)	.17 (**.04**)	.19 (**.03**)

5. RECOMMENDING RESEARCHERS

The second experiment is about recommending researchers. A user gets recommendations of researchers profiled by their publications based on his or her social media profile. Through the experiment, we examine how well the entity scoring functions can distinguish a user from other users. We investigate the effectiveness of the different entity scoring functions, the temporal influence (i.e., do older publications enhance recommendation performance?), the influence of the number of publications and tweets, and the influence of using abstracts for publication profiles (i.e., do the abstracts enhance recommendation performance?)

5.1 Procedure and Metrics

First, we compute similarity scores between a user's social media profile and each of all the publication profiles, using cosine similarity. We rank the publication profiles by similarity scores. Subsequently, we compute the Mean Reciprocal Rank (MRR) defined as shown below:

$$MRR = \frac{1}{|U|} \sum_{u \in U} \frac{1}{rank(u)}, \tag{9}$$

where $rank(u)$ denotes the rank at which u's publication profile appears in the list of all users' publication profiles sorted by similarity scores.

Please note that we aim to investigate how well the entity scoring functions can discriminate a user from other users in this experiment. In a practical recommender system, it is not usual for users to get themselves as recommendation. However, due to lack of the gold standard and difficulty of obtaining it [16], we consider users themselves as a right recommendation in this experiment. We assume that researchers ranked near by him or her have similar interests.

5.2 Results

Entity Scoring Functions. Table 4 illustrates the performance of recommendations for each entity scoring function shown in Section 3.2. We observe that while BM25 and TF-IDF perform best for the computer science dataset, BM25C and BM25HC outperform the other functions for the medicine dataset. Regarding the difference between the two academic domains, the medicine dataset consistently shows better performances. The medicine dataset contains 64 users, fewer users compared to the computer science dataset. The minimum value of reciprocal rank (i.e., 1/64) in the medicine dataset is higher than the one in the computer science dataset (i.e., 1/88). Thus, the medicine dataset gets higher MRR. We investigate the significance of the differences between the entity scoring functions as we do in Section 4.2. It reveals no significances in the medicine dataset. But for the computer science dataset BM25 and TF-IDF have significant differences with Freq, Basic, Bell, and BellLog ("titles": $t(87)$ is in $[3.92, 3.34]$, $p < .05$, $\epsilon = .36$, "all": $t(87)$ is in $[4.21, 3.32]$, $p < .05$, $\epsilon = .35$). Details omitted for reasons of brevity.

Table 4: MRR of different entity scoring functions (SD in parentheses). "all" denotes the titles and abstracts. The overall best results are marked in bold font.

	Computer Science		Medicine	
	title	all	title	all
Freq	.18 (.28)	.21 (.29)	.25 (.36)	.26 (.36)
Basic	.15 (.28)	.17 (.30)	.25 (.36)	.28 (.38)
Bell	.18 (.30)	.21 (.32)	.23 (.34)	.25 (.36)
BellLog	.18 (.28)	.21 (.29)	.25 (.36)	.26 (.36)
TF-IDF	.31 (.35)	**.33** (.37)	.38 (.42)	.38 (.42)
CF-IDF	.22 (.30)	.24 (.32)	.38 (.41)	.38 (.41)
HCF-IDF	.22 (.31)	.22 (.31)	.37 (.41)	.38 (.42)
BM25	**.33** (.38)	.32 (.40)	.33 (.39)	.33 (.39)
BM25C	.26 (.35)	.25 (.34)	**.43** (.44)	.38 (.42)
BM25HC	.24 (.33)	.25 (.35)	.41 (.42)	**.40** (.43)

Temporal Influence. We examine how the older publications affect on recommending researchers. Specifically, we start to measure MRR with publication profiles created from publications in the most recent year, incrementally add publications published in the older years and measure MRR. Figures 3 and 4 illustrate the results of the experiments for the computer science dataset and medicine dataset. For the entity-based scoring functions, we observe that the recommendation performs best when using all publications published after around 2004 in the computer science dataset. For TF-IDF and BM25, the recommendation performs best when considering all publications published after around 2010. When using publications published before 2010, the performance gets worse, especially for BM25. Referring to the medicine dataset, the performance of the recommendation does not vary very much, when older publications are added. But for TF-IDF and BM25, the performance is low when using only publications published in the most recent year.

Influence of the Number of Publications and Tweets. We investigate whether the number of publications and tweets have an influence on the performance. We compute the Kendall's rank correlation coefficient τ and measure correlation between MRR and the number of publications and between MRR and the number of tweets. Tables 5 and 6 show the results. While we observe the moderate correlation between MRR and the number of publications in Table 5, Table 6 indicates that there is almost no correlation between MRR and the number of tweets.

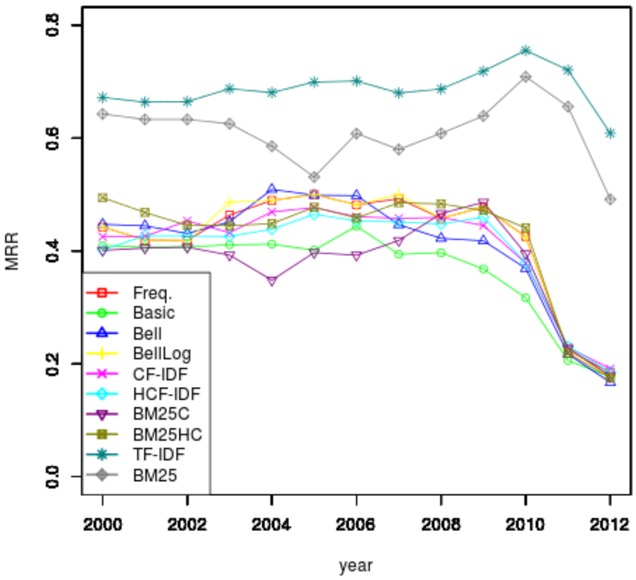

Figure 3: Effectiveness of the older publications for recommending researchers in the computer science dataset. All the publications published after a year shown in the x-axis are used for publication profiling. The y-axis represents the MRR.

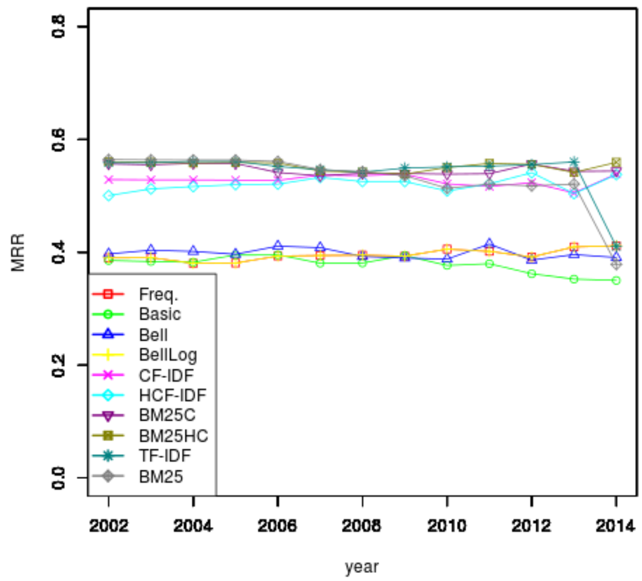

Figure 4: Effectiveness of the older publications for recommending researchers in the medicine dataset. All the publications published after a year shown in the x-axis are used for publication profiling. The y-axis represents the MRR.

Influence of using Abstracts. In Table 4, we observe that abstracts have a positive influence for the entity scoring functions Freq, Basic, Bell, and BellLog (which base on spreading activation theory). On the other hand, for the functions involving the statistical methods, abstracts have no influence or negative influence.

6. DISCUSSION

Regarding the entity scoring functions, while Basic generally produces the largest similarity scores between a user's social media profile and the corresponding publication profile (see Table 1), the statistical methods TF-IDF, BM25, BM25C, and BM25HC demonstrate better performance for recommending researchers (see Table 4). It indicates that although similarity scores between profiles weighted by the statistical methods are small, they can distinguish a user's publication profile from others. Thus, for information retrieval or recommender systems, it is better to employ the statistical methods, including TF-IDF, BM25, BM25C, and BM25HC. Referring to the result shown in Table 4, we can state that while TF-IDF and BM25 perform best in the computer science dataset, BM25C and BM25HC outperform the others in the medicine dataset. A possible reason is the richness of the domain-specific knowledge base. While the CCS for the computer science dataset contains only 2,299 entities, MeSH for the medicine dataset has 27,300 entities. Thus, MeSH covers much more terms and may better enable to extract sufficient entities to represent user interests.

In terms of the temporal influence, we observe that there is no negative influence by adding the older publications for user profiling (see Figures 1 and 2). In addition, publication profiles covering only the latest years are not sufficient to get high similarity scores with a social media profile. On the other hand, Figure 3 shows a negative influence on recommending researchers, when using the old publications (i.e., paper published around 2004 and before). Thus, we should take into account temporal aspects for recommender systems. Referring to the two academic domains, we observe less temporal influence in the medicine dataset in both experiments (see Figures 2 and 4). A possible reason is that researchers working in the field of medicine might be more likely not to change their professional interests than computer scientists and the terminology in medicine is much more stable and less agile like in computer science where new "buzzwords" are emerging every six months.

Regarding the influence of the number of publications, we observe a moderate correlation in both experiments. In both experiments, the correlations between performances and the number of tweets are weaker than the ones between performances and the number of publications. In addition, users disseminate not only tweets relevant to professional interests but also ones unrelated to them on Twitter. Thus, the influence of the number of tweets regarding professional interests may vary depending on individual users. Thus, we investigate the average percentage of a user's tweets that contribute to the professional profile, i.e., how many tweets contain at least one entity from the considered domain. For the computer science dataset, we observe that on average 10.44% of the users' tweets (SD: 4.20) contain at least one entity in the CCS. In contrast, with MeSH we find on average 48.18% of the users' tweets (SD: 11.61) that contain at least one entity. Thus, tweets from the medicine dataset contain much more domain-specific entities than ones from the computer science dataset. A possible reason is that MeSH stores much more entities than the CCS. Thus, MeSH can pick entities from many tweets. On the other hand, publication titles and abstracts are texts focusing on professional interests. Thus,

Table 5: Kendall's τ between MRR and the number of publications. "all" denotes the titles and abstracts. The p-values in parentheses are marked in bold font if $\leq .05$.

	Computer Science		Medicine	
	title	all	title	all
Freq	.31 (**.00**)	.36 (**.00**)	.49 (**.00**)	.48 (**.00**)
Basic	.27 (**.00**)	.37 (**.00**)	.53 (**.00**)	.54 (**.00**)
Bell	.29 (**.00**)	.37 (**.00**)	.51 (**.00**)	.50 (**.00**)
BellLog	.33 (**.00**)	.38 (**.00**)	.49 (**.00**)	.48 (**.00**)
TF-IDF	.33 (**.00**)	.37 (**.00**)	.37 (**.00**)	.45 (**.00**)
CF-IDF	.24 (**.00**)	.31 (**.00**)	.41 (**.00**)	.44 (**.00**)
HCF-IDF	.28 (**.00**)	.36 (**.00**)	.43 (**.00**)	.45 (**.00**)
BM25	.36 (**.00**)	.44 (**.00**)	.47 (**.00**)	.53 (**.00**)
BM25C	.20 (**.00**)	.24 (**.00**)	.43 (**.00**)	.48 (**.00**)
BM25HC	.23 (**.00**)	.28 (**.00**)	.45 (**.00**)	.50 (**.00**)

Table 6: Kendall's τ between MRR and the number of tweets. "all" denotes the titles and abstracts. The p-values in parentheses are reported like above.

	Computer Science		Medicine	
	title	all	title	all
Freq	.07 (.36)	.00 (.99)	-.01 (.89)	.00 (.98)
Basic	.11 (.13)	.03 (.70)	-.05 (.59)	.00 (.96)
Bell	.10 (.20)	.00 (.96)	-.05 (.62)	-.02 (.82)
BellLog	.10 (.20)	.00 (.99)	-.02 (.79)	-.01 (.92)
TF-IDF	-.01 (.91)	-.03 (.65)	.01 (.94)	.03 (.72)
CF-IDF	.09 (.42)	.02 (.91)	.07 (.97)	.05 (.87)
HCF-IDF	.10 (.35)	.03 (.82)	.04 (.87)	.02 (.69)
BM25	.01 (.85)	.01 (.92)	-.01 (.95)	.00 (.97)
BM25C	.12 (.28)	.06 (.84)	.06 (.98)	.03 (.49)
BM25HC	.16 (.22)	.08 (.91)	.00 (.62)	.01 (.35)

we observe weaker correlation with the number of tweets, compared to the number of publications.

Referring to the influence of using abstracts, the results of both experiments show that the abstracts slightly improve the performance, compared to using only titles. Using spreading activation does not affect improvement of the performance by abstracts. Therefore, the titles generally contain sufficient amount of information to create a user profile and provide proper recommendations.

7. CONCLUSIONS

In this paper, we presented the results from two experiments on profiling professional interests and recommending researchers. We compared different scoring functions and investigate the influence of older data in both experiments. We conducted the experiments with the datasets in the field of computer science and medicine. The results revealed that while the functions using spreading activation produced large similarity scores between a user profile and publication profile, the scoring functions with the statistical methods performed best for recommendation. In terms of the temporal influence, the older data have almost no influence on the performance in the medicine dataset. However, in the computer science dataset, while there is a positive influence on user profiling, there is a negative influence in recommending researchers when adding too old data.

8. REFERENCES

[1] F. Abel, Q. Gao, G.-J. Houben, and K. Tao. Analyzing temporal dynamics in Twitter profiles for personalized recommendations in the social web. In *WebSci*, pages 1–8. ACM, 2011.

[2] F. Abel, Q. Gao, G.-J. Houben, and K. Tao. Analyzing user modeling on Twitter for personalized news recommendations. In *UMAP*, pages 1–12. Springer, 2011.

[3] F. Abel, E. Herder, and D. Krause. Extraction of professional interests from social web profiles. In *AUM*, pages 1–6, 2011.

[4] J. Chen, R. Nairn, L. Nelson, M. Bernstein, and E. Chi. Short and tweet: experiments on recommending content from information streams. In *CHI*, pages 1185–1194. ACM, 2010.

[5] A. M. Collins and E. F. Loftus. A spreading-activation theory of semantic processing. *Psychological Review*, 82(6):407–428, 1975.

[6] T. De Pessemier, S. Dooms, T. Deryckere, and L. Martens. Time dependency of data quality for collaborative filtering algorithms. In *RecSys*, pages 281–284. ACM, 2010.

[7] F. Goossen, W. IJntema, F. Frasincar, F. Hogenboom, and U. Kaymak. News personalization using the CF-IDF semantic recommender. In *WIMS*. ACM, 2011.

[8] G. Grosse-Bölting, C. Nishioka, and A. Scherp. Generic process for extracting user profiles from social media using hierarchical knowledge bases. In *ICSC*, pages 197–200. IEEE, 2015.

[9] P. Kapanipathi, P. Jain, C. Venkataramani, and A. Sheth. User interests identification on Twitter using a hierarchical knowledge base. In *ESWC*, pages 99–113. Springer, 2014.

[10] J. Letierce, A. Passant, J. Breslin, and S. Decker. Understanding how Twitter is used to spread scientific messages. In *WebSci*, pages 1–8. ACM, 2010.

[11] C. D. Manning, P. Raghavan, and H. Schütze. *Introduction to information retrieval*, volume 1. Cambridge university press Cambridge, 2008.

[12] S. E. Robertson, S. Walker, and M. Beaulieu. Okapi at TREC-7: Automatic ad hoc, filtering, VLC and interactive track. In *TREC-7*, pages 253–264. NIST, 1999.

[13] G. Salton, A. Wong, and C.-S. Yang. A vector space model for automatic indexing. *Communications of the ACM*, 18(11):613–620, 1975.

[14] J. P. Shaffer. Modified sequentially rejective multiple test procedures. *Journal of the American Statistical Association*, 81(395):826–831, 1986.

[15] W. Shen, J. Wang, P. Luo, and M. Wang. Linking named entities in tweets with knowledge base via user interest modeling. In *KDD*, pages 68–76. ACM, 2013.

[16] R. Zafarani and H. Liu. Evaluation without ground truth in social media research. *Communications of the ACM*, 58(6):54–60, 2015.

[17] Q. Zheng and H. H. Ip. Effectiveness of the data generated on different time in latent factor model. In *RecSys*, pages 327–330. ACM, 2013.