Profiling vs. Time vs. Content: What does Matter for Top-k Publication Recommendation based on Twitter Profiles?

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

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

ABSTRACT

So far it is unclear how different factors of a scientific publication recommender system based on users' tweets have an influence on the recommendation performance. We examine three different factors, namely profiling method, temporal decay, and richness of content. Regarding profiling, we compare CF-IDF that replaces terms in TF-IDF by semantic concepts, HCF-IDF as novel hierarchical variant of CF-IDF, and topic modeling. As temporal decay functions, we apply sliding window and exponential decay. In terms of the richness of content, we compare recommendations using both full-texts and titles of publications and using only titles. Overall, the three factors make twelve recommendation strategies. We have conducted an online experiment with 123 participants and compared the strategies in a within-group design. The best recommendations are achieved by the strategy combining CF-IDF, sliding window, and with full-texts. However, the strategies using the novel HCF-IDF profiling method achieve similar results with just using the titles of the publications. Therefore, HCF-IDF can make recommendations when only short and sparse data is available.

1. INTRODUCTION

The social media platform Twitter is popular among scientists to share and discuss their professional thoughts and interests [14]. Thus, they are a natural resource for building up a user's professional profile and using it for recommending scientific publications. Recommending scientific publications based on a user's social media items has several advantages: First, users receive recommendations based on their current and ongoing professional interests. In contrast, systems like Google Scholar and Sugiyama et al. [26] recommend scientific publications based on a user's publication record. It can take up to two years (for conferences) or longer (for journals) until a paper is taken into consideration by the recommender system. Second, content-based profiling from a user's social media items mitigates the well-known cold-start problem observed in collaborative filtering systems [11]. The cold-start problem refers to the initial situation where a recommender system yet does not know anything about a user's interests. Collaborative filtering systems need to analyze a large amount of user activities in order to provide reasonable recommendations. In contrast, content-based recommender systems like our work make recommendations based on similarity scores between a user profile and candidate items. Therefore, they can generate recommendations based on a single user profile already.

There is various research on user profiling from social media items [5, 21, 24, 29] and recommending scientific publications [15, 26, 28]. However, it is unclear how different profiling methods affect the recommendation performance. In addition, the age of social media items as well as scientific publications has an influence on profiling [24, 21]. But again, it has not been compared. Finally, we investigate whether it is possible to make reasonable recommendations when using only the publications' titles, i. e., when only short and sparse information about the candidate items is available. We have conducted an online experiment to evaluate these three factors of top-k recommendations of scientific publications based on a user's social media profile. In detail, the factors are:

(i) Profiling Method: The first factor is the *Profiling Method*, where we use Concept Frequency Inverse Document Frequency (CF-IDF) [7] as baseline. CF-IDF is a modification of TF-IDF where term frequencies are replaced by frequencies of semantic concepts. In an experiment with 19 participants, Goossen et al. have shown that CF-IDF outperforms TF-IDF for news article recommendations [7]. Recently, we have extended the statistical strength of CF-IDF with the semantics provided by a hierarchical knowledge base [19]. The resulting Hierarchical CF-IDF (HCF-IDF) model is capable of revealing semantic concepts that are not explicitly mentioned in texts but still are highly relevant. This is achieved by applying a spreading activation over a hierarchical knowledge base, which is typically provided as domain-specific taxonomy. Please note that we also considered using BM25 and TF-IDF as profiling method. However, our earlier work showed that HCF-IDF performs better for user profiling from social media items [19]. As third method, we apply Latent Dirichlet Allocation (LDA) [2, 1], a state-of-the-art topic modeling method. LDA is a generative machine learning approach and thus does not require any prior information such as a knowledge base.

(ii) Decay Function: As the second factor, we investigate two temporal *Decay Functions*. They are based on

JCDL '16, June 19 - 23, 2016, Newark, NJ, USA

ISBN 978-1-4503-4229-2/16/06…$15.00

DOI: http://dx.doi.org/10.1145/2910896.2910898

the idea that the importance of information declines gradually as time passes. We compare sliding window [24] and exponential decay [21, 26]. Both decay functions have been used in the past for user profiling [24, 21, 26]. But so far no comparative study was carried out.

(iii) Document Content: The third factor defines the richness of *Document Content* used for profiling candidate items (i.e., scientific publications). We compare the use of full-texts and titles of scientific publications for profiling versus profiling only based on titles.

We compared twelve recommendation strategies making use of different combinations of the three factors described above. For the experiment, we have recruited $n = 123$ participants who are posting about their professional interests on Twitter. For each strategy, the participants have received recommendations of five publications from a large corpus of $|D| = 279,381$ scientific publications in the broader field of economics. We used rankscore [4] to measure the recommendation performance. We also computed Mean Average Precision (MAP), Precision, Mean Reciprocal Rank (MRR), and normalized Discounted Cumulative Gain (nDCG), which show similar results and documented in the TR [20].

The results are very interesting: The strategy that employs the profiling method CF-IDF and the decay function Sliding window with both titles and full-texts achieves the overall best recommendation performance. Although the strategy using CF-IDF shows the highest performance, it has a drawback that it requires full-texts of scientific publications. Thus, it is remarkable that the strategies with HCF-IDF can achieve comparable results using only titles. We observe no significant difference between the best performing strategy and strategies with HCF-IDF. Thus, we conclude that the use of the spreading activation function over the hierarchical knowledge base enables HCF-IDF to compensate for the sparseness when only titles are available due to e.g., legal reasons to hinder the use of full-texts. Please note, there is no lack in domain-specific hierarchical knowledge bases such as the one used in the experiment for economics. In fact, these knowledge bases are freely available for many domains[1]. Furthermore, they are manually crafted by domain experts and thus are of high quality.

In addition, we have applied a correlation analysis between the recommendation performance and the number of tweets a participant has published, the number of concepts extracted from these tweets, the number of concepts extracted per tweet, and the percentage of tweets containing at least one concept respectively. Our results show no significant correlations in any strategies. Thus, the methods are robust against the amount of tweets.

Subsequently, we review related work in Section 2. Section 3 introduces the problem definition. In Section 4, we describe the three experimental factors used in o4ur recommender system. We present the experiment setup and procedure in Section 5. The results are presented in Section 6 and discussed in Section 7 before we conclude the paper.

2. RELATED WORK

Recommender systems are categorized into content-based recommender and collaborative filtering [11]. Collaborative filtering requires analyzing a large amount of user activities in order to predict items to other users [29]. In contrast, we focus on content-based recommender, which suggest items based on similarity scores between a user profile and candidate items. A content-based recommender can make recommendations based on data from a single user already. Thus it does not suffer from the cold start problem. Recommender systems for scientific publications mostly employed user profiles based on publications [26, 27] or clicks [15]. Instead, we create user profiles based on social media items.

Many works have extracted user interests from social media platforms [5, 21, 24, 29]. Chen et al. [5] studied a recommender system incorporating Twitter, which recommended URLs based on a user's tweets and follower-followee relationships. In order to find out the best recommendation strategy, they evaluated twelve strategies from three factors: content sources, topic interest models for users, and social popularity. Referring to the factor content sources, Chen et al. showed that profiling based on one's own tweets performed better than based on tweets by one's followees. Hence, we build up user profiles from social media items produced by the users themselves.

In the past years, profiling methods based on semantic concepts (i.e., ontology-based profiling) extraction have been developed [7, 16]. They extract semantic concepts from texts, using a structured knowledge base, e.g., DBpedia. Goossen et al. [7] proposed CF-IDF, as an extention of TF-IDF. CF-IDF counts frequencies of a concept instead of a term. Their news arcticle recommendation experiment with 19 participants demonstrated that CF-IDF outperforms TF-IDF. Lu et al. [16] proposed a recommender system for tweets based on what a user tweeted. They constructed user profiles represented as a set of weighted Wikipedia concepts that correspond to Wikipedia articles. The experiment demonstrated that concept-based approaches outperform TF-IDF. Other works employed a hierarchical structure of a knowledge base for profiling [12, 18, 16] and demonstrated their effectiveness. These approaches can reveal user interests that are not explicitly mentioned in the texts, using a structure of a knowledge base and spreading activation. In particular, Middleton et al. [18] constructed user profiles based on a hierarchical knowledge base using spreading activation for a recommender system of scientific publications. Their user experiment compared a profiling method using the structure of a hierarchical knowledge base and a method not using the structure. The result demonstrated superiority of using the hierarchical knowledge base. Topic modeling such as LDA [2] is one of the most popular profiling methods. It is used in the context of social media [10] but particularly suited for document profiling.

Time-aware user profiles are constructed based on the assumption that the degree of user interests declines as time passes. The decline of user interests is modeled by a decay function. In the past, the decay functions sliding window [24] and exponential decay [21, 26] have been employed for user profiling. However, they have not been compared so far like we do in this work.

3. PROBLEM DEFINITION

We address the problem of taking the social media stream as input in order to recommend items such as scientific publications the user might be interested in. The problem can be decomposed into three parts: (1) First, we need to extract the professional interests that a user exposes through his social media stream and represent the interests in a user

[1] http://www.w3.org/2001/sw/wiki/SKOS/Datasets

Table 1: Symbol Notation

u	a user
i	a social media item
I_u	the set of u's social media items
c	a concept
C	the set of concepts
d	a candidate item (scientific publication)
D	the set of candidate items
t_i, t_d	the time stamp of i and d, respectively
P_u	u's user profile
P_d	d's document profile
Φ	a profiling function
w'	a weighting function (not considering temporal decay)
f	a decay function
w	a weighting function that extends w' with temporal decay
σ	a similarity function

Table 2: Three factors and their choices for the experiment spanning in total $3 \times 2 \times 2 = 12$ strategies

Factor	Possible Design Choices		
Profiling Method	CF-IDF	HCF-IDF	LDA
Decay Function	Sliding window		Exponential decay
Document Content	All (title + full-text)		Title

profile. (2) Likewise, we profile candidate items (i.e., scientific publications) and represent them in a way that they are comparable with the user profile. (3) We need a ranking function to compute the top-k items based on similarity scores between the user profile and each candidate item. In the following, we formalize the three steps required to create a recommender system based on a user's professional interests extracted from the social media stream. Symbols used in this paper are summarized in Table 1.

(1) User profiling from social media items. We consider I_u as set of social media items i produced by user u. A social media item $i \in I_u$ has a certain time stamp t_i. Subsequently, P_u, the user profile of the user u, is created over a set of concepts C by assigning a specific weight for each concept $c \in C$. Generally speaking, a concept c is a key subject in a dedicated field, coming from a given domain-specific knowledge base C. For instance, "financial crisis" is a concept in the field of economics. We construct P_u by employing different user profiling functions Φ and we compare them. Formally, user profiles are defined as:

$$P_u = \Phi(I_u, C) := \{(c, w(c, I_u)) \mid \forall c \in C\} \qquad (1)$$

Here, w is an arbitrary weighting function that returns a weight of a concept c in a user's social media stream I_u. Thus, it determines how important a concept c is for the user u. Profiling functions Φ and weighting functions w are described in Sections 4.1 and 4.2. Specifically, we describe weighting functions w' that do not consider temporal decay in Section 4.1 and provide weighting functions w which extend w' with temporal decay in Section 4.2.

(2) Profiling candidate items. We have a set of candidate items D. A candidate item $d \in D$ has a time stamp t_d, indicating its published year. To determine the similarity scores between a user profile P_u and each candidate item $d \in D$, we need to construct profiles of candidate items in a way that they are comparable with the user profile. Formally, we represent a candidate item d as a profile $P_d = \Phi(d, C) := \{(c, w(c, d)) \mid \forall c \in C\}$. Since our candidate items are scientific publications, we refer to this process document profiling.

(3) Ranking candidate items. We rank candidate items based on similarity scores between the user profile P_u and a document profile P_d. A similarity function σ takes as input a user profile P_u and document profile P_d. It is defined as $\sigma(P_u, P_d) \to [0, 1]$. The similarity function is applied to all candidate items $d \in D$. Finally, the top-k most relevant items (i.e., documents whose similarity scores with P_u are ranked in the top-k) are recommended to the user u. The similarity functions σ are described in Section 4.3.

4. EXPERIMENTAL FACTORS

According to the three factors (i)-(iii) stated in the introduction, we form the design space of our experiment. We illustrate the design space in Table 2, where each cell is a possible design choice we can make in one of the three factors. Subsequently, we detail the factor *Profiling Method* in Section 4.1 and the factor *Decay Function* in Section 4.2. Further, we describe similarity functions σ in Section 4.3. The factor *Document Content* investigates whether full-texts of scientific publications enhance the recommendation performance compared to using only titles.

4.1 Profiling Method

We investigate three methods for user profiling and document profiling. For each method, we define a weighting function w' that gives a certain weight to each concept c. The final weighting function w taking temporal decay into account is described in Section 4.2.

CF-IDF: Compared to the traditional TF-IDF, CF-IDF (Concept Frequency Inverse Document Frequency) counts frequencies of a semantic concept instead of term frequencies [7]. Semantic concepts or short concepts are stored in an external knowledge base. Each concept has a unique resource identifier (URI) and one or more labels describing the concept [2]. The concept's labels are treated as synonyms. As an example, the concept "clothing industry" has the URI `http://zbw.eu/stw/version/latest/descriptor/13128-2` and is defined in the thesaurus STW, a domain-specific knowledge base for economics (described in Section 5.3). The concept has not only the label "clothing industry" but also the synonymous labels "garment industry" and "apparel industry". We count the label frequency, i.e., the number of times the label appears, in the social media items and candidate items. Subsequently, we calculate the concept frequency, i.e., the number of times the concept appears, by summing up the frequencies of the labels referring to the concept. For instance, if the labels "clothing industry" and "garment industry" appear twice and once in a text, the total frequency of the concept referring to "clothing industry" is three.

For the social media items I_u of the user u, CF-IDF is computed along with Equation 2.

$$w'_{cf\text{-}idf}(c, i) = cf(c, i) \cdot \log \frac{|I_u| + |I_r|}{|\{i \in I_u \cup I_r : c \in i\}|}, \qquad (2)$$

where $cf(c, i) = \frac{\text{the number of times concept } c \text{ appears in } i}{\text{the number of times all concepts appear in } i}$. The denominator $|\{i \in I_u \cup I_r : c \in i\}|$ counts the number of social media items that contain a concept c. I_r is a set of random social media items.

[2] https://www.w3.org/DesignIssues/LinkedData.html

We employ a set of random social media items I_r, because it allows to better distinguish the relevant concepts in the user's social media items I_u, as Chen et al. [5] and Lu et al. [16] did for TF-IDF. For instance, assuming there are two social media items from a user u and both include the concept "currency competition". Although "currency competition" should have a high weight in the user profile, in this case IDF and a final CF-IDF score would be 0 because "currency competition" is common in a user u's social media items. The random social media items are sampled from public microblog postings. In our case, they are obtained from the public Twitter stream using the Twitter API.

We have conducted a simple pre-experiment to empirically determine the optimal amount of random tweets to be used in the profiling method in the context of our experiment of recommending economics publications. Given this pre-experiment, we set the size of random social media items to five times of $|I_u|$. In more detail, we applied different sizes of I_r, starting from 0 to 1000 random tweets. For 26 Twitter accounts, we computed the IDF scores for user profile over $I_u \cup I_r$ and compared it using cosine similarity with the user profile computed only over I_u. The Twitter accounts were taken from a list of famous economists[3] that are frequently tweeting. We ensured that the set of random tweets I_r is disjoint do the user's tweets, i. e.. $I_r \cap I_u = \emptyset$. Particularly, we looked into the changes of the cosine similarity while adding more random tweets. We observed the changes in the IDF scores became stable after about a factor of five w.r.t. to $|I_u|$. The changes indicate the influence of the IDF scores to user profile. Using this technique is effective as the IDF score ensures that too generic concepts do not get too high weights in the user profiling. Those generic concepts are at the upper levels of the hierarchy of the domain-specific knowledge base. In our case those concepts are like "product" and "economics". Please note that the factor may depend on the domain of economics considered in this paper and that a different factor may be chosen for other domains.

Regarding document profiling, CF-IDF is computed as defined in Equation 3. The computation is basically identical with the one for user profiling shown in Equation 2. The difference is that CF is computed over single documents and IDF is computed over the document collection.

$$w'_{cf\text{-}idf}(c,d) = cf(c,d) \cdot \log \frac{|D|}{|\{d \in D \ : c \in d\}|} \tag{3}$$

HCF-IDF: The novel profiling method HCF-IDF (Hierarchical CF-IDF) [19] extends CF-IDF by using a hierarchical knowledge base, where the concepts are hierarchically organized in a taxonomy. HCF-IDF can reveal concepts that are indirectly mentioned in texts by applying a spreading activation over the hierarchical knowledge base. Figure 1 shows an example where a user's profile includes the concept "social recommendation". Due to the hierarchical structure of the knowledge base, also the concepts "web searching" and "world wide web" are activated and obtain non-zero weights even if they are not mentioned. Different from the profiling methods using spreading activation [12, 18], HCF-IDF avoids to provide too high weights to generic concepts like "economy", as it employs IDF. Specifically, HCF-IDF combines the statistical strength of CF-IDF with semantics provided by the hierarchical knowledge base. We compute HCF-IDF along with Equation 4.

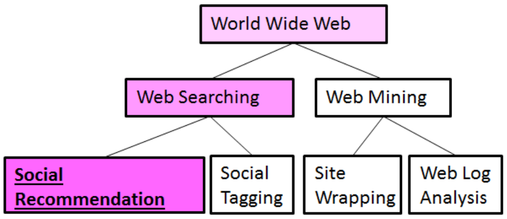


Figure 1: An example of HCF-IDF

$$w'_{hcf\text{-}idf}(c,i) = BL(c,i) \cdot log \frac{|I_u| + |I_r|}{|\{i \in I_u \cup I_r : c \in i\}|} \tag{4}$$

$BL(c,i)$ denotes the spreading activation function BellLog from Kapanipathi et al. [12]. It returns a weight of a concept c in a social media item i and is defined below:

$$BL(c,i) = cf(c,i) + FL(i) \cdot \sum_{c_j \in C_l(c)} BL(c_j,i), \tag{5}$$

where $FL(c) = \frac{1}{\log_{10}(nodes(h(c)+1))}$. $h(c)$ returns the level where a concept c is located in the knowledge base and *nodes* provides the number of concepts at a given level in a knowledge base. For example, in Figure 1 h("web searching') returns 2 and $nodes(h($"web searching'$) + 1)$ returns 4. $C_l(c)$ returns the set of concepts located in one level lower than the concept c. In Figure 1 the function C_l("world wide web") returns "web searching" and "web mining".

For scientific publications, weights are computed as defined in Equation 6. The computation is basically identical with the one for user profiling as shown in Equation 4. The difference is that BL is applied over single documents and IDF is computed over the document collection.

$$w'_{hcf\text{-}idf}(c,d) = BL(c,d) \cdot log \frac{|D|}{|d \in D : c \in d|} \tag{6}$$

LDA: As third profiling method, we use LDA [2, 1], an unsupervised topic modeling method. LDA identifies latent topics in a document collection, where each document is represented as a probability distribution over topics, while each topic is again represented as a probability distribution over a number of words. Please note that for user profiling, we treat the set of social media items I_u published by a user u as one *single* social media document in this profiling method. It is known that topic models that treat a user's microblog postings as one combined social media document outperform topic models computed over single postings of a user for recommendation tasks [10]. We first create a topic model for the entire document collection D (using the parameters and tools described in detail in Section 5.3). Subsequently, we run LDA with the given topic model for the document collection D and infer a probability distribution over topics for the user's social media document I_u.

Again, we use the same notation of concepts c as introduced above: Each topic generated by LDA is treated as a concept $c \in C$. The weight of a concept c is defined by $w'_{lda}(c, I_u) = p(c \mid I_u)$ for user profiles and $w'_{lda}(c,d) = p(c \mid d)$ for document profiles, where $p(c \mid d)$ and $p(c \mid I_u)$ denote the probability of the concept (i.e., topic) c in the social items I_u and document d, respectively.

[3] http://www.huffingtonpost.com/2012/11/13/economists-twitter_n_2122781.html

4.2 Decay Function

We compare two decay functions f, namely sliding window and exponential decay. In the past, both functions have been used in recommender systems [24, 21, 26]. However, so far they have not been empirically compared. The profiling functions w' described in the previous section are combined with a decay function f in order to obtain a final weight w. The final weights are computed by Equation 7 for the set of social media items and Equation 8 for the candidate items.

$$w(c, I_u) = \sum_{c \in i : i \in I_u} f(t_i) \cdot w'(c, i) \tag{7}$$

$$w(c, d) = f(t_d) \cdot w'(c, d) \tag{8}$$

Please note that when employing LDA, the decay functions can only be applied on the candidate items, because we treat the user's social media items as one single document.

Sliding Window: There are two kinds of sliding window functions, whose window size is defined by (a) the number of items [13] and (b) the period of time [25]. The approach (a) is employed to identify relatively short-term features (e.g., user interests from web browsing histories) [13], while the approach (b) is used to identify long-term features [25]. We aim at extracting a user's professional interests, which are rather long-term. Thus, we take the approach (b) and use only social media items and documents that are younger than a given threshold point in time $thresh$. The sliding window function can be represented as Equation 9.

$$f_{sw}(t) = \begin{cases} 1 & for\ t \geq thresh \\ 0 & for\ t < thresh \end{cases} \tag{9}$$

For user profiles, we set the threshold based on the work by Orlandi et al. [21]. They found out that the half life time is $thresh_{social} = 250\ days$. For document profiles, Sangam et al. [22] observed that the half-life time of the scientific publications in the field of social science is 9.04 $years$. In our experiment, we use a dataset of scientific publications in economics (see Section 5.3), which has a large overlap with social science. Thus, we set $thresh_{doc} = 9.04\ years$ [22] and remove scientific publications published more than 9.04 $years$ ago from the candidate items.

Exponential Decay: The exponential decay function is defined as shown in Equation 10.

$$f_{exp}(t) = e^{-(t_{current} - t)/\tau}, \tag{10}$$

where $t_{current}$ denotes the current time and τ is a positive number presenting mean-life [21]. For user profiles, we set $\tau = 360\ days$ based on Orlandi [21]. Since Sangam et al. [22] found out that the mean-life of scientific publications in social sciences is 13.05 $years$, we set $\tau = 13.05\ years$ for document profiles.

4.3 Similarity Functions

We calculate the similarity scores between a user profile P_u and each document profile P_d. We cast a user profile P_u and document profiles P_d to a user profile vector $\vec{p}_u$ and document profile vectors $\vec{p}_d$, respectively. Each element in the vectors corresponds to a weight of a concept c.

Temporal Cosine Similarity: We employ the temporal cosine similarity function described in Equation 11 for the profiling methods CF-IDF and HCF-IDF.

$$\sigma_{tcossim}(P_u, P_d) = f(t_d) \cdot \frac{\vec{p_u} \cdot \vec{p}_d}{||\vec{p_u}|| \cdot ||\vec{p}_d||}, \tag{11}$$

It extends the cosine similarity by the function $f(t_d)$, which results in higher similarity score to newer documents. $f(t_d)$ is a decay function from Equation 9 or Equation 10. t_d is time stamp of a scientific publication d. i.e., the year at which d was published.

Dot Product: For LDA, we employ the dot product computed as $\sigma_{dp}(p_u, p_d) = \vec{p}_u \cdot \vec{p}_d$. Since LDA represents documents as probability distribution, it is more reasonable to use Kullback-Leibler divergence (KL divergence). However, the dot product outperforms cosine similarity and Kullback-Leibler divergence (KL divergence) when representing documents using LDA [9].

5. EVALUATION

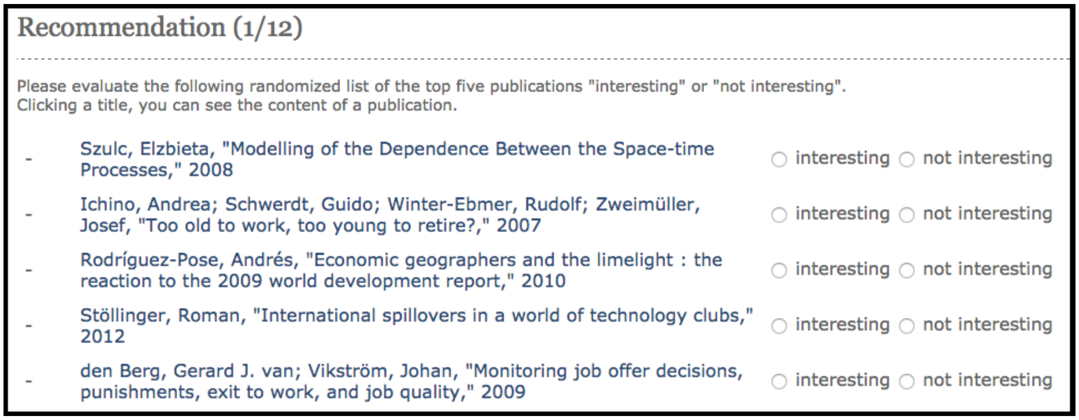
We conducted an online experiment with $n = 123$ participants in order to identify the best strategy for a recommender system along the factors described in Section 4. As social media platform, we choose Twitter as it is widely used in scientific communities [14]. We design our experiment following the experiment setup and procedure of Chen et al. [5]: Each participant obtains top-5 recommendations for each of the twelve strategies formed from the three factors. The recommendation performance of each strategy is measured by the rankscore [4]. Below, we describe the details of our experiment procedure and participants. Subsequently, we explain the dataset and the knowledge base used in the experiment. Finally, we introduce our evaluation metric.

5.1 Procedure

The participants are invited to a web application implementing the twelve recommendation strategies. First, participants input their public Twitter handles and e-mail address. Then, the participants' tweets are retrieved from the Twitter API. Subsequently, user profiles are created from the tweets using each of the three profiling methods and two decay functions. Based on the user profiles, personalized top-k recommendations of scientific publications are generated for each of the twelve strategies. We set the number of recommendations per strategy $k = 5$ along with Chen et al. [5]. After computing the recommendations, the participants receive an e-mail invitation to assess the recommendations. The participant go through all of the twelve strategies like as Chen et al. [5]. Thus, we apply a repeated measures design. Each participant obtains $12 \cdot 5 = 60$ recommendations in total throughout the experiment.

Prior to starting the experiment, participants are informed about the task of the experiment, i.e., rating the recommended publications based on relevance to their research interests, and confirmed consent. On each of the subsequent pages, the participants see a list of five recommendations produced by one of the twelve strategies. An example screenshot of the evaluation page is shown in Figure 2.

For each recommended scientific publication, the participants see its bibliographic information, i.e., authors, title, and year of publication. In addition, participants can look into the original PDF files by clicking on a link attached to



Recommendation (1/12)

Please evaluate the following randomized list of the top five publications "interesting" or "not interesting".
Clicking a title, you can see the content of a publication.

- Szulc, Elzbieta, "Modelling of the Dependence Between the Space-time Processes," 2008 ○ interesting ○ not interesting
- Ichino, Andrea; Schwerdt, Guido; Winter-Ebmer, Rudolf; Zweimüller, Josef, "Too old to work, too young to retire?," 2007 ○ interesting ○ not interesting
- Rodríguez-Pose, Andrés, "Economic geographers and the limelight : the reaction to the 2009 world development report," 2010 ○ interesting ○ not interesting
- Stöllinger, Roman, "International spillovers in a world of technology clubs," 2012 ○ interesting ○ not interesting
- den Berg, Gerard J. van; Vikström, Johan, "Monitoring job offer decisions, punishments, exit to work, and job quality," 2009 ○ interesting ○ not interesting

Figure 2: Screenshot of our experiment web page showing a randomized list of top-5 recommendations for the first of twelve strategies (which again are randomly ordered). For each recommendation the participants could assess the bibliographic record as well as click on the full-text document. The participants rated each recommended publication as "interesting" or "not interesting"' based on their research interests.

the bibliographic record. In order to avoid bias, the participants go through the twelve strategies in random order. For each strategy, the participants receive one list of five recommendations. The five recommendations in the lists are again shown in random order to the participants to avoid the well-known ranking bias. Typically, participants assume that top-ranked recommendations are essentially more relevant [3, 5]. Thus, again prior to starting the experiment we have explicitly informed the participants that we have randomized the order of the items in the top-5 lists. However, the actual ranks of the recommendations as well as their positions where the recommended items appeared on the participants' screen are stored in the database for later analyses. Participants evaluate each recommendation as "interesting" or "not interesting" by clicking on radio buttons next to the publication records like Chen et al. [5]. Please note, the participants had to evaluate all recommended items.

At the end of the experiment, we collect the demographic information of each participant, including gender, age, highest academic degree, major, years of profession, and current employment status (academia/industry). Finally participants could state free comments regarding the experiment.

5.2 Participants

We recruited $n = 123$ participants through mailing lists, tweets, and word-of-mouth on the Internet. Initially 160 participants registered their Twitter handles and email address for our experiment. Among them, 134 participants started the experiment after receiving the e-mail invitation. From these 134 participants, only eleven dropped in the course of assessing the recommendations in the twelve strategies. Thus, finally we obtain evaluations for all strategies from $n = 123$ participants. From these, 27 participants are female. The average age of the participants is 32.83 years (SD: 7.34). Regarding the highest academic degree, we have acquired 21 with a Bachelor, 58 have a Master, 32 a PhD, and 12 are lecturers/professors. While 83 participants work in academia, 40 work in industry. Tweets of the participants were retrieved via Twitter API. We only collected tweets in English as the scientific publications are also in English. The participants published on average 1096.82 English tweets (SD: 1048.46). The maximum and minimum numbers of tweets are 3192 and 2, respectively. Twitter users who have not produced any tweets in the last 250 *days* could not register and participate in the experiment, since we use a 250 *days* threshold for the decay function Sliding window (see Section 4.2). Five Twitter users could not participate in the experiment for this reason.

The participants spent on average 517.54 seconds to complete the assessment of the $5 \times 12 = 60$ recommendations (SD: 376.72). This does not include the time spent to register for the experiment, read the instructions, and filling out the final questionnaire. As incentive, each participant received the information about his most similar economist among 26 famous economists[4] and the top-5 dominant semantic concepts in their tweets after the experiment. In addition, the participants could opt-in to a raffle for one of two Amazon vouchers worth of 50 €.

5.3 Dataset Preparation

We use a large-scale dataset of scientific publications in the field of economics as candidate items and a high-quality taxonomy as a knowledge base for profiling methods.

Dataset of Scientific Publications. We collaborate with the providers of EconBiz[5], a portal for scientific publications in economics managed by ZBW, the German National Library of Economics. From this portal, we obtained 1 million URLs of open access publications and extracted full-texts and metadata (i .e., authors, title, year of publication) of 413, 098 scientific publications. Finally, we determined the document language[6] and got 279, 381 scientific publications in English, which were used in this experiment.

Knowledge Base in Economics. The ZBW also maintains and further develops the hierarchical knowledge base STW[7], a thesaurus specialized for the field of economics. The STW is freely available and is of high quality due to its manual maintenance by domain experts. The knowledge

[4] http://www.huffingtonpost.com/2012/11/13/economists-twitter_n_2122781.html

[5] http://www.econbiz.de/

[6] https://code.google.com/p/language-detection/

[7] http://zbw.eu/stw/version/8.12/about.en.html

base is poly-hierarchically organized with six levels. It contains $6,335$ semantic concepts and $11,679$ labels. The hierarchically organized concepts are connected with each other via $14,875$ edges. In order to extract as many labels as possible, we enhanced the original STW with DBpedia redirects[8]. From DBpedia redirects we can retrieve the synonymous labels for a concept. STW contains $2,692$ concepts that have both a DBpedia mapping and one or more DBpedia redirects. As an example, for the concept "Telecommunications industry" in the thesaurus, we obtain the DBpedia redirects "Telecommunications operator" and "Telephone companies" and use them as synonymous labels referring to the concept "Telecommunications industry". Finally, our extended STW contains $6,335$ concepts and $37,733$ labels. This extended STW is used for the profiling methods CF-IDF and HCF-IDF. For CF-IDF, we ignore the edges between concepts.

Processing of the tweets and publications. For the profiling methods CF-IDF and HCF-IDF, we extract semantic concepts from the participants' tweets and the scientific publications by matching the texts with the labels from the extended STW (i. e., a gazetteer-based approach). Before processing, we lemmatize both the tweets and the scientific publications using Stanford Core NLP[9] and remove stop words. Regarding the tweets, some of them contain hashtags indicating topics (e .g., #election) and user mentions (e .g., @UNICEF). We remove only the symbols # and @ from the tweets as Feng et al. [6] observed that the combination of the tweets' textual content with the hashtags and user mentions made the highest performance for tag recommendation.

This process extracts only the users' professional interests from tweets and helps to avoid noise (i .e., topics not relevant to professional interests in economics). A participant has published on average 1096.82 tweets (SD: 1048.46). On average $1,214.93$ concepts (SD: 1181.43) are contained in a participant's tweets and 1.07 concepts (SD: 0.31) are contained per tweet. Regarding CF-IDF and HCF-IDF, we calculate the ratio of the number of tweets containing at least one concept and the total number of tweets the user has published. This indicates the percentage of tweets that have contributed to creating the user profile. On average, 62.24% of the tweets (SD: 13.55) that a participant has published contain at least one concept in economics. These tweets are assumed to be relevant to the professional interests.

LDA. For constructing profiles by LDA, we use JGibbLDA[10]. We first run LDA to generate the topic model based on the given document set D. Following Blei et al. [1], we lemmatize the scientific publications using Stanford NLP Core. Subsequently, we remove stop words and words that appear in fewer than 25 scientific publications. We optimized the number of topics K regarding the maximum mean log likelihood of words given topics as suggested by Griffiths et al. [8]. We experimented with $K = 20$, 50, 100, 200, 500, 1000, and 5000 and obtained the highest log likelihood for $K = 100$. All topic models were computed over 500 iterations. Regarding the further parameters for LDA, we set $\alpha = 0.5$ and $\beta = 0.1$ as suggested by Griffiths et al. [8]. To infer a topic distribution over a user's tweets, we run LDA again using the topic model for the document set D with 200 iterations. Prior to this, we prepare the tweets of a user u in a single social media document as described in Section 4.1.

5.4 Evaluation Metric

In order to assess the recommendation performance, we compute the rankscore [4] as used by Bostandjiev et al. [3] and introduced by Jannach et al [11]. Rankscore posits that each successive item in a list is less likely to be viewed by users with an exponential decay, as defined in Equation 12.

$$rankscore' = \sum_{d \in hits} \frac{1}{2^{\frac{rank_d - 1}{\theta - 1}}} \tag{12}$$

θ denotes a viewing halflife parameter controlling the speed of the exponential decay. As suggested by Breese et al. [4], we set $\theta = 5$. $hits$ refers to the set of documents d evaluated as "interesting" and $rank_d$ denotes the rank of a recommended item d in a list. Please note $rank_d$ denotes the actual rank stored in the database different from the position where a item d appears in the list (cf. Section 5.1). The normalized rankscore is computed by $rankscore = rankscore'$ $/rankscore_{max}$, where the maximum rankscore $rankscore_{max} = \sum_{j=1}^{k} \frac{1}{2^{\frac{j-1}{\theta-1}}}$. Here, k is the number of the recommended items. We set $k = 5$. We also computed Mean Average Precision (MAP), Precision@5, Mean Reciprocal Rank (MRR), and normalized Discounted Cumulative Gain (nDCG). Overall, the results are similar to the rankscore and thus omitted for reasons of brevity. The interested reader may refer to the details in the appendix [20].

6. RESULTS

In this section, we document the results of the experiment[11] and conduct the statistical analyses. We set a significance level of $\alpha = 5\%$ for all statistical tests (please do not confuse with α for LDA in Section 5.3).

6.1 Quantitative Analyses

We first report the best performing strategy among the twelve strategies. Subsequently, we analyze the influence by the experimental factors followed by investigating the correlations between the recommendation performance and the numbers of tweets written by a user. Finally, we analyze the performance related to the number of times the participants clicked on the full-text of a publication.

Best performing strategy. Table 3 documents the average rankscores of the twelve strategies sorted in decreasing order. Overall, the best performing strategy is the strategy CF-IDF × Sliding window × All. We apply a one-way repeated-measure ANOVA in order to identify if there are significant differences between the strategies. For using ANOVA, we first need to verify whether the variances of the rankscores of the twelve strategies are equal. This is done by using Mauchly's test, which reveals a violation of sphericity in the strategies ($\chi^2(65) = 435.90$, $p = .00$). It may lead to positively biased F-statistics and increases the risk of false positives. To reduce this risk, we apply a Greenhouse-Geisser correction of $\epsilon = .61$ and run the one-way repeated-measure ANOVA. It reveals a significant difference in the rankscores of the strategies ($F(6.60, 805.33) = 21.98$, $p = .00$). To assess the pair-wise significant differences between

[8] http://oldwiki.dbpedia.org/Downloads39#redirects

[9] http://nlp.stanford.edu/software/corenlp.shtml

[10] http://jgibblda.sourceforge.net/

[11] The anonymized experimental data is available from: http://dx.doi.org/10.7802/1224

the twelve strategies, a post-hoc analysis is conducted. We have applied Shaffer's modified sequentially rejective Bonferroni procedure (Shaffer's MSRB procedure) [23] that takes into account the number of different experiment conditions, i. e., the number of recommendation strategies. The result of the post-hoc analysis is presented in Table 4. The vertical and horizontal dimensions of the Table 4 show the eleven-by-eleven comparison of the twelve strategies. As one can see, we observe various significant differences between the strategies ($p < .05$, marked in bold font). For example, while we observe a significant difference between the strategies CF-IDF × Sliding window × Title and HCF-IDF × Sliding window × All ($t(122) = 4.77$, $p = .00$), there is no significant difference between the strategies CF-IDF × Exponential decay × Title and LDA × Sliding window × Title ($t(122) = 2.43$, n.s., $p = .41$).

Table 3: Rankscores of the strategies in decreasing order. M and SD denote mean and standard deviation, respectively.

	Strategy			Rankscore
	Profiling Method	Decay Function	Content	M (SD)
1.	CF-IDF	Sliding window	All	.59 (.33)
2.	HCF-IDF	Sliding window	All	.56 (.34)
3.	HCF-IDF	Sliding window	Title	.55 (.33)
4.	HCF-IDF	Exponential decay	Title	.52 (.30)
5.	CF-IDF	Exponential decay	All	.51 (.32)
6.	HCF-IDF	Exponential decay	All	.49 (.30)
7.	CF-IDF	Exponential decay	Title	.41 (.29)
8.	CF-IDF	Sliding window	Title	.39 (.27)
9.	LDA	Exponential decay	Title	.35 (.31)
10.	LDA	Sliding window	Title	.33 (.31)
11.	LDA	Exponential decay	All	.32 (.30)
12.	LDA	Sliding window	All	.27 (.33)

Difference in experiment factors. Subsequently, we analyze the results with respect to each experimental factor. To this end, we first apply Mendoza's test [17] to check for violations of sphericity against the factors. Mendoza's test is an extension of Mauchly's test to adopt to multi-way repeated-measure ANOVA. It shows significances with the global ($\chi^2(65) = 435.90$, $p = .00$) and the factors *Profiling Method* ($\chi^2(2) = 12.21$, $p = .00$), *Profiling Method* × *Decay Function* ($\chi^2(2) = 20.02$, $p = .00$), and *Profiling Method* × *Document Content* ($\chi^2(2) = 8.61$, $p = .01$). Subsequently, we run a three-way repeated-measure ANOVA with a Greenhouse-Geisser correction of $\epsilon = .60$ for the global and $\epsilon = .91$ for the factors *Profiling Method*, $\epsilon = .87$ for *Profiling Method* × *Decay Function*, and $\epsilon = .93$ for *Profiling Method* × *Document Content*. Table 5 shows the results of the ANOVA with F-ratio, effect size η^2, and p-value. The effect size is small when $\eta^2 > .02$, medium when $\eta^2 > .13$, and large when $\eta^2 > .26$. The analyses reveal significant differences in all three factors and their contributions except the factor *Decay Function*. For all factors with significant differences, we apply again a post-hoc analysis using Shaffer's MSRB procedure with respect to each factor. In terms of the factor *Profiling Method*, the post-hoc analysis reveals significant differences between all pairs of HCF-IDF, CF-IDF, and LDA (details of the post-hoc analysis are omitted for the reasons of brevity and documented in our TR [20]). Although the strategy CF-IDF × Sliding window × All performs best as shown in Table 3, the best *Profiling Method* is HCF-IDF as it performs under all other factors better than CF-IDF and LDA. Regarding the factor *Document Content*, "All" outperforms "Title" ($F(1, 122) = 5.18$, $p = .02$). Regarding the factor *Profiling Method* × *Decay Function*, the result suggests that the strategies with the Exponential decay function perform better than those with the Sliding window function when LDA is employed. In addition, there are significant differences among the three profiling methods when a decay function is fixed. In both decay functions, HCF-IDF performs best, followed by CF-IDF, and LDA. Referring to the factor *Profiling Method* × *Document Content*, the result indicates that All is a better choice than Title, when CF-IDF is employed. In profiling methods HCF-IDF and LDA, the factor *Document Content* makes no significant difference. It indicates that HCF-IDF does perform well when only titles of candidate items are available. In addition there are significant differences among the profiling methods when a choice of *Document Content* is fixed. In those cases, HCF-IDF always outperforms others. In terms of the factor *Decay Function* × *Document Content*, All is a better choice than Title, when Sliding window is used.

Correlation of recommendation performance with the number of tweets, the number of concepts, the number of concepts per tweet, and the percentage of tweets containing at least one concept. We computed Pearson's r and Kendall's τ between the users' mean rankscores and each of the number of tweets, concepts, concepts per tweet and the percentage of tweets containing at least one concept. A correlation may show a dependency that could influence the recommendation performance. The results show no significant correlation: As stated in Section 5.3, a participant has published on average 1096.82 tweets (SD: 1048.46). There is no significant correlation with the rankscores ($r(121) = .04$, n.s., $p = .62$ and $\tau = .00$, n.s., $p = .98$). Referring to the number of concepts, on average $1,214.93$ concepts (SD: 1181.43) are contained in a participant's Twitter stream. The correlation coefficients are non-significant ($r(121) = .05$, n.s., $p = .60$ and $\tau = -.01$, n.s., $p = .94$). Regarding the number of concepts per tweet, a participant's tweet contains on average 1.07 concepts (SD: 0.31) with again no significant correlation to the rankscores ($r(121) = -.05$, n.s., $p = .59$ and $\tau = -.02$, n.s., $p = .71$). Regarding the tweets that contribute in computing the user profiles for the methods with CF-IDF and HCF-IDF, we calculate the percentage of the number of tweets containing at least one concept and the number of tweets for each user. On average, 62.24% of the tweets (SD: 13.55) that a participant has published contain at least one concept, with no significant correlation ($r(121) = -.04$, n.s., $p = .67$ and $\tau = -.03$, n.s., $p = .73$)

6.2 Questionnaire Feedback

At the end of the experiment, the participants were asked to rate: "How easy it was to make the decisions whether a recommended publication is interesting". Using a 5-point Likert scale, where values between 1 and 5 refer to very difficult to very easy, the result is fairly high with an average of 3.68 (SD: 0.88). Regarding question "Whether the participants noticed a difference among the twelve strategies", the result is similarly high with an average of 3.46 (SD: 1.20). In the free text feedback, one participant denoted that the recommender system failed to pick up his primary field de-

Table 4: Post-hoc analysis with pairwise p-values over the twelve strategies using Shaffer's MSRB procedure. The p-values are marked in bold font if $p < .05$, which indicates a significant difference between the two strategies. Strategies are sorted by rankscores as shown in Table 3.

				All	Title	Title	All	All	Title	Title	Title	Title	All	All
				Sliding window	Sliding window	Exponential decay	Exponential decay	Exponential decay	Exponential decay	Sliding window	Exponential decay	Sliding window	Exponential decay	Sliding window
				HCF-IDF	HCF-IDF	HCF-IDF	CF-IDF	HCF-IDF	CF-IDF	CF-IDF	LDA	LDA	LDA	LDA
				2.	3.	4.	5.	6.	7.	8.	9.	10.	11.	12.
1.	CF-IDF	Sliding window	All	.99	.97	.72	.22	.12	**.00**	**.00**	**.00**	**.00**	**.00**	**.00**
2.	HCF-IDF	Sliding window	All		.99	.99	.99	.99	**.00**	**.00**	**.00**	**.00**	**.00**	**.00**
3.	HCF-IDF	Sliding window	Title			.99	.99	.99	**.00**	**.00**	**.00**	**.00**	**.00**	**.00**
4.	HCF-IDF	Exponential decay	Title				.99	.99	**.01**	**.00**	**.00**	**.00**	**.00**	**.00**
5.	CF-IDF	Exponential decay	All					.99	**.04**	**.00**	**.00**	**.00**	**.00**	**.00**
6.	HCF-IDF	Exponential decay	All						.12	**.02**	**.00**	**.00**	**.00**	**.00**
7.	CF-IDF	Exponential decay	Title							.99	.99	.41	.28	**.01**
8.	CF-IDF	Sliding window	Title								.99	.84	.61	**.03**
9.	LDA	Exponential decay	Title									.99	.99	.72
10.	LDA	Sliding window	Title										.99	.99
11.	LDA	Exponential decay	All											.88

Table 5: Three-way repeated-measure ANOVA with Greenhouse-Geisser correction with F-ratio, effect size η^2, and p-value.

Factor	F	η^2	p
Profiling Method	58.40	.48	**.00**
Decay Function	1.17	.01	.28
Document Content	5.18	.04	**.02**
Profiling Method × *Decay Function*	4.63	.04	**.01**
Profiling Method × *Document Content*	17.09	.14	**.00**
Decay Function × *Document Content*	4.69	.04	**.03**
Profiling Method × *Decay Function* × *Document Content*	3.35	.03	**.04**

spite having tweeted about that field. Apart from this, we received many positive comments (e.g., interesting, useful).

7. DISCUSSION

The strategies with HCF-IDF perform almost equally well compared to the best performing strategy CF-IDF × Sliding window × All. There is no significant difference between them as described in Table 4. The strong advantage of HCF-IDF is that it reaches its performance already when using only the titles of the scientific publications. The reason is that spreading activation over the hierarchical knowledge base used in HCF-IDF successfully reveals concepts that are not explicitly mentioned in the texts. CF-IDF works well when full-texts are available. Referring to LDA, the recommendation performance of the strategies with LDA is overall low, even if full-texts are available. A possible reason is that LDA cannot construct accurate user profiles because of the shortness and sparseness of social media items. Without accurate user profiles it is impossible to make good recommendations, even if full-texts are available. In fact, a slight correlation between the rankscores of LDA and the number of tweets is observed [20]. It indicates that participants with more tweets receive better recommendations. Please note as documented in [20], rankscores are almost exact same values with Precision@5 and nDCG. Although rankscores are slightly different with MAP and MRR, the order of performance of strategies are almost identical. Thus, the arguments described in this paper do not be influenced by differences among those evaluation metrics.

Our dataset covers scientific publications in the broader field of economics. Thus, although the dataset is obtained from a portal of economics literature, it contains scientific publications from various fields including, e.g., social sciences, political sciences, and information sciences. In the experiment, 31 of 123 participants do not have a major in economics. We have conducted an ANOVA test to identify whether the recommendation performance is significantly different for participants from economics and those not in economics. The result shows that majors make no significant difference ($F(1, 121) = 0.01$, n.s., $p = .94$). Thus, we assume that our approach may be transferred to other domains. Furthermore, there are a lot of domain-specific hierarchical knowledge bases in other domains freely available such as Medical Subject Headings (MeSH) for medicine and ACM Computing Classification System (ACM CCS) for computer science. An overview of freely available hierarchical knowledge bases is maintained by the W3C as cited in the introduction. The knowledge bases are of similar structure to the STW used in this paper. They are of high quality as they are manually crafted by domain experts. Therefore, HCF-IDF can be easily applied to other fields. Our approach could be integrated with other social media platforms (e.g., Facebook, LinkedIn), where users generate short and sparse texts. In addition, HCF-IDF is robust against the number of tweets a user published, because there is no correlation between the number of tweets and the rankscores of the strategies with HCF-IDF.

Our results may potentially be influenced by the amount of time that each participant spent for evaluating the $5 \times 12 = 60$ recommended publications by the twelve strategies in the experiment. However, they spent on average 517.54 seconds (SD: 376.72) to complete the evaluation of the 60 recommendations. In addition, we randomized the order of the strategies presented to the participants to counterbalance any influence on the order of the strategies. Thus, we think that our results are not influenced by it. Another potential threat to the validity of our results could be the procedure how we recruited the participants. We believe that the risk is low since we collected enough participants regarding each demographic factor (as shown in Section 5.2). Regarding the demographic factors, we found significant differences only for the participants' highest academic degree and participants' gender (details are documented in the TR [20]). However, they do not affect the order of the recommendation performance of the different strategies.

8. CONCLUSIONS

This paper contributes to content-based recommender systems for scientific publications based on user profiles extracted from social media platforms. We have constructed twelve different recommendation strategies along three factors, namely profiling method, decay function, and document content. The online experiment revealed that titles of scientific publications are sufficient to achieve competitive recommendation results when employing the profiling method HCF-IDF. Thus, the spreading activation over the hierarchical knowledge base enables HCF-IDF to extract a sufficient number of concepts from titles to compute competitive recommendations. This is an important result as full-texts are not always available, e. g., due to legal reasons.

Acknowledgement. This research was co-financed by the EU H2020 project MOVING (http://www.moving-project.eu/) under contract no 693092. We like to thank the anonymous participants of our study to support this research.

9. REFERENCES

[1] D. M. Blei and J. D. Lafferty. Dynamic topic models. In *ICML*. ACM, 2006.

[2] D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent dirichlet allocation. *JMLR*, 3, 2003.

[3] S. Bostandjiev, J. O'Donovan, and T. Höllerer. TasteWeights: a visual interactive hybrid recommender system. In *RecSys*. ACM, 2012.

[4] J. S. Breese, D. Heckerman, and C. Kadie. Empirical analysis of predictive algorithms for collaborative filtering. In *UAI*. Morgan Kaufmann, 1998.

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

[6] W. Feng and J. Wang. We can learn your# hashtags: Connecting tweets to explicit topics. In *ICDE*. IEEE, 2014.

[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] T. L. Griffiths and M. Steyvers. Finding scientific topics. *NAS*, 101, 2004.

[9] T. J. Hazen. Direct and latent modeling techniques for computing spoken document similarity. In *the Spoken Language Technology*. IEEE, 2010.

[10] L. Hong and B. D. Davison. Empirical study of topic modeling in Twitter. In *SOMA*. ACM, 2010.

[11] D. Jannach, M. Zanker, A. Felfernig, and G. Friedrich. *Recommender systems: an introduction*. Cambridge University Press, 2010.

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

[13] M. K. Khribi, M. Jemni, and O. Nasraoui. Automatic recommendations for e-learning personalization based on web usage mining techniques and information retrieval. In *ICALT*. IEEE, 2008.

[14] J. Letierce, A. Passant, J. G. Breslin, and S. Decker. Understanding how twitter is used to spread scientific messages. In *WebSci*. Web Science Trust, 2010.

[15] Y. Li, M. Yang, and Z. M. Zhang. Scientific articles recommendation. In *CIKM*. ACM, 2013.

[16] C. Lu, W. Lam, and Y. Zhang. Twitter user modeling and tweets recommendation based on Wikipedia concept graph. In *AAAI Workshops*, 2012.

[17] J. L. Mendoza. A significance test for multisample sphericity. *Psychometrika*, 45(4), 1980.

[18] S. E. Middleton, D. C. De Roure, and N. R. Shadbolt. Capturing knowledge of user preferences: ontologies in recommender systems. In *K-CAP*. ACM, 2001.

[19] C. Nishioka, G. Große-Bölting, and A. Scherp. Influence of time on user profiling and recommending researchers in social media. In *i-KNOW*. ACM, 2015.

[20] C. Nishioka and A. Scherp. Profiling vs. time vs. content: What does matter for top-k publication recommendation based on twitter profiles? - an extended technical report. http://arxiv.org/abs/1603.07016.

[21] F. Orlandi, J. Breslin, and A. Passant. Aggregated, interoperable and multi-domain user profiles for the social web. In *I-SEMANTICS*. ACM, 2012.

[22] S. L. Sangam and S. S. Mogali. Obsolescence of literature in the field of social sciences. *PEARL*, 7(3), 2013.

[23] J. P. Shaffer. Modified sequentially rejective multiple test procedures. *J. of the ASA*, 81(395), 1986.

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

[25] S. J. Soltysiak and I. B. Crabtree. Automatic learning of user profiles - towards the personalisation of agent services. *BT Tech. J.*, 16(3), 1998.

[26] K. Sugiyama and M.-Y. Kan. Scholarly paper recommendation via user's recent research interests. In *JCDL*. ACM, 2010.

[27] K. Sugiyama and M.-Y. Kan. Exploiting potential citation papers in scholarly paper recommendation. In *JCDL*, pages 153–162. ACM, 2013.

[28] C. Wang and D. M. Blei. Collaborative topic modeling for recommending scientific articles. In *KDD*. ACM, 2011.

[29] Z. Zhao, Z. Cheng, L. Hong, and E. H. Chi. Improving user topic interest profiles by behavior factorization. In *WWW*. IW3C2, 2015.