Bachelor thesis
Information Science

Radboud University Nijmegen

Profiling knowledge workers using open online profiles

Author:
J.R. (Joël) Cox
joel.cox@student.ru.nl
Student No. s4023390

First supervisor/assessor:
dr. S. (Suzan) Verberne
s.verberne@cs.ru.nl

Second assessor:
prof. dr. ir. W. (Wessel) Kraaij
w.kraaij@cs.ru.nl

June 28, 2013

Abstract

This thesis explores the creation of user terminology models containing the interests, topics and expertise of a knowledge worker. These models are generated from existing, open profiles found on the web, specifically user profiles from Twitter, LinkedIn and ArnetMiner.

In order to correct for the sparseness of these profiles, information extracted from the network, related to the user, is used to enrich the profiles. The models are generated by using a frequency based scoring function, together with a background corpus.

The generated models are analyzed for overlap between networks and evaluated by their owners. Terms are rated for relevancy, specificity and whether the term belongs to the user's professional or private profile.

Ultimately, the generated models proved to be of high average precision, but the overlap between models was quite low. Terms included in LinkedIn profiles were rated with the highest specificity, terms from Twitter models the lowest.

Academic profiles only contained professional terms, while terms from Twitter models were evenly distributed between private and professional. Including information from a user's network didn't show any improvement in the quality of the model.

Contents

Chapter 1

Introduction

Knowledge workers face enormous amounts of information every day, all with different levels of relevancy to the current task the user is performing. The SWELL project[1] aims to develop tools for helping knowledge workers in their daily processes. An example of such a tool is the filtering of email messages.

In order to provide such a tool to the user, a model about the user's interests, topics and expertise has to be created. The practice of user profiling is widely used across the web, for things such as:

- Personalized search
- Recommendation engines
- Targeted advertising

The construction of such a model either relies on implicit or explicit user information. The latter method expects the user to provide the information for the model themselves, which can be quite time-consuming. Without the corporation of the user, the generation of an explicit user model is impossible [5]. The other method – implicit data collection – collects data without explicit action of the user through agents installed on the user's computer, for instance by collecting computer interactions [8].

This thesis combines both approaches by using existing online profiles (explicit information) and – as explained later – information retrieved from the different networks these online profiles are situated in (implicit information). This drastically reduces the information the user has to supply in order to have a profile generated; only the unique identifier of the user on such a network has to be provided. We formulate the following research question: *How accurate are the user terminology models extracted from existing online profiles?*

[1] `http://www.swell-project.net/`

In order to explore this question, a number of user models will be created using data from three different online networks, specifically:

- Twitter
- LinkedIn
- ArnetMiner (scientific publications)

Because of the different nature of these networks, differences between the different user profiles and thus the models are expected. We therefore formulate the following hypotheses about the resulting user models:

1. Models generated from the ArnetMiner network will result in a highly specialized model, matching the research interests of a user.
2. Models generated from the Twitter network will result in a model dominated by personal interests, rather than professional interests.
3. Models generated from the LinkedIn networks will results in models that include both private as well as professional interests, because resumes often include volunteer experience, related to personal interests.

A major difference between these different networks is the data that can be extracted from these networks. Twitter is – by definition – a network that limits interactions to 140 characters. LinkedIn profiles typically provide information that would be included in a person's resume. Academic papers are relatively long documents, but may not be easily accessible due to licensing issues.

Because of the limitations of these networks, we expect these profiles to be rather sparse. In order to correct for this, we enrich the profiles with data extracted from other nodes within the network. The inspiration for this was the work of Kostoff et al. [9], who used abstracts of citing papers to create a model of the cited paper.

These considerations lead to the following sub questions:

1. How similar are the models created from the different networks?
2. How specific are the models created from the different networks?
3. Does including information from adjacent nodes in the network produce better profiles?

Chapter 2

Preliminaries

Text mining is a broad research field that covers topics such as data mining, linguistics, information retrieval, machine learning and various other research topics. The high level goal of this field is to use “large [online] text collections to discover new facts and trends” [6]. Various techniques have been developed to attain the different goals in this field. This thesis focuses on a technique called frequency term analysis.

This technique is built on the idea that important terms within a certain corpus are more frequently used than terms of lesser importance. The frequency of a term alone is not a good indicator of its importance. For example, in this thesis the word ‘create’ may be frequent, but is is not very descriptive for this thesis. Therefore, a scoring function is used to assign scores to the different terms extracted from the corpus, based on different variables.

Additionally, this thesis builds upon the work of the social network analysis research community, which is closely related to sociology, but also relies on the more formal graph theory. This study is mainly involved with researching the relationships between different nodes within (social) networks [14].

Finally, we borrow some terminology from the computer privacy and identity management research field. Hypothesis (1) states that profile models may vary across different networks because subjects may represent different identities across these networks.

An example of this is that a person might not expose a political preference in a professional setting, while exposing this preference in a personal setting. This phenomenon was described by Clauß and Köhntopp [3] and called ‘partial identities’.

Chapter 3

Research

In order to research the stated research question(s) and hypotheses, profile models were generated for a selected group of knowledge workers. These models were analyzed on their own, as well as with the help of their owner. This chapter will go into the way the profiles were retrieved, how the models were generated and the different kinds of analysis that were performed.

3.1 Data collection

In order to retrieve the information needed for composing the corpora for the different users, APIs provided by Twitter[1] and LinkedIn[2] were used. Collecting information about academic publications proved to be more difficult. Ultimately, ArnetMiner[3], a data mining system for creating an academic social network [17], was used to obtain paper titles.

Not only the profiles of the knowledge workers were retrieved, but also the profiles connected to the user to enrich the user's profile. Table 3.1 gives an exact overview of the data fields that were retrieved from the different networks.

The textual data was then tokenized into unigrams and decoded from unicode to ASCII. Characters that are not supported by ASCII were ignored. A list of English and Dutch stop words were used to filter common words. All input was consistently transcoded to ASCII.

No differentiation between language was made during the data collection, nor during the term scoring process. Manual inspection showed that a majority of the profiles were provided in English, apart for the Twitter data. Ideally, language detection will have corrected for this, or Dutch profiled excluded from the analyzed corpora.

[1] https://dev.twitter.com/docs/api/1.1

[2] https://developer.linkedin.com/documents/profile-api

[3] http://arnetminer.org/RESTful_service

Table 3.1: The different data fields retrieved from each network.

Network	Subject	Data field	Note
Twitter	User	`tweet.text`	The last 500 tweets, excluding replies
Twitter	Followed by user	`user.description`	
LinkedIn	User	`profile.{industry, headline, summary, specialties,interests, skills, educations, three-current-positions, three-past-positions}`	
LinkedIn	Connections of user	`profile.{industry, headline, summary, specialties, positions}`	Limited by `r_basicprofile` API permission
Academic publications	User	`publication.title`	All papers associated by ArnetMiner
Academic publications	Co-authors of user	`publication.title`	

In total 10 LinkedIn, 8 Twitter and 6 academic profiles were analyzed, provided by 13 separate users. An anonymized list of types of profiles supplied by users can be found in table 3.2. An aggregated overview can be found in table 3.3.

The majority of the users were sourced from TNO, an independent research organization. The profiles can thus be qualified as profiles belonging to knowledge workers, the target demographic of the SWELL project.

Only profiles specified on these networks were collected, rather than personal websites. This was due to the relative ease of retrieval and the explicit irelations specified on these networks. Potentially these relations could have been specified on personal websites using XFN[4].

[4] `http://gmpg.org/xfn/`

Table 3.2: The networks supplied by individual users.

User	LinkedIn	Twitter	Academic
1	Provided	Provided	Provided
2	Provided	Provided	-
3	Provided	-	Provided
4	-	Provided	-
5	Provided	Provided	Provided
6	Provided	-	-
7	-	Provided	-
8	Provided	-	-
9	Provided	Provided	Provided
10	Provided	Provided	-
11	Provided	-	-
12	-	-	Provided
13	Provided	Provided	Provided

Table 3.3: Aggregated overview of networks supplied by users.

Networks	Frequency
LinkedIn	10
Twitter	8
Academic	6
LinkedIn $\wedge$ Twitter	6
Twitter $\wedge$ Academic	4
Academic $\wedge$ LinkedIn	5
LinkedIn $\wedge$ Twitter $\wedge$ Academic	4

3.2 Term scoring

In order to find the most important terms in a profile, we want to assign a score to each term. Only unigrams will be considered.

During the explorational phase of this thesis an algorithm proposed by Hiemstra et al. was used [7]. This was later changed to a point-wise Kullback-Leibler divergence-based algorithm due to early results of a term scoring comparison study [19].

Point-wise Kullback-Leibler divergence is used as a measure that describes the change between two language models; how much has a language model diverged from the model it is compared with?

We used parts of an algorithm proposed by Tomokiyo and Hurst [18]. Their algorithm consists out of two parts, namely 'informativeness' ("how well a phrase captures or illustrates the key ideas in a set of documents") and 'phraseness', a sequence of terms belonging to a meaningful phrase [19]. Because this thesis only analyzes unigrams, only the 'informativeness' aspect of the algorithm was used.

$$P(p||q) = \sum_{x} p(t) \log \frac{p(t)}{q(t)}$$

Both the functions p and q return the probability of a term t given a corpus. p represents the function for the foreground corpus, while q represents the function for the background corpus.

The foreground corpus is the profile text from the user, while the background corpus contains a corpus of a large contemporary English texts.

While this scoring method showed promising results, there were cases where the generated model was very prone to hapax legomena (terms only occurring one time within context for a single corpus). This was due to the sparseness of the profiles supplied by the user.

In order to increase the size of the corpus an approach comparable to Kostoff et al. [9] was taken. Rather than just taking the user's corpus (C_u) into account, information from the network was added to the corpus (C_n). For example, LinkedIn profiles were enriched by the (partial[5]) profiles from direct connections. Table 3.1 gives an exact overview of the data fields that were used.

However, the initial results of this technique showed a lot of noise and non-relevant terms. Because the extra terms were added to the original corpus, the weight of the supporting corpus was of equal importance of the terms from the user's profile. In practice this means that terms that are shared between nodes within the network will show up in the user's model, even tough the term may not have occurred in the profile of the user itself.

To eliminate this noise the supporting corpus (C_n) is treated as a separate corpus. Rather than adding all terms to the user corpus, only the terms that are present in the user corpus are counted in the frequency analysis.

$$r(t) = \frac{(tf(t, C_u) + tf(t, C_n) * found(t, C_u))}{|C_u|}$$

[5] The LinkedIn API restricts the retrieval of certain information from first degree connections.

t is the term for which the score is computed. The $found$ function only evaluates to 1 when the term is found in the respective corpus, therefore canceling out the additional term frequency if it's not included in the user corpus. The tf function returns the frequency of a term within a corpus.

An example of a model can be found in table 3.4. The model consists out of 10 terms, ordered by the scoring function specified above, combined with the algorithm of Tomokiyo and Hurst to compute the score of a single term:

$$r(t) \log \frac{r(t)}{q(t)}$$

Table 3.4: Top 10 terms from LinkedIn models generated for user 1. The scores of the terms in the user corpus clearly shows the sparseness of the corpus.

Term	Score (C_u)	Term	Score (C_n)
extraction	0.715	phd	0.614
nlp	0.715	retrieval	0.560
humanities	0.715	linguistics	0.454
retrieval	0.715	computational	0.398
linguistics	0.715	postdoctoral	0.216
phd	0.715	lecturer	0.216
postdoctoral	0.715	applications	0.207
visiting	0.715	extraction	0.202
classification	0.654	wolverhampton	0.175
why-questions	0.654	nlp	0.149

3.3 Model overlap

To measure the similarity of the models as stated in research question (1), the top-n terms from each model were cross-referenced with the other models generated for the user. The retrieved terms are placed in a set and then compared to another set of terms from another profile. The formula below indicates the amount of overlap and is a slight rewrite of the Jaccard similarity coefficient [15]; the sets we compare are always of the same length.

$$2 * \frac{|M_1 \cap M_2|}{|M_1| + |M_2|}$$

Because the order of the terms is not taken into account in the overlap, two models can look distinctively different when viewed as a ranked list. The tables below show the average overlap of all analyzed profiles.

Table 3.5: Average overlap in analyzed profiles for top 20 terms.

$n = 20$	Twitter	Twitter supported	LinkedIn	LinkedIn supported	Academic	Academic supported
Twitter	-	0.756	0	0.017	0.053	0.053
Twitter supported	0.756	-	0.025	0.042	0.066	0.066
LinkedIn	0	0.025	-	0.640	0.100	0.100
LinkedIn supported	0.017	0.042	0.640	-	0.080	0.090
Academic	0.053	0.066	0.100	0.080	-	0.800
Academic supported	0.053	0.066	0.100	0.090	0.800	-

The data in table 3.5 shows a high degree of overlap between a model and the supported model from the same network, ranging from 64% to 80% overlap. Overlap between the different networks is significantly lower, with LinkedIn and academic networks overlapping between 8% and 10%. The Twitter models overlap the least with other models.

Expanding the analysis to include the top 40 terms rather than 20 shows little change; generally inter-network overlap decreases, while inner-network has increased for both LinkedIn and the academic network.

Individual user models were manually inspected for these characteristics and confirmed the above findings. Figure 3.1 shows the overlap between terms for all three networks for a single user. An example model of the same user was displayed in table 3.4. There was only one term which was found in the top 40 of all three networks. Other than that, there are no terms that are included in both the top 40 Twitter and academic models for this user.

3.4 User evaluation

In order to evaluate the quality of the models, personalized surveys were created by taking the 20 highest scoring terms for each models. The user models as well as the network supported models were evaluated.

Table 3.6: Average overlap in analyzed profiles for top 40 terms.

$n = 40$	Twitter	Twitter supported	LinkedIn	LinkedIn supported	Academic	Academic supported
Twitter	-	0.759	0.013	0.025	0.028	0.035
Twitter supported	0.759	-	0.038	0.050	0.035	0.035
LinkedIn	0.013	0.038	-	0.855	0.068	0.073
LinkedIn supported	0.025	0.050	0.855	-	0.068	0.073
Academic	0.028	0.035	0.068	0.068	-	0.925
Academic supported	0.035	0.035	0.073	0.073	0.925	-

Terms that were included in multiple models only occurred once in the survey; users were asked to evaluate 120 terms at most (three networks, two models per network) if all three networks were supplied. In practices this number came down to ~70 terms, due to the term overlap. Users were not told which term was extracted from which network. All terms were then ordered alphabetically.

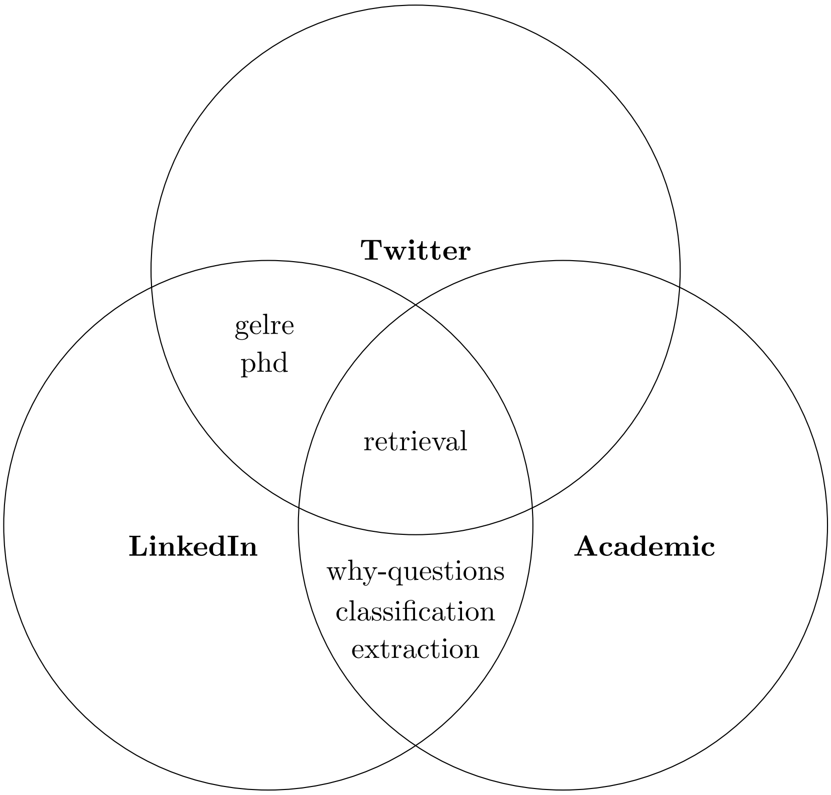
For each term the user was asked whether they judged the term to be relevant to their online profile; does the term tell anything about their online profile? If this was the case, the user was asked to rate the terms on specificity using a scale ranging from 1 to 5 (1 being a very general term, 5 being a very specific term). For example, 'researcher' is a more general term than 'biologist'. Finally, the user marked the term as being part of the user's professional or private profile.

For the evaluation of the ranked term lists for each profile the Average Precision measure [12] was used, a common measure in Information Retrieval. All terms were extracted from the survey, grouped by model and ordered by their original score.

$$\frac{\sum_{k=1}^{n}(P(k) * relevant(k))}{n_c}$$

The *relevant* function evaluates to 1 only if the term was deemed relevant by the user. The P functions returns the precision of the term at that position; the number of relevant terms before the current term, divided by the rank of the current term. n_c represents the number of relevant terms evaluated.

Figure 3.1: Terms found in all three networks, generated from the network supported corpus of user 1.



3.5 Results

The difference in Average Precision between the model extracted from the user corpus (C_u) and the model extracted from the network supported corpus (C_n) is quite low. Looking closer, user corpus models perform better in the case of LinkedIn and academic, contrary to Twitter, which might be explained by the higher level of noise found in tweets, such as URLs and usernames. Another explanation might be that user that the C_u corpus contains tweets, while the C_n corpus also contains user bios, which contain higher quality terms

Initial analysis of table 3.8 might let the reader believe that Twitter profiles are of high specificity, however the amount of non-relevant terms are not included in these numbers. The corrected numbers give a more accurate representation of the specificity; terms that were judged as non-relevant were assign a specificity score of 0.

There weren't any surprises in the data in table 3.9 about whether terms

Table 3.7: Average precision of different models as rated by the user.

	Twitter	LinkedIn	Academic
C_u	0.555	**0.802**	**0.801**
C_n	**0.583**	0.770	0.777

belonged to a user's professional or private profile; most of the data aligned with hypotheses (1, 2 and 3). Twitter terms contain the fewest professional terms. LinkedIn models proved to be predominantly professional. The academic profile only includes professional terms, naturally.

One user made an interesting remark after filling out the survey:

> "I noticed that a lot of terms weren't only relevant to my professional profile, but also to my personal profile. I wasn't able to indicate this in the survey."

This remark makes it clear that the separation between profiles is not always binary, but may be represented by a scale. Thus, partial identities overlap.

Table 3.8: Average specificity (1-5) of different models as rated by the user. Corrected numbers also take non-relevant terms into account.

	Twitter	LinkedIn	Academic
C_u	**3.197**	3.076	**2.838**
C_n	3.103	**3.135**	2.776
C_u corrected	1.219	2.03	**1.856**
C_n corrected	**1.319**	**2.085**	1.788

Table 3.9: Propotion of terms belonging to the professional profile as rated by the user.

	Twitter	LinkedIn	Academic
C_u	0.525	0.856	1
C_n	0.515	0.857	1

Chapter 4

Related Work

4.1 Online network user profiling

As stated in the introduction, user profiling on online networks is common practice. User models are created by the companies that run decentralized networks, as well as third-parties that either use data provided by the user, or data exposed by the network.

A notable work is that of Lops et al. [11] which introduces a paper recommendation system, based on the 'Specialties', 'Interests', and 'Groups and Associations' data entities provided by LinkedIn profiles. Each term and user is then represented in a vector space, using a normal *td-idf* score. Vectors of adjacent users in the network are then added to the user's vector. The recommendation engine calculates the similarity between the user's and the paper's vector in order to recommend the appropriate papers.

An almost similar approach was taken by Abel et al. [1], but rather than just using the user supplied texts (in this case a 140-characters tweet), URLs that are included in the user's tweets are retrieved and in turn analyzed [2]. Additionally, the user model is further enriched by using entity recognition. The user model is again represented in a vector space, as are the articles which are recommended to the user.

Tang et al. [16] took a different approach to finding interests of researchers; probabilistic topic modeling. This method of topic detection relies on statistical models to analyze terms in large bodies of texts and how they are interconnected.

4.2 Term scoring

Term scoring methods are often used in Information Retrieval to score the relevance of a document given a certain query. [13].

Term frequency-inverse document frequency (*td-idf*) is a measure which takes the frequency of a term in a single document into account, weighted by the number of documents that contain the term.

This thesis takes a different approach where a corpus is compared to a background corpus, after which the difference of term distribution in computed. The background corpus should be of sufficient size in order to accurately reflect the term distribution of an average corpus [4]. The foreground corpus is the corpus that is actually analyzed.

4.3 Collaborative filtering and triangulation

A common way to enrich datasets is done using a technique called collaborative filtering. This technique leverages the activity of other users in order to drive recommendations for a given user [10].

When datasets are extracted from networks, this becomes easier because relationships between nodes in a network are often made explicit, rather than inferred from user interaction. Examples of these explicit relationships are friend (mutual relationship), follow (single-direction relationship) and group (many-to-many relationships) relationships.

These ad hoc groups are often formed across common interests and thus make it possible to use the data generated by these peers to enrich the profile of the user.

An example of this from of triangulation was given by Kostoff et al. [9], in which paper abstracts that cited the original paper where split in unigrams, bigrams and trigrams after which frequency analysis was performed, amongst other techniques. This way of trans-citation analysis proved to be very successful way to detect the general theme of an article.

Chapter 5

Conclusions and future work

In this thesis we explored the generation of user terminology models using open profiles and a frequency based scoring function. These models were evaluated by their owners as well as analyzed by the author.

Overall, all the models were of reasonable quality, scoring between 0.55 and 0.802 in the average precision measurement, on average. A clear distinction between the different networks was found, Twitter models scored significantly lower.

The overlap between the different models generated for the networks proved to be minimal. This however doesn't mean that all the users represent a different identity on the different networks per se, but can possibly be attributed to the the type of media and the scoring function.

Models generated from Twitter profiles showed to be the least helpful. These models were of less quality in terms of average precision, specificity and contained a lot of terms that were linked to personal profiles, but not the majority as predicted in hypothesis (2). Hardly any overlap between Twitter models and other models was found.

Both LinkedIn and ArnetMiner were of high quality and high specificity, confirming hypothesis (1). LinkedIn showed very few personal terms (less than 14%), but did include some like stated in hypothesis (3).

Evaluation by the users didn't show any difference in quality between the models generated from the user corpus and the network supported corpus. While the quality of the models remained the same, the amount of data used for generation of the network supported models was multiple times larger. This did help with the granularity of the term scores.

While the explored method of enriching corpora with network didn't show any significant change in the quality of the model according to the user evaluation, this algorithm did manage to cope with a larger corpus while maintaining the quality of the model. Variations of this approach may be worth exploring.

Every entity in the corpora was split into unigram tokens, rather than taking into bigram and trigram phrases. Including these longer phrases might yield better results, because phrases are better understood out of context [19].

Bibliography

[1] Fabian Abel, Qi Gao, Geert-Jan Houben, and Ke Tao. Analyzing user modeling on twitter for personalized news recommendations. *User Modeling, Adaption and Personalization*, pages 1–12, 2011.

[2] Fabian Abel, Qi Gao, Geert-Jan Houben, and Ke Tao. Semantic enrichment of twitter posts for user profile construction on the social web. In *Proceedings of the 8th extended semantic web conference on The semanic web: research and applications - Volume Part II*, ESWC'11, pages 375–389, Berlin, Heidelberg, 2011. Springer-Verlag.

[3] Sebastian Clauß and Marit Köhntopp. Identity management and its support of multilateral security. *Computer Networks*, 37(2):205 – 219, 2001. Electronic Business Systems.

[4] Fred J Damerau. Generating and evaluating domain-oriented multi-word terms from texts. *Information Processing & Management*, 29(4):433–447, 1993.

[5] Susan Gauch, Mirco Speretta, Aravind Chandramouli, and Alessandro Micarelli. The adaptive web. chapter User profiles for personalized information access, pages 54–89. Springer-Verlag, Berlin, Heidelberg, 2007.

[6] Marti A. Hearst. Untangling text data mining. In *Proceedings of the 37th annual meeting of the Association for Computational Linguistics on Computational Linguistics*, ACL '99, pages 3–10, Stroudsburg, PA, USA, 1999. Association for Computational Linguistics.

[7] Djoerd Hiemstra, Stephen Robertson, and Hugo Zaragoza. Parsimonious language models for information retrieval. In *Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 178–185. ACM, 2004.

[8] Saskia Koldijk, Mark van Staalduinen, Stephan Raaijmakers, I van Rooij, and W Kraaij. Activity-logging for self-coaching of knowledge workers. In *2nd workshop on information access for personal media archives*, 2011.

[9] Ronald N. Kostoff, J. Antonio del Río, James A. Humenik, Esther Ofilia García, and Ana María Ramírez. Citation mining: Integrating text mining and bibliometrics for research user profiling. *Journal of the American Society for Information Science and Technology*, 52(13):1148–1156, 2001.

[10] Danielle H. Lee and Peter Brusilovsky. Using self-defined group activities for improvingrecommendations in collaborative tagging systems. In *Proceedings of the fourth ACM conference on Recommender systems*, RecSys '10, pages 221–224, New York, NY, USA, 2010. ACM.

[11] Pasquale Lops, Marco de Gemmis, Giovanni Semeraro, Fedelucio Narducci, and Cataldo Musto. Leveraging the linkedin social network data for extracting content-based user profiles. In *Proceedings of the fifth ACM conference on Recommender systems*, pages 293–296. ACM, 2011.

[12] Zhu M. Recall, precision and average precision. *Department of Statistics and Actuarial Science, University of Waterloo, working paper*, 2004.

[13] Gerard Salton and Christopher Buckley. Term-weighting approaches in automatic text retrieval. *Inf. Process. Manage.*, 24(5):513–523, August 1988.

[14] John Scott. Social network analysis. *Sociology*, 22(1):109–127, 1988.

[15] Pang-Ning Tan, Michael Steinbach, and Vipin Kumar. *Introduction to data mining.* 2005.

[16] Jie Tang, Limin Yao, Duo Zhang, and Jing Zhang. A combination approach to web user profiling. *ACM Trans. Knowl. Discov. Data*, 5(1):2:1–2:44, December 2010.

[17] Jie Tang, Jing Zhang, Limin Yao, Juanzi Li, Li Zhang, and Zhong Su. Arnetminer: extraction and mining of academic social networks. In *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*, KDD '08, pages 990–998, New York, NY, USA, 2008. ACM.

[18] Takashi Tomokiyo and Matthew Hurst. A language model approach to keyphrase extraction. In *Proceedings of the ACL 2003 workshop on Multiword expressions: analysis, acquisition and treatment - Volume 18*, MWE '03, pages 33–40, Stroudsburg, PA, USA, 2003. Association for Computational Linguistics.

[19] Suzan Verberne, Maya Sappelli, and Wessel Kraaij. Term extraction for user profiling: evaluation by the user. In *Proceedings of the 21th International Conference on User Modeling, Adaptation and Personalization (UMAP 2013).*