Profiling knowledge workers using open online profiles

Joël Cox Suzan Verberne

Institute for Computing and Information Sciences, Radboud University Nijmegen

Abstract

In this paper we investigate the accuracy of user terminology models extracted from open online profiles. In our project, user models are used for the profiling of knowledge workers in order to assist them with information tasks such as email filtering and professional search. We created terminology models from profiles on LinkedIn, Twitter and ArnetMiner (scientific publications) using a term scoring function based on Kullback-Leibler Divergence. The resulting term profiles were evaluated by their owners. Overall, all the models were of reasonable quality, scoring between 0.55 and 0.80 Average Precision. We analyzed the overlap between the models, the subjects' rating for specificity of the models and the distinction between personal and professional interests. We experimented with the potential of the network context by adding information from connected users to the model. However, this did not improve the quality of the model. In future work, we plan to compare the user models created from online profiles with user models created from local documents in their ability to improve personalized information filtering.

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 applications that assist knowledge workers in their daily processes. Examples of applications are the automatic filtering of email messages and the personalization of search results. In order to provide such tools to the user[2], a model of the user's interests, topics and expertise has to be created.

The construction of a user model either relies on implicit or explicit user information.This paper presents the result of an exploratory study in which explicit and implicit online information is combined: We use existing online profiles (explicit information) and 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. From online profiles, we extract user models in the form of lists of keywords (terms) that represent the user's online content. We formulate the following research question: *What is the quality of the user terminology models (in the form of lists of keywords) extracted from open online profiles?*

In order to answer this question, we created user models using data from three different online networks: Twitter, LinkedIn and ArnetMiner (scientific publications) and we asked the owners of the profiles to judge the relevance and the specificity of the terms. We use these judgments as ground truth for the evaluation of our method for constructing the models.

Because of the different nature of these networks, differences between the different user profiles for each network are expected, for two reasons: First, users may represent different identities across these networks.

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

[2] We will use the term user and knowledge worker interchangeably in this paper

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 [4] as 'partial identities'. Second, there is a big difference in the information that can be extracted from the three networks. Twitter limits interactions to 140 characters per utterance. LinkedIn profiles typically provide information that would be included in a CV, therefore including names of institutions and schools. Scientific publications are relatively long documents, but may not be easily accessible due to licensing issues. In addition, we expect the 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. [6], who used abstracts of citing papers to create a model of the cited paper.

We formulate four sub questions that we will answer in this paper:

1. How much overlap is there between the models created from the different networks?
2. Does including information from adjacent nodes in the network produce better profiles?
3. How specific are the models created from the different networks?
4. To what extent can we distinguish professional from personal identities by modelling the user profiles of one user?

2 Related Work

Information filtering is based on concepts, methods and techniques from different research areas [5]. In this section, we will first describe earlier attempts to create user models from online profiles (Section 2.1) and next we describe two papers about enriching user profiles through collaborative filtering (Section 2.2).

2.1 User modelling with online profiles

User modelling is a field of AI that is concerned with gathering information about a user and then using that information to adapt a system to the user [10]. The goal of user modelling in our project is to filter information (e-mails, search results) based on its relevance for the user. In this paper, we focus on the user's terminology: we extract user models in the form of lists of terms that represent the user's online content.

Lops et al. [8] introduce 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. 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 vector and the paper's vector in order to recommend the appropriate papers.

An almost similar approach was taken by Abel et al. [2], but instead of LinkedIn, Twitter data was used, including the content of the URLs from the user's tweets [3]. Additionally, the user model was 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. [12] use probabilistic topic modeling for finding interests of researchers. This method relies on statistical models to analyze terms in large bodies of texts and how they are interconnected.

A survey by Abdel-Hafez and Xu [1] gives a clear overview of recent approaches to user modelling on social networks.

2.2 Collaborative filtering and triangulation

One possible approach to enriching sparse datasets is collaborative filtering. This technique uses the activity of other users in order to generate user-specific recommendations [7]. User relations are often formed by common interests and thus make it possible to use the data generated by these peers to enrich the profile of the user. Kostoff et al. [6] extracted terms from citing papers to describe the topics of the cited paper. This way of trans-citation analysis proved to be very successful way to detect the general theme of an article.

Table 1: The 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
ArnetMiner	User	`publication.title`	All papers harvested by ArnetMiner
ArnetMiner	Co-authors of user	`publication.title`	

3 Methodology

In order to answer our research question(s), profile models were generated for a selected group of knowledge workers. The majority of these subjects 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.

3.1 Data collection

In order to retrieve the information needed for composing the corpora for the different subjects, APIs provided by Twitter[3] and LinkedIn[4] were used. Collecting information about academic publications proved to be more difficult since there is no API available for Google Scholar due to licensing restrictions. Ultimately, ArnetMiner[5], a data mining system for creating an academic social network [13], 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 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. No differentiation between languages was made during data collection, nor during the term scoring process. Manual inspection showed that a majority of the profiles were provided in English, except for some of the Twitter profiles.

In total 10 LinkedIn, 8 Twitter and 6 ArnetMiner profiles were analyzed, provided by 13 separate subjects. An overview of the collected data can be found in table 2.

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

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

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

Table 2: Aggregated overview of networks supplied by the subjects.

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

3.2 Term scoring

The goal of term scoring in user profiling is to find the terms that are the most descriptive for a user's corpus. In this work, we restrict ourselves to unigrams.[6] As term scoring algorithm, we implemented pointwise Kullback-Leibler divergence as proposed by Tomokiyo and Hurst [14]. Their algorithm consists of two parts: 'informativeness' (how informative is the term for the corpus) and 'phraseness', (how tight are the words in a sequence of multiple words). Both phraseness and informativeness are estimated using a language modelling approach. Because we only analyze unigrams, we only apply the 'informativeness' aspect of the algorithm, measuring the difference between the language model of the user and the language model of a background corpus. We chose the Corpus of Contemporary American English as the background corpus, which is free to use and is easy to process because the developers provide a word frequency list. Applying the informativeness language model, we weigh the probability of a term t in the user corpus ($r(t)$) with the probability of the term in the background corpus ($q(t)$):

$$p(t) = r(t) \log \frac{r(t)}{q(t)} \tag{1}$$

For the estimation of $r(t)$ both the user corpus C_u and the supporting network corpus C_n are taking into account, but the terms that do not occur in C_u are disregarded:

$$r(t) = \frac{(\text{count}(t, C_u) + \text{count}(t, C_n) * \text{found}(t, C_u))}{|C_u| + |C_n|} \tag{2}$$

The $count$ function returns the count of the term within a corpus. 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 present in the user corpus. The resulting scores are normalized so that the highest score becomes 1. An example of a model can be found in Table 3. The top-10 terms are shown, ordered by the outcome of Equation 1.

3.3 User evaluation

In order to evaluate the quality of the models, personalized surveys were created by taking the 20 highest scoring terms for each model. The user models as well as the network supported models were evaluated. Terms that were included in multiple models only occurred once in the survey; subjects were asked to evaluate 120 terms at most (three networks, two models per network) if all three networks were supplied. In practice this number came down to around 70 terms on average, due to the term overlap. Subjects 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. If this was the case, the user was asked to rate the terms on specificity using a scale ranging from 1 to 5

[6] We will later extend to bi- and trigrams. In [15] we found that longer terms (with two or three words) are more often considered relevant by the profile owner than unigrams, so we expect that better models can be created if we extend the terms to multi-words.

Table 3: 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 ($C_u + C_n$)
extraction	1.0	phd	1.0
nlp	1.0	retrieval	0.912
humanities	1.0	linguistics	0.74
retrieval	1.0	computational	0.647
linguistics	1.0	postdoctoral	0.352
phd	1.0	lecturer	0.352
postdoctoral	1.0	applications	0.337
visiting	1.0	extraction	0.329
classification	0.892	wolverhampton	0.286
why-questions	0.892	nlp	0.243

(1 being a very general term, 5 being a very specific term). For example, 'researcher' is a more general term than 'biologist'. In addition, the user marked each term as being relevant for the user's professional or personal profile. With the user assessments, the ranked term lists for each profile were evaluated using Average Precision [9]:

$$\frac{\sum_{k=1}^{n}(P(k) * relevant(k))}{n_c} \tag{3}$$

The $relevant$ function evaluates to 1 only if the term was deemed relevant by the user. The P function returns the precision of the ranked list at position k. n_c represents the total number of relevant terms in the list.

4 Results

4.1 Overlap between the models created from the different networks?

Table 4: Average overlap in analyzed profiles for top 20 terms. In C_u, only the term counts in the user profile itself are taking into account. In $C_u + C_n$, the term counts in profiles from connected users have been added.

$n = 20$	Twitter C_u	Twitter $C_u + C_n$	LinkedIn C_u	LinkedIn $C_u + C_n$	ArnetMiner C_u	ArnetMiner $C_u + C_n$
Twitter C_u	-	0.756	0	0.017	0.053	0.053
Twitter $C_u + C_n$	0.756	-	0.025	0.042	0.066	0.066
LinkedIn C_u	0	0.025	-	0.640	0.100	0.100
LinkedIn $C_u + C_n$	0.017	0.042	0.640	-	0.080	0.090
ArnetMinder C_u	0.053	0.066	0.100	0.080	-	0.800
ArnetMinder $C_u + C_n$	0.053	0.066	0.100	0.090	0.800	-

To measure the overlap between 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 coefficient [11]; the sets we compare are always of the same length.

Table 5: Variance overlap in analyzed profiles for top 20 terms.

$n = 20$	Twitter C_u	Twitter $C_u + C_n$	LinkedIn C_u	LinkedIn $C_u + C_n$	ArnetMiner C_u	ArnetMiner $C_u + C_n$
Twitter C_u	-	0.222	0	0.003	0.015	0.015
Twitter $C_u + C_n$	0.222	-	0.009	0.007	0.012	0.012
LinkedIn C_u	0	0.009	-	0.484	0.02	0.025
LinkedIn $C_u + C_n$	0.003	0.007	0.484	-	0.013	0.017
ArnetMinder C_u	0.015	0.012	0.02	0.013	-	0.265
ArnetMinder $C_u + C_n$	0.015	0.012	0.025	0.017	0.265	-

$$2 * \frac{|M_1 \cap M_2|}{|M_1| + |M_2|} \tag{4}$$

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. Table 4 shows the average overlap of all analyzed profiles. The data in Table 4 show a high degree of overlap between the model C_u and the network-supported model $C_u + C_n$ from the same network, ranging from 64% to 80% overlap. Overlap between the different networks is lower, with LinkedIn and ArnetMinder networks overlapping between 8% and 10%. The Twitter models overlap the least with other models.

4.2 Including information from adjacent nodes in the network

Table 6: Average Precision of different models as rated by the user.

	Twitter		LinkedIn		ArnetMiner	
	Average	Variance	Average	Variance	Average	Variance
C_u	0.555	0.604	**0.802**	0.077	**0.801**	0.093
$C_u + C_n$	**0.583**	0.634	0.770	0.060	0.777	0.056

The results of the user profile evaluation are in Table 6. The difference in Average Precision between the model extracted from the user corpus (C_u) and the model extracted from the network supported corpus ($C_u + C_n$) is small; a paired t-test ($n = 24$) shows that this difference is not significant ($P = 0.56$). User corpus models perform better in the case of LinkedIn and ArnetMiner, compared to Twitter; the difference between the averages for ArnetMiner and Twitter is significant on the 0.05-level ($P = 0.047$ according to a t-test for independent samples).

4.3 How specific are the models created from the different networks?

Table 7: Average specificity (1–5) of different models as rated by the user. Terms that were judged as non-relevant were assigned a specificity score of 0.

	Twitter	LinkedIn	ArnetMiner
C_u	1.43	2.26	2.07
$C_u + C_n$	1.65	2.34	1.84

The results for the model specificity are in Table 7. Terms that were judged as non-relevant were assigned a specificity score of 0. The results show that LinkedIn models were judged as the most specific and Twitter models as the least specific. The differences between specificity scores for Twitter on the one hand and LinkedIn or ArnetMiner on the other hand are both significant with $P < 0.0001$; the difference between LinkedIn and ArnetMiner is significant on the 0.05-level with $P = 0.031$ according to a t-test for independent samples.

4.4 Distinguishing distinguish professional from personal identities

Table 8: Proportion of terms belonging to the professional profile as rated by the user.

	Twitter	LinkedIn	ArnetMiner
C_u	52.5%	85.6%	100%
$C_u + C_n$	51.5%	85.7%	100%

The results of the distinction between professional and personal interests are in Table 8. Twitter terms contain the fewest professional terms. LinkedIn models proved to be predominantly professional. The ArnetMiner profile only includes professional terms. 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. In other words, professional and personal identities seem to overlap.

5 Conclusions and future work

In this preliminary research we explored the generation of user terminology models using open profiles and a frequency based scoring function for a small group of knowledge workers. These models were evaluated by their owners. Overall, all the models were of reasonable quality, scoring between 0.55 and 0.80 Average Precision. The overlap between the different models generated for the networks proved to be minimal. This however does not necessarily mean that user models represent different identities of the user on different networks, but can possibly be attributed to the the type of media.

Models generated from Twitter profiles were judged to be the least in quality and in specificity. Twitter profiles contain many terms that were relevant to the user's personal interest, but not as many as we would have expected. Hardly any overlap between Twitter models and the other models was found. Both LinkedIn and ArnetMiner were of high quality and high specificity, which is consistent when taking the kind of network into account. Evaluation by the subjects did not show a large difference in quality between the models generated from the user corpus and the network supported corpus. While the quality of the models remained similar, 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.

In future work, we plan to take into account multi-word phrases (bi- and trigrams) in addition to unigrams. In previous work we already showed that for other term profiling tasks, multi-word terms were generally assessed as more informative than unigrams [15]. Finally, we plan to compare the user models created from online profiles with user models created from local documents in their ability to improve personalized information filtering, in particular professional search.

Acknowledgements

This publication was supported by the Dutch national program COMMIT (project P7 SWELL).

References

[1] Ahmad Abdel-Hafez and Yue Xu. A survey of user modelling in social media websites. *Computer and Information Science*, 6(4):p59, 2013.

[2] 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.

[3] 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.

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

[5] Uri Hanani, Bracha Shapira, and Peretz Shoval. Information filtering: Overview of issues, research and systems. *User Modeling and User-Adapted Interaction*, 11(3):203–259, 2001.

[6] 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.

[7] 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.

[8] 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.

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

[10] Michael F McTear. User modelling for adaptive computer systems: a survey of recent developments. *Artificial intelligence review*, 7(3-4):157–184, 1993.

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

[12] 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.

[13] 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.

[14] 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.

[15] 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)*, 2013.