Generic Process for Extracting User Profiles from Social Media using Hierarchical Knowledge Bases

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

Chifumi Nishioka, Ansgar Scherp
Leibniz Information Centre for Economics and Kiel University, Germany
Email: chni@informatik.uni-kiel.de, asc@informatik.uni-kiel.de

***Abstract*—We present the design and application of a generic approach for semantic extraction of professional interests from social media using a hierarchical knowledge-base and spreading activation theory. By this, we can assess to which extend a user's social media life reflects his or her professional life. Detecting named entities related to professional interests is conducted by a taxonomy of terms in a particular domain. It can be assumed that one can freely obtain such a taxonomy for many professional fields including computer science, social sciences, economics, agriculture, medicine, and so on. In our experiments, we consider the domain of computer science and extract professional interests from a user's Twitter stream. We compare different spreading activation functions and metrics to assess the performance of the obtained results against evaluation data obtained from the professional publications of the Twitter users. Besides selected existing activation functions from the literature, we also introduce a new spreading activation function that normalizes the activation w.r.t. to the outdegree of the concepts.**

I. Introduction

Social media platforms such as Twitter are used to connect people in professional context and to share professional thoughts [1]. Thus, the question arises if it is possible to learn a user's professional interests model from his social media activities. Previous work by Abel et al. [2] compared the performance of extracting professional interests from different platforms (Delicious, LinkedIn, and Twitter). They found out that the data contained a lot of noise and that the performance largely depended on the size of the social media profiles. On the other hand, Kapanipathi et al. [3] developed an approach for the extraction of cross-domain interests from Twitter based on their own hierarchical knowledge-base created with a lot of effort and applying different spreading activation functions, in order to reveal user interests which are not mentioned directly in the Twitter stream.

In this work, we present a generic approach for the extraction of professional interests from social media. To this end, we combine and extend the idea of Abel et al. [2] on extracting professional interests from social media with applying an existing hierarchical knowledge-base with spreading activation functions from the work of Kapanipathi et al. [3]. In order to reduce noise that Abel et al. [2] observed, we make use of an external domain-specific knowledge base as a background knowledge for named entity detection. Thus, different from Kapanipathi et al., we concentrate on a specific domain based on the assumption that a single Twitter profile would not cover various different professional domains. Having such a taxonomy readily at hand can be assumed as they are freely available for many professional fields like computer science, social sciences, economics, agriculture, medicine, and so on.[1] We focus on the field of computer science and make use of the ACM Computer Classification System (CCS)[2]. We employ Twitter as a social media source because many scientists use it to disseminate their professional thoughts [1]. Different from Abel et al. [2], we do not use the co-authorship relations as evaluation data for the experimental study, but assess the performance of our approach by linking the created social media profiles with the user's professional publication lists. In our case in the field of computer science, we make use of the data provided from the DBLP Computer Science Bibliography[3] as evaluation data. We compare different existing spreading activation functions and introduce a new activation function for extracting professional interests using a domain-specific knowledge base. Furthermore, we apply different measures to assess the performance of the obtained results.

II. General Process and Formalization

For extracting and assessing professional interests from social media, we define a generic process as illustrated in Figure 1. We aim at a fully automatic approach that makes use of an existing knowledge base in the domain under investigation. We use the knowledge base for named entity detection from the social media data as well as some given evaluation data. Spreading activation allows to further activate some "hidden" concepts that could not be directly observed in the social media data and evaluation data, respectively. Below, we describe the single steps of our generic process in detail:

(1) Named Entity Detection: The domain-specific hierarchical knowledge base (e. g., the ACM Computer Classification System), can be seen as a graph. Every concept retains its relations to higher order concepts (*generalization*) and lower order concepts (*specialization*) and sometimes has synonyms, alternate dictions, or abbreviations. Named entities are detected from text sources and mapped to a concept in the background knowledge graph. Thus, a named entity is considered a concept modeled in the background knowledge base. As shown in Figure 1, this step is applied to both the social media items (left hand side) and the evaluation data (right hand side). Each concept is given a score by the number of appearances in text sources, based on the assumption that concepts which appear frequently are more important for a user than others.

[1]The W3C provides an extensive list of taxonomies, see http://www.w3.org/2001/sw/wiki/SKOS/Datasets, last access: August 24, 2014

[2]http://www.acm.org/about/class/class/2012, last access: August 24, 2014

[3]http://www.informatik.uni-trier.de/~ley/db/, last access: August 13, 2014

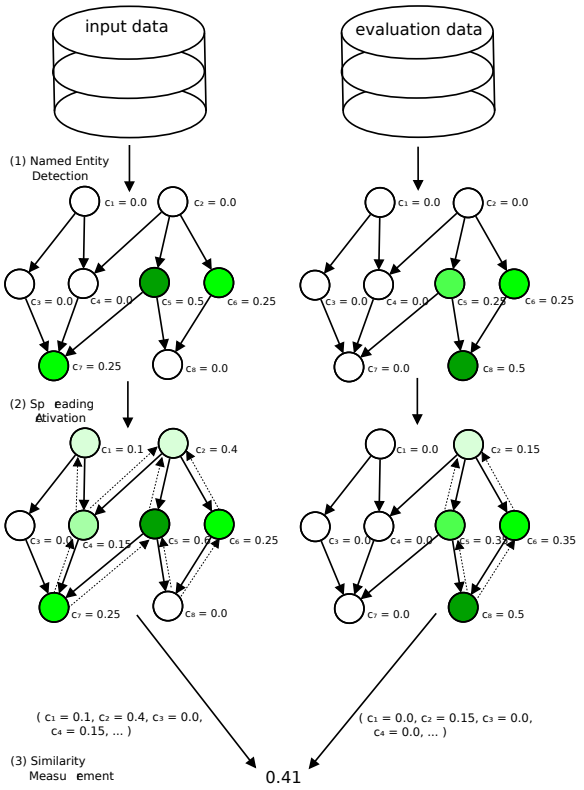

Fig. 1. Using domain-specific background knowledge, the (1) named entity detection extracts concepts defined in the background knowledge from both the social media (left hand side) and the evaluation data (right hand side). Each concept is given a score in the hierarchical knowledge graph, determining its importance in the social media and evaluation data, respectively, Subsequently, (2) spreading activation is applied on the concepts of the hierarchical knowledge graphs extracted from the social media and the evaluation data. In the final step, (3) the graphs are transformed into a vector and a similarity measure is applied to compare the professional interests extracted from the social media with the evaluation data.

(2) Spreading Activation: The concepts extracted from the social media data and the evaluation data are assigned a score based on some spreading activation functions. The functions spread the scores to higher order concepts in a hierarchical graph. Thus, higher order concepts that not appeared in text sources (their score was 0) get scores. With other words, spreading activation reveals concepts, which are not mentioned explicitly but are highly related to the user's interests. The concepts extracted from the social media data and the evaluation data form a graph as shown in Figure 1. Subsequently, the graphs are vectorized in order to apply similarity measures.

(3) Similarity Measurement: In this step, the professional interests vector from the user's social media data is compared with the vector obtained from the evaluation data. To this end, we apply some similarity measures like Jaccard similarity and cosine similarity. It aims to assess to which extend the social user profile reflects the evaluation data, i. e., the users' professional publications.

The background knowledge base is represented as hierarchical graph. Thus, it can be formalized as $HG = (V, E, L, \mu)$, where V denotes a set of nodes (each node represents a concept), E stands for a set of ordered pairs of nodes that represent edges between those nodes, and μ denotes a node labeling function: $\mu : V \rightarrow \wp(L)$, where L is the set of labels, a node can have. Thus, a node can have multiple labels, which reflects the fact that concepts defined in external knowledge bases often retain synonyms, alternate dictions, or abbreviations. For nodes and their labels, the following constraint applies: $\forall v, k \in V : v \neq k \Rightarrow \mu(v) \cap \mu(k) = \emptyset$. It represents that it is impossible for nodes to share labels. Thus, each label is uniquely assigned to one node in the graph.

III. Spreading Activation Functions

We use spreading activation functions to enrich the hierarchical knowledge graph: For nodes whose score is > 0, a spreading activation function is applied. This propagates the score to the higher order concepts. It gives a score to the concepts that were not mentioned explicitly and overall enhances the semantic richness of the graphs. Below, we describe several spreading activation functions used in the work by Kapanipathi et al. [3] as well as introduce a new branch-normalized activation function.

The *Basic Spreading Activation* function is

$$a_i = a_i + a_j \times D, \tag{1}$$

where a_i is a score of the higher-level node (that receives the score), a_j is a lower-level node and D denotes the decay parameter.

The distribution of concepts across the different levels of a hierarchical graph may follow a bell curve as shown by Kapanipathi et al. [3]. Therefore the authors suggest using a normalization function such as *Bell Activation* (see Equation 2) or *Bell Logarithmic Activation* (see Equation 3):

$$a_i = a_i + a_j \times F_i, \tag{2}$$

where F_i is defined as: $F_i = \frac{1}{nodes_{(h_i+1)}}$, and

$$a_i = a_i + a_j \times FL_i, \tag{3}$$

where FL_i is defined as: $FL_i = \frac{1}{log_{10} nodes_{(h_i+1)}}$. In both cases, h_i denotes the depth of node i in the hierarchical graph and $nodes_h$ denotes the total number of nodes at this depth.

While Bell (Logarithmic) Activation normalizes the activation with respect to the count of nodes for the current level of activation, our new *Branch-normalized Activation* function normalizes the value of an activated node by the count of its higher-level nodes. Thus, nodes with many higher order concepts do not gain more "influence" than concepts that have few ancestors. Branch-normalized activation is defined as follows:

$$BN_i = \frac{1}{|higher\text{-}order\text{-}nodes_i|}, \tag{4}$$

where $higher\text{-}order\text{-}nodes_i$ is the set of higher order nodes for a given node i. Equation 5 shows the activation function as a whole.

$$a_i = a_i + a_j \times D \times BN \tag{5}$$

IV. Profiling Twitter Users in Computer Science

The baseline of our experiments is the extraction of professional interests using no spreading activation function. Below, we first describe the dataset used and subsequently provide an overview of the applied similarity functions, before we present and discuss the results in the subsequent sections.

A. Datasets

a) Hierarchical Knowledge Base: For the extraction of concepts, ACM's Computer Classification System (CCS) was utilized. The ACM CCS contains 2299 concepts ($|V| = 2299$). It contains concepts in the field of computer science as well as their relations, alternative spellings, and dictions. The number of labels is $|L| = 11385$. Thus, on average a concept has 4.95 labels ($SD = 3.59$). The CCS consists in total of six hierarchy levels. The number of nodes (i. e., concepts) over the different levels follows a normal distribution. Thus, it forms a bell shape (cf. discussion in Section III).

b) Twitter Data: A group of users was collected by searching Twitter for A*-rated[4] computer science conference hashtags. A*-rated conferences were chosen because of their high number of participants and importance for the scientific community. We used only those hashtags that were officially used and propagated on the conference homepages or official conference Twitter profiles (26 conferences). We queried the Twitter API for each of the 26 conferences and extracted users who used one of those hashtags in at least one of their tweets. Subsequently, we filtered the obtained user list and kept only these users who also appeared on DBLP. To avoid ambiguity, we dismissed all Twitter users whose actual full names match with more than one DBLP entry. Through this procedure, we identified the Twitter accounts and corresponding DBLP records of 157 computer scientists. We retrieved the tweets for all 157 users. In total, these are $96,437$ tweets. On average, a user published 614.24 tweets (SD$= 403.13$).

c) DBLP Data: For the evaluation data, we retrieved the users' scientific publication lists from DBLP. We used the extended DBLP data set, *AMiner Citation Network Dataset*[5] to obtain the titles as well as abstracts of the publications. In total, we have obtained $4,111$ titles. Thus, on average a user has 26.18 publications. For nearly one third of the scientists in our data set, we also obtained the abstracts of their publications. Altogether, we were able to obtain 825 abstracts.

B. Similarity Measures

Jaccard Similarity is used as first similarity measure. It estimates the *semantic closeness* between the social media vector and the evaluation vector, which are computed as introduced in Section II. Jaccard similarity is defined in Equation 6, where s_i denotes the number of times a certain concept is found in the social media data and e_i represents the number of times a certain concept is found in the evaluation data set.

$$similarity_{jac} = \frac{\sum_{s_i \in S, e_i \in E} min(s_i, e_i)}{\sum_{s_i \in S, e_i \in E} max(s_i, e_i)} \tag{6}$$

With *Precision at k (P@k)*, we measure to which extend the professional profile extracted from the social media data contains concepts that are part of the evaluation data. Concepts are ranked by their scores. It is calculated per user and defined as:

$$P@k = \frac{|S_k \cap E_k|}{|E_k|}, \tag{7}$$

where S_k is the set of the top k concepts in the ranked list of concepts extracted from the social media data for a specific user. E_k is the set of the top k concepts extracted from the corresponding evaluation data. Finally, P@k is averaged over all users.

Cosine Similarity calculates—metaphorical spoken—the *angle* between the social media vector S and the evaluation data vector E. It is interpreted like the Jaccard similarity.

$$similarity_{cos} = \frac{S \cdot E}{\parallel S \parallel \parallel E \parallel} \tag{8}$$

While P@k only considers the existence of a concept in the k-sized, ranked list of the data, the Average Precision (AvgP) takes the rank of the concepts into account as shown in Equation 9. Here, $number\text{-}of\text{-}concepts$ stands for the number of concepts that were considered, i. e., k.

$$AvgP = \frac{\sum_{k=1}^{n}(P@k \times rel(k))}{number\text{-}of\text{-}concepts}, \tag{9}$$

where $rel(k)$ equals 1 if the concept found at rank k is relavent and 0 otherwise. $P@k$ is used as defined above.

The *Mean Average Precision (MAP)* is calculated by summing up the average precision values for all users and dividing it by the number of users. Thus, MAP is an indicator not only for how well the social data reflects the professional interests, but also how well the importance is reproduced. Its definition is given below. Please note, $AvgP(u)$ is the $AvgP$ as defined in Equation 9 for a given User u.

$$MAP = \frac{\sum_{u=1}^{U} AvgP(u)}{|U|}, \tag{10}$$

One might assume that the highest ranked concept in the ranked list of evaluation data represents the most significant professional interest of a user. In order to evaluate this most important concept, we use the *Mean Reciprocal Rank (MRR)*:

$$MRR = \frac{1}{|U|} \sum_{i=1}^{|U|} \frac{1}{|rank_i|} \tag{11}$$

U denotes the set of users. For each user $u \in U$, the most significant professional interest is extracted from the evaluation data set, i. e., the concept with the highest score. Next, the rank for this concept in the social media profile is identified. The MRR is the summed up fraction of 1 by the rank for all users, divided by the total number of users. In other words, the MRR shows the mean rank of the most important concept per user, averaged over the set of all users U.

Rankscore is a measure based on the assumption that a user might have less interests in elements that appear in a lower rank. Equation 12 describes how to calculate the rankscore for a specific user:

$$rankscore_u = \sum_{i \in hits_u} \frac{1}{2^{\frac{rank_i - 1}{\alpha - 1}}}, \tag{12}$$

where α denotes a *viewing halflife* parameter which controls the speed of the decay. Following the suggestion of Breese et al. [4], we use $\alpha = 5$. Furthermore, $hits_u$ refers to the concepts c_i found in our evaluation data and $rank_i$ stands for its rank

[4]CORE ranking, see http://103.1.187.206/core/, last access: August 31, 2014

[5]http://arnetminer.org/citation, last access: August 31, 2014.

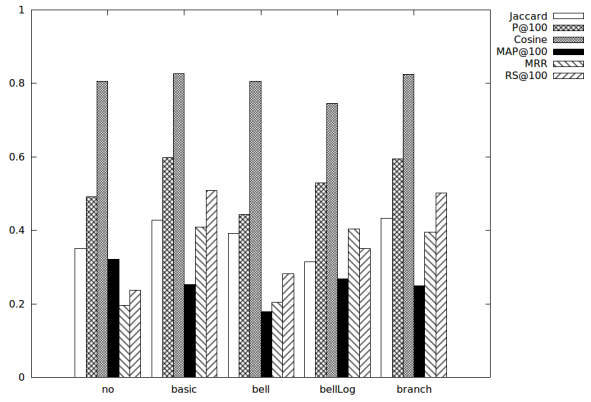

Fig. 2. Measures for different spreading activation functions.

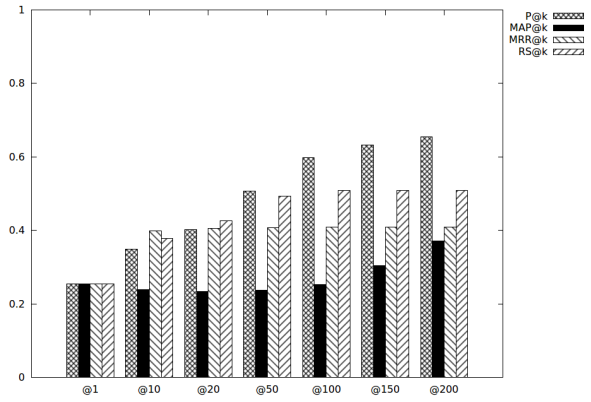

Fig. 3. Influence of different values of k on the measures P@k, MAP@k, MRR@k, and RS@k using basic activation with a decay factor of 1.0.

of a concept c_i in the social media items. When a concept from the evaluation data is not found in the social media items, we simply ignore it. Following Breese et al. [4], the score is normalized. The interpretation of rankscore is similar to MRR. However, in contrast to MRR, the rankscore does not only consider the top concepts but takes the corresponding ranks of all detected concepts into account. This reveals deeper insights into the accuracy of the ranking of concepts compared to MRR. We applied a variation of the rankscore that we call *Rankscore@k (RS@k)*, where we perform a cut-off at k concepts (e. g., 10, 20, 50 or 100).

V. Results

Figure 2 shows the baseline results of using no spreading activation function (left) and the four activation functions with respect to the evaluation measures Jaccard, P@100, cosine similarity, AP@100, MRR, and RS@100. As one can see, the spreading activation functions have an influence on the similarity measures. The strength of this influence differs from measure to measure. For example, the spreading activation functions show less impact on P@100 and cosine similarity compared to, e. g., MRR and RS@100. While for some measures like MRR and RS@100, basically any spreading activation function increases the scores, some measures like MAP@100 show overall lower scores when applying spreading activation. Again, for other measures like Jaccard and cosine similarity, it depends on which activation function is used. For instance, cosine similarity is slightly higher when the basic spreading activation or the branch-normalized activation is used and is lower when the Bell logarithmic activation is applied. We obtain the highest values for all measures except P@100 when applying the basic spreading activation function.

For those measures that deal with a selection of the top k concepts, i. e., P@k, MAP@k, MRR@k, RS@k, we additionally investigated the influence of considering different values of k. We choose to investigate the basic spreading activation function as it showed the best performance for all of the @k-measures. The results of applying different values of k between 1 and 100 on the basic activation function with a decay factor of 1.0 are shown in Figure 3. As one can see, P@k and RS@k increase with a higher value of k. They perform best at $k = 100$. MAP@k decreases slightly, but stays overall stable.

VI. Discussion

The use of spreading activation functions for concept extraction mostly enhances the performance of the investigated evaluation measures. We assume that the higher order concepts revealed by spreading activation functions further shape the already extracted user profile and produce overall higher similarity scores. Thus, this coincides with the work by [3], who also observed an effect using spreading activation but used their own cross-domain knowledge base. In contrast, we applied a readily available and high-quality knowledge base in the domain of computer science. In general, the approach to extract professional interests described in this paper is feasible and practical. A potential influence to our experiment might be the sampling of the users from popular and high-ranked conferences. However, the presented approach for extracting professional interests from social media is neither dependent on the popularity of the conferences nor their rankings. We use the conferences and their official hash tags only to find users on Twitter. Once a relevant user on Twitter has been identified and his or her publication list on DBLP is found, our measures are solely dependent on the social media items authored by the scientist as well as his or her publications. Another potential impact might be that users tweet the title of their own publications together with a link. However, the impact of such dissemination activities to the overall results can be considered low as tweeting one's own publication title is likely to happen only once.

References

[1] J. Letierce, A. Passant, J. Breslin, and S. Decker, "Understanding how Twitter is used to spread scientific messages," *Web Science Conference*, 2010.

[2] F. Abel, E. Herder, and D. Krause, "Extraction of professional interests from social web profiles," in *Proceedings of International Workshop on Augmenting User Models with Real World Experiences to Enhance Personalization and Adaptation (AUM)*. Springer, 2011, pp. 1–6.

[3] P. Kapanipathi, P. Jain, C. Venkataramani, and A. Sheth, "User interests identification on Twitter using a hierarchical knowledge base," in *The Semantic Web: Trends and Challenges*. Springer, 2014, pp. 99–113.

[4] J. S. Breese, D. Heckerman, and C. Kadie, "Empirical analysis of predictive algorithms for collaborative filtering," in *Proceedings of the Fourteenth conference on Uncertainty in artificial intelligence*. Morgan Kaufmann Publishers Inc., 1998, pp. 43–52.