Leveraging the LinkedIn Social Network Data for Extracting Content-based User Profiles

Pasquale Lops
Dept. of Computer Science
University of Bari Aldo Moro,
Via E. Orabona, 4 - Bari, Italy
lops@di.uniba.it

Marco de Gemmis
Dept. of Computer Science
University of Bari Aldo Moro,
Via E. Orabona, 4 - Bari, Italy
degemmis@di.uniba.it

Giovanni Semeraro
Dept. of Computer Science
University of Bari Aldo Moro,
Via E. Orabona, 4 - Bari, Italy
semeraro@di.uniba.it

Fedelucio Narducci
Dept. of Computer Science
University of Bari Aldo Moro,
Via E. Orabona, 4 - Bari, Italy
narducci@di.uniba.it

Cataldo Musto
Dept. of Computer Science
University of Bari Aldo Moro,
Via E. Orabona, 4 - Bari, Italy
cataldomusto@di.uniba.it

ABSTRACT

In the last years, hundreds of social networks sites have been launched with both professional (e.g., LinkedIn) and non-professional (e.g., MySpace, Facebook) orientations. This resulted in a renewed information overload problem, but it also provided a new and unforeseen way of gathering useful, accurate and constantly updated information about user interests and tastes. Content-based recommender systems can leverage the wealth of data emerging by social networks for building user profiles in which representations of the user interests are maintained.

The idea proposed in this paper is to extract content-based user profiles from the data available in the LinkedIn social network, to have an image of the users' interests that can be used to recommend interesting academic research papers. A preliminary experiment provided interesting results which deserve further attention.

Categories and Subject Descriptors

H.3.1 [**Information Storage and Retrieval**]: Content Analysis and Indexing—*Dictionaries, Indexing methods, Linguistic processing*; H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval—*Information Filtering*

General Terms

Algorithms, Experimentation

Keywords

Content-based Recommender Systems, Social Networks, LinkedIn

RecSys'11, October 23–27, 2011, Chicago, Illinois, USA.

1. MOTIVATIONS AND RELATED WORK

Each researcher registering to a large scientific congress generally takes a look to the conference program on-site, and the risk is that she feels completely overloaded by the long lists of talks and presentations listed. Which presentations should be worth to attend?

Usually the most interesting talks might be guessed from their titles and authors or by taking a quick look to the papers in the conference proceedings. Recommender systems address these issues because they have the effect of guiding users in a personalized way to interesting objects in a large space of possible options.

Systems for the recommendation of academic research papers are based on very different strategies. Quickstep and Foxtrot adopt an ontological approach to user profiling [9]. They represent user profiles in terms of a research paper topic ontology, by unobtrusively monitoring behaviour and by adopting a relevance feedback mechanism. Research papers are classified using ontological classes and collaborative recommendation algorithms are used to recommend papers seen by similar people on their current topics of interest.

Other interesting approaches are based on the use of citation graphs. In [8], a collaborative filtering approach to paper recommendation is adopted, by using the citation graph between papers to create the ratings, while in [3], the PaperRank algorithm is presented. It adopts a random walk based scoring strategy, which can be used to recommend papers according to a small set of user selected relevant articles.

Another interesting approach has been proposed in [2]. The idea is that researchers from the same area tend to be interested to the same articles, so it is possible to improve search results by using recommendations based on previous searches performed by other people with similar interests.

In a previous work we developed the *Conference Participant Advisor* service to suggest papers to be read and talks to be attended by a conference participant [10]. The service was built using a content-based recommender system exploiting a Bayesian learning method, and a WordNet-based Word Sense Disambiguation procedure [4] to learn semantic user profiles exploited for deliverying personalized conference programs. The system is trained by a set of *interesting*

and *not interesting* research papers in order to learn the researcher's profile.

The above mentioned approaches are proved to be accurate, but they have some limitations:

- need of both positive and negative examples of researchers' interests;
- cold start problem is poorly addressed, i.e. the system is not able to provide an accurate profile for new researchers.

For example, let us suppose to adopt a strategy to extract profiles by the analysis of papers written by a specific researcher: how to build the profile for a researcher who has not co-authored any paper? Is it possible to acquire interests from already available sources, such as social networks?

The goal of this paper can be formulated in form of a research question as follows: *Is it possible to exploit information coming from social networks, such as LinkedIn, for modeling researchers' interests in user profiles which can be used to provide accurate recommendations?*

Some recent works argued the benefits of interweaving public profile data on the Web [1], and the usefulness of users' self-defined social relations to increase the quality of the recommendation process in collaborative filtering systems [6], thus we are confident to obtain similar results in content-based recommender systems as well.

2. EXTRACTING USER PROFILES FROM LINKEDIN

2.1 The LinkedIn Social Network

LinkedIn[1] is a business-oriented social networking site launched in May 2003, with more than 90 million registered users (January 2011), spanning more than 200 countries worldwide. A new member joins LinkedIn approximately every second. The purpose of the site is to allow registered users to maintain a list of contact details of people they know and trust in business, called *connections*. The list of connections determines the user's social graph.

Each registered user provides both personal and professional data as free text. Among professional data, we exploited *Specialties* (professional skills), *Interests*, and *Groups and Associations* the user is member of, which have been proved very important sources for improving collaborative filtering systems [7]. Figure 1 depicts an example of professional data extracted from LinkedIn.

2.2 The LinkedIn Extractor system

We developed the *LinkedIn Extractor* system, which processes the information extracted from LinkedIn for building the researcher profile. The idea is to exploit both professional data of the researcher and those of the connections in her social graph, i.e. her "colleagues". Indeed, data extracted by the connections have revealed a valuable source of information, since social networks grow around common interests [7]. This is also claimed in recent studies on homophily, which shows that friendship and interests are strongly interlinked: having even a few common interests makes friendship significantly more likely, and being friends

[1]http://www.linkedin.com/

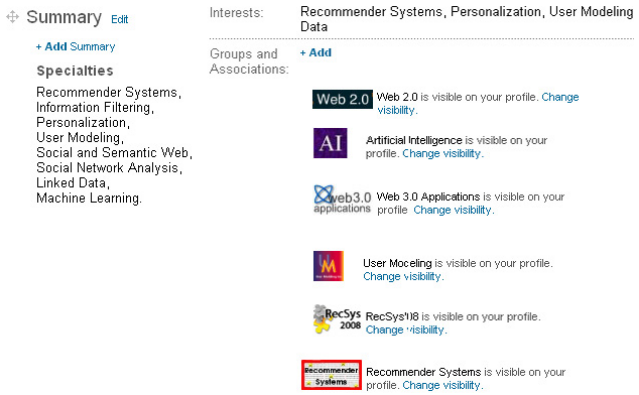

Figure 1: Specialties, Interests and Groups and Associations in a LinkedIn profile

also makes a pair of users more likely to share common interests [5]. Inspired by these results, we decided to include in the researcher profile the information coming from her social graph.

The architecture of the system is depicted in Figure 2. Let u_a be the active user (for whom recommendations must be provided), and C_a be the set of users in the u_a social graph. The user u_a approves the connection of the LinkedIn Extractor to the LinkedIn social network using the OAuth protocol[2] in order to protect her account credentials. Then, data extraction can be performed. In the following, the whole process for building the profile for u_a is described by defining the role of each component of the architecture:

- *Data Extractor*: it exploits the LinkedIn APIs for collecting the text describing the *Specialties*, *Interests*, and *Groups and Associations* of u_a. The raw text coming from those three types of data is included in a single document. The same process is performed for all users in C_a, in order to create a corpus of documents. This strategy allows the adoption of the vector space retrieval model for the internal representation of users.
- *Indexer*: it performs basic natural language processing operations (tokenization and stopword elimination) on the document collection provided by the previous module, in order to build the dictionary, i.e. the feature space. Each user u_i is represented as a vector $\overrightarrow{u_i}$ in an n-dimensional space, where each dimension corresponds to a keyword in the dictionary:

$$\overrightarrow{u_i} = \langle w_{1i}, w_{2i}, \ldots, w_{ni} \rangle \quad (1)$$

  w_{ji} in the vector is the standard TFIDF score of the keyword k_j in the document d_i associated with u_i. This representation strategy allows also to measure the closeness between two users by simply computing the cosine similarity $sim(\overrightarrow{u_i}, \overrightarrow{u_j})$ between the corresponding vectors.
- *Profiler*: it builds the profile $\overrightarrow{p_a}$ of u_a as a vector of weights obtained by adding the vectors of connections in C_a to $\overrightarrow{u_a}$. According to this strategy, the importance of a keyword k_j in the profile of the active user

[2]http://oauth.net

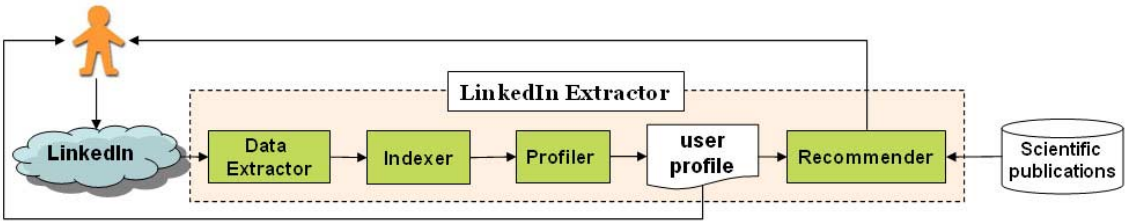

Figure 2: The conceptual architecture of the LinkedIn Extractor system

is computed by taking into account not only the weight of k_j in the vector associated with u_a, but also in the vectors of similar connections. This allows to enrich the profile with data coming from her neighborhood, in the same way as collaborative filtering algorithms do. The complete formula is as follows:

$$\overrightarrow{p_a} = \alpha * \overrightarrow{u_a} + (1 - \alpha) * \sum_{k=1}^{|C_a|} sim(\overrightarrow{u_a}, \overrightarrow{u_k}) * \overrightarrow{u_k} \qquad (2)$$

Parameter α is responsible for shaping the resulting vector in a direction closer to the active user or to her connections, while $sim(\overrightarrow{u_a}, \overrightarrow{u_k})$ ensures that the contribution of connection u_k to the profile of the active user is proportional to its similarity. $\alpha = 1$ corresponds to a profiling strategy which neglects the data of the neighborhood, while $\alpha = 0$ allows to rely just on the data coming from the connections. Other values of α allow to take into account both professional data of the user, and that coming from her connections. Finally, the resulting weights are normalized in the [0,1] interval.

- *Recommender*: it adopts a *ranked list* approach to select relevant documents in a target collection. Cosine similarity is computed between the user profile and all documents in the target collection, provided that they are represented as vectors. Documents are ranked in descending order according to the computed value and then the top-k documents are selected as recommendations. In the context of recommending scientific papers to a conference participant, the target document collection might be the set of papers accepted for presentation at that conference.

3. PRELIMINARY EVALUATION

The goal of the preliminary evaluation was to measure the accuracy of profiles extracted from LinkedIn data in the task of recommending relevant papers to a researcher.

The set of profiles to be evaluated was obtained by inviting 45 researchers in the area of Computer Science to connect to the LinkedIn Extractor service, in order to allow the extraction of the information concerning their interests and connections. After one month, 22 researchers completed the procedure, hence a set of 22 profiles was available for the evaluation.

In order to avoid to require a feedback about the recommendations produced by the system, for each researcher, a test set containing both *relevant* and *not relevant* papers (50% - 50%) was automatically built by selecting papers from the DBLP XML dump[3], containing 1.4 million publications. The set of *relevant papers* is built by considering those cited in the papers co-authored by the researcher (active citations), and those citing the papers co-authored by the researcher (passive citations). The set of *not relevant documents* is built by randomly selecting the rest of articles in the DBLP dump, from which we removed also papers co-authored by the researcher and those published in journals or conferences proceedings to which the researcher has contributed.

For each paper in the test set, title and abstract are extracted from the DBLP dump and processed by the *Indexer* in order to have a vector representation. For each researcher involved in the experiment, the articles in the test set are ranked by the *Recommender* according to the similarity to the researcher profile, as described in the previous section.

The measure adopted to evaluate the accuracy of the ranked list was R-Precision, that is the precision of the system at the R-th position in the ranking, where R is the number of relevant papers for each researcher. R-Precision is computed as the ratio between the number of relevant papers in the top-R positions of the list and R.

Two parameters are used in the experiment: α, which controls the importance of keywords occurring in the researcher data with respect to those in the data of her connections, and the threshold T, introduced for selecting the most similar connections in the social graph. In equation (2), only connections whose similarity with the active user exceeds T are included in the computation of the profile. Results of the experiment are presented in Table 1.

Table 1: Performance of the LinkedIn profiles in terms of R-Precision. Connections in a user's social graph with a similarity greather than T are selected.

Profiling strategy	T = 0.05	T = 0.15	T = 0.25
α=1	0.71	0.71	0.71
α=0	0.37	0.39	0.47
α=0.6	0.51	0.53	0.70

The first outcome is that the highest precision is achieved whether researchers profiles are built on their own LinkedIn data without exploiting connections ($\alpha = 1$). Results are encouraging even though the amount of text extracted from LinkedIn is limited and the vocabulary adopted by users is usually very restricted. We also investigated the performance of profiles built by relying exclusively on data coming from the connections ($\alpha = 0$). In this case the precision improves whether the most similar connections of the active

[3] http://dblp.uni-trier.de/xml/

user are taken into account: the higher the threshold T, the higher the profile accuracy. The same observation can be made whether both data of the active user and her connections contribute to the profile ($\alpha = 0.6$). In particular, precision comparable to that of profiles built with $\alpha = 1$ is achieved by setting $T = 0.25$, while lower values for T lead to a significant decrease. This is a clear indicator that the way of selecting neighbors from the users' social graph is important for obtaining accurate recommendations.

The main result observed is that accurate profiles are obtained either without taking into account connections or by including in the profiling process only the most similar connections of the active user. We need to further investigate this aspect, in particular whether neighborhood selection methods other than cosine similarity could lead to an improvement of system performance. The tuning of the parameter α is also required.

Despite the low number of users involved in the experiment, the results of this preliminary evaluation is encouraging. We have shown that data extracted from the LinkedIn social network are a reliable source of information for representing a researchers' interests, even though they contain very simple data. This might represent a possible solution to the *cold-start* problem, for providing recommendations to new researchers – for whom a set of publications from which to learn a profile might not be available – or for modeling their *evolving interests*. Indeed, if a system just relies on the bibliography of the researcher for building her own profile, it might not be able to model new interests, not yet substantiated by scientific publications.

4. CONCLUSIONS AND FUTURE WORKS

We investigated the problem of establishing whether professional information extracted from social networks, such as LinkedIn, is helpful for building researcher profiles which are exploited to suggest relevant academic research papers. We presented LinkedIn Extractor that includes in the profile of a LinkedIn user both her own professional data and those coming from her connections. An evaluation in the context of recommending relevant scientific papers to Computer Science researchers showed the potential of the approach and raised some issues for future investigations. Two main outcomes were observed:

- LinkedIn professional data are reliable for modeling user interests since the accuracy of the resulting profiles is acceptable;
- the whole social graph of a researcher introduces noise into the recommendation process, but a subset of accurately selected colleagues might contribute positively to her profile.

This leads us to foresee further investigations on neighborhood selection methods other than cosine similarity. Moreover, more sophisticated techniques might be adopted by the *Indexer* in order to create a space of concepts, rather than simple keywords, for the representation of both documents and profiles. For this purpose, we plan to exploit the ACM Computing classification hierarchy, and our previous work for semantic indexing of documents [4].

Another future improvement is to introduce in the profiling strategy new data available in LinkedIn, such as *Events*, *Publications*, and *Discussions* within groups.

A more extensive user study is ongoing, that combines LinkedIn profiles with those extracted by analyzing the bibliography of researchers, obtained by the list of publications on DBLP. The main problem is the unwillingness of people to provide access to her own personal data. For this reason we are automatically gathering a massive amount of public LinkedIn profiles of selected authors from the DBLP dump (LinkedIn APIs allow for this), in order to test how well such profiles may perform in the specific recommendation task.

5. REFERENCES

[1] F. Abel, N. Henze, E. Herder, and D. Krause. Interweaving public user profiles on the web. In *User Modeling, Adaptation, and Personalization, UMAP 2010*, volume 6075 of *Lecture Notes in Computer Science*, pages 16–27. Springer, 2010.

[2] N. Agarwal, E. Haque, H. Liu, and L. Parsons. Research paper recommender system: A subspace clustering approach. In *Advances in Web-Age Information Management*, volume 3739 of *Lecture Notes in Computer Science*, pages 475–491. Springer, 2005.

[3] M. Gori and A. Pucci. Research paper recommender systems: A random-walk based approach. In *Web Intelligence*, pages 778–781. IEEE Computer Society, 2006.

[4] G.Semeraro, M. Degemmis, P. Lops, and P. Basile. Combining Learning and Word Sense Disambiguation for Intelligent User Profiling. In Manuela M. Veloso, editor, *IJCAI 2007, Proceedings of the 20th International Joint Conference on Artificial Intelligence, Hyderabad, India, January 6-12, 2007*, pages 2856–2861. Morgan Kaufmann, 2007.

[5] H. W. Lauw, J. C. Shafer, R. Agrawal, and A. Ntoulas. Homophily in the digital world: A livejournal case study. *IEEE Internet Computing*, 14(2):15–23, 2010.

[6] D. H. Lee and P. Brusilovsky. Social networks and interest similarity: the case of CiteULike. In M. H. Chignell and E. Toms, editors, *Proceedings of the 21st ACM Conference on Hypertext and Hypermedia, Canada*, pages 151–156. ACM, 2010.

[7] D. H. Lee and P. Brusilovsky. Using self-defined group activities for improving recommendations in collaborative tagging systems. In X. Amatriain, M. Torrens, P. Resnick, and M. Zanker, editors, *Proceedings of the 2010 ACM Conference on Recommender Systems, Barcelona, Spain*, pages 221–224. ACM, 2010.

[8] S. M. McNee, I. Albert, D. Cosley, P. Gopalkrishnan, S. K. Lam, A. M. Rashid, J. A. Konstan, and J. Riedl. On the recommending of citations for research papers. In *Conference on Computer Supported Cooperative Work*, pages 116–125, 2002.

[9] S. E. Middleton, N. R. Shadbolt, and D. C. De Roure. Ontological User Profiling in Recommender Systems. *ACM Trans. on Information Sys.*, 22(1):54–88, 2004.

[10] G. Semeraro, P. Basile, M. Degemmis, and P. Lops. Discovering User Profiles from Semantically Indexed Scientific Papers. In *From Web to Social Web: Discovering and Deploying User and Content Profiles*, volume 4737 of *Lecture Notes in Computer Science*, pages 61–81. Springer, 2007.