Semantic-based Expert Search in Textbook Research Archives

Marco Pavan[1] and Ernesto William De Luca[2]

[1] University of Udine,
Via delle Scienze 206, Udine, Italy.
marco.pavan@uniud.it
[2] Georg-Eckert-Institut - Leibniz-Institute for international Textbook Research,
Celler Straße 3, Braunschweig, Germany.
deluca@gei.de

Abstract. Expert finding and the identification of similar professionals are important tasks for many services provided by companies and institutions. Nowadays, the rapid growth of web services and social and professional networks, allowed different kind of users to share personal data and increased the amount of information available. Most of research works focus on a limited set of users, characterized by the same kind of main activities, e.g., researchers, or exploit external knowledge, such as predefined ontologies. An heterogeneous environment, with possible lack of information, and not well structured data, puts forward new challenges, to address the problem of adapting user profiling and consequently expert search. In this paper, we first provide a general perspective on studies on expert finding, similar people identification and social recommender systems, highlighting some critical issues related to information extraction and user profile definition. We then present a first attempt to create an expert search system to support users in Library and Archiving Communities, such as researchers, students, authors, in the field of Textbook Research, in finding other experts to get in contact with or to start a cooperation. This is organized in three phases: first, we collect information about users in order to build structured profiles; then, we build a Community Knowledge Graph (CKG) which defines relationships and weights among terms that occur in the profile sections, emphasizing information shared by users in the entire analyzed community. As third step, we exploit the CKG structure to define similarity values among users based on weights got by their common terms in the CKG, and their distances in the graph. We conjecture that the CKG allows to model users emphasizing new semantic aspects of relationships among profile elements, and helps to improve the similarity computation and the expert search. We present a preliminar experimental evaluation on real users.

Keywords: Semantic Enrichment, Expert Search, User Modeling

1 Introduction

In recent years, the increasing pervasiveness of social network platforms has made users able to share personal data, also for professional purposes, in order to find positions and collaborations offering competence and knowledge. Also several companies and institutions have strong interest in exploiting that kind of information to find people with particular skills and expertise that fit their needs. With these premises it is clear how important is addressing the expert search problem and exploiting new sources of potentially interesting information, therefore, accurate expert search systems enable users and companies to quickly find the right desirable people without being overwhelmed by irrelevant information or losing too much time seeking on several web platforms.

There has been an increasing interest among researchers in improving the search results of expert finding, and several solutions had been proposed, even recently [8], [12], [13], [20]. However, there are still many problems when adopting those solutions in real situation.

Most of works in expert search focus on a single assumption about expert similarity (i.e., researchers are similar if they have similar papers, or similar topics; people on Q&A platforms are similar if their questions and answers cover similar topics, etc.). While this is working well on specific platforms for specific tasks, a more general and adaptable framework could be needed to take in care multiple possible sources for user similarity, i.e. research papers if available, tweets if users use Twitter, etc., and to allow us to build an adaptive system able to handle different search tasks. In real situations users could have different needs and different approaches on expert search. Someone could be an expert who is looking for partners with same skills, thus, she could try to search profiles similar to herself; on the other hand, an inexperienced user, such as a student, might need general knowledge about a topic, with no particular requirements on user expertise, just with the sufficient background or interest. Yet others, could be interested in finding users that have collected a lot of information about a topic, thus, not related to work or skills, but with great interest on a subject, such as a particular sport, political party, etc. On this basis, it is clear how a more general profiling could help, and how information about users obtained from heterogeneous information sources could enrich the profile, providing different insights on user similarity. Recent studies [11], [10], [1], [17] demonstrate that information from social networks can be exploited to improve accuracy of recommendations and user similarities, therefore it seems reasonable to continue on this research direction.

Another critical issue is how to get the correct knowledge base, that helps to connect texts to semantic entities, such as categories or topics, and allows us to properly model users and build their profiles. Research works, focused on a single domain, exploit specific ontologies or databases to address this problem, but obviously this approach is limited to a set of users of the same type, for instance researchers in a specific field, employees in a particular company, or customers interested in specific products.

One of the major challenges of digital archiving for Textbook Research is how to deal with changing technologies and changing user communities, which necessitate tools to formalize, detect and measure knowledge evolution. Semantic representations of contextual knowledge about cultural heritage objects and users, especially in Textbook Research, will enhance organization and access of data and knowledge, because usually the relationships among them are not emphasized or even identified.

In this paper, we focus on a novel semantic-Blser approach for expert search. We create a Community Knowledge Graph (CKG) used for semantic enrichment. The graph is dynamically built during the users set analysis and allows to define user similarities by comparing the relationships among terms and entities. In particular, we collected structured information similarly to a CV, and taking into account the terms occuring together in different fields to emphasize their relationships to other users sharing the same content. Then we build the knowledge behind their profiles. Indeed, we observed that some users who work on similar projects might have similar biographical data, or share same interests. On the other hand, students with interests on particular topic, could be connected to experts through similar profile content. With a specific ontology or database the community analysis is limited and it is not possible to properly model a general user who could seek for experts with different profile but with some intersection.

The novelty of our approach is the semantic enrichment of the profiles based on a community graph in order to improve the results of similarities scores among users. We exploit the CKG structure, with its weighted relationships, and we compare users considering the graph distances of their common terms in the CKG.

The paper is structured in 6 sections. After the introduction, we discuss related work in Section 2. We focus on the problem statement in Section 3 and describe our approach in Section 4. Section 5 describes our preliminary evaluation and Section 6 concludes the paper and shows our future work.

2 Related Work

The expert search problem has arised great interest among researchers, and several groups addressed the main problems behind this important topic, such as lack of information about users, or defining supporting knowledge needed to build models and improve the results. Gollapalli et al. [7] propose a graph-based model for expertise retrieval with the objective of enabling search using either a topic or a name. El-korany [5] proposes a novel cascaded model for expert recommendation using aggregated knowledge. He exploits social networks in order to extract useful contents for building a vector space model. With this approach he computes the relevance of contents respect to a specific query and applies PageRank algorithm to rank candidate experts. Fang et al. [6] investigate how to merge and weight heterogeneous knowledge sources to improve the expert finding process. Cetintas et al. [4] propose a discriminative probabilistic model

that identifies latent content and graph classes for people with similar profile. Moreira et al. [12] exploit unsupervised rank aggregation methods. They combine multiple estimators of expertise, derived from the textual contents, from a graph-structure of the citation patterns, and information extracted from user profiles. Yang et al. [20] utilize Normalized Google Distance (NGD) to enhance the relevance between initial query and extended query, and to improve the accuracy of the search results of the expert finding system.

Other researchers focus on specific platform, such as Twitter, to address specific expert search problem. Stankovic at al. [19] propose a method for suggesting potential collaborators for solving challenges online, based on their competence and interests. Yet others focus on issues related to collaborative filtering. Spaeth et al. [18] explore text-mining techniques to improve classical collaborative filtering methods for matching people who are looking for expert advice on a specific topic. Gujral et al. [8] try to go further and proposed a knowledge prototype for expert knowledge synthesis. Plumbaum et al. [16] propose a personalized recommendation application by combining semantic and structured information available in research communities. In particular they exploit encyclopedic knowledge sources, a large news article dataset, and collected implicit user feedback.

All of these related works have significant value, given the importance of all the issues addressed to improve expert search systems, and it is clear how one of the most important issues underlying these systems is the user modeling and the related eventual enrichment process. In this direction other authors recently presented their work. Abel et al. [1] propose a model based on Twitter posts linked to related news articles to identify activities. Noureddine et al. [13] address the problem of researchers profiling, exploiting structured and unstructured data from different heterogenous web sources. They propose an ontology-based architecture that correlates information coming from several sources. Also other researchers exploited multiple external sources to address the profile enrichment problem. Orlandi, in two works [14], [15] proposes a semantic approach for interlinking social websites and provenance management on the Web of Data; and a methodology for the automatic creation and aggregation of interoperable and multi-domain user profiles of interests using semantic technologies. The author build user profiles based on qualitative and quantitative measures about user activities across social sites. Song [17] demonstrates how integrating multiple social networks as external sources outperforms the use of only a single source, by proposing a conceptual volunteering decision model. Mizzaro et al. [11], [10] exploit Twitter and Wikipedia as external sources for enriching short texts and build a network-based user model to improve similarity scores. Al-Kouz et al. [2] analyze the users' social graph and the users' interactions with attention on posts and group memberships to model user interests and fields of expertise.

Another important issue often present in research works related to expert search is the use of a supporting knowledge base, usually defined with an ontology. Some researchers focused on ontology-based user interests modeling and matching. Cena et al. [3] propose an approach for propagating user interests on in ontology-based user models, in order to solve the cold-start problem in rec-

ommender systems by exploiting the ontological structure of the domain. Koh et al. [9] present an iconcept-matching approach to measure degrees of similarity among users. They exploit Kullback-Leiber distance to measure similarity on users represented by concept hierarchy.

3 Problem statement

To better understand what are the main problems and difficulties emerging with expert search tasks and related user profiling, we list a set of conceptual problems presented by the current state-of-the-art solutions. Many expert search systems, and even recommender systems, as first step, need to extract the starting information about users in order to build the profile. Most of existing systems rely on a single data source, with the risk to have incomplete information about users. Depending of the nature of that source, the obtained information could be related on only one aspect of the users, i.e. only their work expertise or skills, or purchases preferences; or it could be very schematic and represented by very short texts, therefore likely with poor knowledge about them. So two first conceptual problems are:

P1a: Poor profile problem - lack of information. With a single data source, the user profile could be incomplete and focused only on certain aspects, therefore ignoring a more complete user overview.

P1b: Poor profile problem - short texts. With a single data source, the user profile could be composed of short texts that make difficult the information extraction, due to their brevity that does not provide sufficient word occurrences.

Moreover, the texts that compose the initial user data, most of time are not structured, and users are represented as bag-of-words, with no information about what kind of data could be related to personal data, expertise, or general interests. We can define this problem as:

P2: Structural problem - unstructured information. Extracted user data do not have structure that allow us to identify what kind of information we have. This issue makes difficult the process of dividing information in sections to properly model users under several aspects.

Very recent works [11], [10] have highlighted how an enrichment process can overcome the lack of information and improve the results for several purposes. Some other new works in the literature [13], [14], [15], [17] introduced multi-source enrichment approaches in order overcome these problems. However some techniques exploit ontologies or other predefined knowledge bases as supporting structure. This approach is useful if the set of user to analyzed has characterized by the same interests, or they work in the same field, or even they share similar CV, but in heterogeneous environments such as Seek&Offer job platforms or Q&A services, a fixed knowledge base could be not perfect suited for that purpose. Moreover, the dynamic aspect of web platforms, where active communities

are not always the same, arise the need of having dynamic also the knowledge support, evolving over time. On this basis we define other two problems:

P3a: Supporting knowledge problem - predefined and domain-dependent semantics. Ontologies and any supporting knowledge bases are usually crafted for specific domains, therefore they might not be well suited for any set of users.

P3b: Supporting knowledge problem - fixed semantic structure. Ontologies and any supporting knowledge bases are usually fixed structures, therefore not able to change over time in adaptable frameworks.

Another important issue in expert search systems is related to user similarity computation, used for defining the scores between couples of users and obtain a ranked list of suggestions that meet the needs expressed by the requesting user. Most of the state-of-the-art proposed systems use a single function to compare the user models they build, but in heterogeneous environments could be helpful to emphasize only some aspects of users, in order to better fit the expressed need, taking in care what kind of user is the requester, or even analyzing for what purpose is the request itself.

P4: Similarity computation problem - Only one similarity score. The computation of only one similarity score between two users is a global value that does not take in care the several aspects of a user profile. Moreover, it does not give different weights to part of profile that need more or less emphasis, based on the current request.

Recent research works [11], [13], [16] addressed problems P1 and P2 exploiting information from external sources to get useful additional data, therefore in this paper we focus on problem P3, related to the supporting knowledge base in heterogeneous environments.

4 Proposed approach

In the following we describe our approach organized into three main phases. During the first phase we collect information about users in order to build structured profiles; the second phase consists in building a graph composed of all texts from all users, with attention on enriching the network with semantic relationships among entities obtained by the profiles structure; in the last phase we compute similarity scores among users exploiting the network structure that allows us to consider distances among terms which represent users.

Figure 1 shows an overall representation of our approach. Each phase is described in full details in the following sections.

4.1 Step 1: Structured user profiles

We have chosen a real situation to analyze as case study, in order to then investigate on the effectiveness of our approach. We started a pilot study at the

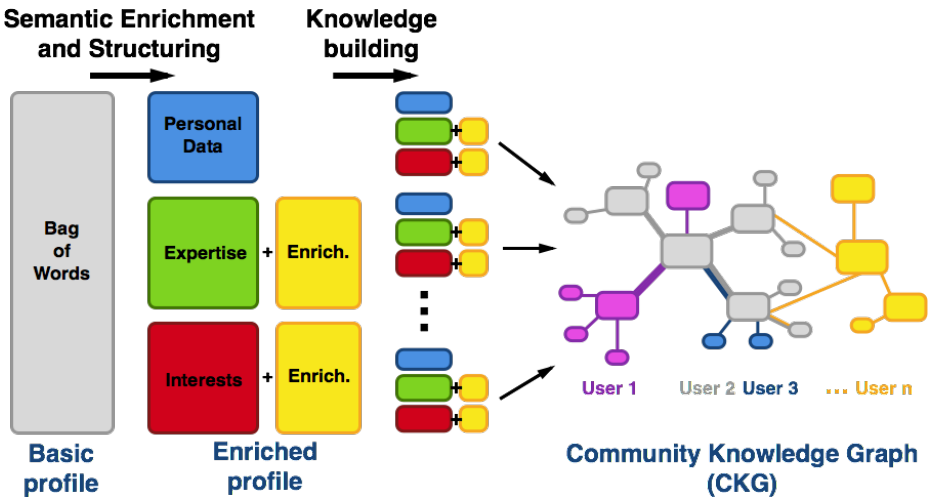

Fig. 1. Overview of the semantic approach

Georg-Eckert-Institute (GEI), to collect data about people who work in the international textbook research field, but with different roles.

A pilot study is a standard scientific tool for testing a research question in a "soft" way, allowing scientists to conduct a preliminary analysis before starting a full-blown experiment. In our case, we decided to start a survey in order to analyse how we can help researchers in finding other researchers that are interested in a very specific research area, namely Textbook Research. We wanted to find out how we can create semantic user profiles being different from other research portals like "research gate" or "academia".

We selected a sample of 32 users distributed as follows: 33% men, 67% women (different nationalities); 18% with age under 30, 55% between 31 and 40, 27% more than 40; 70% graduates, 30% doctors; 67% with activities related to research. We collected data related to biographical information, their current position, projects where they are/were involved in, their interests, etc., all for researchers, authors and students/trainees. We then structured information by grouping terms which compose the collected textual descriptions in a tree structure with a set of pre-selected entities. It is clear how an heterogeneous set of users could led us to have different profiles with different kind of shared interests or expertise have been collected into the community knowledge graph.

Figure 2 shows an example of such a structured user profile with the related entities and terms.

To test the feasibility, equipment and methods, we started the pilot study for finding out what are the important relations that should be collected for enriching the base profile within semantic information that can be derived from the needs of the researchers we asked to participate. For this study we focus on digi-

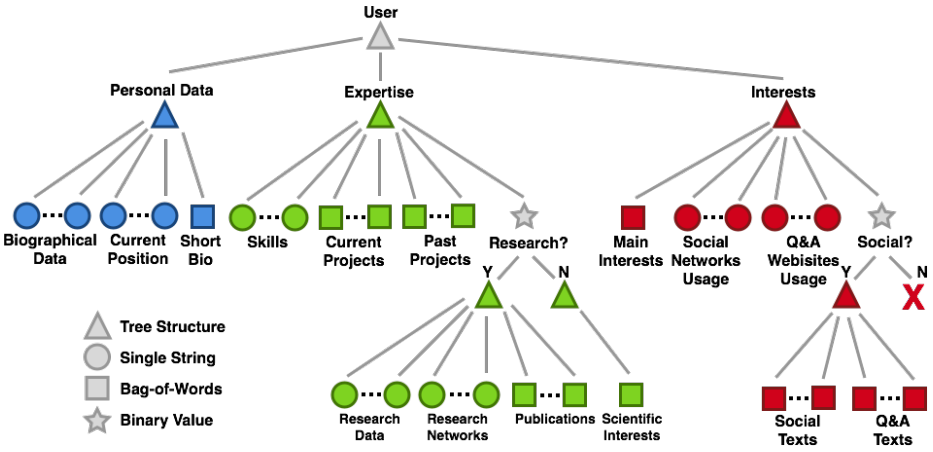

Fig. 2. Semantically enriched user profile

tal archives field and in particular on the case study of Textbook Research, that does not involve only researcher, but, as mentioned before, people with different backgrounds and interests. Therefore, this preliminary test is also important for training inexperienced users, explaining them how to interact with the system, what is the knowledge we extract, in order to get useful information to find out what are the sub-group of experiments we have to take into account after this pilot phase. Pilots are rapidly becoming an essential pre-cursor to many research projects, in order to reduce costs. At the same time we could find out how to set the next experimental phase, because the users gave us feedback of unclear statements or questions.

4.2 Step 2: Community Knowledge Graph

As second step we build a knowledge graph in order to have a structure that represent the entire community. We define this structure as Community Knowledge Graph (CKG). Based on the structure defined for user profiles we keep the same entities as nodes and we add directed edges in order to connect words used by users to those entities which represent the important aspects that characterize those users. We also add a new set of nodes in order to represent users and connect them to the words they used. In this way it is possibile to analyze the network by user. We scan all users' data in order to dynamically build the CKG and semantically enrich it with higher weights on edges where the relationship between a word and an entity is repeated. With this approach we emphasize the important relationships inside a community facing the problems P3 described in Section 3.

More formally, let $CKG = (V, E)$ the Community Knowledge Graph, where $V = \{v_1, v_2, ..., v_n\}$ is the set of all entities and words extracted from user profiles,

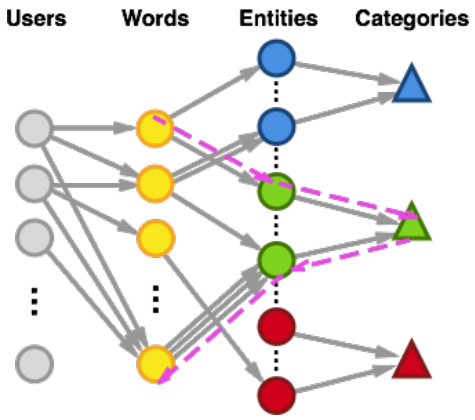

Fig. 3. Community Knowledge Graph (CKG)

and $E = \{e_1, e_2, ..., e_m\}$ is the set of undirected edges. We say that there exists $e_{v_i,v_j} \in E \Leftrightarrow v_j$ *isRelatedTo* v_i. The property "*isRelatedTo*" could be intended in several ways in order to define multiple semantic relationships. The *Term-Entity* relation is used for words belonging to the same concept, the *Entity-Entity* relation represents concepts related to a more general category of information which characterize users and the *User-Term* relation connects users with the words they use.

Figure 3 shows an example of CKG where it is possible to see how users are connected to the words they use, words to Entities, and finally to the more general entities that we call Categories. During the building process, if some words or entities are already present in the model the edges weights are increased in order to emphasize those relationships[3]. With this methodology we can build the real structure that users' aspects have. For instance, if a lot of people with a specific skill work in a particular company, we can have that strong relationship as high weight for the edge that connect the word with company name and the entity which represent the skills.

4.3 Step 3: Similarity computation

The creation of the CKG lets us comparing different users extracting information from their user profiles, defining similarity scores between them. The network structure allows us to have connections between words and entities based on the actual relationships found in the original texts, therefore we extract only the sharing nodes of two users, and we measure a similarity score based on the weights of the relationships computed as in the following:

Starting from two analyzed nodes v_i and v_j we follow the path in the graph until they share a common node. If this is the case, then we compute the similarity score based on the valid relationships, otherwise we discard it. In the

[3] For an easier reading, in Figure 3 multiple edges are shown.

case there are multiple paths we consider the path with the highest total edges weight, in order to emphasize the strong relationship between the two nodes that CKG has extracted.

In Figure 3, the purple dashed line highlights an example of selected path for the weighted distance. Formally, Let $F_{i,j} \subset E$ be the set of edges into the selected path from v_i to v_j defined with our rule previously described, we define $w(v_i, v_j)$, the weight of the relationships between v_i and v_j, as follows:

$$w(v_i, v_j) = \frac{w(F_{i,j})}{p_{len}(v_i, v_j)}$$

where $w(F_{i,j})$ is sum of edges weights into the respective set of edges, and $p_{len}(v_i, v_j)$ is the length of the path from v_i to v_j.

By repeating this process for all couples of nodes we can get a global score for user similarity.

Let $V_{u_a} \subset V$ and $V_{u_b} \subset V$ be respectively the sets of nodes of hypothetical user u_a and user u_b. We define the similarity score as follow:

$$sim(u_a, u_b) = \sum w(v_i, v_j)$$

with $v_i, v_j \in V_{u_a} \vee v_i, v_j \in V_{u_b}$.

5 Preliminary Evaluation

In order to study the effectiveness of our approach, we have set up a preliminary experimental evaluation. We have run three different algorithms on the dataset composed of the 32 users described in section 4.1. The first one, called *bow-s*, is a classic *bag of words* approach which exploit only the information users provide about their biography, therefore with no additional information about skills, current projects, interests. It simply counts the term frequency in order to build the user profile. The second one, called *bow*, is the same *bag of words* approach but using all information available about users, not only the "short bio". The third one, called *pei*, is our approach based on the semantic-based enrichment process that uses our CKG. We have run all algorithms 32 times in order to get the top ten suggested users for each one of the analyzed user. For the evaluation we have chosen a sample of 6 testers: 3 internals, therefore people who are working at the GEI and who know all users profiles and activities inside the institute; and 3 external, people who do not know them, but with access to the dataset with all text inserted.

Table 1 shows first results obtained by each approach, using the Precision metric at several levels, with resulting ranked lists with 3, 5 and 10 elements.

Table 2 displays the Discounted Cumulative Gain scores, using the Precision metric at several levels, with the same granularity levels.

Analyzing the results, we can notice that *bow* is less effective than *bow-s*, even if it exploit more text and information. The *pei* approach outperforms the

Table 1. Precision scores

Algo	P@3	P@5	P@10
bow-s	0.29	0.28	0.30
bow	0.12	0.18	0.30
pei	0.48	0.41	0.40

Table 2. Discounted Cumulative Gain scores

Algo	DCG@3	DCG@5	DCG@10
bow-s	0.72	0.96	1.50
bow	0.28	0.54	1.33
pei	1.48	1.82	2.59

other approaches and confirms the assumption that semantic enrichment can increase the performance of the retrieval system personalizing the results.

By looking at the Precision scores on Table 1 it is possible to notice how *bow-s* tends to keep the scores around a certain value, differently than *bow* that increases value as the resulting ranked list get longer. This issue highlights how the use of additional information could help on systems who provide longer lists of results but with no high precision on top ranked elements. Our proposed approach *pei* overcomes this issue by providing good top ranked results using more semantic information.

Table 2, with DCG scores, shows how all algorithms provide more gain for users, as the resulting ranked list get longer, and it is possibile to see how our technique *pei* outperforms the others at all levels.

6 Conclusions and future work

In this paper, we provided a first perspective on studies on expert finding, similar people identification and social recommender systems, highlighting some critical issues related to information extraction and user profile definition. We presented our semantic-based enrichment approach for expert search that bases on a Community Knowledge Graph, which defines relationships and weights among textbook researchers. The semantic relations help in finding the different experts that could be of interest and could be recommended.

For future work, we plan to expand our approach and investigate how the network structure could be exploited to improve the results, and to run the next evaluation with detailed user tests. Moreover, we want to explore the possibility to extend our approach to resources that users interact with, such as textbooks, manuscripts, or other cultural heritage objects, in order to improve semantic search and information retrieval tasks in digital archives and libraries, considering the relationships among them and with users.

References

1. F. Abel, Q. Gao, G. j. Houben, and K. Tao. Semantic enrichment of twitter posts for user profile construction on the social web. ESWC, June 2011.
2. A. Al-Kouz, E. W. De Luca, and S. Albayrak. Latent semantic social graph model for expert discovery in facebook. IICS, June 2011.
3. F. Cena, S. Likavec, and F. Osborne. Propagating user interests in ontology-based user model. In *Artificial Intelligence Around Man and Beyond*, volume 6934 of *Lecture Notes in Computer Science*, pages 299–311. September 2011.
4. S. Cetintas, M. Rogati, L. Si, and Y. Fang. Identifying similar people in professional social networks with discriminative probabilistic models. SIGIR, July 2011.
5. A. El-korany. Integrated expert recommendation model for online communities. *IJWesT*, Vol. 4, October 2013.
6. Y. Fang, L. Si, and A. P. Mathur. Discriminative probabilistic models for expert search in heterogeneous information sources. *Journal of Information Retrieval*, Vol. 14, April 2011.
7. S. D. Gollapalli, P. Mitra, and C. L. Giles. Ranking experts using author-document-topic graphs. JCDL, July 2013.
8. M. Gujral and S. Chandra. Beyond recommenders and expert finders, processing the expert knowledge. *International Journal of Computer Science Issues*, Vol. 11, January 2014.
9. W. Koh and L. Mui. An information theoretic approach to ontology-based interest matching. IJCAI, August 2001.
10. S. Mizzaro, M. Pavan, and I. Scagnetto. Content-based similarity of twitter users. ECIR, March 2015.
11. S. Mizzaro, M. Pavan, I. Scagnetto, and M. Valenti. Short text categorization exploiting contextual enrichment and external knowledge. SoMeRA, SIGIR, July 2014.
12. C. Moreira, B. Martins, and P. Calado. Using rank aggregation for expert search in academic digital libraries. *CoRR*, January 2015.
13. H. Noureddine, I. Jarkass, H. Hazimeh, O. A. Khaled, and E. Mugellini. Carp: Correlation-based approach for researcher profiling. SEKE, July 2015.
14. F. Orlandi. Multi-source provenance-aware user interest profiling on the social semantic web. UMAP, July 2012.
15. F. Orlandi, J. Breslin, and A. Passant. Aggregated, interoperable and multi-domain user profiles for the social web. I-SEMANTICS, September 2012.
16. T. Plumbaum, A. Lommatzsch, E. W. De Luca, and S. Albayrak. Serum: Collecting semantic user behavior for improved news recommendations. UMAP, 2011.
17. X. Song. Enrichment of user profiles across multiple online social networks for volunteerism matching for social enterprise. SIGIR, July 2014.
18. A. Spaeth and M. C. Desmarais. Combining collaborative filtering and text similarity for expert profile recommendations in social websites. In *UMAP*, volume 7899 of *Lecture Notes in Computer Science*, pages 178–189. 2013.
19. M. Stankovic, M. Rowe, and P. Laublet. Finding co-solvers on twitter with a little help from linked data. In *The Semantic Web: Research and Applications*, volume 7295 of *Lecture Notes in Computer Science*, pages 39–55. 2012.
20. K. Yang, Y. Lin, and C. Chuang. Using google distance for query expansion in expert finding. ICDIM, IEEE, September 2014.