Latent Semantic Social Graph Model for Expert Discovery in Facebook

Akram Al-Kouz, Ernesto William De Luca, Sahin Albayrak

DAI-Labor
TU-Berlin
Ernst-Reuter-Platz
10587 Berlin
{akram.alkouz,ernesto.deluca,sahin.albayrak}@dai-labor.de

Abstract: Expert finding systems employ social networks analysis and natural language processing to identify candidate experts in organization or enterprise datasets based on a user's profile, her documents, and her interaction with other users. Expert discovery in public social networks such as Facebook faces the challenges of matching users to a wide range of expertise areas, because of the diverse human interests. In this paper we analyze the social graph and the user's interactions in the form of posts and group memberships to model user interests and fields of expertise. The proposed model reflects expertise and interests of users based on experimental analysis of the explicit and implicit social data in Online Social Networks (OSNs). It employs social networks analysis, text mining, text classification, and semantic text similarity techniques to analyze and discover the latent semantic social graph model that can express user's expertise. The proposed model also considers the semantic similarity between user's posts and his groups, Influence of friendship on group's membership, and Influence of friendship on user's posts. Experiments on the Facebook data show significant validity of the proposed model.

1 Introduction

When looking for professional knowledge, we usually turn to people we know they are experts in a specific topic to ask for reliable and quick information or recommendations [BC03]. With the advent of OSNs a range of expert finding systems are emerging to help in locating experts [KN08]. Expert finding systems try to utilize OSNs structure and data to reach the candidate users in an efficient way. The structure of OSNs provides interpersonal communication among users leading to a large repository of personal data. Schneier in [Bs10] provides taxonomy of six categories based on the initiator source of data, the destination of data, and access control of data. From our point of view part of personal data in social networks is explicit data and the other part is implicit data.

Explicit data can be extracted directly from user's profile, such as geographic location and school attended Not all users provide these attributes, thereby reducing the usefulness of such social services [AB10]. Classical expert finding systems utilize user profile information and associations in explicit social graph to map a query to a candidate expert. The explicit social graph can also be used to propagate the likelihood from possible experts to other candidates. We think such model is not enough to reflect the real interests of the user in an environment such as Facebook, because neither the profiles data can really express current user interests, since it is not necessarily contains a correct, a complete and up to date information. Nor the influence of friendship is suitable to assume that all the friends in explicit social graph share the same interests, since friendship in Facebook is a binary relation.

Implicit data is latent in the social behavior of user must be derived and retrieved using a social discovery mechanism. Social discovery mechanism needs to analyze and utilize the explicit social graph as well as the implicit social graph of the user. Explicit social graph formed of user's friendship and group's membership that explicitly mentioned in the users profile. On the other hand, implicit social graph or activity graph can be predicted from user's behavior and interaction patterns, activity graph is the network that is formed by users who actually interact using one or many of the methods provided by the social network site [BA09]. From user's behavior and interaction patterns in Facebook, we have the observation that users interact about topics. To discover those topics we need to classify the user's textual production in Facebook to predefined ontology organized in a hierarchical structure of topics.

In this paper we aim to detect the implicit social graph in an efficient way to enhance the performance of expert finding systems. We analyze the explicit and implicit social graph to model user interests and fields of expertise. We used Facebook as a case study of OSNs. Based on the groups that Facebook user is member in, and the textual posts that he published in the last three months, we assumed that groups and posts can express the interests and expertise of the users in different levels, user's groups and posts have been classified to a predefined topics based on a general ontology derived from dmoz[1] open directory project. The intersection of the group's topics and post's topics has been examined to infer a valid model to represent the user expertise.

We tried to answer the following key research questions:

1. Can the group's membership reflect user's expertise and interests?
2. Can the user's posts be used to express his expertise?
3. How is the content of user' posts semantically related to the content of user's groups?
4. What is the suggested model to represent the latent semantic social graph?
5. What is the influence of friendship on group's membership and posts?

We found that group membership of a user can reflect his interests and expertise, but it maps user's groups to relatively wide range of topics. Results weakly proved the

[1] http://www.dmoz.org

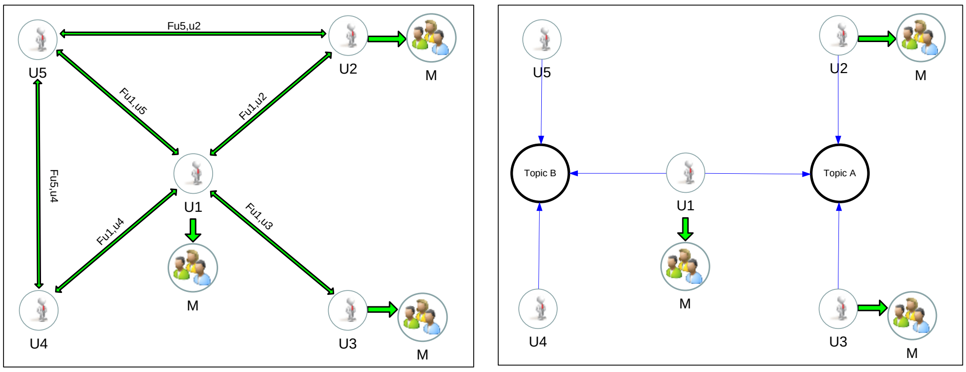
assumption that posts of a user can reflect his interests and expertise, because it maps user's post to wide diverse number of topics. Empirical experiment showed that intersection of group's topics and post's topics can dramatically reduce the number of candidate experts in our semantic implicit social graph. Friendship influence on user's behavior in terms of group membership and posts he published is low.

The rest of this paper is organized as follows: We defined the problem in section 2. We discussed some related work in section 3. In section 4 we described the architecture and dataset used in the evaluation. We introduced our approach and evaluation in section 5. Finally, we conclude and highlight some future work.

2 Problem Definition

Expert finding system designers rely on the analysis of the explicit social graph in OSNs to match the candidate expert to a specific query. User profile contains static data such as interests, group's membership, and friendship of the user. Facebook explicit social graph can be represented as a graph in which there is set of users $U = \{u_1, \dots u_N\}$. Each user u_i is represented by a node. Each user has friends represented by matrix $F = [F_{ui, uj}]_{N \times N}$. Each friendship relation $F_{ui, uj}$ is represented by an edge in the graph. Facebook has set of groups $G = \{g_1, \dots g_M\}$, the user's membership in groups is expressed by the matrix $M = [M_{u, g}]_{N \times M}$. Where $M_{u, g}$ denotes user u is a member in group g, M is associated with each node, as demonstrated in Fig 1(a). Fu et al. in [YR07] proposed expert propagation process, which utilizes user profile information and associations in explicit social graph to discover and propagate the expertise from possible experts to other candidates. We think such model is not enough to reflect the real interests of the user, because neither the profile data can really express user interests, because it is not necessarily contains a correct, a complete and up to date information. Nor the influence of friendship is suitable to assume that I and my friends share the same interests, since $F_{ui, uj}$ is a binary value in Facebook. This kind of assumption leads to inefficient or inaccurate expert recommendation in expert finding systems.

From user's behavior and interaction patterns in Facebook, our assumption is that users interact about topics, either through their posts or through the groups they are members in. This kind of interaction forms a community centered around topics, this community can be represented by implicit social graph as illustrated in Fig 1(b). We can notice that user's explicit social graph can has more than one latent implicit social graphs, in the example showed in Fig 1(b) one implicit social graph centered around Topic A, and the other centered around Topic B. In this research we aim to detect the implicit social graph in an efficient way to enhance the performance of expert finding systems.



(a) Explicit Social Graph, a group membership M attached to a user node U, which connected by friendship edges Fui,uj

(b) Implicit Social Graph, formed of semantic topics and users centered around those topics

Figure 1: Explicit Social Graph and Implicit Social Graph

3 Related Work

In this section we introduce some related work on expert finding in OSNs. Many efforts have been devoted to expert finding in different contexts. User profile has been widely investigated to classify users to different expertise topics based on the static information mentioned in their profiles [DK03]. Furthermore, most of research in this field focused on discovering the candidate expert in organization or enterprise dataset. The Spree[2] expert finding system provides online tool to enable users in organization to search for experts in a certain area, user's queries can be matched with expert's profile which was automatically generated from user's related documents [FC07]. However, the proposed system did not deal with user generated data in OSNs.

ExpertRank algorithm was introduced in [JJ09] to investigate the expertise users displayed in online communities by integrating discussion thread contents and social network extracted from user interaction. In one hand, it used the thread content information to find the most relevant experts to a specific query. On the other hand, it employed the expert network to improve expert finding performance. Although dataset used is very large, it did not represent real online community, and it did not propose a mechanism to build an integrated user's profile that reflects user's expertise.

LDA model proposed in [JS08] studied expert finding in Yahoo Answer[3], the content of user's documents was analyzed to discover the latent topic of interest. The influence of

[2] http://www.askspree.de

[3] http://answers.yahoo.com

social interaction have not been utilized which make it not real online community analysis.

Different ontology engineering approaches and comprehensive knowledge base explored in [MD09] to extract interests and relations between interests. Wikipedia used to find interest definition, latent semantic analysis to find similarities between interests, in another approach Wikipedia category graph used to extract relationships between interests. Different approaches showed good efficiency in building a hierarchy over user interests. But still user's interests extracted from his profile.

Our work is different from all the above in that it used Facebook as the largest real online community, it utilized both the explicit and implicit social graph, it used the user's creatures in the form of post and groups to categorize users into different expertise, more details are introduces in sections 5.1 and 5.2, it introduced a semantic model discussed in detailed in section 5.3 to reduce the number of candidate expertise to answer a specific query.

4 Architecture and Data Set

Recent developments in OSNs such as Facebook Graph API have led to a large incremental in social services [NE10]. Facebook Graph API[4] allows you to easily access all public information about an object in Facebook. Developers can create smarter applications that leverage the social networking aspects of its users. In this paper we utilized Facebook Graph API as application platform to extract the data set and build a prototype. The prototype is a Facebook application integrating Spree expert finding framework with Facebook. Facebook users can interact with Spree using their own credentials.

Five different Facebook user accounts have been used to extract the data set, each user account with an average of 200 friends, average number of groups a user member in was 13, most of them described in English, in average 30 posts was retrieved for each one of the 1000 friends regardless of the text language of the post. Many challenges and limitations was faced during data extraction, such as restrictive Facebook data access policy, limited number of posts content can be retrieved using Graph API, multi lingual nature of the retrieved textual content which make it unable to be classified by Spree classifier, and missed profiles of users who restricted access to their personal data. As the size of our data set is relatively small, we cannot generalize the result we have, because we do not know till now if the sample five users and their 1000 friends we used are representative sample or not.

[4] http://developers.facebook.com/docs/

5 Our Approach and Evaluation

Classical expert finding systems utilize ontologies to classify textual content in user's profiles to predefined topics by matching documents into some ontology entries. Ontology based text classification used widely in social semantic data discovery to categories documents into topics [TS08]. In our research we employed the Spree expert finding system developed at DAI-Labor[5] to match queries to experts.

Spree uses a hierarchal ontology tree of topics where each node has a sub-topic or area of knowledge and the fields of expertise represented by user's profiles is a sub-tree of this ontology. By default Spree uses a Naive Bayes text classifier. For each node, this classifier estimates the likelihood that the corresponding n-gram distribution generated the n-gram sequence observed in the input text. The m most likely nodes are then considered valid classifications where m is a system parameter. Classifications are always complete in the sense that if a category is assigned to a given text also all parent categories are considered valid classifications. Classifications, therefore, always appear as sub-tree of the taxonomy [GA09].

The user query can be mapped to a sub-tree, matching query to the corresponding expert becomes a graph matching problem [FC07]. In our prototype SpreeBook we integrated Spree with Facebook through Facebook Graph API. Then we modified Spree classifier to be able to classify the posts and the group's description of Facebook users into topics. Next we match candidate experts to topics. The fundamental idea of the matching algorithm is to represent experts and user questions as serialized vectors of nodes v(T) ∈ S(T) where S(T) ⊂ RN is the ontology space. The values of v(T) are set to 0 or 1. Once all registered experts and an incoming question have been mapped to sub-trees, the similarity between an expert and the question can be calculated as a weighted dot product [GA09].

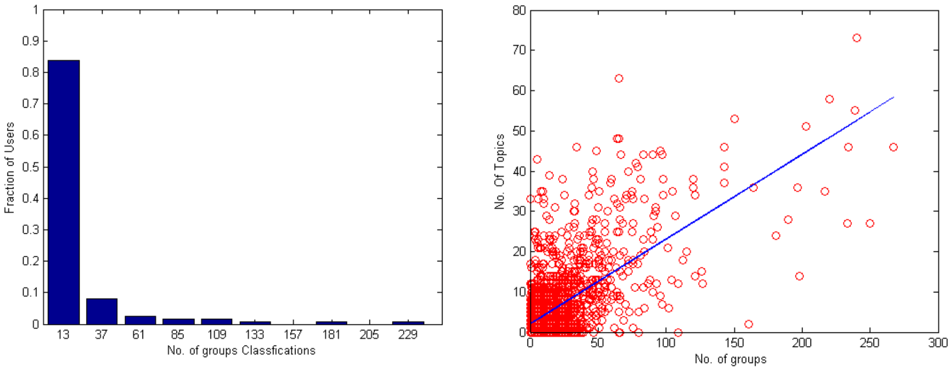
5.1 Matching experts based on user's groups membership

SpreeBook used to classify the description of user's groups to set of topics $GT = \{gt_1, \dots gt_X\}$. The user membership matrix M used to match a query Q to a corresponding user' group topic gt_i then to the candidate user u_i, where ui ∈ U. We extracted the groups of the friends of the five Facebook accounts mentioned previously of almost 13000 groups. Therefore, we classified the description of the groups of every individual user using SpreeBook.

In order to evaluate our approach we raised the first research question "Can the group's membership reflects user's expertise and interests?". To answer this question we used

[5] http://dai-labor.de

the tabulate[6] function to find the frequency table of the number of topics each user classified to. The frequency table used to plot Fig 2(a). Where x axis represents the number of the group's topics, Y axis represents the fraction of users classified to each unique topic. As noticed in Fig 2(a) around 80 percent of users in the sample data set have less than 13 topics and around 10 percent of users have topics between 13 and 37, while the rest 10 percent of users have wide diverse topics. Fig 2(b) shows that the total number of the groups of each user and his number of classified topics increases almost linearly. Another thing to notice is that most of users tend to have relatively small number of groups, generally less than 50 groups, which can be classified to a little number of topics around 10 topics. Based on these observations we can say that group's membership of a user can reflect his interests and expertise, but still it maps user's groups to relatively wide range of topics. A semantic model to reduce the number of matched topics is introduced in section 5.3.



(a) Fraction of users in each group classification topic

(b) Number of groups and number of its classification topics for each user

Figure 2: The Reflectivity of User's Expertise Based on His Group's Membership

5.2 Matching experts based on user's posts

If we trace the online activities of a large set of friends, we are likely to find that much of the information we receive is not useful or interesting because humans are diverse creatures. In other words, we lose precision[7]. Implicit social graph focuses on analyzing features of user's interaction, such as posts, messages, comments, pictures, and opinion expression methods to discover interaction patterns. We will consider the posts feature in this stage of our research for simplicity. In Facebook each user has creatures as a set of posts $P = \{p_1, \dots p_Y\}$. We classified user's posts to set of topics $PT = \{pt_1, \dots pt_Z\}$.

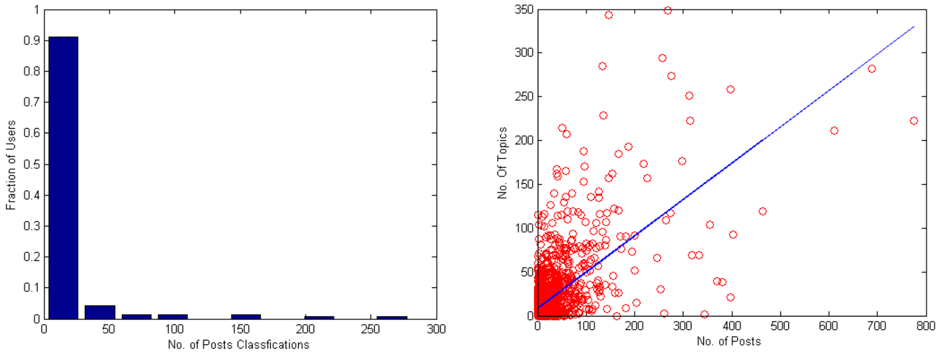
[6] http://www.mathworks.com/help/toolbox/stats/tabulate.html

[7] http://soe.stanford.edu/research/ate/asktheexpert.html

The user's posts set P used to match a query Q to a corresponding user' post topic pt_i then to the candidate user u_i, where ui ∈ U. This raised up our second research question "Can the posts of the user be useful to express his expertise?".

In order to answer this question and evaluate our approach, the posts of the friends of our five Facebook accounts have been extracted and classified using SpreeBook of almost 30000 posts, then we calculated the frequency table of the number of topics each user classified to, based on his posts. Fig 3(a) shows that around 80 percent of users have topics of interest between 1 and 25, the rest users have wider diverse topics. As Fig 3(b) shows that the total number of the post of each user and his number of classified topics act almost the same as his group's classifications in terms of linearity. While the less posts less classification phenomena is noticeable. This can weakly prove the assumption that posts of a user can reflect his interests and expertise, because it maps user's post to wide diverse number of topics as twice of the group's topics. Thereby, we need to reduce the number of topics. In section 5.3 we discuss the details of our proposed model to reduce the number of matched topics.

Further manual analysis of the data set shows that this diversity of post's topics related to the quality of extracted data, since we use the News Feed Channel[8] in Facebook. News Feed Channel has several sources of contents, some of them are automatically generated, which leads to a diverse topics of posts. To enhance the quality of extracted data a smart data crawler to extract and analyze tempo-spatial hotspots in virtual worlds was proposed in [AE10]. Even though, this crawler works for virtual worlds, the same concept can be applied in the case of Facebook, if we considered the user's wall as a spatial hotspot.



(a) Fraction of users in each post classification topic

(b) Number of post and number of its classification topics for each user

Figure 3: The Reflectivity of User's Expertise Based on His Posts

[8] http://developers.facebook.com/docs/channels

5.3 Semantic relatedness between user’s groups and posts

To map a user in Facebook to topics of expertise in efficient way, we need to reduce the number of topics matched by SpreeBook classifier. Even groups based classification performed the posts classification by almost twice, the number of matched topics still big. We need a model to reduce the number of matched topics. We used Cosine similarity measurement to discover the intersection classifications between groups and post as in the following formula:

$$Sim(GT,PT) = \sum_{u=1}^{N} \frac{GT.PT}{|GT||PT|},\ ui \in U\ . \qquad \textbf{(1)}$$

We conducted an experiment to find the intersection between user’s interests based on groups and his interests based on posts. The results of the experiment shown in Fig 4 explain that around 70 percent of users have no similarity between posts and groups, and most of the rest 30 percent of users have weak similarity between posts and groups. Therefore, we can say statistically user’s posts and groups are not necessarily similar. Further analysis of our data set have been made to explain this phenomena, by manual reviewing of user’s posts and groups, we found that most of the user’s posts are short sentences and do not contain enough key words which makes the classifier to retrieve a wide diverse of topics. On the contrary of that, group’s description is professional and long enough for the classifier to work efficiently.

We can infer that, users with similar posts and groups topics are experts in those topics. Fig 4 shows that there are few fraction of users with similarity between posts and groups higher than 0.5. Empirical experiment showed that 0.5 cosine similarity is a suitable threshold that can dramatically reduce the number of candidate experts in our semantic implicit social graph, because it reduce the number of matched topics for each user. Which make each user has a small range of expertise topics.

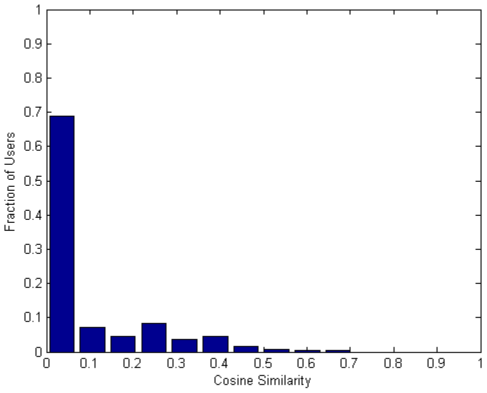


Figure 4: Similarity between group's topics and post's topics for each user in explicit social graph

Expertise in our implicit social graph based on the categorization of the user's groups can be expressed by the matrix $\mathrm{EG} = [\mathrm{E}_{\mathrm{u}, \mathrm{gt}}]_{\mathrm{N} \times \mathrm{X}}$. Where $\mathrm{E}_{\mathrm{u}, \mathrm{gt}}$ denotes expertise of user u in groups topic gt. On the other side expertise based on user's posts categorization represented by this matrix $\mathrm{EP} = [\mathrm{E}_{\mathrm{u}, \mathrm{pt}}]_{\mathrm{N} \times \mathrm{Z}}$. Where $\mathrm{E}_{\mathrm{u}, \mathrm{pt}}$ denotes expertise of user u in posts topic pt. Expertise matrix in implicit social graph expressed by $E = EG \cap EP$. This leads to the illation that implicit social graph can be detected from the intersection topics between groups and posts, which is the answer of our fourth research question.

6 Conclusion and Future Work

This paper presented an approach for expert finding in Facebook based on the classification of the user's groups and posts. In this research paper we analyzed the explicit social graph and the user's interactions in the form of posts and group's membership to discover the implicit social graph (latent semantic social graph). SpreeBook classifier based on general predefined ontology topics used to match experts to topics. Matching experts based on group's membership or posts of the user showed that group's membership or posts of a user can reflect his interests and expertise but still it maps user's groups or posts to relatively wide range of topics. Examining Semantic relatedness between user's groups and posts leads to the illation that implicit social graph can be detected from the intersection topics between groups and posts.

Five Facebook accounts were used to extract the data set. Therefore, our data set is relatively small. Future work is planned to make the prototype available online to be able to extract more data. Utilizing user's comments as well as posts can enhance mapping accuracy. Building a multilingual ontology based classifier can override the problem of multilingualism in Facebook posts, since SpreeBook is based on dmoz project, which is a monolingual based ontology. Future research need to be conducted to formalized a model for expertise propagation in Facebook utilizing the explicit and implicit social graphs.

Bibliography

[BC03] Borgatti, S.P.; Cross, R.: A Relational View of information seeking and learning in social networks, Management Science, vol. 49, no. 4, pp. 432-445, 2003,.

[KN08] Kate, Ehrlich; N. Sadat, Shami: Searching for expertise, in the twenty-sixth annual SIGCHI conference on Human factors in computing systems (CHI '08), pp. 1093-1096, New York, NY, 2008.

[Bs10] B., Schneier: A Taxonomy of Social Networking Data, Security & Privacy, IEEE , vol. 8, no. 4, p. 88, July 2010.

[AB10] Alan, Mislove; Bimal, Viswanath; Krishna ,P. Gummadi; Peter, Druschel: You are who you know: inferring user profiles in online social networks, in Proceedings of the third ACM international conference on Web search and data mining, pp. 251-260, New York, 2010.

[BA09] Bimal, Viswanath; Alan, Mislove; Meeyoung, Cha; Krishna, P. Gummadi: On the evolution of user interaction in Facebook, in Proceedings of the 2nd ACM workshop on Online social networks, pp. 37-42, Barcelona, 2009.

[DK03] D. Seid ; A., Kobs: Expert Finding Systems for Organizations: Problem and Domain Analysis and DEMOIR Approach, Jurnal of Organizational Computing and Electronic Commerce, vol. 13, no. 1, 2003.

[FC07] F. Metze, C. Bauckhage; T. Alpcan; K. Dobbrott; C. Clemens: The "Spree" Expert Finding System, in International Conference on Semantic Computing, 2007. ICSC 2007., p. 551. Irvine, CA, 2007.

[JJ09] Jian, Jiao; Jun, Yan; Haibei, Zhao; Weiguo, Fan: ExpertRank: An Expert User Ranking Algorithm in Online Communities, International Conference on New Trends in Information and Service Science, 2009. NISS '09., p. 674, Beijing, 2009.

[JS08] Jinwen, Guo; Shengliang, Xu; Shenghua, Bao; Yong, Yu: Tapping on the potential of q&a community by recommending answer providers, in Proceeding of the 17th ACM conference on Information and knowledge management, pp. 921-930, New York, 2008.

[MD09] Mandar, Haridas; Doina, Caragea: Exploring Wikipedia and DMoz as Knowledge Bases for Engineering a User Interests Hierarchy for Social Network Applications, in Move to Meaningful Internet Systems, Berlin, 2009.

[NE10] N., Klym; M.J., Montpetit; E., Blain: Building Social Services, in Consumer Communications and Networking Conference (CCNC), 2010 7th IEEE , p. 1, Las Vegas, NV , 2010.

[TS08] Traian, Rebedea; Stefan, Trausan-Matu; Costin-Gabriel, Chiru: Extraction of Socio-semantic Data from Chat Conversations in Collaborative Learning Communities, in Proceedings of the 3rd European conference on Technology Enhanced Learning: Times of Convergence: Technologies Across Learning Contexts, pp. 366-377, Berlin, 2008.

[GA09] Gerald, Eichler; Andreas, Lommatzsch; Thomas, Strecker; Danuta, Ploch; Conny, Strecker; Rober,t Wetzker: From Community towards Enterprise - a taxonomy-based search for experts, in Proceeding of the 9th International Conference on Innovative Internet Community Systems I2CS 2009, Jena, Germany, 2009.

[AE10] Akram. Al-Kouz; Ernesto, William De Luca; Jan, Clausen; Sahin, Albayrak: The Smart-TSH-Finder: Crawling and Analyzing Tempo-Spatial Hotspots in Second Life, in Workshop "Knowledge Discovery, Data Mining, Maschinelles Lernen 2010" der Fachgruppe KDML, Kassel, 2010.

[EK09] E., Gilbert; K., Karahalios: Predicting tie strength with social media, in CHI ’09, 2009, pp. 211–220.

[AE10] Alan, Said; Ernesto, W. De Luca; Sahin Albayrak: How social relationships affect user similarities, in ACM IUI'10 Workshop on Social Recommender Systems, Hong Kong, 2010.