Contact Recommendations from Aggegrated On-Line Activity

Abigail Gertner, Robert Gaimari, Justin Richer, and Thomas Bartee

The MITRE Corporation
202 Burlington Road, Bedford, MA 01730
{gertner,jricher,tbartee}@mitre.org

Abstract. We describe a system for recommending people based on similar interests and activities as part of a company-wide social networking site. Our contact recommendation service aggregates input from multiple on-line data sources and combines them using a Bayesian network to generate a rating of the overall match between two users. The system is running as part of an experimental social networking site at MITRE. We present the results of two experiments in which we evaluated the performance of the recommender algorithm and user interface.

1 Introduction

The use of social networking platforms in enterprise settings, including industry and government, is growing rapidly. One of the primary stated uses for these tools is to help workers connect with each other within their organizations. This is particularly attractive in the case of large corporations and government agencies whose organizational structure as well as geographic separation can make it very difficult to know who you should be talking to.

Modeled on the popular internet social networking sites, such as Facebook and LinkedIn, enterprise social networking tools generally include contact lists, allowing users to connect with the people they know. These tools also often include a suite of social utilities such as blogs, wikis, bookmarks, tags and file sharing. By connecting with people via the contact list, a user can then keep track of their contacts' activities on the site, providing a *social filter* on the available information.

Social networking sites become more useful and more attractive the more connected their users are. One way that sites can make it easier for users to create and build their online networks is by recommending other users to connect to. For example, Facebook and LinkedIn both provide suggestions of people to connect to, primarily based on existing network connections (friend of a friend relationships).

In this paper we present a contact recommendation tool that is designed to help workers in a large distributed enterprise environment make connections with others who share similar interests or work activities. We generate contact recommendations by aggregating information about users from diverse data sources within the company. The contact recommendations are implemented as a stand-alone web service which is designed to be integrated with a social networking front end. We show how the recommendations appear in our company's internal social networking tool and then discuss two experiments we did to evaluate the recommendations.

2 Related Work

The literature on recommender systems has primarily focused on recommending items to users. Much of this work is based on *collaborative filtering* [1] – a technique that clusters people according to their item preferences and then recommends items that other similar users have liked.

There are several research projects that have looked at recommending people to each other. ReferralWeb [2] had the goal of finding existing chains of relationships between people by mining on-line documents such as co-authored papers and organizational charts. The *Do You Know?* system from IBM [4] also attempts to find people who are already known to the user in order to suggest that they be added to their social network. *Do You Know?* is implemented using SONAR [5], a social network aggregation tool that is probably the most similar to our own contact recommendation tool, in that it brings together evidence from multiple data sources to form its recommendations. The primary differences are the way we do the aggregation, and the fact that the IBM tool is attempting to identify existing social relationships, while we are primarily concerned with recommending people who are *not* known but possibly should be.

Terveen and McDonald [3] coined the term "social matching" to refer to systems that try to recommend and connect people each other. They outline the problem space in a series of claims, such as the need for explicit user models and the application to on-line social networks.

Another related area of research is the field of expertise finding [6]. This typically involves keyword searches for an expert who can answer a question or help solve a problem. In contrast, our contact recommender is looking for people who may be appropriate to form a longer term connection with based on common interests, not necessarily based on their expertise on a single topic.

3 Contact Recommendation Implementation

Our contact recommendation service is implemented as part of MITREverse, an experimental social networking system that is deployed inside the MITRE firewall. MITREverse is built on the Elgg open source social networking platform [7], which includes the basic social network features of friend lists, activity streams and message boards, as well as providing additional social tools such as groups, blogs, bookmarks and file sharing. The contact recommendation system will soon be deployed to Handshake, MITRE's external social networking and collaboration site.

The original insight behind the contact recommendation service was that by bringing together information about people from multiple places on the network, we could form a more accurate picture of what their interests are and what activities they are engaging in. As with many large companies, over the past five years or so MITRE has been making a number of social media tools available to its employees on our intranet. These include blogs, wikis, email lists, microblogging and social bookmarking tools. Some of these tools are official corporate supported offerings, and others are grassroots efforts started by individual employees. Many of these tools have built-in APIs that allow other programs to easily access and re-use their data. By aggregating the data from these different services together, the contact recommender creates a multi-dimensional view of what users have in common with each other.

On MITREverse and Handshake we use seven data sources to compute the similarity scores: use of the same tags in onomi, MITRE's social bookmarking site, shared bookmarks in onomi, shared membership in internal email lists, co-editing of pages on the corporate-wide wiki, membership in groups on MITREverse/Handshake, friend of a friend relationships, and use of the same tags on MITREverse/Handshake. All of the data sources are accessed via public (to all MITRE users) APIs. We chose to avoid privacy concerns by basing our recommendations only on data that would be accessible to anyone looking at the recommendations.

The Handshake implementation is a little more complicated because Handshake actually lives outside the MITRE firewall and has both MITRE and non-MITRE members. Recommendations on Handshake use data sources that are both inside the MITRE firewall (onomi, wiki, email lists) as well as outside the firewall on Handshake (groups, friends and tags). We therefore use two servers, one internal and one external, to generate the recommendations and push them out to the Handshake site. Since only MITRE employees can use the inside the firewall services, they are the only ones who will get recommendations based on those data sources.

The contact recommender works by first generating a similarity score between each pair of users for each data source being considered. Since each data source may have a different type of user data, the data sources may use different metrics for computing the similarity scores. For instance, in the case of social bookmarking tags and MITREverse tags we use the cosine similarity of tag frequency vectors to compare two users' collections of tags. In the case of data sources in which there is a simple binary association between users and items, such as mailing list memberships, we use the Jaccard similarity coefficient (the size of the intersection divided by the size of the union) as the measure of similarity between two users. If there is not enough information about a user for one of the data sources, no scores will be computed for that data source for user pairs involving that user.

After the similarity scores are generated for the individual data sources, an overall score is generated for the match between each pair of users based on the combination of those scores. The overall match is represented as a rating from zero to five, which is displayed on the user interface as a set of zero to five stars, as shown in Figure 1. The icons beneath the stars represent the data sources that were used to generate the recommendation, and clicking on the recommendation will take the user to a detailed explanation page for the recommendation.

Fig. 1. The display of a single recommendation

There are several possibile approaches to combining the individual data source scores into an aggregated rating. The most straightforward solution is to use the average score of the data sources, possibly weighted according to which data sources are considered more important. This is the approach taken by the *Do You Know?* system [4]. However, there are often cases where there is a strong match between two users on one or two of the data sources and a weak match on the rest. Using an average over all data sources would cause the overall score in these cases to be low, whereas we believed that people with a strong match in even one area would be likely to benefit from knowing each other. Therefore we decided to use a model that works more like an OR relationship – if *any* of the data sources scores is high, the resulting aggregated score will be high.

The model we are using is a causal probabalistic model called the Noisy-MAX [8], which is used in Bayesian networks to model a multi-valued variable whose value depends on the maximum value of its causal influences. Figure 2 shows a graphical representation of the Noisy-MAX model for aggregating similarity scores. The definition of the Noisy-MAX says that the inferred value of the node representing the outcome variable (in this case, the strength of the match in question, from zero to five) is determined by the maximum value produced independently by that node's inputs. The "noise" built into the probabilistic relationship means that the more inputs there are with a high similarity score, the more likely the overall similarity is to have a high value.

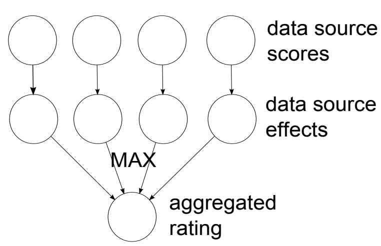

Fig. 2. The Noisy-MAX model for aggregating scores

The Noisy-MAX takes advantage of the fact that the causal influences on the effect node are considered to be independent of each other. In this case, the causal influences are the individual data sources and the effect is the aggregated rating. Using this independence assumption, it is possible to define the relationship between the causes and the effect with far fewer parameters than a full conditional probability table. For each link from a data source to the combined rating, the parameters needed specify the probability that the combined rating will be equal to each possible value (0-5 stars), given the value of the data source and assuming that all other data sources are absent. Currently the parameters for the noisy-MAX are estimated subjectively but we are working to derive values from actual user judgments of matches between users.

We expect that it is the case that some data sources are more influential than others in determining a good match between users. Furthermore, each user may prioritize the various data sources differently. Therefore it is important to be able to weight the inputs in order to adjust their effect on the aggregated rating, as was done in [4] with the weighted average of the input scores. We have modified our implementation of the Noisy-MAX to include a weight parameter for each input data source, so that the influence of the individual data sources on the aggregated score can be adjusted according to the corresponding weights. Since different users may have different priorities for the data sources, we plan to allow them to adjust these weights via the user interface, although this feature is not yet implemented.

The contact recommender runs nightly to update its recommendations. With 1819 users on Handshake, the update takes about twelve minutes. However much of this time is spent downloading the full user data from the external (outside of Handshake) data sources and so that time will not increase as Handshake gains additional users.[1] The noisy-MAX computation to combine the data source scores currently takes about 1 minute and forty seconds. Extrapolating that value to apply it to all possible pairs of MITRE's approximately 7000 employees, the update would take 8.35 hours, so it will still be possible to update the recommendations once a day.

4 Initial Evaluation of Recommendations

As a preliminary evaluation of the accuracy of the contact recommendation ratings we looked at the correspondence between the ratings and the actual friend relationships that currently exist in Handshake. We hypothesized that for pairs of users who are connected to each other in the social network, the recommendation rating should be higher than for pairs of people who are not connected. This is not a perfect measure because, first, there may be people who know each other

[1] Several of the web services we connect to include API calls that allow us to retrieve all user data in a single call. We do this even though many of the users are not current Handshake users because we include non-Handshake users in our recommendations in order to encourage current users to invite their recommended contacts onto the site.

who have not yet connected on the site and, second, people may have friends on the site who they don't have much in common with. However, both of these disadvantages actually make it less likely that a difference would be detected. If we can see a difference in the ratings between the friend pairs and the non-friend pairs, it would give us an initial confirmation that our recommendations are doing the right thing.

At the time of the evaluation, there were 1,819 total user accounts on the Handshake site, making 3,306,942 possible friend relationships (since recommendations are not symmetric, we consider the relationship of A to B separately from the relationship of B to A). Out of these possible friend relationships there are 11082 actual friend relationships in the site. The contact recommender found enough information in at least one data source to generate recommendations for 457,183 of the possible friend relationships. Table 1 summarizes the number of recommendations generated for existing friend and non-friend pairs. Clearly, friends are much more likely to be recommended than non-friends.

Table 1. Total number of recommendations generated for friend and non-friend pairs

	Friends	Not Friends
Recommendation	10485	446698
No Recommendation	597	2849162

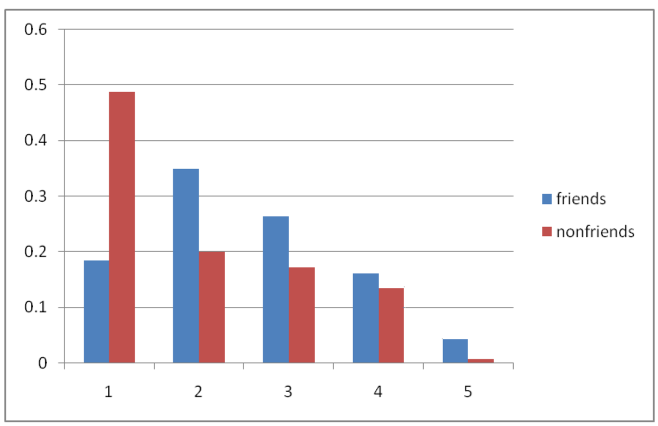

Fig. 3. Percentage of friend and non-friend user pairs assigned each rating, from one to five stars, using the noisy-MAX function

We then compared the ratings (number of stars) that were generated for the friend pairs with the ratings that were generated for the non-friend pairs. The mean recommendation rating (taken from the 1-5 star ratings) for the friend relationships is 2.52, while the mean for the non-friends is 1.97. While this is a small difference, it is encouraging given our caveats about the friend relationships not being a completely reliable correlate of match strength. Figure 3 shows the percentage of the friend and non-friend pairs that were given each of the five possible ratings by the contact recommender. The non-friend pairs are 2.6 times as likely to be given a rating of one star as compared to the friend pairs. Conversely, the friend pairs are 1.6 times as likely to get a rating of two or higher and 5.9 times as likely as non-friends to get a five star rating.

5 Human Subject Evaluations

Two evaluation experiments were run with volunteers from the MITRE employee population. The goal of the first experiment was to validate the similarity metrics used by the contact recommender by seeing what metrics people would use on a similar task. The second experiment looked more directly at the effectiveness of the recommendations generated by the system.

5.1 Experiment 1: Evaluation of Similarity Metrics

There are several assumptions built in to the way the contact recommendation system compares two users, so we wanted to test whether these assumptions would be intuitive to people using the tool. First, for each data source a metric is needed to assess how similar two users are on the basis of that data source. Second, an algorithm is needed for the aggregation of the data source similarity scores into an overall similarity rating. As discussed above, we chose to use the noisy-MAX function to aggregate the scores, but there are other possible options, such as a weighted average of the scores.

Our goal for this evaluation was to present human judges with a sequence of pairs of synthetic "users," with information about those users' shared data for each data source, and ask them to judge the strength of the recommendation for one user to the other for each individual data source and for the overall recommendation. To make this task tractable we chose to limit the data sources to binary data only. For example, we included information about the number of shared tags, but not the frequency that each tag was used by each user. For the purposes of the experiment, we implemented an application that would generate sequences of user pairs based on the actual usage statistics for the data sources we were using, so that the numbers presented to the subjects would be realistic. We also implemented a web application to display the user pair information to the subjects and collect their responses.

Method Eleven subjects participated in the experiment and rated an average of 43 user pairs in half an hour. For each data source and each pair of synthetic users, the subjects were given the number of items each user had in their profile, and how many they shared in common. For example: "user 1 belongs to 10 mailing lists, user 2 belongs to 13 mailing lists, and they have 4 mailing lists in common." Subjects were asked to rate the strength of the recommendation based on the amount of overlap in each of the five data sources independently, and then to rate the overall combined rating of the recommendation. Each user pair was presented in terms of the first user receiving the recommendation and the second user being recommended, so the ratings did not necessarily have to be symmetrical. Finally, the subjects were asked to rank each data source in terms of its importance to the recommendation task.

Each subject also answered a short questionnaire about how they rated the user pairs. The survey questions fell into three categories: open-ended questions about their rating strategies, questions about their similarity ratings for the individual data sources (5 point likert scale), and questions about how they computed the combined similarity score (5 point likert scale).

Results There were two main results from the questionnaire. First, there was no general consensus about what strategy to use. Each subject came up with their own idiosyncratic strategy for rating the user pairs. Second, all of them agreed that some data sources were much more important than others in rating the recommendations. However, they did not agree at all on which data sources these were. Some of them favored the more content-based data sources, such as the wiki and social bookmarks, while other favored the social network (friend of a friend). For the questions about what factors were most important in comparing the two users, the results are shown in Table 2. All of the factors we asked them about were judged to be fairly important, but the percentage of overlap and the data source got the highest ranking.

Table 2. How did these factors rate in your decision for each pair? (1=not at all, 5=very high)

common count column entry	percentage overlap	cardinality of column entry	data source
4.18 (s.d. 0.98)	4.72 (s.d. 0.46)	3.9 (s.d. 0.83)	4.36 (s.d. 0.81)

There was less agreement about the strategy for combining the scores to produce the final rating for each pair. Some subjects strongly prefered the strategy of looking at the data source with the highest (maximum) similarity score. Others prefered to take the average of all the data sources. Again, the strongest agreement was on the importance of individual data sources (but not which ones).

In addition to the questionnaires, we also looked at the correlation of the actual ratings produced by the subjects with several alternative similarity metrics. For the pairwise data source ratings, we compared the subjects ratings to four different metrics as follows. Let $|A|$ be the number of items in user A's profile for a given data source, $|B|$ be the number of items in user B's profile for the same data source and $|C|$ be the number of items they have in common for that

Table 3. How did these factors rate in your decision for the final *combined* recommendation strength for each pair? (1=not at all, 5=very high)

maximum of individual ratings	minimum of individual ratings	average of individual ratings	importance of each data source
2.9 (s.d. 1.45)	2.27 (s.d. 1.19)	3.37 (s.d. 1.10)	4.18 (s.d. 1.40)

data source. The metrics we considered were: $\frac{|C|}{min(|A|,|B|)}$ (%min), $\frac{|C|}{max(|A|,|B|)}$ (%max), $\frac{|C|}{|A|}$ (%user1 - the person being recommended to), and $\frac{|C|}{(|A|+|B|)}$ (also known as the Jaccard similarity coefficient).

The %max and Jaccard coefficient are very highly correlated with each other so their correlation coefficient with the user ratings is very similar (Pearson correlation coefficient of 0.69 and .67 respectively). The correlation of the user ratings with the other two metrics was lower: 0.56 for the %user1 metric and 0.54 for %min. This implies that people took both users into account when determining their ratings, and did not consider the number of items held by the user being recommended for as more important.

We then looked at the correlation of the overall combined ratings with three possible strategies for combining the ratings: the average value of all ratings, the maximum value of the ratings, and the minimum value of the ratings. We also looked at the correlation of the combined ratings with the ratings of each data source, to see if any data sources seemed to be more important overall in determining the combined rating. The correlation coefficients are shown in Table 4. Average and max were the most highly correlated, with max being somewhat higher (0.77 vs 0.68). In fact, there was one subject who always chose the max as the value of their combined rating. The minimum value was not correlated at all with the overall score.

For the individual data sources, none of the correlations was extremely high, but onomi tags and onomi URLs (social bookmarks) were the highest, with friends of a friend being close to those two. It seems that the combination of the data sources was in fact more important than any individual source, at least in the aggregate, which agrees with our observation from the questionnaires that the subjects did not agree on which data sources were the most important.

Table 4. Pearson correlation of subjects' overall rating of user pairs' similarity with candidate metrics for computing recommendation rating

average	max	min	onomi tags	onomi URLs	mailing lists	wiki	friend of friend
0.68	0.77	0.18	0.50	0.45	0.33	0.29	0.41

5.2 Experiment 2: Evaluation of Recommendations in Handshake

For our second experiment, the goal was to assess the recommendation system in the context that the recommendations would actually be delivered. We were assessing both the usability of the interface as well as the quality of the recommendations. We wanted to get feedback from people who were likely to use the system, so we recruited people who were active Handshake users, based on their number of connections on the site, but also had a reasonable number of strong recommendations available for them to look at.

Method Eleven subjects participated in this experiment. We began with a questionnaire to determine the subjects' overall familiarity and usage of social networking tools. All of them were active users of social media, particularly Facebook and LinkedIn. They admitted to using these tools anywhere from "continuously" to "weekly". Only two said they would consider themselves a "power user." All of them used social networks to interact both with friends and family and with colleagues. All but three had used recommendations on Facebook or LinkedIn to find connections. Their usage patterns on Handshake were somewhat less frequent, although most of them said that they check in with Handshake on a weekly basis. Two of them claimed to be Handshake power users.

In the second part of this experiment, we asked each subject to look at their top ten recommendations in Handshake. Because one of the goals of the experiment was to test the intuitiveness of the user interface, we gave them very little introduction to the tool itself. We asked them to look at it and tell us what they thought it was for and how they should use it. We then asked them to answer a series of questions about each of their top ten recommendations. Finally we asked the subjects to answer a short post-experiment questionnaire about their impressions of the contact recommendation tool.

Results All of the subjects were easily able to understand what the recommendations were for and to understand the basic functionality of the user interface. The responses to the 110 recommendations (10 recommendations each for 11 subjects) is summarized in Table 5. We are encouraged to see that most of the recommendations are for people that the subjects did not know and often had not heard of before, and yet they said they would be interested in getting to know that person about half the time. Subjects were more reluctant to say that they would connect (the Handshake version of "friend") the person right away, but as the contact recommender is intended to make people aware of others with similar interests, we would not expect them to take the step of actually connecting without further interaction beforehand. Many of them said that they would need more information before connecting and stated that it would be extremely important for the users being recommended to have a completed profile on the Handshake site so that they could find out more about that person before connecting to them.

In addition to the information presented in the table, we also asked them to rate each recommendation on a scale of 1-5, for which their average rating

Table 5. Subject responses to recommendations on Handshake

	Do you know this person?	Have you heard of them?	Would you like to meet them?	Would you connect to them on Handshake?
yes	33	18	48	18
no	75	55	46	26
maybe	2	1	16	1
need more information				65

was 2.79 (standard deviation=1.4). There was a fairly strong divergence between subjects in how they rated their recommendations. Seven of the eleven gave an average rating of three or higher to their recommendations, while three of them had an average rating below two. More investigation is needed to determine what would be needed to improve the recommendations for those users.

Finally, we asked them why they thought the person was being recommended to them. In the vast majority of cases (65 out of 110, they said that it was because they had shared topics of interest in common). Two recommendations were explained by shared project work, four by shared organizational affiliation, and two by shared sponsor relationships. We also asked if these reasons were still relevant today, and for 81 out of 110 they were, 21 of them were not, and 8 were unsure.

On the post-experiment questionnaire, we looked at the usability of the contact recommendation tool and the usefulness of individual features of the tool. Tables˜reftable:postques-usability and ˜reftable:postques-features show the results from that questionnaire. In general, the ratings were quite high for usefulness, usability and intuitiveness. Accuracy was rated a bit lower, and with more variability. Of the individual interface features, the only one that was not rated over four for usefulness were the tabs, which can be used to filter the recommendations down to only those that involve a particular data source. Since we didn't ask the subjects to perform a task where this would be important, it is not surprising that they did not rate that feature as highly.

Table 6. Assessments of recommendation tool general usability (mean and standard deviation)

Usefulness	Ease of Use	Intuitiveness	Accuracy
4.27 (0.86)	4.55 (0.50)	4.18 (0.57)	3.45 (1.16)

Some of the subjects were confused by the ordering of the recommendations. Since the similarity for the data sources is computed as a percentage of both users' total number of items, it sometimes is the case that a match with a smaller number of items will be ranked higher than a match with a larger number,

Table 7. Assessments of recommendation tool features (mean and standard deviation)

Ordering	Stars	Icons	Details	Tabs	Adding	Profiles
4.55 (0.99)	4.36 (0.48)	4.45 (0.78)	4.64 (0.88)	3.73 (1.05)	4.64 (0.64)	4.82 (0.38)

because the person who is being recommended has a smaller total number of items, making the percentage value higher. This would argue for an asymmetric way of rating the recommendations (considering just the percentage of the person being recommended for and not the person being recommended), although this strategy was not used by any of our subjects in the first experiment.

6 Discussion and Future Work

We have described a contact recommendation tool that looks at data available about users based on their on-line activities and uses that information to generate recommendations for other people with similar interests. This capability will soon be deployed company wide on our social networking site called Handshake.

We presented the results of two experiments. The first looked at the similarity metrics used to compare users in order to generate the recommendations, and the second was aimed at evaluating the actual recommendations generated by the system as well as the user interface for presenting them. The assessment of similarity metrics found that there is a great deal of disparity between people as to how they measure the similarity or relevance of others. In general, each person has their own priorities for the different data sources, so we believe that it is important to give users control over this aspect of the recommendations. We are doing this already in making available the tabs for filtering recommendations by data source, and we would like to also make it possible for users to define the weights that are used by the algorithm itself. For combining the recommendations, the most common strategy was the max, although it was closely followed by the average. We feel that this is enough support to continue using the Noisy-MAX model for combining data source scores.

The subjects in our second experiment in general were very enthusiastic about the potential of the contact recommendation tool to help them find people on Handshake, something that many of them felt was quite difficult to do at present. This experiment also pointed out the importance of having information available on-line, such as rich user profiles and connections to other user-created data, in order to evaluate the relevance of a potential connection. Furthermore several subjects suggested that looking at the recommendation tool would motivate them to update their own profiles so that the recommendations they were receiving would be more accurate and relevant to them.

References

1. Herlocker, J. L., Konstan, J. A., Borchers, A., Riedl, J.: An Algorithmic Framework for Performing Collaborative Filtering. In: Proceedings of Research and Develop-

ment in Information Retrieval, ACM, New York (1999)
2. Kautz, H., Selman, B., Shah, M.: Referral Web: combining social networks and collaborative filtering, Communications of the ACM 40(3), pp 63–65 (1997)
3. Terveen, L. and McDonald, D. W.: Social Matching: A Framework and Research Agenda. ACM Transactions on Computer-Human Interaction, 12(3), pp. 401–434, (2007)
4. Guy, I., Ronen, I., Wilcox, E.: Do You Know? Recommending People to Invite into Your Social Network, In: Proceedings of the 13th international conference on Intelligent user interfaces, pp. 77–86. ACM, New York (2009)
5. Guy, I., Jacovi, M., Shahar, E., Meshulam, N., Soroka, V., Farrell, S.: Harvesting with SONAR: the value of aggregating social network information. In: Proceedings of the twenty-sixth annual SIGCHI conference on human factors in computing systems (CHI '08), pp. 1017–1026. ACM, New York (2008)
6. Maybury, M., D'Amore, R., House, D.: Automating Expert Finding. International Journal of Technology Research Management. 43(6): 12-15 (2000)
7. Elgg open source social networking platform, `http://www.elgg.org/`
8. Díez, F., Druzdzel, M.: Canonical probabilistic models for knowledge engineering, Technical Report CISIAD-06-01. UNED, Madrid, (2006)