Independent Study Report

Shuguang Han, shh69@pitt.edu

School of Information Science, University of Pittsburgh

Supervisor: Peter Brusilovsky

Part of this report has been published in SBP 2013 (Social Computing, Behavioral-Cultural Modeling and Prediction):

Shuguang Han, Daqing He, Peter Brusilovsky, and Zhen Yue. 2013. Coauthor prediction for junior researchers. In Proceedings of the 6th international conference on Social Computing, Behavioral-Cultural Modeling and Prediction (SBP'13), Ariel M. Greenberg, William G. Kennedy, and Nathan D. Bos (Eds.). Springer-Verlag, Berlin, Heidelberg, 274-283.

Identifying Potential Coauthors for Junior Researchers

Abstract. Research collaboration can bring in different perspectives and generate more productive results. However, finding an appropriate collaborator can be difficult due to the unawareness of implicit social connections. Link prediction is a related technique for collaborator discovery; but its focus has been mostly on the core authors who have relatively more publications. We believe that the fresh researchers are actually more willing to have help in identifying potential collaborators. In this paper, we focus on the coauthor prediction problem for junior researchers. Previous works on coauthor prediction found that the local network features can outperform global network features because of less noise information. This may not be true for junior researchers because the data sparseness of our targeted users. In our experiments, we found a significant improvement by simply combing local network feature and global network feature, comparing to use them separately. Besides network features, the content-based features were also considered in information retrieval community for better locating people with certain expertise, which can also be used in our case for coauthor finding. To integrate multiple resources, we further developed a regularization based approach, in which we found it outperforms the simple linear combination method. This method was then adopted in PeopleExplorer, a people finding system that leverages content relevance, local network features and global network features.

Keywords: Coauthor prediction; link prediction; social network; expert search

1 Introduction

Identifying and maintaining potential collaboration relations are critical in a researcher's academic life [18] because collaboration can bring together diverse expertise to the same research problem and generate more influential results. The link prediction techniques developed in social network research community [12] can help predict future collaboration and make researchers aware of the possible coauthors. However, most of the research works considered only the core authors [12,20] who have at least a certain number of publications both in the training dataset and the testing dataset (three in [12], and five in [20]). Considering the skewed distribution between the number of authors and the number of publications [13], the selection criteria will cut off a large proportion of authors. The conclusion from the core authors may not be useful for the rest authors, because predicting from sparse data is more difficult [15]. Besides, the prediction in current works is in the global level, in which top-k ranked pairs among the entire candidate pairs are selected as the predicted links (k is the number of links in the testing dataset). In the global level prediction, there is no control of generating prediction for particular individuals; however, we believe that it is it is more useful if the prediction is for individuals, especially for junior authors. They usually don't have sufficient coauthors, and are more eager to form new connections.

Data sparseness is recognized as a major problem for the prediction of coauthors for junior researchers. To relieve data sparseness, content information was commonly used. Related techniques in expert search [1] utilize content information to find relevant experts. In the recommender system domain, a hybrid method combining both content information and social network feature was often used to solve the cold start problem [14,11,19]. In previous research that considered social network information, either the local network features (e.g. the direct connections) or global network features (e.g. shortest path) were used respectively. Our experiments showed that combing the local and global network features significantly improve the prediction performance, no matter content information was added or not.

Since multiple features are used in this task, a following question is how to combine them effectively. Linear combination is a simple solution, but it is difficult to scale different scores and tune the parameters. An alternative method is to treat the prediction as a binary classification problem [16,20,3] based on multiple features. However, to train a binary classifier, we need to use both positive subjects (real coauthor pairs) and negative subjects (real non-coauthor pairs). Negative subject sampling is difficult because not observing a coauthor link does not imply two authors not are real non-coauthor pair. It may because the coverage of the dataset is limited.

To sum up, the focus of this paper is to predict coauthors for junior researchers. Multiple features including local network features, global network features and content features are considered to improve the prediction performance. In the remainder sections of this paper, we first review related work in section 2. Then, in section 3, a new approach to combine multiple evidences using regularization framework is proposed. Then, we described the datasets and evaluation metrics in section 4. In addition, Empirical results analysis is discussed in section 5 and then we summarize our findings and propose future directions. In section 6, we adopted the new regulariza-

tion based method in a working system named PeopleExplorer for better supporting finding potential collaborators.

2 Related Works

In the literature, coauthor prediction has been modeled as a similarity measuring problem, a recommendation problem, or a classification problem. When viewed as a similarity measuring problem, the similarities between any two authors are calculated, and then the author pairs are ranked and those in top positions are chosen as the predicted links [12]. The core of this approach is to define the vertex similarity [5]. Authors with high vertex similarity are assumed to have high probabilities of collaboration. Network topological features are usually used to measure the vertex similarity. Both local network measures such as common neighbor, Jaccard similarity, Adamic/Adar, preferential attachment, and global network measures such as the shortest path, simRank, and Katz index have been used before. All of these measures are mentioned and compared in [12].

Coauthor prediction can also be viewed as a personalized recommendation problem. The Collaborative Filtering (CF) method was extended in [19] for people-to-people recommendation; however, CF suffers from the cold-start problem when data is sparse. This problem is particularly important in our task because the junior researchers are usually lacking of coauthor information. A hybrid method that combines both social network information and content information can be adopted to relieve the data sparseness. The combination can be a simple linear combination [11,4], a regularization based combination [14], or a filtering based combination [17].

Other researchers [3,20,16] found that besides local and global network topological features, other features can also help improve the prediction performance. For example, the authors' keywords matching, the publication classification code matching [3,16,11] and the meta-path in heterogeneous information networks [20] were all found useful. In order to combine multiple features, the coauthor prediction was modeled as a binary classification problem.

The expert search in the information retrieval domain is also a related work. Related techniques of expert search were not well-studied until TREC's expert finding task [6], in which researchers are required to build an algorithm and rank candidates based on their relevance to the user issued queries. The widely adopted method for expert search is to construct expert profiles using the their previous publications or co-occurrence texts [1]. Expert search didn't model users' social context, which make it less useful than social network based method [11]. However, combining the expert profiles and social context information performs better than using them separately. However, as Terveen and McDonald [21] pointed out, it fundamentally changes the nature of the problem when the returned results are people rather than documents. People are social creatures; "assessing" people is significantly more complex than assessing web documents. As a result, they proposed the concept of "social matching" to emphasize the "social" dimension. "Social matching" systems such as Referral

Web [22] and Expertise Recommender [23] are able to return highly relevant candidates who are also socially related to the information seekers.

3 Methodology

3.1 Problem Definition

The prediction task is formalized as follows: we divide the dataset into the training dataset $\mathbb{D}$ and the testing dataset $\mathbb{D}'$. The division criteria are described in section 4. The test documents $\mathbf{D}' \subseteq \mathbb{D}'$ are defined as those documents with junior researchers as the first authors. Each document d in $\mathbf{D}'$ is further presented by a triple: $< u_1, \mathbf{u} - u_1, \mathbf{m} >$, which indicates the authors of d: u_1 is the first author (u_1 is a junior researcher), $\mathbf{u}$ represents all the authors of the document and $\mathbf{m}$ represents the metadata such as title and/or abstract. The junior researchers are defined as those people who published at least one first-author paper in $\mathbb{D}'$, and at least one but no more than five papers in $\mathbb{D}$. Our goal is to predict the collaborations between u_1 and the rest of the authors $\mathbf{u} - u_1$. However, if u_1 and any author in $\mathbf{u} - u_1$ are coauthors in $\mathbb{D}$, then that coauthor link is not included in our prediction because we are predicting the new coauthor links. $\mathbf{m}$ is used to simulate u_1's topic interest in document d, and we assume that u_1 has already known this information before he/she wants to build connections with authors $\mathbf{u} - u_1$.

3.2 Baseline Models

We adopted two link prediction measures as the baselines, i.e. the Adamic/Adar index, and the Katz index to represent the best practices using the local network topology features and the global network topology features. We also adopted the Balog Model 2, which is served as the best practice in content-based method. Besides, we considered the standard Collaborative Filtering algorithm which has been found as an effective method in recommendation systems. We adopt the similarity measuring approach for link prediction, the core of which is to rank candidate ca based on his/her similarity with author u_1.

The **Adamic/Adar** index [10] (**AA**) is a typical local network feature based method. In our task, we compute the similarity between candidate ca and u_1, i.e. $s(ca, u_1)$, using Formula (1). $\Gamma(z)$ denotes a set of neighbors of author z, and $|\Gamma(z)|$ denotes the size of $\Gamma(z)$.

$$S(ca, u_1) = \sum_{z \in \Gamma(ca) \cap \Gamma(u_1)} \frac{1}{\log |\Gamma(z)|} \quad (1)$$

The **Katz** [9] (**Katz**) index takes into account of the global network structure. It is defined as the summarization of all paths between candidate ca and u_1, which is

computed using Formula (2). $\text{Path}^{l}_{ca,\text{u}_1}$ is all the length l path between u_1 and ca. β is the damping factor that controls the weight of the path.

$$S(ca, \text{u}_1) = \sum_{l=1\ldots\infty} \beta^l \cdot |\text{Path}^{l}_{ca,\text{u}_1}| \quad (2)$$

In the content-based baseline model Balog Model 2 (ES) [1] the content similarity is calculated between the topic interest of ca and that of u_1 in paper d using Formula (3). The topic interest is represented by the bag-of-words in $\boldsymbol{m}$ and it is used to mimic user query in **ES**. $p(\boldsymbol{m}|d)$ is estimated using the standard language modeling approach in information retrieval, and $p(ca|d)$ is the association between author ca and document d. In this paper, we used the uniform association for multi-authored papers, and each author receives the same weight of association regardless of author order.

$$S(ca, \text{u}_1) = p(\boldsymbol{m}|ca, \text{u}_1) \propto \sum_d p(\boldsymbol{m}|d)p(ca|d) \quad (3)$$

The fourth baseline is the user-based Collaborative Filtering (CF) algorithm [2]. The traditional scenario of CF consists of users, items and users' ratings on items. However, in the case of people-to-people recommendation, the user and item are both people and there are no explicit ratings on items. In order to apply the CF into the coauthor prediction, we treat people as both the user and the item, and the number of papers two people coauthored as the people's rating on each other, i.e. users' ratings on items. Using the simple average weighted aggregation, the similarity between u_1 and ca is calculated using Formula (4), in which $\boldsymbol{C_k}$ is k most nearest neighbors of u_1. $\text{r}_{u',ca}$ is u''s ($u' \in \boldsymbol{C_k}$) rating on ca, i.e. the number of coauthored papers of u' and ca. $w(u_1, u')$ measures the similarity of rating on items between user u_1 and u', which is calculated by the cosine similarity of their coauthors (see Formula (5).). κ is the normalized term.

$$S(ca, \text{u}_1) = \kappa \sum_{u' \in \boldsymbol{C_k}} w(u_1, u')\text{r}_{u',ca} \quad (4)$$

$$w(u_1, u') = cosine(\Gamma(u_1), \Gamma(u')) \quad (5)$$

3.3 Multiple Objective Optimization using Regularization

Each of the baseline models only considered one type of feature. Since the combination of multiple features has proven to be useful in many works [11,4,20,3,16] , a following problem is to combine multiple features more effectively. The simple linear combination works only when features in the combination are independent. As mentioned in [12], when β is small, Katz is very similar to the neighborhood based approach such as AA, which means these two features are not independent to each other. Therefore, here we propose to use a regularization based approach as suggested in paper [7].

Our first regularization based combination approach is named as **AAN**, in which local network feature based method AA is set as the base, and the objective is to combine features from global networks and/or content information. For each document d in $\mathbb{D}'$, we need to rank ca for u_1 based on their similarity score vector $\boldsymbol{S}$. $\boldsymbol{S}$ is initialized as a zero vector. $\boldsymbol{S}$ is updated according to an objective function Ω_1 defined in formula (6), in which $\boldsymbol{S}^*$ denotes the final score vector, $\boldsymbol{I} - \mathbf{M}$ (M is the adjacent matrix of coauthor networks) is the difference matrix, and $\|\cdot\|$ denotes the L2 norm of a vector. $\boldsymbol{S}^{*\boldsymbol{T}}(\boldsymbol{I} - \mathbf{M})\boldsymbol{S}^*$ helps propagate local similarity scores through the global network while$\|\boldsymbol{S}^* - \boldsymbol{S_{AA}}\|^2$ ensures the final score $\boldsymbol{S}^*$do not go far away from $\boldsymbol{S_{AA}}$, and μ_a is the importance parameter. To minimize the objective function, we set derivation of Ω_1 to $\boldsymbol{S}^*$ equals to 0, and the closed-form solution is shown in Formula (7). However, solving the inverse of a matrix is time consuming. An alternative method is to use the power iteration method as suggested in [7]. In each iteration, we can update the score $\boldsymbol{S}^*_{\boldsymbol{AAN}}(\mathrm{t})$ using Formula (8) and the final solution for the iteration is $\boldsymbol{S}^*_{\boldsymbol{AAN}}(\mathrm{t}) = \boldsymbol{S}^*_{\boldsymbol{AAN}}(\infty)$.

$$\Omega_1 = \boldsymbol{S}^{*\boldsymbol{T}}(\boldsymbol{I} - \mathbf{M})\boldsymbol{S}^* + \mu_a\|\boldsymbol{S}^* - \boldsymbol{S_{AA}}\|^2, \mu_a > 0 \tag{6}$$

$$\boldsymbol{S}^*_{\boldsymbol{AAN}} = (1 - \alpha)(1 - \alpha\mathbf{M})^{-1}\boldsymbol{S_{AA}}, \alpha = 1/(1 + \mu_a) \tag{7}$$

$$\boldsymbol{S}^*_{\boldsymbol{AAN}}(\mathrm{t} + 1) = \alpha\mathbf{M}\boldsymbol{S}^*_{\boldsymbol{AAN}}(\mathrm{t}) + (1 - \alpha)\boldsymbol{S_{AA}} \tag{8}$$

For the comparison purpose, we also proposed a linear combination model AANL that combines both local and global network feature. We computed two different similarity scores: the Adamic/Adar score $S_{AA}(ca, \mathrm{u}_1)$ and the Katz score $S_{Katz}(ca, \mathrm{u}_1)$. Then, the two scores are combined using Formula (9), in which λ indicates the importance of Katz score.

$$S_{AANL}(ca, \mathrm{u}_1) = (1 - \lambda)S_{AA}(ca, \mathrm{u}_1) + \lambda S_{Katz}(ca, \mathrm{u}_1) \tag{9}$$

In order to introduce the second regularization based combination approach **AANE**, we first define a simple linear combination model AAE (shown in Formula 11) which incorporate content information with AA. **AANE** then incorporate both content and global network information with AA. The objective function Ω_2 in AANE is defined in Formula (11). The closed solution of Formula (11) is Formula (12). The power iteration method can also be used for **AANE** to optimize the objective function.

$$\boldsymbol{S}^*_{\boldsymbol{AAE}} = \gamma\boldsymbol{S}_{\mathbf{ES}} + (1 - \gamma)\boldsymbol{S_{AA}}, \gamma = 1/(1 + \mu_b) \tag{10}$$

$$\Omega_2 = \boldsymbol{S}^{*\boldsymbol{T}}(\boldsymbol{I} - \mathbf{M})\boldsymbol{S}^* + \mu'_a\|\boldsymbol{S}^* - \boldsymbol{S}_{\mathbf{ES}}\|^2 + \mu'_b\|\boldsymbol{S}^* - \boldsymbol{S}_{\boldsymbol{AA}}\|^2, \mu'_a, \mu'_b > 0 \quad (11)$$

$$\boldsymbol{S}^*_{\boldsymbol{AANE}} = (\boldsymbol{I} - \alpha'\mathbf{M})^{-1}(\gamma'\boldsymbol{S}_{\mathbf{ES}} + (1 - \alpha' - \gamma')\boldsymbol{S}_{\boldsymbol{AA}})$$

$$\text{Where, } \alpha' = 1/(1 + \mu'_a + \mu'_b), \gamma' = \mu'_a/(1 + \mu'_a + \mu'_b) \quad (12)$$

4 Dataset and Evaluation Design

The dataset used in this study contains 151,165 ACM hosted conference papers that were published between 2000 and 2011 in the ACM Digital Library. Each paper in the dataset includes a title and an abstract. The authors of these papers were disambiguated using the ACM author identifiers (In the ACM Digital Library, each author is assigned a unique identifier number). In total, there are 209,592 unique authors. Coauthor relations are extracted to create a coauthor network. A link between two authors is added if they co-published at least one paper.

The dataset is divided into three parts according to publishing time for evaluation: T1= [t_{2000}, t_{2003}], T2= [t_{2004}, t_{2007}] and T3= [t_{2008}, t_{2011}]. There are 3,760 papers in T2, and 5,914 papers in T3 that have junior researchers as the first author. These papers were selected for evaluation. T2 is the testing set when using T1 as the training set, while T2 is the training set when using T3 as the testing set. Therefore, as the two dataset used for evaluation are named asT1-T2 and T2-T3.

Two evaluation metrics were used. The first metric is the accuracy in top-10 positions **(WTP),** which examines whether the correct coauthor is ranked within the top-10 positions. However, the exact ranking position information is lost in this case. If two algorithms both can recommend results in top-10 positions, we cannot distinguish their performance using WTP. Therefore, another evaluation metric mean reciprocal rank (**MRR**) [22] was also used as it reflects the exact ranking position.

5 Result Analysis and Discussion

5.1 Parameter Selection

AA and ES were implemented directly as there are no explicit parameters in these two algorithms need to be tuned. For other algorithms, parameters were tuned and the one with best performance were selected. When the performances on WTP and MRR have conflictions, the parameter that has better performance on WTP was chosen.

For the user-based Collaborative Filtering (CF) algorithm, as shown in Formula (4), we tried different values of k (1, 3, 5, 7 and 9), and finally chose 5 because it has the best performance in terms of both MRR and WTP. This means that the 5 nearest neighbors were selected as the similar users. In the Katz index method, we follow the Gauss-Southwell algorithm [8]. A set of damping factors (i.e. the β) values are adopted and compared, including 0.1, 0.05, 0.005, 0.0005. Finally, $\beta = 0.05$ were selected as it is the one with best performance on both MRR and WTP. In AANL, λ is set to be 0.95 because it gives the best performance on both WTP and MRR.

For the rest three models AAN, AAE and AANE, both of the parameters α and γ are ranging from [0,1]. We set different values for the parameters from 0 to 1, with 0.1 as gradient step and chose the one with best performance ones: $\alpha = 0.5$ for T1-T2, and $\alpha = 0.1$ for T2-T3 in AAN; $\gamma = 0.1$ in T1-T2, $\gamma = 0.02$ is in T2-T3 for AAE,$\alpha' = 0.5$, $\gamma' = 0.01$ in T1-T2 and $\alpha' = 0.1$, $\gamma' = 0.002$ in T2-T3 for AANE. In AAE, the ES scores are usually small; therefore, we use a heuristic method to multiple them by 1000 in order to be able to combine with other scores.

5.2 Comparative Evaluation of Eight Models

The result analysis on each metric consists of two parts: a bar chart on how each model performed and a statistical test to reveal the significance of experimental results. Non-parametric test Wilcoxon Signed Ranks was used since the normality was not satisfied. The following results show the comparisons of eight models: AA (Formula 1), Katz (Formula 2), ES (Formula 3), CF (Formula 4), AAE (Formula 10), AANL (Formula 9), AAN (Formula 8) and AANE (Formula 12).

The evaluation results on WTP is shown in Figure 1 and results on MRR is shown in Figure 2. We found that the four proposed hybrid models (AAE, AANL, AAN and AANE) are all significantly better than the single feature based models (ES, CF, AA and Katz) on both WTP and MRR. It may suggest that different features actually reveal different aspects of data, and combing them can improve the performance. The previous studies either only considered the local network feature or only the global network feature. The fact that AAN and AANL performs significantly better than both AA and Katz indicates that combining the local network features and the global network can improve the prediction accuracy. We also found that the regularization based model AAN is significantly better than linear combination models AANL. This indicates that the regularization based approach is a better approach for multiple feature combination compared to the simple linear combination. Among all the eight models, AANE performs the best. This indicates that incorporating all three features together using regularization based approach produce the best predication accuracy.

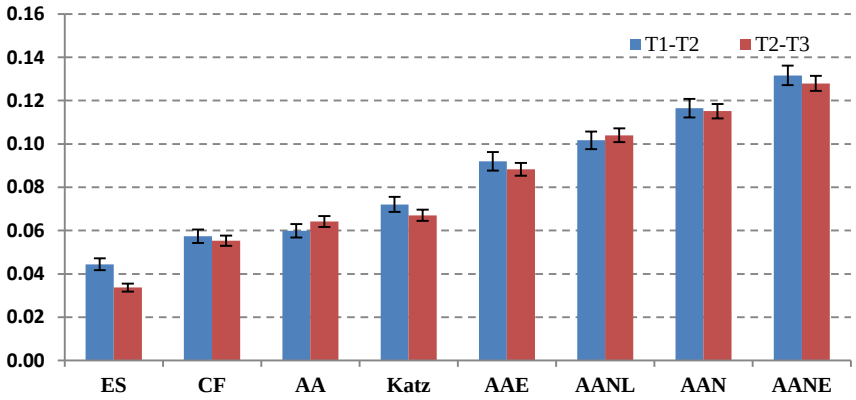

Fig. 1. WTP evaluation with stand errors

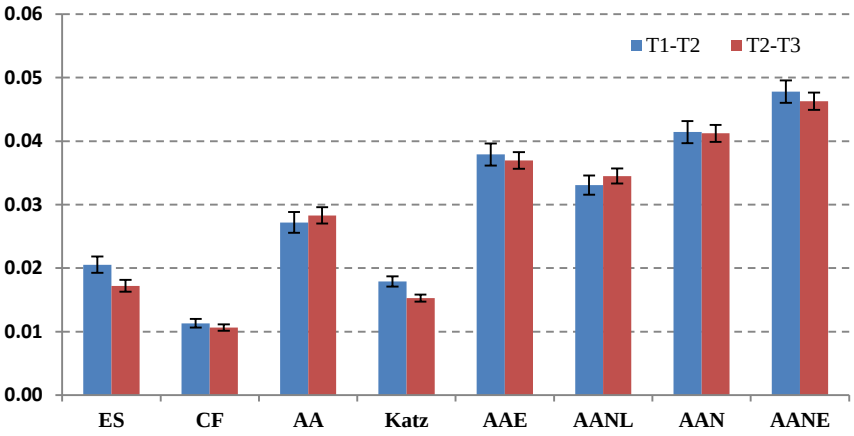

Fig. 2. MRR evaluation with stand errors

Some previous works found that combining content and network feature improves the predication. In our results, AAE is significantly better than content only model (ES) and network only model (CF, AA and Katz), which confirms the previous findings. In addition, we actually found that combing the local network feature and the global network feature helps more than combing the local network features with the content features. This is supported by the fact that AAN and AANL are significantly better than AAE on both WTP and MRR. The reason that AANL is better than AAE only on WTP is that MRR measures the exact rank positions while using content information can avoid ranking the right candidate in extreme low positions. We also found that pure network features based methods, i.e. CF, AA and Katz, are significantly better ($p<0.001$) than the pure content based method ES in terms of WTP evaluation. Content match focuses on finding potential coauthors using topic interest match, but people are unlikely to form coauthor relationships if they are not reachable to each other in the network even though they share similar topic interest. However, ES seems to be superior to CF on MRR, which may suggest that although it is unable to rank the right candidates to the top positions, it is also unlikely to rank the right candidates in extreme low positions.

AA and Katz have conflict performance on WTP and MRR. We found that in the T1-T2 dataset, Katz is significantly better than AA on WTP while it seems to be worse than AA on MRR evaluation in T2-T3. We think that AA suffers from data sparseness problem because it only considers local network feature. However, when only the global network feature is included in Katz, it introduces many noises.

5.3 Conclusion and Discussion

In this section, we look into the coauthor prediction problem for junior researchers who were actually ignored in previous works. The global network, the local network and the content-based feature were found to be useful in previous works for link pre-

diction. We proposed two regularization based models to combine multiple features and optimize them simultaneously. Comparing to the four baseline models that each consider only a single feature, our proposed models performed significantly better on the predication accuracy. A particularly interesting finding is that propagating feature from the local network to the global network improves the performance significantly compared to the model that combine content and local network feature. This indicates that although many previous works focused on combing the content feature and the local network features, they actually didn't take full advantage of the network features by not taking global network feature into account. Most importantly, the results show that our proposed regularization approach is better than simple linear combination and can be easily expanded to multiple features combination. In the next step, we will further explore the propagation method for multiple features combination, such as random walk or belief propagation.

6 Support Coauthor Finding in PeopleExplorer

In order to better support the coauthor finding problem, I developed the PeopleExplorer, an interactive integrated system, which not only predict potential collaborators but also provide the transparency of recommended candidates and let user control different facets, such as the research communities, seniority, affiliations as well as social similarity. A sample interface is shown in Fig. 3 [1]. The social similarity was computed combining both local network features and global network feature. System will show how the use will be connected with the potential candidate by suggesting possible connection path(s). We think the awareness of those connections will be helpful in increasing the awareness of the real connections. Besides, we also accommodate the content-based relevance, as shown in Formula (12).

7 Acknowledgement

Thanks very much for the help from Dr. Daqing He (Daqing He is an associate professor in School of Information Sciences, University of Pittsburgh). He provided useful suggestions and guidance for building up the search system and the algorithms. Thanks for Zhen Yue's help on paper writing and clarifying the research questions and paper editing (Zhen Yue is a PhD student at School of Information Sciences). This paper was presented in SBP 2013. The full text of the paper can also be in this link[2]. The presentation slides can be found using this link[3].

[1] http://crystal.exp.sis.pitt.edu:8080/PeopleExplorer/
[2] http://link.springer.com/chapter/10.1007/978-3-642-37210-0_30
[3] http://www2.pitt.edu/~shh69/mobileinformation.pdf

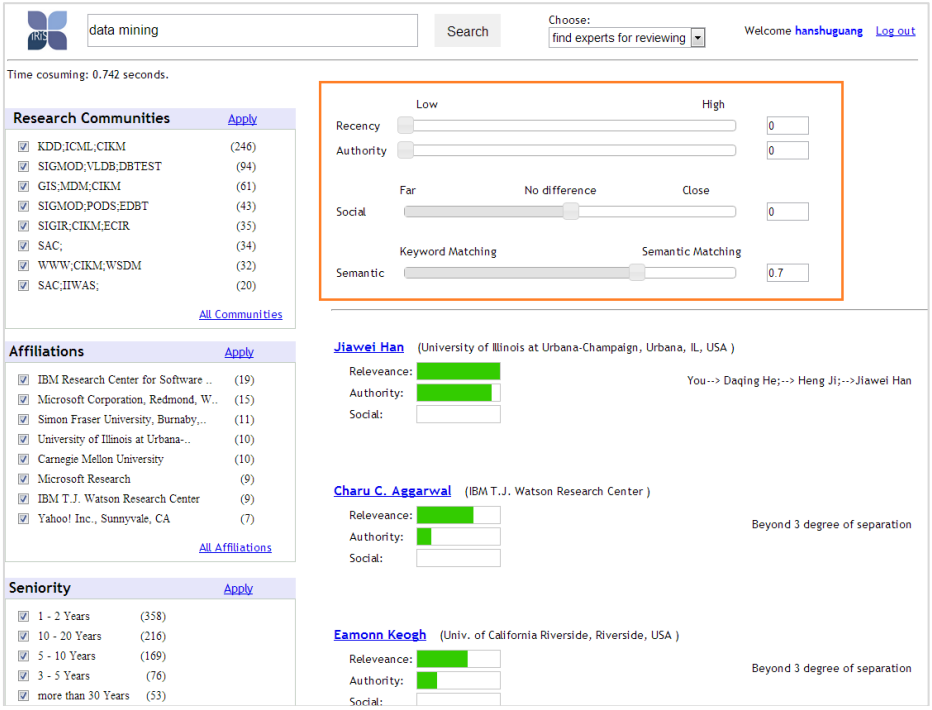

Fig. 3. A sample interface for PeopleExplorer

References

1. Balog K, Azzopardi L, Rijke Md (2006) Formal models for expert finding in enterprise corpora. Paper presented at the Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval.
2. Breese JS, Heckerman D, Kadie C (1998) Empirical analysis of predictive algorithms for collaborative filtering. Paper presented at the Proceedings of the Fourteenth conference on Uncertainty in artificial intelligence, Madison, Wisconsin,
3. Chao Wang VS, Srinivasan Parthasarathy Local Probabilistic Models for Link Prediction. In: Seventh IEEE International Conference on Data Mining, 2008.
4. Chen H-H, Gou L, Zhang X, Giles CL (2011) CollabSeer: a search engine for collaboration discovery. Paper presented at the Proceedings of the 11th annual international ACM/IEEE joint conference on Digital libraries, Ottawa, Ontario, Canada,
5. Chen H-H, Gou L, Zhang X, Giles CL (2012) Discovering missing links in networks using vertex similarity measures. Paper presented at the Proceedings of the 27th Annual ACM Symposium on Applied Computing, Trento, Italy,
6. Craswell N, de Vries, A. P. and Soboroff, I Overview of the trec-2005 enterprise track. In: Proceedings of the 14th Text REtrieval Conference, 2005.

7. Deng H, Han J, Lyu MR, King I (2012) Modeling and exploiting heterogeneous bibliographic networks for expertise ranking. Paper presented at the Proceedings of the 12th ACM/IEEE-CS joint conference on Digital Libraries, Washington, DC, USA,
8. Francesco Bonchia PE, David F. Gleichc, Chen Greifd & Laks V.S. Lakshmanand (2011) Fast Matrix Computations for Pairwise and Columnwise Commute Times and Katz Scores. Internet Mathematics 8 (1-2)
9. Katz L (1953) A new status index derived from sociometric analysis. Psychometrika 18 (1):39-43
10. Lada Adamic EA (2002) Friends and Neighbors on the Web. Social Networks 25:211-230
11. Lee D, Brusilovsky P, Schleyer T (2011) Recommending Future Collaborators using Social Features and MeSH terms. Paper presented at the Proceedings of the 74th Annual Meeting of the American Society for Information Science and Technology
12. Liben-Nowell D, Kleinberg J (2003) The link prediction problem for social networks. Paper presented at the Proceedings of the twelfth international conference on Information and knowledge management, New Orleans, LA, USA,
13. Lotka AJ (1926) The frequency distribution of scientific productivity. Journal of the Washington Academy of Sciences 16 (12):317–324
14. Ma H, Zhou D, Liu C, Lyu MR, King I (2011) Recommender systems with social regularization. Paper presented at the Proceedings of the fourth ACM international conference on Web search and data mining, Hong Kong, China,
15. Ming-Sheng Shang LL, Wei Zeng, Yi-Cheng Zhang and Tao Zhou (2009) Relevance is more significant than correlation: Information filtering on sparse data. EPL, 88 (6)
16. Mohammad Al Hasan VC, Saeed Salem , Mohammed Link Prediction Using Supervised Learning. In: Proc. of SDM 06 workshop on Link Analysis, Counterterrorism and Security, 2006.
17. Paweena Chaiwanarom RI, Chidchanok Lursinsap (2010) Finding potential research collaborators in four degrees of separation. Paper presented at the ADMA,
18. R.L. Kahn DJP (1994) Interdisciplinary collaborations are a scientific and social imperative. The Scientist
19. Xiongcai Cai MB, Alfred Krzywicki, Wayne Wobcke, Yang Sok Kim, Paul Compton, Ashesh Mahidadia (2011) Collaborative Filtering for People to People Recommendation in Social Networks. LNCS 6464:476-485
20. Yizhou Sun RB, Manish Gupta, Charu C. Aggarwal, Jiawei Han Co-Author Relationship Prediction in Heterogeneous Bibliographic Networks. In: International Conference on Advances in Social Networks Analysis and Mining, 2011. pp 121-128
21. Terveen, L. and McDonald, D. W. Social matching: A framework and research agenda. ACM Trans. Comput.-Hum. Interact., 12, 3 2005), 401-434.
22. Kautz, H., Selman, B. and Shah, M. Referral Web: combining social networks and collaborative filtering. Commun. ACM, 40, 3 (1997), 63-65.
23. McDonald, D. W. and Ackerman, M. S. Just talk to me: a field study of expertise location. In CSCW 1998: 315-324