An exploratory study on using social information networks for flexible literature access

Lynda Tamine, Amjed Ben Jabeur, Wahiba Bahsoun

An Exploratory Study on Using Social Information Networks for Flexible Literature Access

Lynda Tamine, Amjed Ben Jabeur, and Wahiba Bahsoun

IRIT SIG-RI, University Paul Sabatier Toulouse III, France
{lechani,jabeur,wbahsoun}@irit.fr

Abstract. It is well known that the fundamental intellectual problems of information access are the production and consumption of information. In this paper, we investigate the use of social network of information producers (authors) within relations in data (co-authorship and citation) in order to improve the relevance of information access. Relevance is derived from the network by levraging the usual topical similarity between the query and the document with the target author's authority. We explore various social network based measures for computing social information importance and show how this kind of contextual information can be incorporated within an information access model. We experiment with a collection issued from SIGIR[1] proceedings and show that combining topical, author and citation based evidences can significantly improve retrieval access precision, measured in terms of mean reciprocal rank.

Keywords: social networks, literature access, experimental evaluation.

1 Introduction

Most popular Web search engines use the content of the Web documents and their hyperlink structures in order to assess the relevance of documents in response to the user's query. This leads to two main drawbacks: the first one is that impersonal results are returned to the user as they don't fit particularly his interests, preferences and more generally his context. The second drawback is that a highly popular resource for a typical topic may dominate the results of another topic in which it is less authoritative. In order to tackle these problems, several solutions have been proposed in both contextual information retrieval (IR) [9,21,4] and web link analysis [16,7,3] research area. Recently, the problems cited above have been addressed by social IR [5,14,11] which is a novel research area that bridges IR and social networks analysis in order to enhance traditional information models by means of social usage of information. With this in mind, we have been inspired by the works in [11,12] and both revised and extended the retrieval models by using both co-author and citation relationships as social contexts features for enhancing particularly the results accuracy of a literature

[1] ACM Special Interest Group on Information Retrieval.

T. Andreasen et al. (Eds.): FQAS 2009, LNAI 5822, pp. 88–98, 2009.

search. Indeed, in our view, some factors extracted from the social network regarding co-author and citation relationships provide clues to identify what is relevant to the subject of related queries. Using the data we collected to understand authors' collaboration within scientific documents, we explore combining topical relevance (closeness between the query and the document) and social relevance (closeness between document' co-authors and target citations) in order to enhance the retrieval accuracy.

To the best of our knowledge, this is the first attempt to verify the assumption that document authorativeness, as measured using related co-author and citation features through a social network, is indeed a contributing factor to relevance estimation in particular within the setting of a literature search task. More precisely, comparatively to previous works, the contributions of the paper are:

- A social network based information access model combining authors' authorativeness and citation.
- An extensive experimental comparison of (1) several relevance measures borrowed from social network analysis in order to show their impact on the search effectiveness regarding two main assumptions of document relevance: most cited and most downloaded viewed as popularity criteria (2) ranking models to show the superiority of our proposed model.

The remainder of this paper is organized as follows: the background and related works will be introduced in section 2 with a focus on the use of social network analysis basis for enhancing information access. Our retrieval approach using evidence from the information network architecture and content is detailed in section 3. The experiments, results and discussion are presented in section 4. Conclusion and future work are given in section 5.

2 Background and Related Works

While being fundamental for the advances and present stage of IR, traditional IR models [19,18] make IR difficult and challenging from the cognitive side, particularly in large scale and interactive environments supporting communities such as bloggers, Wikipedia authors and users, online communities through Facebook, Myspace, Skyblog etc. The main criticism is that, in these approaches, retrieval ignores the influence of user's interactions within his social context on the whole IR process. Thus, the use of social networks theoretical foundations become tractable to achieve several retrieval tasks. In what follows we give an overview of social networks analysis basis and then focus on their use as support for dealing with literature access.

2.1 Social Networks Analysis: A Brief Overview

Social network analysis (SNA) is a research area that attempts to model actor behavior based on his social relations to other members of a group [22]. More practically, SNA views social relationships in terms of nodes V and edges E

within a graph $G = (V, E)$. Nodes are the individual actors within the networks, and edges are the relationships between the actors [Wikipedia]. In particular, social content graphs are specific graphs with two types of nodes: people and content. Social edges depend on the nature of the nodes being connected, they could be categorized into four main types covering several semantic relationships [2]: (1) person to content such as *authored by*, (2) person to person such as *friendship*, (3) content to content such as *hyperlink*, (4) content to person such as *endorsed by*.

An essential tool for the analysis of social networks are centrality measures defined on the graph edges. They are designed to rank the nodes according to their position in the network and interpreted as the importance or relevance of the nodes embedded in a social structure. This can be analyzed thanks to the main following centrality measures:

- *Degree:* degree centrality $C_d(u)$ of a node u is the number of edges directly connected with. High number of direct contact is an indicator of high social activity. $C_d(u)$ is computed as:

$$C_d(U) = \sum_{v \in V} e_{u,v} \tag{1}$$

  where $e_{u,v}$ is the edge between nodes u and v.
- *Closeness:* closeness centrality $C_c(u)$ is the the reciprocal of the total distances from a node to all the other nodes in the network. Closeness expresses the 'reachability' of a node from another one; it can be computed as:

$$C_c(U) = \frac{1}{\sum_{v \in V}} d(u, v) \tag{2}$$

  where $d(u, v)$ is the geodesic distance between nodes u and v measured as the shortest path between them.
- *Betweenness:* betweenness centrality $C_b(u)$ focuses on the ratio of the shortest paths a node lies on. A node having a high betweenness connect most of the nodes in the graph. Betweenness is computed as follows:

$$C_b(U) = \sum_{s \neq u \in V} \sum_{t \neq u \in V} \delta_{s,t}(u)^* \tag{3}$$

  where $\delta_{s,t}(u)$ is the probability that node u is involved between nodes s and t through the network connections, such as:

$$\delta_{s,t}(u)^* = \frac{\delta_{s,t}(u)}{\delta_{s,t}}$$

  where $\delta^*_{s,t}$ is the total number of shortest paths between nodes s and t.

The well known PageRank [16] and Hits [13] are also considered as centrality measures expressing the importance of a node within a social network.

We outline that the computation of centrality measures, particularly betweenness, for large and sparse networks (such as web subgraphs extracted from on line communities) is prohibitive. In order to tackle this problem, SNA literature suggests simpler measures as approximations of betweenness centrality, for instance based only on linkages between the neighbours of each node [6].

2.2 Literature Access

The advent of digital libraries on the web have boosted the availability and accessibility of bibliographic resources and therefore challenged literature access task. This latter has been addressed by a wide range of research approaches that focused, most of them, on the use of citation features as indicators of importance or authority of scientific publications [1]. Citation information and contexts have been used in early stage of IR area according to the principle of bibliographic coupling and co-citation analysis [10,20]. IR access has been improved with citation information at both indexing and retrieval levels [17,8]. In [17] citation information allowed to improve the document descriptors (index) by using terms extracted from citing documents to additionally describe the cited document. In [8], citations are viewed as hyperlinks and link structure, including anchor text used to enhance retrieval.

Recently there was an attempt to address literature search from the social view where the main actors are authors and documents and edges express the authorship relation. To the best of our knowledge, the research works in this range are [12,11]. The authors proposed a model of social IR including: (1) a social network extracted from the bibliographic resource where the main actors are authors and documents and edges express the authorship relation (2) a multiplicative relevance scoring based on the combinaison of query-document similarity and document authority.

In this paper, considering the potential usefulness of citation information, we explore the use of an additional social relation extracted from citation and then attempt to combine linearly a relevance score and a social score within both authorship and citation relations. Furthermore, we undertake an extensive experimental analysis on the impact of centrality measures for expressing authority nodes and leverage them with different assumptions of relevance issued from social endorsement.

3 Combining Topical and Social Relevance over Author and Citation Networks

In this section, we argue that social relations between a bibliographic resource' authors mainly, co-authorship and citation (over the documents) can potentially provide clues to better estimate the relevance of a document in response to a user query. In the rest of this section, we first describe the social network supporting the information access model, then we detail the relevance estimation measure.

3.1 From a Bibliographic Resource to a Social Network Graph

Suppose we have a bibliographic resource containing documents authored by n authors, we build a social network graph $G = (V, E)$ where:

1. The nodes set $V = \{v_i\}_1^n$ represents all the authors identified in the resource.
2. The edges set $E = \{e\{j,k\}\}$. Each edge expresses one of the two main following relations:
   - *an implicit direct social relation* between authors expressing the co-authorship relation. For a pair of co-authors v_j and v_k of at least one document, we plot and undirected edge $e_a\{j,k\}$,
   - *an implicit indirect social relation* between authors expressing the citation relation. For a pair of authors v_j and v_k such as v_j is cited by v_k at least through one document, we plot a directed edge $e_c\{j,k\}$.

Figure 1. illustrates the social graph issued from a sample resource.

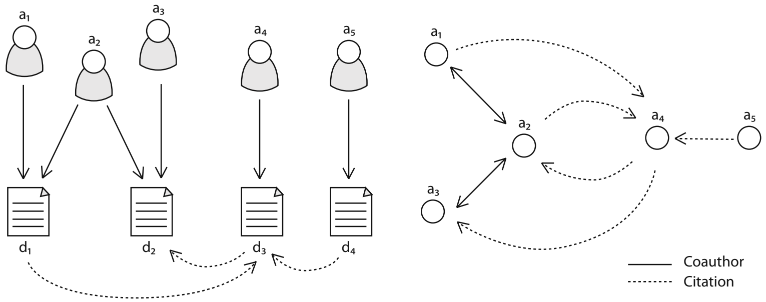

Fig. 1. The social content graph

3.2 Relevance Estimation Using the Social Context

In this paper we address a ad hoc search task initiated by a user query q. The idea of document relevance estimation within a social graph is to derive a more accurate response to the user by combining the topical relevance of document d and the importance of the associated authors regarding the social relations between them. Intuitively, when assessing a document d, a user is likely to assess it as relevant if it covers the query topic and that the corresponding authors are important and close within this topic regarding the overall bibliographic resource topics. According to this, we combine the two scores of relevance as follows:

$$Rel(d) = \alpha * RSV(q,d) + (1-\alpha) * Imp(d) \tag{4}$$

where $\alpha \in [0 \ldots 1]$ is a weighting parameter, $RSV(q,d)^2$ is a normalized similarity measure between query q and document d descriptors, $Imp(d)$ is the importance of document d authored by k authors $\{v_i\}_{i=1\ldots k}$, computed as:

[2] Relevance Status Value.

$$Imp(d) = \sum_{i=1..k} C(v_i) \tag{5}$$

where $C(v_i)$ is a normalized centrality measure (Cf. 2.1).

Table 1 shows an illustration of the normalized importance measure computation of the nodes in the social network presented in figure 1.

Table 1. Authors' importance values using centrality measures

	Degree	*Closeness*	*Betweenness*	*Pagerank*	*Hits*
a_1	0,11	0,25	0	0,14	0,14
a_2	0,33	0,33	4,5	0,39	0,28
a_3	0,22	0,2	0	0,23	0,25
a_4	0,33	0,25	3,5	0,22	0,33
a_5	0	0,13	0	0,02	0

4 Experimental Evaluation

In this section, we describe the dataset used for our experimental evaluation and then detail the experiments we have undertaken in order to achieve the main following objectives: (1) evaluating the impact of several centrality measures on search effectiveness, (2) comparing our model effectiveness Vs. both models using document content solely and those combining content and social co-authorship social relation [11,12].

4.1 Experiments with Importance Measures Schemes

In this experiment, we compared the impact of five (5) importance measures on the search effectiveness: *pagerank*, *hits*, *closeness*, *degree* and *betweenness*. According to this objective, we did not consider here the *RSV* measure (Cf. formula (4)) by setting $\alpha = 0$. Figures 2 gives the MRR measures corresponding to each importance measure considering respectively the most cited Vs. the most downloaded relevance assumption.

We can notice that for the most cited queries, *pagerank* and *hits* measures show better ranking results. This could be explained by the fact that assuming that the number of citations is an indicator of authority, most cited documents would be authoritative entities and this property would be inherited by the corresponding authors. Therefore, improving most cited documents positions in the result set means improving the position of authoritative authors and that is particularly assured by *pagerank* and *hits* measures. Furthermore, most cited documents have often a higher degree value and that explains the good results given by the *degree* measure. Considering the most downloaded documents, we notice that closeness measure results are generally the best ones, being at the top of the social importance measures. We can explain this fact

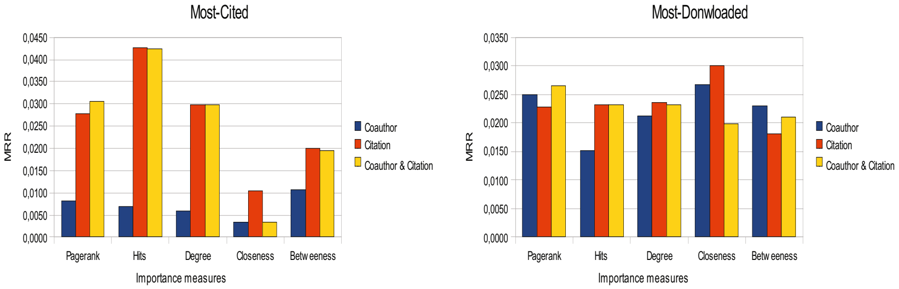

Fig. 2. Importance measures comparison

after the analysis of the most downloaded documents' content. Indeed, we highlight that most downloaded documents are published recently (2008) and related queries contain mainly specific terms belonging to new research topics (social IR, collaborative IR etc). Unlike most cited document's related queries, results set contain restricted number of documents dealing with few topics and authored by a restricted group of authors that usually work together. Authors of most-downloaded documents are central entities in their neighborhood or research topic and already have several collaborations in the same research topic, consequently they have a higher closeness value. In addition, authors of most downloaded documents are not authoritative in the whole data collection but in their neighborhood and they have the opportunity to be cited thanks to their past published documents. This explains the good results given by the *pagerank* and *hits* measures too.

We retained the two best importance measures within each query test collection to tune our model as detailed in the following.

4.2 Experiments with Relevance Models Schemes

We address through these experiments the effectiveness of our model compared to baseline models. In order to achieve this objective, we first tuned our model by varying α parameter (Cf. formula (4)) and then retained the best setting in order to analyse the comparative evaluation.

Parameter Tuning. We studied the impact of the tuning parameter α according to the two relevance assumptions. Figure 3 and Figure 4 show the MRR measures when α is varied for the two best importance measures retained for each test set collection from the experiments detailed above.

We can see that for α greater than 0.5, using social importance measures improve significantly the retrieval performances. Furthermore, it can be seen on MRR curves that the best values of MRR are achieved for α lower than 1 whereas $\alpha = 1$ corresponds to the basic ranking algorithm using topical similarity. This ensures, according to our general motivation, that combining document content based score and auhor's social importance score can improve the final ranking

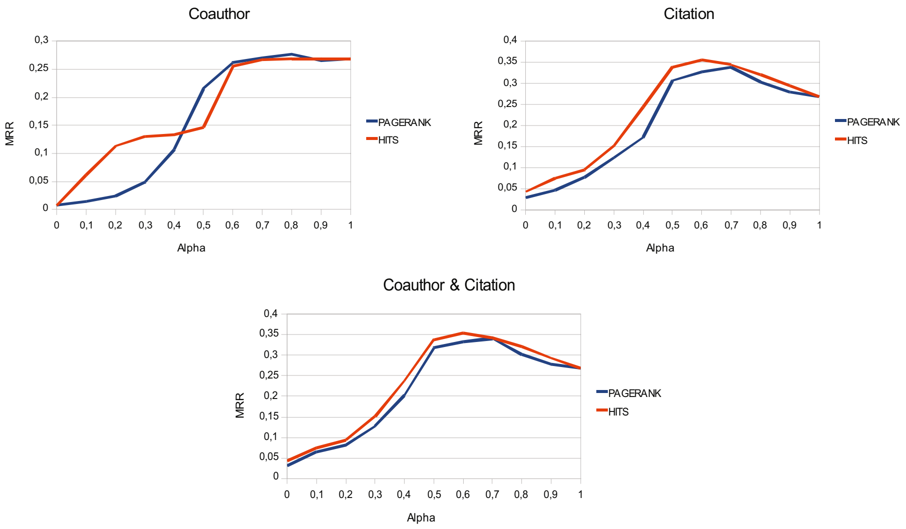

Fig. 3. Parameter tuning using most cited relevance assumption

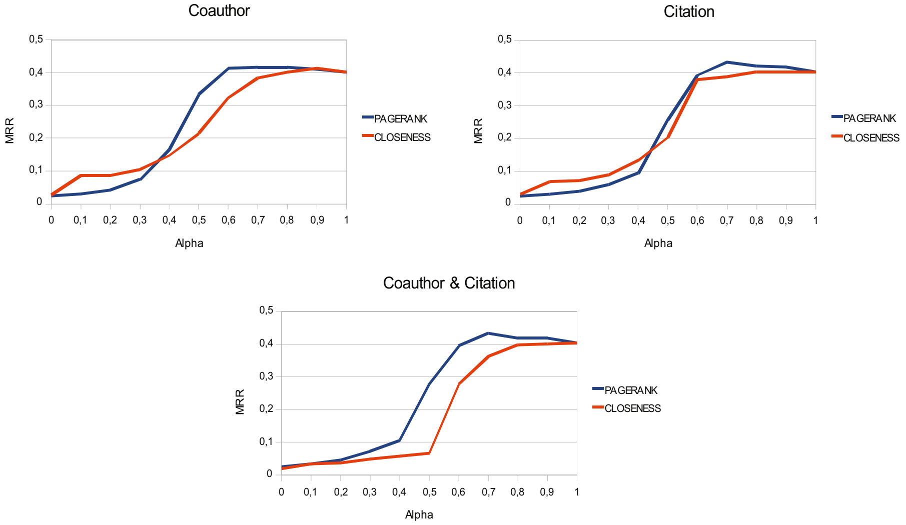

Fig. 4. Parameter tuning using most downloaded relevance assumption

results. We also notice that the best tuning parameter α value is 0.7 for *pagerank* measure, 0.6 for *hits* measure and 0.9 for *closeness* measure.

In addition, we can notice that if α value is greater than 0.5, the *pagerank* based MMR values are close to *hits* in the case of most cited relevance assumption, and that the *pagerank* measure performs better than the *closeness* measure in the case of most downloaded assumption. We retained the best mean value $\alpha = 0.7$ and the best common importance measure *pagerank* for the remaining experiments.

Evaluation of Ranking Models Effectiveness. We compared our proposed retrieval model with two baseline models: (1) RSV vectorial model based on traditional $TF * IDF$ ranking measure and (2) Kirsh's model [12] that adopt a simple method for combining author's based measures (Cf 2.2) and content based measure $RSV(q, d) * Imp_a(d)$ using evidence issued only from co-author social relation.

Furthermore, in order to undertake an extensive experimental study and to guarantee an accurate comparative evaluation, we extended Kirsh' baseline model with evidence issued from citation and both authors and citation and then compared the performances to our model's ones. Table 3 summarizes the obtained results using the three network settings: co-author, citation, co-author and citation networks.

Table 2. Comparative evaluation of retrieval algorithms performances

Co-author Network

Assumption	TF*IDF	Kirsh's Model	Our Model	Improvement	
				% TF*IDF	% Kirsh's Model
Most cited	0,268	0,212	0,270	1%	27%
Most downloaded	0,403	0,324	0,417	4%	29%

Citation Network

Relevance assumption	TF*IDF	Kirsh's Model	Our Model	Improvement	
				% TF*IDF	% Kirsh's Model
Most cited	0,268	0,212	0,338	26%	59%
Most downloaded	0,403	0,324	0,433	7%	34%

Co-author & Citation Network

Relevance assumption	TF*IDF	Kirsh's Model	Our Model	Improvement	
				% TF*IDF	% Kirsh's Model
Most cited	0,268	0,212	0,342	27%	61%
Most downloaded	0,403	0,324	0,433	7%	34%

First, we observe that our model performs better with most cited relevance assumption than with most downloaded relevance one. This can be explained, as already outlined in previous experiments, that query terms extracted according to the most cited assumption are more general and popular, so adding a social importance in the relevance computation of documents would have a significant impact on the ranking results while the most downloaded assumption favors the ranking according to the closeness of a more restricted list of documents. However, in all the cases our model outperforms the two baseline models according to the different network settings. More specifically, comparing our model using only the evidence issued from co-author social relations with Kirsh's model, we notice an improvement about 27% (most cited assumption) and this is a positive argument for our choice of a linear combination of social and content scores for computig the final document score. Improvement is greater with citation network and co-author and citation network achieving 61% (most cited assumption). We conclude that integrating citation social link in the author's network as a social relationship improves the final ranking result compared to co-author network. Finally, we notice that there is a small improvement of both co-author and citation network (59% with most cited assumption) compared to citation network (61% with most cited assumption), in other words combining co-author relationship to citation relationship, and this can be explained by the fact that citation network dominates the co-author network with large number of citation links between authors, as can be expected from the social network characteristics presented in table 2.

5 Conclusion and Future Works

This paper presented an extended social based retrieval model based on evidence issued from co-author and citation social relations extracted from a bibliographic collection. We particularly outlined the effectiveness of our model using several importance based measures compared to state of the art retrieval models. Furthermore, two main social relevance assumptions have been used in order to ensure the soundness of our results. In future, we plan to enhance the social network model by introducing weights in the edges expressing the strengthness of the social relations between authors through both co-authorship and citations relations. Furthermore, we plan to test our retrieval model on larger web collections containing bibliographic documents and blogs and test the impact of more specific relations, issued from social bookmarking, on the retrieval performances.

References

1. Amento, B., Terveen, L., et al.: Does 'authority' mean 'quality'? predicting expert quality ratings of documents. In: Annual ACM Conference on Research and Development in Information Retrieval SIGIR, August 2000, pp. 296–303 (2000), booktitle
2. Amer Yahia, S., Benedikt, M., Bohannon, P.: IEEE Data Eng. Bull. 30(2), 23–31 (2007)
3. Bharat, K., Henzinger, M.R.: Improved algorithms for topic distillation in hyperlinked environments. In: Proceedings of the 21st Annual ACM Conference on Research and Development in Information Retrieval SIGIR, pp. 104–111 (August 1998)
4. Daoud, M., Tamine, L., Boughanem, M.: Towards a graph based user profile modeling for a session-based personalized search. In: Knowledge and Information Systems. Springer, Heidelberg (2009)
5. Dion, G., Shubert, F.: Social information retrieval systems: Emerging technologies and applications for searching the web effectively. Premier Reference Source

6. Everett, M.G., Borgatti, S.P., Krackhardt, D.: Ego-network betweenness. In: Proceedings of the 19th International Conference on Social Network Analysis, Charleston, South Carolina
7. Haliwala, T.H.: Topic-sensitive PageRank. In: Proceedings of the 11th International World Wide Web Conference (2002)
8. Hawking, D., Craswell, N.: The very large collection and web tracks. In: Voorhees, E.M., Harman, D.K. (eds.) TREC: Experiments and evaluation in information retrieval, ch. 9. MIT Press, Cambridge (2005)
9. Jeh, G., Widom, J.: Scaling personalized Web search. In: Proceedings of the 12th International World Wide Web Conference, pp. 271–279 (2003)
10. Kessler, M.M.: Bibliographic coupling between scientific papers. American documentation (14), 10–25 (1963)
11. Kirchhoff, L., Stanoevska-Slabeva, K., Nicolai, T., Fleck, M.: Using social network analysis to enhance information retrieval systems. In: Social network applications conference (2008)
12. Kirsh, M.K.: Social information retrieval, PhD Thesis in Computer Science, Computer science department III, Bonn (March 14, 2003)
13. Kleinberg, J.M.: Authoritative sources in a hyperlinked environment. Journal of the ACM 46(5), 604–632 (1999)
14. Korfiatis, N.T., Poulos, M., Bokos, G.: Evaluating authoritative sources using social networks: an insight from Wikipedia. Online Information Review 30(3), 252–262 (2006)
15. Meij, E., De Rijke, M.: Using prior information derived from citations in literature search. In: Proceedings of Recherche d'Information Assistée par Ordinateur, RIAO (2007)
16. Page, L., Brin, S., Motwani, R., Winograd, T.: The PageRank citation ranking: Bringing order to the web (1998) (Unplished draft)
17. Ritchie, A., Teufel, S., Robertson, S.: Using terms from citations for IR: some first results. In: Macdonald, C., Ounis, I., Plachouras, V., Ruthven, I., White, R.W. (eds.) ECIR 2008. LNCS, vol. 4956, pp. 211–221. Springer, Heidelberg (2008)
18. Robertson, S., Jones, K.S.: Relevance weighting for search terms. Journal of the American Society for Information Science 27(3), 129–146
19. Salton, The SMART Information retrieval system. Prentice-Hall, Englewood Cliffs
20. Small, H.: Co-citation in the scientific literature: A new measurement of the relationship between two documents. Journal of the American Society of Information Science 24(4), 265–269 (1973)
21. Teevan, J., Dumais, T.S., Horvitz, E.: Personalizing search via automated analysis of interests and activities. In: Proceedings of the 28th Annual ACM Conference on Research and Development in Information Retrieval SIGIR, pp. 449–456 (August 2005)
22. Wasserman, S., Faust, K.: Social network analysis: methods and applications. Cambridge Uni. Press, Cambridge