Personalized Document Rankings by Incorporating Trust Information From Social Network Data into Link-Based Measures

Claudia Hess, Klaus Stein

Laboratory for Semantic Information Technology
Bamberg University
{claudia.hess,klaus.stein}@wiai.uni-bamberg.de

Abstract. Recently, two-layer approaches that integrate information from social networks into link-based measures on document reference networks such as the web, scientific publications or wikis gained much attention. In this paper, we extend the framework presented in [1] in order to achieve a stronger personalization of the document recommendations. Therefore, we investigate alternative approaches for bringing trust data from an author trust network "down" to the document reference network and discuss how they can be combined in a comprehensive personalization strategy. In a simulation study, we show the effects of our personalization approach on document recommendations. Moreover, we discuss how the framework presented can be realized in practice.

1 Introduction

Searching for reliable information in the huge amount of webpages, discussion groups, wikis and blogs we often ask ourselves "Can I trust this information?". Trust-based recommender systems such as [2], [3] address this question by analyzing a certain kind of social relationship data: trust relationships between users. Recently several approaches investigated how trust data can be used for improving document rankings. In particular, they combine methods from social network analysis with link-based measures for the analysis of network structures of webpages, wikis or blogs (e. g., [4], [5], [1], [6]). [5], for instance, uses centrality and prestige measures from social network analysis for analyzing wikis. Other approaches integrate information from author trust networks into reference-based measures such as PageRank [7] or HITS [8] in order to improve document recommendations and rankings. Trust networks, a special type of social network in which users express their degree of trust in other users, seem particularly useful because the trust relationships can be used to personalize the results by reference-based measures. For example, if I have high trust in a researcher to write good papers, it is quite likely that I read some of the papers cited in her papers. Using trust relationships for personalization is particularly interesting for non-mainstream users who do not appreciate the recommendations for the average user [9]. This can for example be a researcher who works in some very specific area or applies specific methods which are considered in general as "strange". Papers interesting for this researcher would be ranked by mere reference-based measures very badly because they attract only few links and are hence not displayed among the first search results. This user would need a personalized ranking.

In [1], we presented a framework for integrating trust between authors into reference-based document rankings. We focused on using trust information in order to capture the semantics of references, i. e., to determine whether a reference is supportive or rather depreciatory. So we developed trust-aware measures for the importance of a document, its so called visibility, such as a trust-weighted PageRank. The framework offers some basic personalization. However, there is room for improvement. We discuss in this work the missing aspects for a comprehensive personalization strategy and embed it again in the framework. For a thorough personalization we exploit alternatives for bringing the trust between the authors "down" to the document reference network. Firstly, computing a document ranking for a user U, U's trust in an author A directly influences the visibility of the documents written by A. Secondly, U's trust in A also influences the impact of the citations of A by modulating the reference semantics. We will discuss these two approaches for personalization as well as their combination.

The paper is structured as follows: section 2.1 provides the basics for the integration of author trust into reference-based measures. Section 3 discusses different approaches for personalization. Section 4 presents the simulation study that was carried out in order to evaluate our personalization strategy. In section 5, it is discussed how the framework presented can be realized in practice. Section 6 concludes the work.

2 Trust-based Document Ranking

2.1 Reference-based Visibility Measures

To be able to handle huge document repositories such as webpages or scientific papers, documents must be presented to the user sorted by relevance. Ordinary search engines use structure based ranking algorithms like PageRank or HITS, where the importance (visibility) of a documents is computed by the visibility of the documents citing it. For example, using PageRank the visibility vis_a of a document p_a is computed:

$$\mathrm{vis}_a \quad = \quad (1-d) \; + \; d \sum_{p_k \in R_a} \frac{\mathrm{vis}_k}{|C_k|} \tag{1}$$

where R_a is the set of pages citing p_a and C_k is the set of pages cited by p_k.[1]

The drawback of this approach is that all references are considered to be supportive, i. e., it is assumed that by setting a link, an author wants to confer some authority to the cited document. Of course, this does not hold in all cases.

[1] $(1-d)$ is the base visibility of each document, for further discussion see e. g. [7].

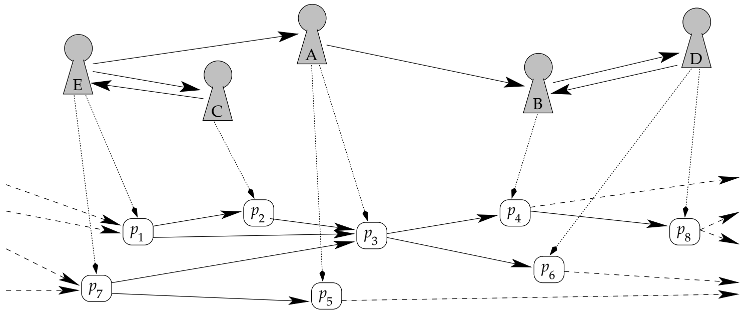

Fig. 1. Combined author trust network and document reference network

For instance, in the context of scientific publications, a paper could discuss cases of scientific fraud and cite some faked papers. For the reader of the paper it is absolutely clear that these faked papers are cited as examples whereas a reference-based visibility measure can distinguish supportive from non-supportive links. To address this problem, information from a second layer, the trust network between the authors of the documents is used.

2.2 Indicating Reference Semantics by Using Trust Information

In order to determine the semantics of references between documents, trust ratings are propagated to the edges in the document network. This means concretely in figure 1 that A's trust rating for B is mapped to all references from A's to B's documents, here to the reference from p_3 to p_5. This gives a hint to the semantics of the citation: if A highly trusts B, then the citation will likely be supportive. It would be contradictory to cite someone in the context of scientific fraud and to assign her at the same time a high trust as author. The other way around, citations to a document by an author considered as distrusted would rather be depreciatory than supportive.

Trust is in general represented as a numerical value, e. g., in $[-1, 1]$ ranging from distrust (-1) and 'no trust' (0) to trust (1). Users express their subjective trust towards other users to be a "good" author, i. e., to provide reliable information of a high quality, not to have any links to spam pages etc. Edges between authors are hence weighted and directed. Technically, the trust edges from the author trust network are mapped to the references in the document network and each reference $p_i \rightarrow p_j$ is associated with a weight $w_{i\rightarrow j}$, which is computed from the trust edges using a mapping function.[2] Now a trust-weighted visibility can be calculated for each document. We modify an existing reference-based visibility function such as PageRank (which is not able to deal with weighted edges) so that documents distribute their visibility according to the edge weight:

$$\mathrm{vis}_a \quad = \quad (1-d) \; + \; d \sum_{p_k \in R_a} \frac{w_{k\rightarrow a}}{\sum_{p_j \in C_k} w_{k\rightarrow j}} \, \mathrm{vis}_k \tag{2}$$

[2] The mapping is needed as trust values range from -1 to 1 and we do not want to have negative weights.

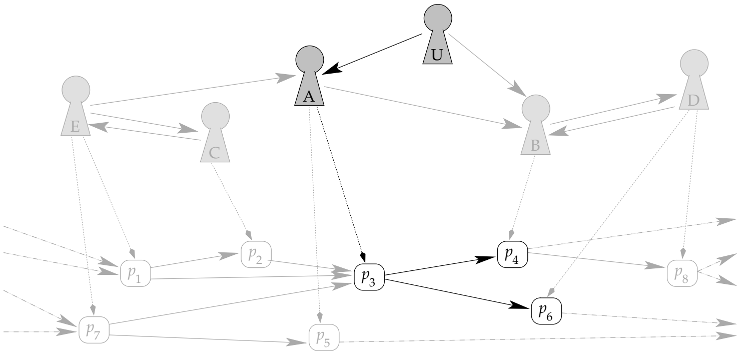

Fig. 2. Personalization by modifying document visibilities by trust

The whole approach is described in detail in [1].

Although trust networks tend to be small in size and in the number of trust ratings, trust propagation offers the possibility to assign edge weights to a considerable part of the references in the document reference network. By trust propagation, additional trust ratings are inferred for indirectly connected users (see e. g., [10], [11]). In figure 1 trust between U and D is calculated on the basis of the trust rating that B, someone about whom U provided directly a trust rating, has assigned to D. Path algebraic trust metrics such as [10] calculate U's trust in D by concatenating the trust values on the path from U to D.

While the approach addresses the question of reference semantics by distinguishing supporting and distrusting links it does not take into account the view of a concrete user, who may trust one author and distrust another one. The question is now how to compute a personalized visibility. To address this question, we look at two alternative approaches for incorporating subjective user trust ratings. As basis, we have a document network with weighted edges whereby the edge weights come from the trust network.

3 Personalization

3.1 Modifying Document Visibilities by Subjective Trust

The first approach for personalizing document rankings is to propagate trust ratings to the document reference network and to modify the visibility of an author's documents by the trust set in her. In the example in figure 2 that means that U's trust in A changes the visibility of p_3 and p_5: depending on U's degree of trust, the personalized visibility of p_3 and p_5 increases or decreases. This is accomplished by modifying the visibility function, e. g., for PageRank[3]:

$$\text{vis}_a \quad = \quad (1-d)(t_{U\to A_a}+1) \;+\; d\sum_{p_k\in R_a}\frac{w_{k\to a}}{\sum_{p_j\in C_k} w_{k\to j}}\,\text{vis}_k \tag{3}$$

[3] We use PageRank throughout this paper, certainly any other visibility function could be used and modified accordingly.

with A_a being the author of document p_a. The idea behind this formula is to modify the base visibility $(1-d)$ of each document by the trust the user has into the author of the document. So the document gets its visibility from the trust the user has in its author and from the (trust-weighted) visibilities of the documents citing it. As $t_{U \to A_a} \in [-1, 1]$ we use $(t_{U \to A_a} + 1)$ to prevent the visibility from becoming negative.[4] If there is no trust edge from U to A_a, a default trust value (usually $t_{\text{default}} = 0$) is used. If the paper has more than one author, the user's trust in the author collective is usually computed by using the average trust; the minimum or maximum can be useful, too and can be configured by the user.

Instead of inserting the user's trust into the visibility formula it can also be used to change the resulting visibility:

$$\text{vis}_a \quad = \quad (t_{U \to A_a} + 1)\left((1-d) \; + \; d \sum_{p_k \in R_a} \frac{w_{k \to a}}{\sum_{p_j \in C_k} w_{k \to j}} \text{vis}_k\right) \tag{4}$$

(4) has a stronger effect on a document's visibility than (3) because it affects the complete reference-based visibility of a document, i.e., the base visibility as well as the visibility support that the document gets from documents citing it. If user U totally distrusts author A_a the visibility vis_a of the document p_a is set to 0, even though it is cited by many highly visible documents. So this approach decreases very effectively the rank of documents written by distrusted authors regardless of their position in the document network.

The trust ratings have an indirect impact, too. As the visibility of a document partly depends on the visibility of the documents citing it, documents cited by a trusted author gain visibility, while documents cited by a distrusted author loose visibility. In figure 2, the visibility of p_3 is changed by U's trust in its author A, and p_3 cites p_4, so the modified visibility is propagated from p_3 to p_4 and to p_6. The indirect propagation of the trust-enhanced visibility has the following semantics: if I consider an author as trustworthy, it is very likely that I follow the links provided in her documents. In contrast, if I distrust an author, it is very likely that I do not follow the links. In this case, it is appropriate that the cited document gets less visibility via the link from the document written by the untrustworthy author.

3.2 Modifying Reference Weights by Subjective Trust

The second approach for personalizing document rankings is to change reference weights by subjective trust and therefore to modulate the support of visibility that a document gives to the documents it cites. As distrusting an author also means distrusting her citations it seems sensible that the trust of an user in an author affects the weights associated with the citations of the author's documents. In figure 3, the weights $w_{3 \to 4}$ of the reference $p_3 \to p_4$ and $w_{3 \to 6}$ of the

[4] Certainly any of the mapping functions presented in [1, section 4.2] could be used instead.

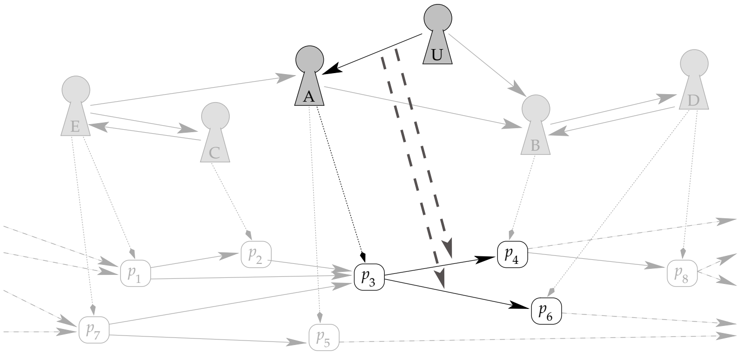

Fig. 3. Personalization by modifying reference weights by trust

reference $p_3 \rightarrow p_6$ are modified by the trust $t_{U \rightarrow A}$ that U has in A:[5]

$$w'_{i \rightarrow k} \quad = \quad \frac{t_{U \rightarrow A} - t_{\min}}{t_{\max} - t_{\min}} \, w_{i \rightarrow k} \tag{5}$$

By using these modified weights with the visibility function (2) references from documents of distrusted authors are suppressed while references from documents of trusted authors are supported.

This approach has the drawback that the personalization only affects the documents cited in the documents of a trusted (or distrusted) author. The visibility of the trusted (or distrusted) author's documents cannot directly be influenced as by the first approach. We therefore propose to combine both approaches for a comprehensive recommendation strategy.

3.3 Combining both Personalization Approaches

As both approaches propagate the user trust information to different parts of the document network, the first approach directly affects the visibility of the documents (nodes) while the second one modifies the weights of the references (edges), they are independent from each other and do not interfere. Therefore they can simply be combined by applying both on one network: the reference weights are modified using function (5) and then the personalized trust-weighted visibility function (3) or (4) is applied to compute the personalized rankings.

4 Simulation Study

In the simulation study we show the effects of the personalization approaches presented in section 3. Each simulation was run 10 times on 10 independent document reference and author trust networks with ≈ 3500 documents each citing

[5] Note that this is *not* the formula we used in [1, section 4.6], as the form given here directly affects the reference weight, which is more appropriate.

	$\mathcal{A}$	$\mathcal{B}$	$\mathcal{C}$	$\mathcal{A}$	$\mathcal{B}$	$\mathcal{C}$	$\mathcal{A}$	$\mathcal{B}$	$\mathcal{C}$	$\mathcal{A}$	$\mathcal{B}$	$\mathcal{C}$
	direct			indirect			direct			indirect		
	function (3)						function (4)					
$U_{\mathcal{B},\overline{\mathcal{C}}}$	0.0%	16.0%	-15.6%	-0.3%	1.8%	1.5%	1.2%	25.0%	-46.6%	-0.5%	1.6%	-6.3%
$U_{\mathcal{C},\overline{\mathcal{B}}}$	0.0%	-16.4%	16.3%	-0.3%	-2.7%	-2.4%	1.2%	-46.8%	25.3%	-0.5%	-6.5%	1.3%
	function (5)+(2)											
$U_{\mathcal{B},\overline{\mathcal{C}}}$	0.0%	-0.2%	0.1%	0.0%	5.0%	-6.3%						
$U_{\mathcal{C},\overline{\mathcal{B}}}$	0.0%	0.1%	-0.1%	0.0%	-6.3%	5.1%						
	function (5)+(3)						function (5)+(4)					
$U_{\mathcal{B},\overline{\mathcal{C}}}$	0.0%	34.8%	-35.1%	-0.6%	7.7%	-7.4%	0.4%	38.9%	-47.5%	-0.8%	7.1%	-8.1%
$U_{\mathcal{C},\overline{\mathcal{B}}}$	0.0%	-36.2%	36.2%	-0.7%	-7.7%	7.6%	0.6%	-46.7%	38.3%	-0.8%	-8.2%	7.0%

Table 1. Results of the simulation (average over 10 independent simulation runs). The numbers give the proportional change of the average position of all documents of authors from group $\mathcal{A}$, $\mathcal{B}$ and $\mathcal{C}$ in the personalized view of user $U_{\mathcal{B},\overline{\mathcal{C}}}$ and $U_{\mathcal{C},\overline{\mathcal{B}}}$ compared to U_0 with the different algorithms. "direct $\mathcal{A}$, $\mathcal{B}$, $\mathcal{C}$" gives the change of the documents *written* by an author of the group while "indirect $\mathcal{A}$, $\mathcal{B}$, $\mathcal{C}$" gives the change of the documents *cited* by documents written by an author of the group.

2 to 7 other documents (the PageRank parameter $d = 0.85$). The documents are written by 100 authors in three groups: group $\mathcal{A}$ with 90 authors and $\mathcal{B}$ and $\mathcal{C}$ with 5 authors each. All authors are neutral to all others $t_{X \to Y} = 0$ in order to eliminate side effects as we want to simulate the effects of personalization.[6] This network is now seen by three different users:

- U_0 is neutral to all authors: $\forall X \in (\mathcal{A} \cup \mathcal{B} \cup \mathcal{C}) : t_{U \to X} = 0$.
- $U_{\mathcal{B},\overline{\mathcal{C}}}$ is neutral to all authors of group $\mathcal{A}$, trusts all authors of group $\mathcal{B}$ and distrusts all authors of group $\mathcal{C}$:
$$\forall A \in \mathcal{A} : t_{U \to A} = 0, \quad \forall B \in \mathcal{B} : t_{U \to B} = 1, \quad \forall C \in \mathcal{C} : t_{U \to C} = -1.$$
- $U_{\mathcal{C},\overline{\mathcal{B}}}$ is exactly contrary to $U_{\mathcal{B},\overline{\mathcal{C}}}$: neutral to all authors of group $\mathcal{A}$, trusts all authors of group $\mathcal{C}$ and distrusts all authors of group $\mathcal{C}$:
$$\forall A \in \mathcal{A} : t_{U \to A} = 0, \quad \forall B \in \mathcal{B} : t_{U \to B} = -1, \quad \forall C \in \mathcal{C} : t_{U \to C} = 1.$$

Now the ranking of all documents is computed using the functions (3) and (4) modifying the document visibility by user trust, as well as the reference weight modifying function (5) in combination with functions (2), (3) and (4). Then the effect on the documents written ("direct") or cited ("indirect") by authors of group $\mathcal{A}$, $\mathcal{B}$ and $\mathcal{C}$ are shown by comparing their average ranking position in the personalized view of user U_0, $U_{\mathcal{B},\overline{\mathcal{C}}}$ and $U_{\mathcal{C},\overline{\mathcal{B}}}$.

Table 1 shows the results of the simulation. In all simulations the ranking of the documents written or cited, respectively, by an author of group $\mathcal{A}$ is nearly not affected because all users neither trust nor distrust authors from this group.

[6] For a simulation of the effects of different trust between authors see [1, section 5].

The personalized ranking of documents written or cited by authors of group $\mathcal{B}$ and $\mathcal{C}$ from $U_{\mathcal{B},\overline{\mathcal{C}}}$ and $U_{\mathcal{C},\overline{\mathcal{B}}}$ compared to the ranking of the "neutral" user U_0 clearly differs, as expected: For function (3) the documents written by authors of the trusted group gain $\approx 16\%$, i. e. in average their position in the sorted list of search results is $\approx 16\%$ better than for user U_0, while documents written by authors of the untrusted group in average loose $\approx 16\%$. Additionally a small indirect effect is found: documents cited by authors of the trusted group gain $\approx 1.6\%$ and documents cited by authors of the distrusted group loose $\approx 2.5\%$. Using function (4) the effect is much stronger: the documents of trusted authors gain $\approx 25\%$, untrusted ones loose nearly 50%.[7] This certainly enhances the indirect effect for distrusted documents ($\approx 6.4\%$), too.

When using personalized reference weights (function (5) with the weighted PageRank function (2), direct effects are neither expected nor found (changes $< 0.2\%$), but the indirect effects are noticeable: documents cited by trusted authors gain $\approx 5\%$, untrusted loose $\approx 6.3\%$.

Combining the node-based and edge-based approaches strengthens the personalization: using function (5) with (3) increases trusted documents and decreases distrusted documents by $\approx 35\%$ (direct). Moreover, it increases (decreases) documents cited by trusted (distrusted) documents by $\approx 7.5\%$. Combining function (5) with (4) gives the largest effect: trusted documents are increased by $\approx 38\%$ while distrusted documents fall down by $\approx 50\%$.[8] The cited documents gain $\approx 7\%$ respectively loose $\approx 8.1\%$.

This shows, that the personalization works as expected and gives the desired results. The combination of both approaches is feasible and useful.

5 The Framework in Practice

We show at the example of ranking scientific publications how the above presented framework can be put into practice. The main question is where to get the trust and the document reference network from. The document reference network with scientific publications can easily be obtained. Citeseer, for example, offers the metadata of all indexed papers for download. Based on the included reference lists, the document network can be built up. Besides of CiteSeer, there are further OAI (Open Archives Initiative) conforming repositories, often maintained at universities, which provide metadata of scientific publications.

A trust network between researchers is more difficult to obtain. Trust networks on the web are in domains such as product reviews (see e.g. epinions) or dating and finding business contacts (see e.g. orkut). A possibility would be to ask users to directly rate other users. However, this requires much effort from the participants' side. Users must not only think about whom they could give a trust rating but for building up a real network some of the rated users must

[7] This is not surprising as $t_{U \to X} = 0$ immediately gives a document visibility of 0 for all documents written by X, so these documents are all at the end of the ranking list.

[8] as more is not possible

indicate trust ratings, too. An alternative are semi-automatically extracted networks. [12] and [13] are two approaches that build social networks by analyzing publicly available data on the web with web-mining techniques. [12] demonstrated their approach by extracting a social network of the Japanese Society of Artificial Intelligence. [13] built a social network of Semantic Web researchers. In the approach presented by [12], the users are the contributors of the last few annual JSAI conferences. Co-occurrences of two users' names in webpages give a hint that there should be an edge between them. In addition, the relevance of an edge is calculated with a measure in the style of the Jaccard coefficient on the results of search engine queries. An edge is set if the Jaccard coefficient is above a threshold. In the next step, labels are assigned to the edges. Based on a content analysis of query results containing the user pair at issue and a set of classification rules, four labels are attributed: coauthors, members of the same institute, colleagues in a project, participants of the same workshop / conference. In their evaluation, [12] could show that they achieved a considerable precision. However, the recall was quite low. [12] derived from the social network a trust network. They calculated a sort of authoritativeness, namely a global trust value for each node with a kind of weighted PageRank. Based on this global trust, individual trust between two users was inferred. Users could then be asked to enhance these automatically calculated values and to add distrust.

Having extracted a social network with web-mining techniques, this network can be connected with the document reference network by matching the users in the trust network to the authors indicated in the metadata of the documents.

6 Conclusion

In the paper we extended a recently presented framework for document rankings with a comprehensive personalization strategy. The personalization is based on a second source of information besides of the document network: a trust network. The trust ratings between authors of documents are used in two ways: first, the requesting user's trust in the author influences the visibility of documents written by this author. Secondly, the weights of references (which give a hint on the semantics of a reference, i.e., whether a reference is supportive or depreciatory) are modified by the requesting user's trust in the citing author. In a simulation study we showed the effects of the trust-weighted visibility functions (applied individually as well as in combination) on the personalization. We can state that the strongest personalization can be achieved by combining both personalization strategies. Last but not least, we conclude that using trust propagation in the trust network as well as visibility propagation in the document network, visibilities can be highly personalized for a considerable fraction of the documents even though trust networks are rather small in size and in the number of trust ratings: our trust network encompassed 100 authors, with the requesting user trusting only 5 and distrusting also only 5 authors, and the personalization showed considerable effects. The next step is of course to put the framework into practice as sketched in this paper.

References

1. Stein, K., Hess, C.: Information retrieval in trust-enhanced document networks. In Ackermann, M., Berendt, B., Grobelnik, M., Hotho, A., Mladenic, D., Semeraro, G., Spiliopoulou, M., Stumme, G., Svatek, V., van Someren, M., eds.: Semantics, Web, and Mining. European Web Mining Forum, EMWF 2005, and Knowledge Discovery and Ontologies, KDO 2005, Porto, Portugal, October 2005, Revised Selected and Invited Papers. LNAI 4289. Springer (2006)
2. Avesani, P., Massa, P., Tiella, R.: A trust-enhanced recommender system application: Moleskiing. In: SAC '05: Proceedings of the 2005 ACM symposium on Applied computing. (2005) 1589–1593
3. Golbeck, J., Hendler, J.: Filmtrust: Movie recommendations using trust in web-based social networks. In: Proceedings of the IEEE Consumer Communications and Networking Conference. (2006)
4. Hess, C., Stein, K., Schlieder, C.: Trust-enhanced visibility for personalized document recommendations. In: Proceedings of the 21st Annual ACM Symposium on Applied Computing, Dijon, France (2006)
5. Korfiatis, N., Naeve, A.: Evaluating wiki contributions using social networks: A case study on wikipedia. In: First on-Line conference on Metadata and Semantics Research (MTSR'05). (2005)
6. García-Barriocanal, E., Sicilia, M.A.: Filtering information with imprecise social criteria: A foaf-based backlink model. In: Fourth Conference of the European Society for Fuzzy Logic and Technology EUSFLAT, Barcelona, Spain (2005)
7. Page, L., Brin, S., Motwani, R., Winograd, T.: The pagerank citation ranking: Bringing order to the web. Technical report, Stanford Digital Library Technologies Project (1998)
8. Kleinberg, J.M.: Authoritative sources in a hyperlinked environment. Journal of the ACM **46**(5) (1999) 604–632
9. Golbeck, J.: Computing and Applying Trust in Web-Based Social Networks. PhD thesis, Faculty of the Graduate School of the University of Maryland (2005)
10. Golbeck, J., Parsia, B., Hendler, J.: Trust networks on the semantic web. In: Proceedings of Cooperative Intelligent Agents, Helsinki, Finland (2003)
11. Ziegler, C.N., Lausen, G.: Spreading activation models for trust propagation. In: Proceedings of the IEEE International Conference on e-Technology, e-Commerce, and e-Service, Taipei, Taiwan, IEEE Computer Society Press (2004)
12. Matsuo, Y., Tomobe, H., Hasida, K., Ishizuka, M.: Finding social network for trust calculation. In: Proceedings of the ECAI 2004. (2004) 510–514
13. Mika, P.: Flink: Semantic web technology for the extraction and analysis of social networks. Journal of Web Semantics **3**(2) (2005)