# Information Retrieval in Trust-enhanced Document Networks

Klaus Stein, Claudia Hess

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

**Abstract.** To fight the problem of information overload in huge information sources like large document repositories, e. g. citeseer, or internet websites you need a selection criteria: some kind of ranking is required. Ranking methods like PageRank use the document reference network structure. However, these rankings do not distinguish different reference semantics. We enhance these rankings by incorporating information of a second layer: the author trust network to improve ranking quality and to enable personalized selections.

## 1 Introduction

The amount of information accessible for everyone is increasing rapidly, mainly driven by computer mediated communication technologies, namely the internet respectively the www. New websites appear, messages are posted, blogs written, papers published etc. No one is able to keep an overview of all these information sources or to find interesting information by herself, so search engines are an important tool to fight the problem of information overload, to "bring order to the web", as the google-programmers Brin and Page [1] call it.

A search engine carries out three important tasks to do its job:[1]

**information gathering:** it crawls the web to collect as much of its content as possible, building a huge repository.

**information selection:** for each search query it selects the subset of corresponding webpages (e. g. webpages containing a given keyword). For common search terms this subset may contain up to many million websites,[2] so a third step is needed.

[1] In praxis the described steps are highly interdependent, the data structures build up in step one are optimized for fast access in step two and three, and selection and ranking may be done in one step.

[2] querying google for "christmas" gives 45 400 000 pages and even "ontology" gives 4 390 000 pages (http://www.google.de/ at July 23. 2005).

**information ranking:** the matching webpages are sorted by some ranking and only the highest ranked pages are presented to the user.

We focus on the task of information ranking. We will use link analysis to rank documents from a set of selected documents. The prominence (the "visibility") of documents such as websites or scientific papers is calculated based on an analysis of the structure of the underlying document collection, in our case the web. In contrast to content-based analysis, web structure mining analyzes references (links or citations) between documents and computes the visibility of a document based on this link analysis.

The basic idea behind these measures is that a webpage, a scientific paper, a journal etc. will be important if cited by many other important pages/papers/journals. Thus its visibility depends on the number of references and the visibility of the referencing documents. This seams feasible (and works well as the success of google shows) but ignores the semantics of the reference: a scientific paper may cite another one because the author finds it useful and supports the work, but also if the author disagrees and wants to give an opposite point of view.[3]

While the semantics of a reference is fairly obvious for the reader of a paper, it is not acessible for a search engine (which simply does not *understand* the text). We claim that we can reincorporate link semantics to some degree by using a second resource: an author trust network. Additionally this will allow to do personalization of rankings based on user's trust.

The rest of the paper is structured as follows: section 2 explains how rankings of documents are computed by exploiting the structure of the document reference network. Section 3 introduces the author trust network which will be combined with the document reference network in the subsequent section, i.e. section 4. A new approach to capture the reference semantics and to compute the adapted visibility is presented in this section. Section 5 summarizes our work and highlights areas for future research.

## 2 Document Reference Network based Ranking

One important measure (besides content-based ratings) to rank webpages or other resources referencing each other is to use the link structure to determine

[3] On the web links of disagreement are seldom set because due to the way search engines rank pages each reference set to a page increases its rank which is not what the disagreeing author normally wants (in context of the discussion around link spam in guestbooks a (controversial) new link attribute `rel="nofollow"` was introduced by Google, MSN Search, Yahoo! and others to mark links not to be counted by ranking algorithms, which also could be used for disagreement links in future).

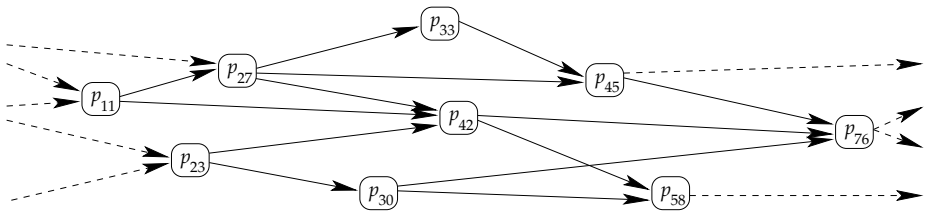


**Fig. 1.** The visibility $\text{vis}_{42}$ of document $p_{42}$ is computed using the visibilities $\text{vis}_{11}$, $\text{vis}_{23}$ and $\text{vis}_{27}$ of $p_{11}$, $p_{23}$ and $p_{27}$: $\text{vis}_{42} = \text{vf}_{p_{42}}(p_{11}, p_{23}, p_{27})$, and $\text{vis}_{42}$ itself contributes to $\text{vis}_{58}$ and $\text{vis}_{76}$. Using PageRank as visibility function $p_{11}$ and $p_{23}$ only count half because they also contribute to $p_{27}$ and $p_{30}$ resp. while $p_{27}$ counts one-third for also contributing to $p_{33}$ and $p_{45}$; for other algorithms this can be different.

the visibility of a certain document. Specifying the structural importance of a page within a document network is a well known problem in social network theory (for an overwiew see [2–5]).[4]

In 1976, Pinski and Narin [6] computed the importance (rank) $r_a$ of a scientific journal $p_a$ by using the weighted sum of the ranks $r_k$ of the journals $p_k$ with papers citing $p_a$. A slightly modified version of this algorithm (the pagerank algorithm) is used by the search engine google [7, 1] to calculate the visibility of webpages (with $\text{vis}_a$ the visibility of a webpage/document $p_a$):

$$\text{vis}_a \quad = \quad (1-\alpha) \quad + \quad \alpha \sum_{p_k \in R_a} \frac{\text{vis}_k}{|C_k|}$$

where $R_a$ is the set of pages citing $p_a$ and $C_k$ is the set of pages cited by $p_k$. Therefore each page $p_k$ contributes by $\frac{\text{vis}_k}{|C_k|}$ to the visibility of $p_a$.

For $n$ pages this gives a linear system of $n$ equations. Solving this equation system is possible but (for large $n$) very expensive, so an iterative approach is used. First all $\text{vis}_i$ are set to some default value and then the new values $r'_i$ are calculated repeatingly until all $\text{vis}_i$ converge[5] (see Fig. 1).

The PageRank algorithm works best for networks with cyclic reference structures (links between webpages, citations between journals). For mostly acyclic structures like citations in scientific papers (where documents are temporally ordered and citations go backward in time) similar but slightly different measures are used, which may additionally incorporate metadata like documents age (see e. g. [8]).

[4] From a mathmatic point of view a document reference network simply is a directed graph with documents (webpages, papers, ...) as vertices and references (links, citations) as edges.

[5] for a discussion of convergence problems in leaves see [1]

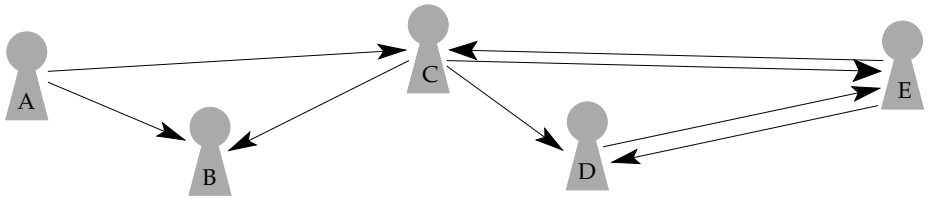


**Fig. 2.** Author trust network

## 3 A Second Source of Information: The Author Trust Network

The information extracted from the document reference network is enhanced with information from a second source of information, the trust network between the authors of the documents. Both networks are independent from each other. Neither the structure of the document network is induced by the trust network nor vice versa. Considering information from an author trust network, it can be distinguished whether a document has a high rank due to its usefulness or because it is very controversially discussed. Authors are connected to others via trust statements indicating that the source of the trust statement has a certain degree of trust in the capabilities of the target to provide 'good' documents, for instance to write excellent scientific papers with a well-elaborated argumentation and a 'sensible' opinion from the source's point of view. Trust statements range from blind trust to absolute distrust represented as numerical values $t \in [t_{\min}, t_{\max}]$, normally $t \in [-1, 1]$. Based on direct trust statements between authors, trust relationships can be interpolated between authors who are indirectly connected. This reflects human behavior in so far as we trust to some extent the friends of our friends. A number of trust metrics have been proposed such as the path algebraic trust metric by Golbeck et al. [9] or the spreading activation strategy by Ziegler and Lausen [10]. Most trust metrics are limited to trust values between 0 (no trust), and 1 (maximum trust). Guha et al. [11] discuss research issues on distrust also concerning the development of a metric that is able to cope with trust and distrust.

Trust networks have attracted much attention in the last years as basis for recommender systems in research as well as in commercial applications. Well-known examples are the Epinions[6] platform for consumer reviews and ebay's[7] reputation system. Theoretical foundations and research projects on trust-based recommender systems are provided for instance by Guha [12], Montaner et al. [13] and Kinateder and Rothermel [14].

Trust data becomes the more and more accessible online. For instance, the Friend-Of-A-Friend (FOAF) vocabulary has been extended to include trust statements[8]

[6] http://www.epinions.com
[7] http://www.ebay.com
[8] see the ontology for trust ratings at http://trust.mindswap.org/ont/trust.owl

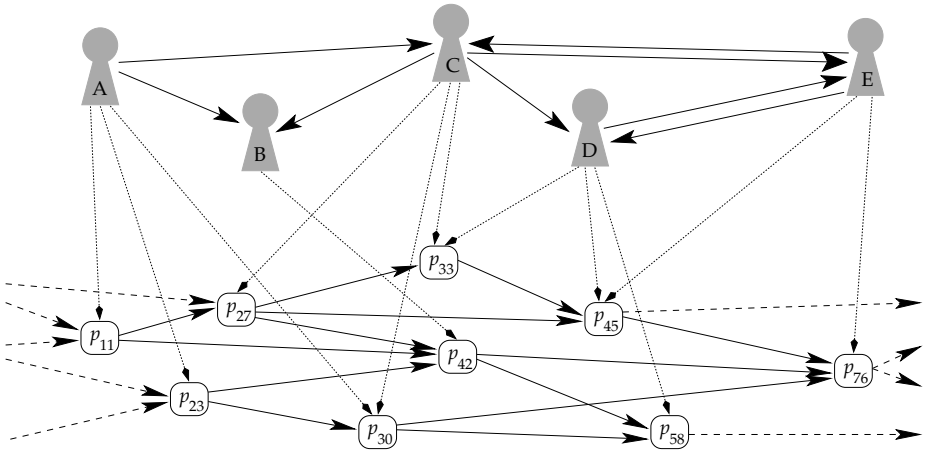


**Fig. 3.** Combining trust and document reference network.

by Golbeck et al. [9]. Users can express in the FOAF files, provided for example at their personal homepages, not only whom they consider as friends but their degree of trust, too.

If we know that author $A$ trusts author $B$ in the sense that $A$ likes $B$'s point of view and the quality of her work we can assume that a reference in a paper of $A$ to a paper of $B$ supports this paper. On the other hand if an author $C$ distrusts an author $D$ (e. g. $D$ is a creationist and $C$ a darwinist) a reference in a paper of $C$ to one of $D$ normally does not support $D$'s paper. This certainly will not hold for all references but works well enough to improve network reference based rankings.

## 4 Trust-based Visibility: A Two-Layer Approach

Figure 3 shows a two-layered network with the author trust network at the top and the document reference network at the bottom.[9] In the document network information is located in the vertices (the documents' visibilities) while in the trust network information is located in the edges (with $t_{A \to B}$ the value of the trust edge from $A$ to $B$).

### 4.1 Propagating Trust to the Document Reference Network

In the first step, the trust values are propagated to the document reference network edges by identifying each document with its author and attributing each

[9] To simplify the example we show an acyclic part of a document network. All algorithms will work on arbitrary graphs, e. g. webpages.

reference with the corresponding trust value (e. g. the edge $e_{11\to 42}$ from $p_{11}$ to $p_{42}$ is attributed with $t_{A\to B}$ for $A$ and $B$ being the authors of the referencing respective referenced document (note that coauthorship maps more than one trust value to one reference).

Now the visibility of a document $p_a$ can be calculated depending on the visibility of the documents referring $p_a$ *and* on the trust attributes of these references using a trust-enhanced visibility function $\mathrm{vf}^t$:

$$\mathrm{vis}_a = \mathrm{vf}^t_{p_a}(R_a, E_a) \qquad \text{with } E_a = \{e_{x\to a} \mid p_x \in R_a\}$$

with $R_a$ the set of documents referencing $p_a$ and $E_a$ the set of attributed edges from $R_a$ to $p_a$. As the next sections show, there are different ways to model how these trust edges contribute to $\mathrm{vf}^t$.

### 4.2 Associating Weights to Document References

The next step to build trust-enhancing visibility functions is to turn the *attributed* into *weighted* edges. This step is necessary in order to deal with negative trust values as used for distrust (note that in most applications dealing with distrust, trust values $t$ are in $t \in [-1, 1]$).

We start with $\bar{e}_{i\to j}$ being the average of all trust values attributed to $e_{i\to j}$ (or a fixed value $\bar{e}_{\text{default}} \in [t_{\min}, t_{\max}]$ if $e_{i\to j}$ is empty). As negative weights cause problems during propagation, the attributes on the edges $\bar{e}_{i\to j}$ cannot directly be used as weights if $t_{\min} < 0$, but we can compute the weight $w_{i\to j}$ from $\bar{e}_{i\to j}$.

This is done by a mapping function $\begin{aligned} I : [-1, 1] &\to [0, 1] \\ w_{i\to j} &= I(\bar{e}_{i\to j}). \end{aligned}$

By choosing different mapping functions, different trust semantics can be established. So one can decide how the trust values influence the weights and whether the impact of negative edges (distrust) should be small or large:

$I_+$: $w_{i\to j} = \Delta + \bar{e}_{i\to j}, \qquad \text{with } \Delta > -t_{\min}$

guarantees non-negative weights, but (depending on $\Delta$) weights may get values greater 1.[10]

$I'_+$: $w_{i\to j} = \dfrac{\Delta + \bar{e}_{i\to j}}{\Delta + t_{\max}}, \qquad \text{with } \Delta > -t_{\min}$

guarantees $w_{i\to j} \in (0, 1]$.

$I_{||}$: $w_{i\to j} = |\bar{e}_{i\to j}|$

highly trusting and highly distrusting references give the same (high) weight,

[10] this is ok for the trust-aware visibility function described in Sec. 4.3 but not for others

$I_\lambda$: $w_{i\to j} = \left\{ \begin{array}{lll} \bar{e}_{i\to j} & & \text{for } \bar{e}_{i\to j} \geq 0 \\ -\lambda \bar{e}_{i\to j} & (\lambda \in (0,1)) & \text{otherwise} \end{array} \right\}$

reduces the influence of distrust references (e. g. with $\lambda = 0.5$ they only contribute half),

$I_0$: $w_{i\to j} = \max\{0, \bar{e}_{i\to j}\}$

distrust references give zero weights.

The weighting functions $I_+$ and $I'_+$ give smallest weights for negative edges, with $I_0$ all negative edges give zero weights, and with $I_\lambda$ and $I_{||}$ they give positive contribution.

Now these weights can be used to create trust-enhanced visibility functions. In the most simple case, weights are directly used to modulate any reference network based visibility function (by multiplying the visibility contributed by $e_{i\to j}$ from $p_i$ to $p_j$ by $w_{i\to j}$) making it trust-aware. Note that the average visibility distributed is lowered by this (for $t_{\min}, t_{\max} \in [-1, 1]$), except for $I_+$.[11]

### 4.3 Weighted PageRank

According to the PageRank, a page $p_r$ referencing $k$ other pages $p_{r_1}$ to $p_{r_k}$ contributes with $\frac{\text{vis}_r}{k}$ to each of the referenced pages. By modulating the contribution by edge weights $w_{r\to r_i}$, we get the contribution[12]

$$\text{vis}_{p_r \to p_{r_i}} = \frac{w_{r\to r_i}}{\sum_{p_{r_j} \in C_r} w_{r\to r_j}} \text{vis}_r \,.$$

Inserting this into PageRank[13] we obtain the trust-aware visibility function[14]

$$\begin{aligned} \text{vf}^t_{p_a}(R_a, E_a) &= (1-\alpha) + \alpha \sum_{p_k \in R_a} \text{vis}_{p_k \to p_a} \\ &= (1-\alpha) + \alpha \sum_{p_k \in R_a} \frac{w_{k\to a}}{\sum_{p_j \in C_k} w_{k\to j}} \text{vis}_k \end{aligned}$$

[11] by introducing reference weights convergence of a given visibility function is not longer granted. Renormalization of all visibilities may hence be necessary in each iteration step.

[12] Note that even for $w_{r\to r_i} > 1$ the fraction is below 1, so $I_+$ can be used as weighting function.

[13] This modification is not restricted to PageRank, other visibility functions are adaptable accordingly.

[14] to increase efficiency the fraction $\frac{w_{k\to a}}{\sum_{p_j \in C_k} w_{k\to j}}$ should be precalculated once for all references of the whole net

Note that by this the amount of visibility distributed from one page is not changed, but now some references gain more and others less. So for a page with only one outgoing reference nothing changes.

The way and the strength to which trust values contribute to the visibility of a document can be customized by changing the weighting function. Note that this algorithm ensures that the same amount of visibility is distributed as for original PageRank, albeit the amount for single edges change.

Using $I_+$ as weighting function, the influence of trust can be finetuned by changing $\Delta$: for $\Delta \rightarrow -t_{\min}$ only references with high trust contribute while with $\Delta \rightarrow \infty$ we get the original PageRank.[15] This definition of $\mathrm{vf}^t$ supports pages important within a certain community.[16] A page gaining many references from within the authors' community (giving high trust values) raises visibility while a page referenced from outside (low trust values, distrust) decreases visibility. This may be feasible to some account but will not fit any users needs. One may claim: "the best ranked papers are those with only supporting references, but for me controversial papers are of greater interest. And anyway, I want to get the important papers of *my* community, not of others."

This motivates to consider two further aspects in the adapted visibility functions. On the one hand, a trust-enhanced visibility function that favors controversially discussed documents is required. On the other, a personalization of the visibility function permits to better match the users' individual needs.

### 4.4 Controversial References

In contrast to the weighted PageRank, the next alternative reflects the assumption that controversially referenced papers are the most interesting papers. The weighted PageRank, examining $\{\bar{e}_{i\rightarrow a} \mid p_i \in R_{p_a}\}$ for a given document $p_a$, tells whether $p_a$ is mainly supported, deprecated or controversial. The following function provides the highest ranks for the most controversially referenced documents (we assume $t_{\min} = -1, t_{\max} = 1$).[17] The new visibility function consists of two components, one for the visibility weighted with trust values and one indicating how controversially a paper is referenced:[18]

$$\widehat{\mathrm{vis}}_a = \sum_{p_i \in R_a} \frac{(1 - \beta + \beta w_{i\rightarrow a})\,\mathrm{vis}_i}{|C_i|},$$

[15] with $I_+$ : $\dfrac{w_{a\rightarrow a_i}}{\sum_{p_{a_j}\in C_a} w_{a\rightarrow a_j}} = \dfrac{\Delta + \bar{e}_{a\rightarrow a_i}}{k\Delta + \sum_{p_{a_j}\in C_a} \bar{e}_{a\rightarrow a_j}}$

[16] with community being authors trusting each other

[17] without distrust ($t_{\min} = 0$) we would not have controversals

[18] We need a weighting function ensuring $w_{i\rightarrow a} \in [0, 1]$, so $I_+$ can not be used here.

$$\widehat{\delta e}_a = \frac{\sum_{p_i,p_j \in R_a} (\bar{e}_{i\to a} \; - \; \bar{e}_{j\to a})^2 \, \mathrm{vis}_i \, \mathrm{vis}_j}{\left(\sum_{p_i \in R_a} vis_i\right)^2}$$

Now $\widehat{\mathrm{vis}}_a$ gives the trust weighted visibility (with trust influence factor $\beta \in [0,1]$), and $\widehat{\delta e}_a$ the degree of controversy. Therefore

$$\widetilde{\mathrm{vis}}_a \quad = \quad \gamma_{\widehat{\mathrm{vis}}} \, \widehat{\mathrm{vis}}_a \; + \; \gamma_{\widehat{\delta e}} \, \widehat{\delta e}_a$$

gives a trust weighted visibility where trust and controversy influence can be finetuned by changing $\beta$, $\gamma_{\widehat{\mathrm{vis}}}$ and $\gamma_{\widehat{\delta e}}$ which can be used to make a reference network based visibility function trust-aware (e. g. $\mathrm{vis}_a = (1-\alpha) + \alpha \, \widetilde{\mathrm{vis}}_a$ for PageRank).

One inaccuracy is left in this model of controversy for we just used average values $\bar{e}_{i\to j}$ for $\widehat{\delta e}$, which would not give a controverse for a referencing document written by two (or) more coauthors with very different trust values regarding the author of the referenced document. You can claim that we simply do not know why the authors decided to set this reference, therefore taking the average is feasible (as done above). Alternatively the discrepancy can be modelled by changing $\widehat{\delta e}$ to not using the average $\bar{e}_{i\to j}$ but all single trust values attributed to $e_{i\to j}$.

### 4.5 Personalized Trust Visibility Rankings

The visibility measure described in the last section points out controversial papers but does not incorporate a personal view. One additional step gives this personalization: assume we have edges $\bar{e}_{i\to k}$, $\bar{e}_{j\to l}$ of high trust with $A$ being author of $p_i$ and $B$ author of $p_j$. A user $U$ trusting $A$ and distrusting $B$ will not agree that both edges count equal. A reference from a personally distrusted author simply should count less. So we modify an edge $\bar{e}_{i\to k}$ by $t_{U\to A}$:

$$\bar{e}'_{i\to k} \quad = \quad \frac{t_{U\to A} - t_{\min} + b}{t_{\max} - t_{\min} + b} \, \bar{e}_{i\to k} \qquad \text{with } b \geq 0$$

Now the influence of each reference directly depends on the trust of the user $U$ in the referencing author $A$ (the degree of this dependence is moderated by $b$ with maximum influence for $b = 0$).

These personalized edges substitute the weighted edges from Sec. 4.2. Personalized edges can be used in any trust-aware visibility function $\mathrm{vf}^t$ including but not restricted to the weighted PageRank algorithm (Sec. 4.3) as well as the controverse-sensitive algorithm of Sec. 4.4.

## 5 Conclusion and Outlook

In the paper we presented a framework for extending visibility functions to include trust information. The structural information from the document reference network which serves as basis for visibility measures is combined with data from a second source of information, the trust network between the authors of the documents. In contrast to visibility functions as used up until now in structural web mining, trust-enhanced visibility functions encompass two novel aspects: on the one hand, they deal with the semantics of the references. A reference can be made due to agreement or disagreement. This is reflected in the proposed visibility functions by considering the trust relationship between authors. Alternative functions have been proposed which permit requesting users to obtain a ranking that corresponds to their information need: papers which are widely agreed on could be favored by a trust-enhanced visibility function or controversially referenced ones by an alternative one. On the other, integrating trust information permits to personalize the ranking.

An alternative consists in integrating trust data in a visibility function propagating trust together with the visibility in the document reference network. Instead of having trust relationships between authors, trust relationships between the readers of documents are considered. As a consequence trust relationships are not propagated to the edges but are included as an additional component in the visibility functions. This alternative is explored in a companion paper.

Having addressed the basic theoretical foundations of trust-enhanced visibility functions in this paper, we currently evaluate the interaction between the document and the trust network in a simulation study[19]. In future work evaluation with real data will complement the simulation. The main problem on real world evaluation is to get a trust network built up independently from a document network. As document network for instance the citeseer document collection could be used.

Another important step in future work is to have a closer look on the author trust network. Currently, trust edges are simply projected from the trust to the document network and $\widehat{\delta e}_a$ measures differences in trust weighted references to $p_a$. In a next step trust edges between referencing authors should be directly taken into account: if author $A$ with paper $p_i$ and author $B$ with paper $p_j$ both cite $p_k$ of author $C$ also the trust edges from $A$ to $B$ and vice versa are interesting. $A$ trusting $C$ *and* $B$ trusting $C$is a more important sign if $A$ and $B$ *distrust* each other. Evaluating all possible triades of authors (trust/distrust edges) in the author trust network is work in process.

[19] using the COM/TE framework [8], http://www.kinf.wiai.uni-bamberg.de/COM/.

## References


1. 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)
2. Watts, D.J.: Six Degrees: The Science of a Connected Age. W. W. Norton & Company, Inc. (2003)
3. Barabási, A.L.: Linked: How Everything Is Connected to Everything Else and What It Means for Business, Science, and Everyday Life. PLUME (2003)
4. Menczer, F.: Growing and navigating the small world web by local content. In: Proceedings of the National Academy of Sciences of the United States of America. Volume 99. (2002) 14014–14019
5. Newman, M.E.J.: The structure and function of complex networks. SIAM Review **45** (2003) 167–256
6. Pinski, G., Narin, F.: Citation influence for journal aggregates of scientific publications: Theory, with application to the literature of physics. Information Processing & Management **12** (1976) 297–312
7. Brin, S., Page, L.: The anatomy of a large-scale hypertextual Web search engine. Computer Networks and ISDN Systems **30** (1998) 107–117
8. Malsch, T., Schlieder, C., Kiefer, P., Lübcke, M., Perschke, R., Schmitt, M., Stein, K.: Communication between process and structure: Modelling and simulating message-reference-networks with COM/TE. JASSS (2005) submitted.
9. Golbeck, J., Parsia, B., Hendler, J.: Trust networks on the semantic web. In: Proceedings of Cooperative Intelligent Agents, Helsinki, Finland (2003)
10. 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)
11. Guha, R., Kumar, R., Raghavan, P., Tomkins, A.: Propagation of trust and distrust. In: WWW '04: Proceedings of the 13th international conference on World Wide Web, New York, NY, USA, ACM Press (2004) 403–412
12. Guha, R.: Open rating systems. Technical report, Stanford Knowledge Systems Laboratory, Standford, CA, USA (2003)
13. Montaner, M., López, B., Lluís de la Rosa, J.: Opinion-based filtering through trust. In Ossowski, S., Shehory, O., eds.: Proceedings of the Sixth International Workshop on Cooperative Information Agents. Volume 2446 of LNAI., Madrid, Spain, Springer Verlag (2002) 188–196
14. Kinateder, M., Rothermel, K.: Architecture and algorithms for a distributed reputation system. In: Proceedings of the First International Conference on Trust Management. Volume 2692., Springer Verlag (2003) 1–16