Beyond the Web: Retrieval in Social Information Spaces

Sebastian Marius Kirsch, Melanie Gnasa, and Armin B. Cremers

Institute of Computer Science III, University of Bonn, Römerstrasse 164,
53117 Bonn, Germany
{kirschs,gnasa,abc}@cs.uni-bonn.de

Abstract. We research whether the inclusion of information about an information user's social environment and his position in the social network of his peers leads to an improval in search effectiveness.
Traditional information retrieval methods fail to address the fact that information production and consumption are social activities. We ameliorate this problem by extending the domain model of information retrieval to include social networks.
We describe a technique for information retrieval in such an enviroment and evaluate it in comparison to vector space retrieval.

1 Introduction

In the late 1990s, the field of information retrieval rose to meet new challenges posed by the ubiquitous nature of the world wide web: Information retrieval in an environment where individual documents are not characterized only by their content, but also by their relationship to other documents. By the means of hyperlinks, a web author can express associations with other authors' documents that may reside anywhere on the web. Successful techniques for this task are primarily characterized by their reliance on the spectral properties of the web graph. Prime examples are the PageRank algorithm [1] and the HITS [2] algorithm, both of which represent first-order approximations of matrices derived from the web graph: The adjacency matrix in the case of PageRank, and the bibliographic and co-citation coupling matrices in the case of HITS. At the core of both algorithms is an acknowledgement of the democratic and social nature of the web: A human author's act of including a hyperlink to another page is an act of social interaction. A hyperlink expresses an endorsement of the page that is linked to. The sum of all hyperlinks is used to determine the relative importance of all pages – as a sum of judgements made by humans. This idea revolutionized the the field of web retrieval and shaped the nature of web retrieval systems for years to come.

However, the nature of the web has changed since the inception of spectral retrieval techniques. Whereas previously, web pages were crafted as individual documents, nowadays many web pages amount to nothing more than user interfaces: interfaces to an underlying database, an underlying information space

that is made accessible via the web. Many of these information spaces model social relationships between their participants in a much more direct manner than one could glean from analyzing the surface hyperlink structure of the interface. A logical next step is directly analyzing the social structure of the information space. This social structure may then be used for the purpose of information retrieval. This paper presents an attempt to leverage social networks for information retrieval, in environments that do not follow the usual presumptions made in web retrieval.

The rest of this paper is structured as follows: Sect. 2 lists related work. Sect. 3 describes social software on the web. Sect. 4 gives an extended information retrieval models which includes social relations, and explains a retrieval technique based on this model. Sect. 5 presents the evaluation of this technique. Sect. 6 concludes the paper.

2 Related Work

Google was one of the first web search engines to incorporate analysis of the web graph into its ranking algorithms. The PageRank algorithm [1] was a novelty among search engines at the time and was quickly singled out among independent observers as the main factor for its success. The impact of PageRank on the quality of Google's search results is not known; as is common for a web search engine, the innards of its scoring algorithm are kept secret. Evidence for the importance of PageRank in web retrieval is still scarce: According to [3], only 11 of 74 submitted runs at the TREC-2004 'Web' track used PageRank, and only one of the top systems used it.

ReferralWeb [4, 5] is a system for mining social relations from the web and exploring social networks. The authors describe it as 'combining of social networks and collaborative filtering'; its focus is extracting a social network from web pages, finding experts for a topic and linking the searcher to the expert by a path in the social network. ReferralWeb differs from other social networking applications because it extracts social links from publicly available information on the web; it does not require the user to sign up with a service and explicitly name his colleagues and collaborators. A formal evaluation of ReferralWeb's effectiveness, as compared to other information retrieval systems, was not conducted to our knowledge.

I-SPY [6] is an experimental meta search engine developed at University College, Dublin, Ireland. I-SPY implements collaborative ranking, borrowing ideas from collaborative filtering: It aggregates relevance judgements from a community of people and uses them in later searches for the same keywords to boost pages which are known to be good. Users are required to join a specific community before executing a query; one user can only be part of one community at a time, requiring the user to change the community as the subject matter of his search changes. I-SPY does not facilitate the formation of a community. It does not use

information about the social relations between its users, and does not facilitate the formation of such relations.

ISKODOR is a prototype system developed at University of Bonn aiming to combine three aspects of user-centred information retrieval: personalization, collaboration and socialization. These principles form three pillars of a new web search paradigm called 'congenial web search' [7]. The framework supports a common representation of documents, queries, and relationships, which form individual context information of a user's search interest. The prototype employs a peer-to-peer architecture in order to share explicit feedback with other users. The user's faith in the service is strengthened, as he himself controls which information is stored and disseminated about him. Whereas personalized and collaborative aspects are already implemented in the prototype, research presented in this paper forms the basis of the social aspect of the system.

3 Social Software and the Web

One of the earliest applications of computer networks were electronic mailing lists and discussion groups. Precursors of the internet, for example BITNET and Usenet, already supported interaction and discussion among groups of users. Interaction between a large number of users is supported on these system at a negligible cost. Recently, the focus of social software has shifted from dedicated platforms to the web; popular examples include the following:

Wikis are a form of collaborative authoring environment that is characterized by the fact that every user can add, edit, and delete content at will. The first wiki was WikiWikiWeb, launched by Ward Cunningham in 1995 as a supplement to the Portland Pattern Repository, a web site about software design patterns. A number of software packages and similar projects followed; the largest wiki is purported to be Wikipedia, an online encyclopedia that employs the wiki principles. The quality of published content varies wildly; some wikis contain nothing more than a few quickly written ideas, others, like Wikipedia, aim for publication-quality content.

Blogs are an internet phenomenon originating in the late 1990s: Websites that continually publish new articles on their front page, written by one individual or a group of people. Blog entries can be tied to their author; linking between entries is supported in the form of comments or so-called 'trackback links', in which the author of another blog refers in his entry to the original entry. Blogs can take many forms, for example personal blogs, topical blogs or corporate blogs. Another typical feature is the so-called 'blogroll': A list of other blogs the author reads regularly. This may be used to determine social links between authors, but it is not universally adopted.

Social networking platforms like Friendster, orkut or openBC are dedicated web applications for the formation of social networks. Users have the ability to name their friends among the users explicitly and advertise them on a special page. Finding paths between two users in the social network is often supported, as are group discussions.

We call these systems 'social information spaces' [8] or 'social software'. The social interactions between users of these systems are hidden beneath the web front-end in databases, and thus are not directly accessible to web search engines. The resulting social network can be seen as the 'deep structure of the web'. Efforts of the Semantic Web initiative [9] aim to provide this information in machine-readable form, for example with the Dublin Core standard [10] for document metadata or the 'Friend of a Friend' standard [11] for expressing the relations between individuals.

With the increasing use of social software, social ties and the structure of the social network become tractable. In such a setting, incorporation of social networks into information retrieval processes is a desirable feature.

4 Models and Techniques for Social Retrieval

Social information retrieval systems are distinguished from other types of IR systems by the incorporation of information about social networks and relationships into the information retrieval process. This feature necessitates an extended model for information retrieval, as well as new techniques that make use of social information.

4.1 Domain Model for Social IR

The traditional models for information retrieval concern themselves with documents, queries, and their relations to each other: A document is relevant to a query, a document references other documents, a query is similar to other queries. Likewise, social network analysis models individuals and their relations with each other. Information retrieval systems traditionally do not model individuals, neither in their role as users of the system, nor as authors of the retrieved documents, and social networks do not incorporate retrievable content.

Social IR combines the models of information retrieval and social networks with each other. By incorporating individuals into the model, we gain a greater insight into their role in the information retrieval and production process (Fig. 1). New associations between the entities become apparent: Individuals appear in their role as information producers or information consumers, queries relate to an individual's information needs, or describe a topic about which an individual possesses knowledge.

A social IR system is characterized by the presence of all three types of entities: documents, queries, and individuals. Most systems will only use a subset of the possible associations between the entities, depending on the domain of the

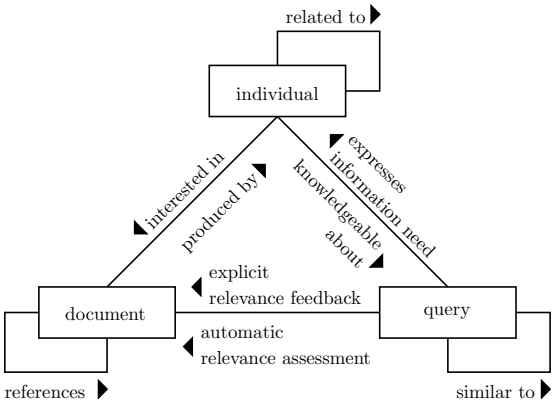

Fig. 1. A domain model for social information retrieval

system. Modeling the relations between individuals is mandatory for a social IR system; all other types of associations are optional, as long as all three entities have an association with at least one other.

This characterization of a social IR system raises the question of suitable domains for such a system. The world wide web in its current state is evidently not a suitable domain: It lacks reliable authorship information, as well as information about social relations between authors. The increasing use of machine-readable metadata – for example in the aforementioned Dublin Core and 'Friend of a Friend' standards – gives hope that this will change in the long run. An attempt at mining social relations from the web is described in [4, 5].

Subsets of the web provide more suitable domains. The entirety of blog sites on the web (often called 'blogosphere') is one such domain: Blog entries can usually be associated with an author, and via comments or so-called trackback links, communication between blog authors can be ascertained, leading to a social network. Such information is usually not available in machine-readable format and has to be extracted using information extraction techniques. Some blogging services, for example LiveJournal [12], already provide it in machine-readable form. Wikis are also an environment that allows to ascertain authorship of a document, usually via the revision history. Interaction between users can be determined by co-authorship, or by discussions on dedicated talk pages; however, this information is often not portable between different wikis. Direct access to the underlying database often makes extraction of this information much easier.

For application of social IR to other domains, availability of the required information needs to be determined beforehand. Specialized techniques may have to be employed in order to extract it. We do however surmise that similar characteristics govern the structure of all social information spaces, and that similar techniques are applicable.

Traditional information retrieval techniques which are based solely on analysing document content, while very successful in many contexts, fail badly when the

information need is underspecified, and when a large number of relevant documents exist. In this sense, social IR can be understood as a formalization of search techniques we commonly use to assess the quality of information – by looking at the author's standing in his community.

We use an associative network as the underlying representation. An associative network is a graph of information items, with unlabeled, weighted, directed or undirected edges ('associations') between nodes. In agreement with the domain model, we use three kinds of nodes: for individuals, documents, and queries.

Definition 1. *For a set of individuals I, a set of documents D, and a set of queries Q the domain is represented by a weighted, directed graph $G = (V, E)$, where $V = I \uplus D \uplus Q$ and $E \subseteq V \times V$. A weight matrix $C \in \mathbb{R}_{\geq 0}^{|V| \times |V|}$ contains the weight of the edges. For edges between individuals $e \in I \times I$, the weight function expresses the strength of a social relationship between two individuals.*

We use this domain for retrieval of documents from keyword queries. This task is the most common task in information retrieval, which ensures comparability with other systems. Systems that store associations between users and queries, or between queries and documents, are mostly found in the experimental field of personalized and collaborative retrieval; they have not found their way into the mainstream of IR yet.

4.2 Techniques for Social IR

The domain model presented in the last section is able to accomodate many aspects of social information retrieval. We concentrate on retrieval of documents from keyword queries in an environment where authorship information is available.

A central idea is that the authority of an author can be inferred from his position in the social network, and that this authority measure can be applied to the documents he authored. Whether a document is relevant to a query can be determined using conventional IR techniques. A social IR system for this task is therefore composed of two parts: An authority measure for individuals in the social network, and a relevance measure for documents as regards queries. Both measures are combined to provide an improved ranking of documents.

In our experiments, we evaluate the use of PageRank as an authority measure for graphs. PageRank [1] is one of the most well-known algorithms for link analysis. In web retrieval, the PageRank algorithm is usually formulated based on a random surfer model: A user starts on a random web page and follows one outlink of this page at random and repeats this process on every page he reaches. Assuming that the link graph consists of a single strongly connected component (ie. there is a path from every page to every other page), the random surfer will eventually visit every page in the web graph. One may consider this sequence of pages as a Markov chain and compute the stationary probability of the random surfer being on a given page at any time. The stationary probability may be computed using linear algebra methods: Let A be the adjacency matrix of the

Table 1. PageRank scores for the coauthorship network of the SIGIR corpus. Scores are normalized and are computed with a teleportation probability of $\epsilon = 0.3$.

rank name	PageRank
1. Bruce W. Croft	7.929
2. Clement T. Yu	4.716
3. James P. Callan	4.092
4. Norbert Fuhr	3.731
5. Susan T. Dumais	3.731
6. Mark Sanderson	3.601
7. Nicholas J. Belkin	3.518
8. Vijay V. Raghavan	3.303
9. James Allan	3.200
10. Jan O. Pedersen	3.135

web graph G. Let M be a row-normalized version of A, that is $(M)_{ij} = \frac{(A)_{ij}}{\sum_k (A)_{ik}}$. Then the PageRank vector $\boldsymbol{r}$ is the maximal eigenvector of

$$\left(\frac{\epsilon}{|V|}\mathbf{1} + (1-\epsilon)M\right)^{\top} ,$$

provided that G is ergodic [13]. ϵ is the 'bias': The probability that the random surfer will teleport to a random page instead of following an outlink.

In order to get an idea of the application of PageRank to a social network, it is instructive to compute the PageRank scores for a well-known social network. We computed PageRank scores for a coauthorship network extracted from 25 years of SIGIR proceedings (from 1978–2003); the ten highest-ranking authors are listed in Tab. 1.

In social IR, we apply the PageRank algorithm to the social network, ie. the graph $G[I]$. We compute a PageRank score r_i for every node i in the social network. We ignore the fact that several disconnected components may exist in the social network: Since they are small compared to the giant component, they can be expected to contribute little to the document set, which means that documents produced by individuals not in the giant component will only be relevant for very few of the expected queries. We use a bias of $\epsilon = 0.3$, further ameliorating the problem.

The score r_i is then assigned to the documents:

$$\forall d \in D \forall i \in I : (i,d) \in E \Rightarrow r_d = r_i$$

If a document has more than one author, one has the option of either accumulating the PageRank scores ($r_d = \sum_{(i,d)\in E} r_i$), or of chosing either the maximum, minimum, or average of the PageRank scores of the authors. If the edges between nodes for individuals and document nodes are non-uniform in weight, one can also incorporate this weight information when transferring PageRank scores from authors to documents.

As a relevance measure for documents as regards a query, we employ a modified vector-space model. For a query q, the text retrieval component produces a set of relevant document $D_q \subset D$ as well as a score $\mathrm{rel}(q, d)$ for every document. The inclusion of r_d does not affect the result set D_q; it only influences the ranking of the documents, enabling the user to find relevant documents more quickly.

There are several models for combining PageRank with a text retrieval system. The simplest method is to sort the documents $d \in D_q$ by their PageRank score, and present those with the highest r_d to the user first. However, this method only works when a high precision of the result set is ensured [1].

A very simple method of combining PageRank and relevance scores is

$$r_d \cdot \mathrm{rel}(q, d) \quad .$$

For our purposes, this method has the advantage of not having tunable parameters, and being invariant to normalization. We choose this method for our experiments.

5 Evaluation

We evaluate the techniques in a known-item retrieval setting and compare them to the baseline technique using the metrics average rank and inverse average inverse rank (IAIR). A known-item retrieval setting reduces the amount of manual labour required and allows a semi-automatic selection of items. By comparing with a baseline technique on the same index, we eliminate external factors that may account for differences in performance; this allows us to gauge the impact of social retrieval techniques on retrieval performance. We use a modified vector-space model as the baseline.

5.1 Setting

For evaluation, we use a mailing list archive from the years 2000–2005; the archive contains 44108 messages written from 1834 different email addresses. For evaluation, two different subsets of the corpus are used, one containing messages from 2000–2005, and one from 2004. We construct a full-text index from the message body, after removing quoted parts.

In addition to the full-text index, an associative network is constructed from the messages:

- An author node is constructed for each email address. No effort is made to reconcile different email addresses of one person.
- Every message is linked to its author, and every author is linked to his messages.
- Authors are linked to each other based on how often they respond to one another's messages.

The extracted social network displays characteristics typical for social networks: It exhibits a high degree of clustering and short average shortest path lengths, making it a 'small-world network' [14]. 70% of all authors are part of a giant weak component, and the degree distribution follows a power law.

5.2 Choosing Query Terms

For choosing appropriate query terms for known-item retrieval, the following strategy is used: From the subject lines of email messages, frequent bi- and trigrams are extracted. Subject lines are a good indicator of user information needs, as many threads on a mailing list start with a question, and the question is usually summarized in the subject. Bi- and trigrams are especially apt candidates, because 'real-world' queries have been found to average between two and three words [15].

Selecting n-grams by frequency alone is suboptimal, as some frequent n-grams correlate highly with the author of the containing messages. In order to remove these n-grams, the mutual information of the occurence of a specific n-gram in the subject line and the author of the messages is determined. A desirable n-gram for use as a query phrase therefore has a low mutual information with the author, and a high document frequency at the same time. We sort n-grams by mutual information divided by the frequency and use the n-grams with the lowest score for evaluation:

$$\text{score}(n\text{−gram}) = \frac{I(n\text{−gram}, \text{author})}{\text{df}(n\text{−gram})}$$

For each of the ten queries, one message is chosen as the 'known item', the objective of this search: Only messages from 2004 are considered as relevant, and only those messages are assessed that actually contain the sequence of query terms in the subject line. The criteria for relevance are selected to mimic a searcher looking for an item he has seen before.

The items to be retrieved are chosen by an expert in the subject matter, and by a complete novice. Using two different relevance assessments allows us to evaluate whether a social IR system caters more to novice users who desire more general results of high quality, but know next to nothing about the authors, or expert users who may have more specific interests, and can judge a person's authority within the community without assistance of the social IR system.

5.3 Results

Results of the evaluation are summarized in Tab. 2. For items chosen by an expert searcher, the combination of PageRank and the vector-space model performs better than the vector-space model alone for four of ten queries on the 2004 corpus; in one case, the result is a draw. While the average rank of the found documents increases for PageRank search, the inverse average inverse rank decreases: The average rank increases by $21.7\% \pm 2.4$, but the inverse average

Table 2. Known-item retrieval on mailing list data. Columns labelled 'VS' contain ranks from vector-space search, columns labelled 'PR×VS' contain ranks scored by pagerank times vector space score. Rows '$\overline{\text{rank}}$ change' and 'IAIR change' contain the change compared to the baseline method 'VS' in percent.

method:	VS	PR×VS	VS	PR×VS
searcher:	expert	expert	novice	novice
		on messages from 2004:		
$\overline{\text{rank}}$:	14.75 ± 0.25	17.95 ± 0.05	17.5 ± 0.3	15.2 ± 0
$\overline{\text{rank}}$ change [%]:		$+21.7 \pm 2.4$		-13.1 ± 1.5
IAIR:	7.548 ± 0.032	7.082 ± 0.010	4.670 ± 0.013	4.599 ± 0
IAIR change [%]:		-6.2 ± 0.5		-1.5 ± 0.3
		on messages from 2000–2005:		
$\overline{\text{rank}}$:	24.4 ± 0.3	41.45 ± 0.05	39.35 ± 0.35	39.6 ± 0
$\overline{\text{rank}}$ change [%]:		$+69.9 \pm 2.3$		$+0.6 \pm 0.9$
IAIR:	8.787 ± 0.040	6.697 ± 0.012	4.962 ± 0.013	7.86 ± 0
IAIR change [%]:		-24.6 ± 0.5		$+58.4 \pm 0.4$

inverse rank decreases by $6.2\% \pm 0.5$. This means that some documents are found considerably later than with vector-space search, but for those documents in the earlier parts of the result list, PageRank combined with vector space performs better. This effect is even more pronounced on the 2000–2005 corpus, where the average rank increases by $69.9\% \pm 2.3$, but the inverse average inverse rank decreases by $24.6\% \pm 0.5$. On the 2000–2005 corpus, the combination performs better for six out of ten queries.

For the novice searcher, results are less pronounced. On the smaller corpus from 2004, both the average rank and inverse average inverse rank decrease (average rank by $13.1\% \pm 1.5$, IAIR by $1.5\% \pm 0.3$), whereas on the larger corpus, the average rank is unchanged, but the IAIR increases sharply (by $58.4\% \pm 0.4$.) On the smaller corpus, PageRank times vector space performs better for five out of ten queries, with one draw; for the larger corpus, it performs better for four out of ten queries, also with one draw.

This mirrors the results from [1], who report that 'the benefits of PageRank are the greatest for underspecified queries' and that 'for more specific searches where recall is more important, the traditional information retrieval scores and the PageRank should be combined.' The very nature of the known-item retrieval task places an emphasis on recall, since the objective is finding one *specific* document instead of just one of several that satisfy the information need.

6 Conclusion

We research how to integrate social networks in the information retrieval process and whether this integration leads to a performance improvement. Several applications of the internet are identified as social media, for example wikis, blogs, or mailing lists.

We propose a model for social information retrieval, which integrates the domains of social network analysis and information retrieval. Meaningful associations become apparent which are not part of the traditional models. We define social information retrieval as a retrieval process which includes a well-defined subset of the constituents of the social IR model.

We apply graph-based techniques to social networks, using them outside their traditional domains within information retrieval, namely web retrieval. We thereby extend the state of the art in graph-based retrieval techniques.

The commonly cited benefits of social software, for example improved communication among group members or emergence of communities, is important but intangible. We aim to derive tangible benefits from the application of social networks, namely improved retrieval performance – by providing retrieval techniques which are tailored to the emerging field of social software. We believe that these tangible benefits will accelerate the adoption of social software.

The main limitation of social IR follows from its domain model: it is only applicable where a social network is present in the domain, or can be derived. Furthermore, the quality of the social network is crucial. Limitations of other graph-based retrieval methods also apply to social information retrieval. Commonly cited limitations of PageRank are that its benefits are greatest for underspecified queries with many relevant results.

Evaluation of the prototype system was performed using non-standardized corpora and evaluation scenarios. For comparing the prototype system with current and future information retrieval systems, standardized corpora and evaluation scenarios must be constructed. Standardized scenarios also permit to tune the system for a particular retrieval task.

We chose not to base our evaluation on a web-based social information space, because of the associated problems of scale, and the difficulties in extracting suitable information. Instead, we use a mailing-list archive as an example of a social information space which lends itself readily to evaluation. The expected transferability of our results to other information spaces needs to be ascertained in further experiments.

An important next step is the integration of social IR in the ISKODOR prototype developed at University of Bonn, in order to implement the third pillar of the congenial web search paradigm [7].

We conclude that social network analysis is an important tool for information retrieval.

Acknowledgements

We would like to thank Alan F. Smeaton for graciously providing a dataset containing 25 years of publications from the SIGIR conference proceedings, on which preliminary experiments were performed.

References

1. Page, L., Brin, S., Motwani, R., Winograd, T.: The PageRank citation ranking: Bringing order to the Web. Technical report, Stanford University (1999)
2. Kleinberg, J.M.: Authoritative sources in a hyperlinked environment. Journal of the ACM **46**(5) (1999) 604–632
3. Craswell, N., Hawking, D.: Overview of the TREC-2004 Web track. In Voorhees, E.M., Buckland, L.P., eds.: Proceedings of the Thirteenths Text REtrieval Conference (TREC 2004). Number 500-261 in NIST Special Publications, Gaithersburg, MD, U. S. National Institute of Standards and Technology (2004)
4. Kautz, H., Selman, B., Shah, M.: The hidden web. AI Magazine **18**(2) (1997) 27–36
5. Kautz, H., Selman, B., Shah, M.: Referral web: combining social networks and collaborative filtering. Commununications of the ACM **40**(3) (1997) 63–65
6. Freyne, J., Smyth, B.: An experiment in social search. In Nejdl, W., De Bra, P., eds.: Adaptive Hypermedia and Adaptive Web-Based Systems, Third International Conference, AH 2004, Eindhoven, The Netherlands, August 23-26, 2004, Proceedings. Volume 3137 of Lecture Notes in Computer Science. Springer (2004) 95–103
7. Gnasa, M., Won, M., Cremers, A.B.: Three pillars for congenial web search. Continuous evaluation for enhancing web search effectiveness. Journal of Web Engineering **3**(3&4) (2004) 252–280
8. Lueg, C., Fisher, D., eds.: From Usenet to CoWebs. Interacting with social information spaces. Springer (2003)
9. Berners-Lee, T., Hendler, J., Lassila, O.: The semantic web. Scientific American **284**(5) (2001) 34–43
10. Dublin Core Metadata Initiative: DCMI Metadata Terms. http://dublincore.org/documents/dcmi-terms/ (2005)
11. Brickley, D., Miller, L.: FOAF vocabulary specification. http://xmlns.com/foaf/0.1/ (2005)
12. Six Apart Ltd.: LiveJournal bot policy. http://www.livejournal.com/bots/ (2006)
13. Flake, G.W., Tsioutsiouliklis, K., Zhukov, L.: Methods for mining web communities: Bibliometric, spectral, and flow. In Poulovassilis, A., Levene, M., eds.: Web Dynamics. Springer Verlag (2004) 45–68
14. Watts, D.J., Strogatz, S.H.: Collective dynamics of 'small-world' networks. Nature **393** (1998) 440–442
15. Silverstein, C., Marais, H., Henzinger, M., Moricz, M.: Analysis of a very large web search engine query log. SIGIR Forum **33**(1) (1999) 6–12