

Using social network analysis to enhance information retrieval systems

Lars Kirchhoff, Katarina Stanoevska-Slabeva, Thomas Nicolai & Matthes Fleck

Lars Kirchhoff

Institute for Media and Communication Management
University of St. Gallen
lars.kirchhoff@unisg.ch

Katarina Stanoevska-Slabeva

Institute for Media and Communication Management
University of St. Gallen
katarina.stanoevska@unisg.ch

Thomas Nicolai

Institute for Media and Communication Management
University of St. Gallen
thomas.nicolai@unisg.ch

Matthes Fleck

Institute for Media and Communication Management
University of St. Gallen
matthes.fleck@unisg.ch

Abstract

Although there is an increasing interest about social networks in general, there is little attention about the application of social network analysis to information retrieval systems. Recent studies (Borgatti et al. 2003; Cross et al. 2001) suggest that a social network of a person has a significant impact on his or her information acquisition. Therefore the paper proposes the application of available social network data in the context of information retrieval systems. An outline of the research design for the exploration of meaningful sources for social network extraction and the impact of meaningful social network analysis methods and measures in the context of information retrieval systems will be given. The evaluation of these methods and measures is conducted on ScientificCommons.org, a search platform for open access publications with more than 21 million publications and 8.5 million extracted authors and their co-author network.

Introduction and motivation

It is an ongoing trend that people increasingly reveal very personal information on social network sites in particular and in the World Wide Web in general (Boyd et al. 2007; Donath et al. 2004; Hangwoo 2006; Joinson 2001; Ralph et al. 2005). As this information becomes more and more publicly available from these various social network sites and the web in general, the social relationships between people can be identified. This in turn enables the automatic extraction of social networks. This trend is furthermore driven and enforced by recent initiatives such as facebook's connect¹, MySpace's data availability² and Google's FriendConnect³ by making their social network data available to anyone.

Furthermore the current development of the World Wide Web, termed as "Web 2.0" by O'Reilly (O'Reilly 2005), enables increasingly more people to publish information without profound technical knowledge. Blogs for example have gained a lot of attention in recent years. The whole blogosphere including more than 70 million blogs (Sifry 2007) forms a reasonable body of information and knowledge. Additionally, hypertext links made between blogs have been described as conversation, affiliation, or readership, implying a form of implicit social structure (Adamic et al. 2003; Flake et al. 2000; Gibson et al. 1998; Herring et al. 2005; Kumar et al. 2004; Marlow 2006). That means that the publicly available information is increasingly annotated with author information which allows the extraction of social networks, too.

These recent developments described above, together with increasing computing power and an increased amount of freely available scientific publication data in diverse databases, has led to a dramatic growth in interest for social network analysis (SNA) and in network analysis in general as shown in figure 1 (Borgatti et al. 2003). However, there is little attention about the application of SNA for use in information retrieval systems. Recent studies (Borgatti et al. 2003; Cross et al. 2001) suggest that the social network of a person has a significant impact on his/her information acquisition. Additionally SNA offers methods that enable the identification of important persons within social networks, who could have a significant influence on the importance of certain information.

¹ <http://developers.facebook.com/>

² <http://developer.myspace.com/community/>

³ <http://www.google.com/friendconnect/>

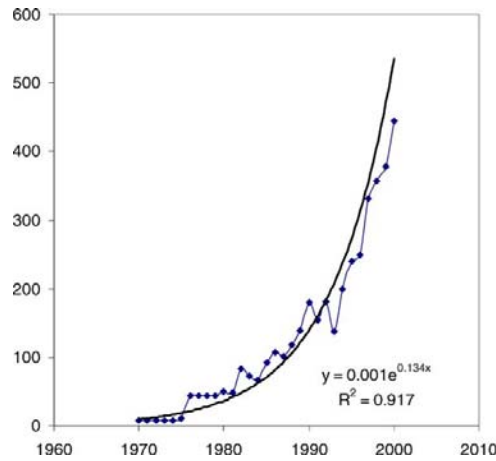


Figure 1: Growth of sociological publications indexed with social network in the abstract (source: Borgatti et al. 2003)

Therefore the paper proposes the application of available social network data in the context of information retrieval systems. An outline of the research design for the exploration of meaningful sources for social network extraction and the impact of meaningful SNA methods and measures in the context of information retrieval systems is presented. An evaluation of these methods and measures is conducted on ScientificCommons.org, a search platform for open access publications with more than 21 million publications and 8.5 million extracted authors and their co-authorship network.

The contribution of this paper is based on an analysis of online information sources in terms of their usability for the extraction of social networks and a research framework for the analysis and application of social network methods to information retrieval systems. The research framework was applied to the co-authorship network of scientific publications. The co-authorship network was used to compute different centrality measures of the authors, which then in turn have been used to refine the relevance ranking of publications within information retrieval systems. The performance of the different rankings based on the different centrality measures has been evaluated by the measurement of the click-through performance in the search results.

The content of the paper is structured as follows. At first, an overview about the research approach is given, followed by an analysis and classification approach of information sources for a meaningful extraction of social networks. Based on that analysis an introduction of the centrality measures and the model for application of these measures to the information retrieval model will be given. Lastly the evaluation method will be explained. The paper will be concluded with a summary and an outlook.

Research approach

The research process consists of four steps based on the design science research process (Järvinen 2000; Nunamaker et al. 1991). Figure 2 illustrates the structure of the research approach used in this paper. At first, the theoretical foundations for the usage of online information sources for the extraction of social networks are outlined. Secondly, a model for a social network enhanced information retrieval system was developed. Therefore an evaluation of methods that are applicable to information retrieval system has been conducted. On the basis of the chosen methods a model for the relevance computation of a document within an information retrieval system was developed. At third, the model was implemented as part of the search engine that runs at ScientificCommons.org. The fourth step is an evaluation whether and how the relevance ranking based on centrality measures has an impact on the relevance ranking of the documents of the information retrieval system.

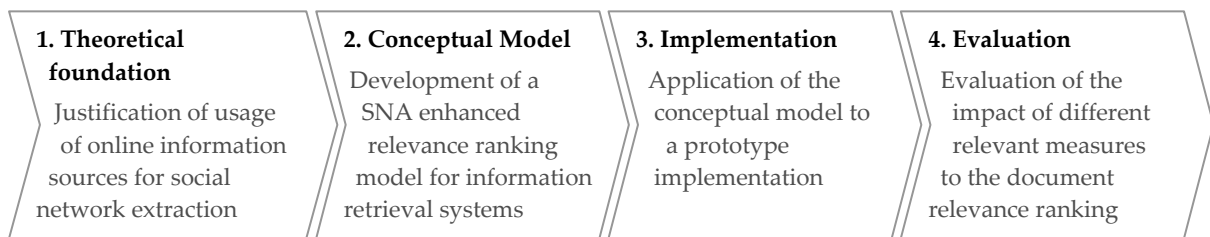


Figure 2: Research Approach

Online social network extraction

Definition of information sources for online social networks extraction

Traditionally SNA is conducted by collecting data via questionnaires, interviews, observations, archival records, experiments, which is mainly a manual process of gathering and analysing available information (Wasserman et al. 1994). In contrast the World Wide Web and diverse databases provide an increasing number of information sources, which exhibit characteristics that are interesting to SNA. Despite these developments it is necessary to consider these information sources carefully to whether they reflect social structures or not. Wellman (2001) proposes very broadly that

“a computer network is a social network”

and argues that humans use technologies to create communities. Within these communities supportive and sociable relationships are built, which form sustainable community ties. Social ties can be created and maintained via various different media types, including face-to-face contact, meetings, telephone, writing and other means of communication. Park (2003) supports this and argues that nodes in communication networks are the same as nodes in traditional social networks and that the content of social relation are the communication exchanges or the information flow. Following this he distinguishes among various types of networks: the social network, the computer network, the internet network and the hyperlink network. In figure 3 the relations between the different networks are shown. Obviously there is a relation of inclusion in a way that one network is included in another one with social networks as the most general case.

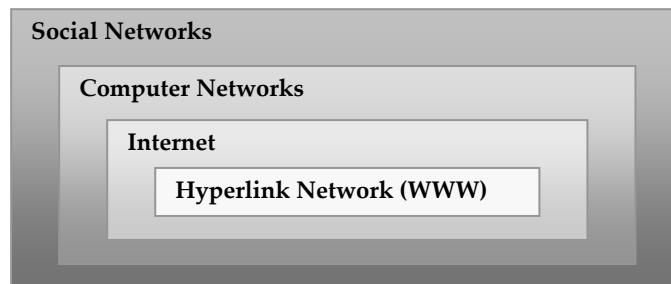


Figure 3: Relation between social networks and hyperlink networks (source: Park 2003)

Therefore computer mediated communication (CMC) leads to computer-mediated social networks (Garton et al. 1997; Jackson 1997). In fact, one of the first studies on computer mediated communication was conducted in order to study how computer networks modify acquaintanceship and friendship (Freeman 1984). Within the last decade research in the field of computer mediated communication has focused largely on the communication characteristics of diverse media types on the internet, ranging from the World Wide Web (WWW), emails, mailing lists, usenet newsgroups, chats, multi-user text-based role-playing environments (MUDs), multimedia environments and conferencing to message boards and Internet forums (Petróczi et al. 2006; Rice 1994). These computer supported social networks can create a sense of community belonging (Wellman et al. 2003). Additionally it is shown that computer mediated communication does not necessarily lead to a loss of intimacy in personal relations and that personalization is not a function of the medium, but rather relates

to the duration of the relationship and possible further communication (Walther 1994; Walther 1995).

Despite the fact that hyperlink networks can be seen as a type of social networks, it is however necessary to distinguish between two main approaches in hyperlink network analysis (Park et al. 2003): *Webometrics* and *Hyperlink networks analysis (HNA)*. Webometrics applies graph theory related techniques without paying attention to social aspects of the hyperlinks. Some of the methods have been implemented in search engines to calculate the most important web pages based on the hyperlink structure (Brin et al. 1998). The roots of webometrics can be found in information science, whereas hyperlink network analysis is originally derived from SNA (Park et al. 2003). Hyperlink network analysis in the sense of social networks, casts hyperlinks between web pages as social and communicational ties and applies techniques from SNA to them in order to analyze specific roles and structures in the network. In this work we follow the hyperlink network analysis approach.

In addition to the above mentioned arguments that the World Wide Web is a hyperlink network and as such can be seen as a special kind of social network, it should be emphasized that both, links and texts, on web pages in the World Wide Web reflect social interactions of users in the real world. This notion allows the extraction of social network structures not only from the hyperlinks but also from available textual information on web pages in the World Wide Web (Adamic et al. 2003). This means that the hyperlink network of web pages is only a part of a greater and more comprehensive social network that can be extracted. In this paper this broader notion of social networks within the World Wide Web is considered as an additional layer on top of the hyperlink network layer.

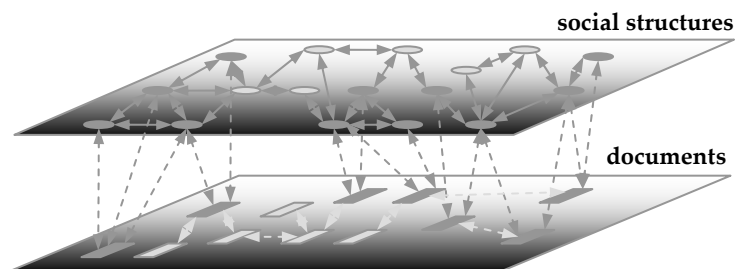


Figure 4: Connection between hyperlink structures and social network structures

Classification of information sources for social network extraction

With the definition described above a variety of different types of social connections between the nodes (actors) of this additional social layer shown figure 4 are possible:

- *Explicit direct social connection*

Hyperlinks that are explicitly meant to reflect social relations. An example for that are social network sites like Facebook, LinkedIn, MySpace or Orkut with millions of users who maintain a personal profile with a Friends list in order to interact and communicate with them (Boyd et al. 2007). The "public display of connection" is perceived as an important identity signal and is being used to maintain impression management (Boyd 2004; Boyd et al. 2007; Donath et al. 2004). Another example is the Friend-of-a-Friend (FOAF) protocol, that explicitly expresses relationships between persons (Finin et al. 2005; Mika 2004; Mika 2005).

- *Explicit indirect social connection*

Hyperlinks within any web page which link to other web sites. This connection is related to the above mentioned hyperlink network, but differs for instance in a way that the emergence of blogs has lead to more personal/social meaning of links (eg. blogroll) (Herring et al. 2005; Herring et al. 2007; Martino et al. 2006).

- *Implicit direct social connection*

Connections extracted from textual information found on a web page, which clearly indicate a social relation between the different actors. An example for such a connection is co-authorship in scientific publications, as they increasingly become more accessible through the World Wide Web. The co-authorship found in these documents can be used to extract a social network of authors (Barabási et al. 2002; Glänzel et al. 2004; Newman 2004; Otte et al. 2002).

- *Implicit indirect social connection*

Connections extracted on the content level of a web page, which might indicate a vague social relation between the actors. An example is the collaborative filtering from amazon. Krebs (2000) argued that in choosing a book as a focal node and assuming that books indirectly represent people who buy them, the resulting emerging network can be seen as social network. Another example are citation networks within scientific

publications (Johnson et al. 2007; Lawrence et al. 2006; Li-Chun et al. 2006; Neuhaus et al. 2006; Redner 1998; Vázquez 2001; White et al. 2004).

Problems related to social network extraction

Apart from the distinction of different information sources to extract and create social networks from the World Wide Web there are several problems related to the extraction of social networks from the various information sources available on the World Wide Web. First, a general problem is the identification of persons because of different naming standards or same names for different persons. Second, the social context and the type of social interactions of the authors within these information sources need to be carefully analyzed in order to obtain a meaningful understanding of the underlying social network structure.

Author and relation identification

The extraction of social networks crucially depends on the successful recognition of person names. Names are the most essential information to identify an individual that act as a network node and are consequently needed to identify the corresponding relations between network nodes. The problems of person extraction are manifold as names are ambiguous in many ways. First, the name of a person can be written differently in various contexts due to abbreviations, misspellings or pseudonyms. Second, a name can belong to more than one person. This ambiguity problem has been addressed in different research fields using different methods such as record linkage, duplicate record detection and elimination, merge/purge, data association, database hardening, citation matching, name matching, and name authority work in library cataloguing practice (Hui et al. 2004). To address misspelling or abbreviations errors probability methods, hidden Markov models and Support Vector Machines (SVM) have been suggested (Hui et al. 2004; Skounakis et al. 2003; Takasu 2003).

Context and weighting of social interactions

It is widely accepted that different social interactions can yield to different social networks and a different network structure leads to different effects on the involved individuals (Burt et al. 1985). Therefore a weighting of the relations by the means of the different suggested types of information sources and the resulting connection types is suggested by Kautz, Selman & Shah (1997). Additionally to the mentioned direct and indirect types of connection that may constitute a social network, these connections may have different meanings in

terms of interpersonal relations. Friends, colleagues, family members, team mates or participants are only some relation types a connection between actors may exhibit. The Friend-of-a-Friend (FOAF) protocol offers more than 30 kinds of relationships⁴, which can be assigned to a connection. These different kinds of relationships lead to different social networks of an actor. A person may be central in the social network of a research community while he is not in the local community.

Such overlapping social networks have been studied in SNA. Simmel (1955) was the first who discussed the theoretical implications of a persons' various social networks (which he called social circles). He argued that the different social networks (circles) are fundamental in defining a social identity (Wasserman et al. 1994). Recent research on large-scale, complex networks has shown that overlaps are significant. Furthermore the overlap of different sub networks within a network is a relevant network property (Palla et al. 2005).

Applicable methods of social network analysis

One of the main motivations for the use of social network data in information retrieval systems is to compute the importance of a person in the social network that is extracted from the documents and to use this relevance to compute the importance a document. In this work it is assumed that information from people with a special position and role within a social network are more valuable. Consequently SNA methods that measure the centrality of an actor are interesting for the application to information retrieval system and will be analysed within this work.

Centrality Measures

Usually centrality is measured as a property of the single node within the network to evaluate the 'reachability' or the 'importance' of this node. Centrality can also be measured as a property of the edges within a network and also for the whole network. Nevertheless the node centrality is most important in this work. There is a number of centrality indices such as degree, eccentricity, centroid of a graph, median of a graph, stress centrality, shortest path centrality, flow betweenness, vitality, current-flow betweenness, current-flow closeness, random walk betweenness, random walk closeness, eigenvector centrality, bargaining

⁴ <http://vocab.org/relationship/> [Accessed 14/06/2008]

centrality, PageRank, and HITS. A comprehensive overview of most of the different centrality indices is given in Brandes & Erlebach (2005). However, only some of them are relevant for the research of social networks. The ones that are important and relevant for this work will be explained in more detail in the following.

Degree

The most simple centrality measure is the degree centrality $C_D(u)$ of a node u which is the number of edges directly connected to a node. The degree centrality can be seen as a measure for the social activity of a person. That means a high degree respectively a high number of direct contacts indicates a high social activity within the network. The degree of a node can be measured by directed and undirected edges.

$$C_D(u) = \sum_{v \in V} e_{u,v}$$

In the case of undirected edges every edge is considered for the degree. If directed edges are taken into account it is possible to distinguish between incoming and outgoing edges and therefore incoming degree $C_D^-(u)$ and outgoing degree $C_D^+(u)$. Normalization of degree centrality $C'_D(u)$ can be achieved by using the number of maximal possible connections of a node as relation to the degree of a node.

$$C'_D(u) = \frac{\sum_{v \in V} e_{u,v}}{(n-1)}$$

Closeness

The closeness centrality $C_C(u)$ defines a measure that indicates how close a node is to every other node in the network. Therefore the closeness centrality measure does not rely on the direct connections solely, but on the indirect connections between nodes also. Therefore the closeness is a measure for the efficiency of person to reach any other person in a network, but also for their independence within the network. The most commonly used mathematical definition is the reciprocal of the total distances from a node to all other nodes in the network

$$C_C(u) = \frac{1}{\sum_{v \in V} d(u,v)}.$$

The distance described here is the shortest path between a pair of nodes, which is also called geodesic distance. Again, for the normalization the result will be divided by the number of maximum number of possible connections of a node $1/(n-1)$.

$$C_c(u) = \frac{(n-1)}{\sum_{v \in V} d(u,v)}.$$

Betweenness

The betweenness centrality $C_B(u)$ is a measure for the ability of a node to control the flow of communication. Therefore the betweenness centrality does not only measure direct and indirect connections between two nodes, but rather analyzes connections with three nodes involved⁵. For every pair of nodes (s, t) the number of shortest paths σ_{st} between these nodes is measured at first. Then it will be analyzed if any of the paths contain the node (u) . The ratio

$$\delta_{st}(u) = \frac{\sigma_{st}(u)}{\sigma_{st}}$$

defines the fraction between s and t that contains u and can be interpreted as probability that u is involved in the communication between s and t . The betweenness centrality is then defined as the sum of all shortest paths between any node pairs, in which u is involved.

$$C_B(u) = \sum_{s \neq u \in V} \sum_{t \neq u \in V} \delta_{st}(u)$$

The shortest paths ending or starting in u are explicitly excluded from the measurement, as the betweenness centrality is a measure for the control of communication.

Computation of centrality measures in large graphs

All-pairs shortest-path problem

The number of documents in information retrieval systems and especially information retrieval system for web content tend to be very large. Hence the social networks extracted from these documents are assumed to be rather large. It is a known problem that the computation of the closeness and the betweenness centrality measure becomes very prohibitive for large graphs with a high number of nodes and edges. This is due to the fact,

⁵ This index assumes that all communication is conducted along the shortest paths only.

that for the computation of these measures it is necessary to compute all node pairs shortest path first. Algorithms that solve this all-pairs shortest-path (APSP) problem are known to have a complexity of $O(nm + n^2 \log n)$, where n is the number of nodes and m is the number of edges in a graph. That means that for large graphs the computation becomes very slow and even with fast matrix multiplication complicated and impractical (Aingworth et al. 1999). Eppstein & Wang suggested an algorithm for fast approximation of centrality (Eppstein et al. 2004) to solve the problem in linear time with an additive error of $\epsilon \Delta$.

Another problem of betweenness centrality is the stability the measures in dynamic networks, where new nodes and edges are attached or removed to the network very frequently. The addition or removal might cause a great perturbation of the betweenness centrality values. A solution to that problem is suggested by Carpenter, Karakostas, & Shallcross (2002). In their proposal they consider only paths between s and t that are no longer than $(1 + \epsilon)d(s, t)$. The resulting centrality betweenness is called ϵ -betweenness centrality (Carpenter et al. 2002; Scott et al. 2003). This betweenness centrality measure is more sensitive to the relative quality of the centrality of a node within the network. Additionally the proposed approach may improve the running times for dynamic APSP as not all-pairs shortest-path are used for the calculation anymore.

Dealing with insufficient connectivity

An important assumption for all centrality measures except degree centrality is that the network is completely connected, that means that every node can be reached from every other node within the network. In case of a disconnected graph the computation of centrality measures can become problematic, because the computation of the distances between a node pair with the nodes in two disconnected subgraphs is not possible. The distance would be undefined and could not be used for computation.

A very naive approach to deal with disconnected networks would be to restrict the computation of the centrality indices to the subgraphs. This approach is not very reasonable in most cases, because it does not take the size of a network into account. Therefore a node in a small network might be as central as a node in a larger network, which intuitively is not true in most cases. Another simple and common way to deal with disconnected subgraphs is to multiply the centrality values with the size of the subgraph. This solution would be reasonable if the centrality measures behave proportional with the network size, but a

experiments by Poulin, Boily, & Mâsse (2000) suggest that this is not the case for all centrality indices.

To overcome this shortcomings repair mechanisms that use either inverse path length or arbitrary fixed values for the distance of two unconnected nodes are proposed. In the latter case the closeness centrality measure can be defined as:

$$C_C(u) = \frac{\sum_{v \in V} \sum_{w \in V} d(v, w)}{\sum_{v \in V} d(u, v)}$$

where $d(u, v)$, the distance between two unconnected nodes is set to k . It is important to find an appropriate value for k , because the choice of k has an essential influence on the centrality measure of the nodes.

Local centrality measures

Computation problems and issues with insufficient connectivity are significant problems when dealing with large social networks such as the one extracted from a large document collections in an information retrieval system with millions of documents. Furthermore centrality measures such as closeness and betweenness centrality of large graphs with millions of nodes are just not very meaningful.

To overcome these problems we propose the usage of local centrality measures such as the ego network betweenness suggested by Everett and Borgatti (2005). With a local centrality approach the measurement of the closeness and betweenness centrality is being conducted by analysing the ego network of a node only. In their work Everett and Borgatti (2005) demonstrate that a strong relationship exists between the ego network betweenness and the whole network betweenness measures. They further show that the ego network betweenness is a good approximation in two situations. One situation is if all actors have similar betweenness measures and the second is if they have highly different measures. The latter is very likely in real world network data. Hence it is assumed in this work that social networks extracted from large document collections of information retrieval systems tend to behave the same way.

Model for social network enhanced information retrieval system

Overview

The principle model of an information retrieval system enhanced by the document relevance ranking based on centrality measures of social networks is illustrated in figure 5. It illustrates the three elementary steps that are conducted in order to calculate the relevance ranking of a document. In the following these three steps will be discussed in detail in the case of the information retrieval system that is implemented at ScientificCommons.org.

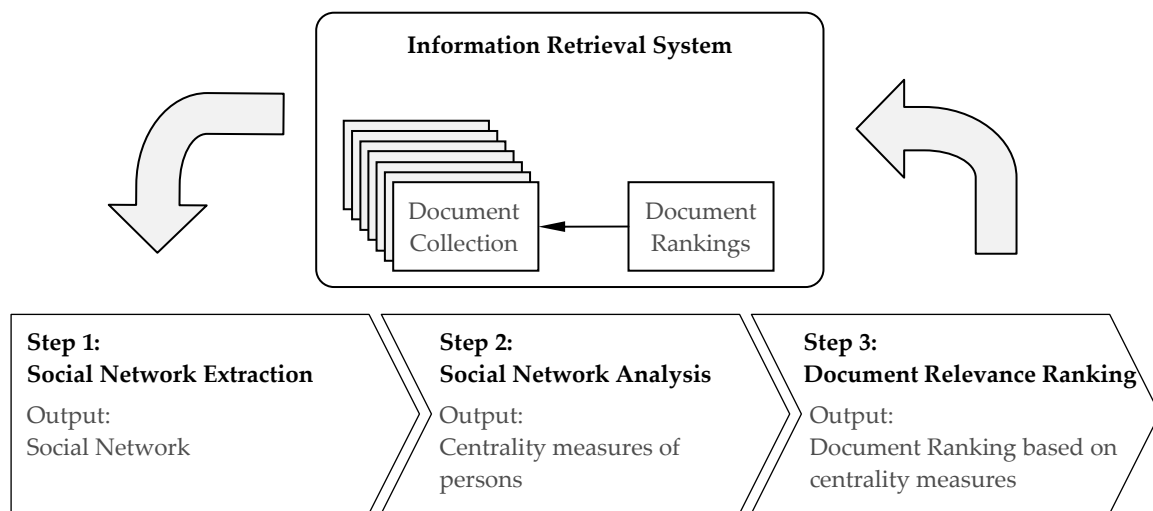


Figure 5: Principle model of the social network enhanced information retrieval system

ScientificCommons.org is a search engine for Open Access publications. It harvests publication metadata and full text documents from more than 900 digital academic repositories worldwide. Currently it stores more than 21 million publication metadata sets and more than 600.000 full text documents. The goal of ScientificCommons.org is to provide a comprehensive and simple access to academic content published in these repositories.

Step 1: Social network extraction

The document collection of ScientificCommons.org consists of publication metadata sets. These publication metadata sets are harvested via the Open Archive Initiative Protocol for Metadata Harvesting (OAI-PMH). OAI-PMH is a XML based protocol that allows the structured distribution of information about academic publications in a variety of known

metadata standards. The most widespread metadata standard is Dublin Core which is used by the majority of digital academic repositories. It is a very basic standard that consists of only 15 core elements: title, date, identifier, creator, subject, language, type, publisher, description, format, rights, source, relation, contributor, coverage. It is not necessary that all 15 core elements must be used in a Dublin Core metadata entry. However the title, the date and the identifier are set in all datasets in every digital repository and the majority of metadata sets contain the creator field. An evaluation of the more than 900 digital academic repositories has shown that more than 85% of all metadata sets contain a creator field and more than 20% contain a contributor field. In figure 6 the distribution of the usage of the different fields is shown. The high number of available creator fields is the reason why the OAI-PMH protocol is valuable for the extraction of author names from the publication metadata sets.

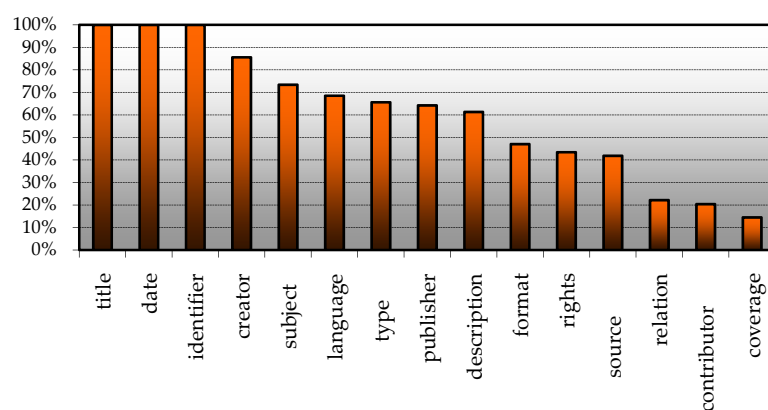


Figure 6: Overview of usage of Dublin Core metadata fields

In many cases academic publications are written by two or more authors. The co-authorship of publications can be seen as a two-mode social network and more specific as an affiliation network. In that affiliation network the authors form the first mode and the publications the second mode (event, affiliation). All authors of a publication form the subset of actors related to a publication. In the case of ScientificCommons.org more than 8.5 million authors could be extracted with more than 25 million relations to the harvested metadata sets.

Step 2: Social network analysis

In order to measure the centrality indices mentioned above a transformation of the two-mode network is necessary at first. This transformation includes the creation of an adjacency

list of the authors based on the co-authorship and the application of a weight to each of the author relation stored in the adjacency list. In case of ScientificCommons.org more than 41 million relationships between the 8.5 million authors could be identified.

The weighting w_e is necessary to normalize the strength of relationships between authors of publication with different numbers of co-authors. For instance it is not meaningful to give a relation between co-authors of publication with 500 and more co-authors the same weighting as a relation created from a publication with two co-authors. Therefore different weighting algorithms have been applied to measure the effect on the calculations of the centrality indices. Two simple weighting algorithms have been implemented so far. The first w_{e_A} is based on the numbers of authors n_a of a publication and the second w_{e_c} is based on the number of possible connections n_c between the authors of a publication. Both calculate the weight as the reciprocal value of the number of either authors or connections:

$$a) w_{e_A} = \frac{1}{n_a}; \quad b) w_{e_c} = \frac{1}{n_c};$$

More advanced algorithms may be implemented later. Based on transformed two-mode network and the different centrality measures have been calculated.

Step 3: Document relevance ranking

At the end the calculated centrality measures of the authors are taken to calculate a relevance score for each document. This calculation includes two steps:

1. *Calculation of a document relevance score based on the centrality measures*

The simplest calculation of the document relevance based on the centrality measure R_C is to simply take the average value of each author associated with the publication.

$$R_C = \frac{\sum_{i \in A} C_i}{n_A}$$

2. *Calculation of the relevance score based on the query*

The ranking of search results can not solely based on the relevance of authors, because it is still important that the keywords used in a query match with the document. Therefore a phrase proximity calculation to the document text and a statistical analysis of the word frequency is conducted for each query. This score R_Q is then used together with R_C to calculate the final rank:

$$R_F = R_Q + dR_C$$

In this equation d is a dumping factor to normalize both rankings.

Evaluation methods for the proposed system

An analysis on the search behavior on ScientificCommons.org seems to be the most appropriate method for evaluating the proposed information retrieval system. For this reason the click-through performance of search results is measured in two ways. First, the ratio of *search queries* to *search clicks* is measured while a lower ratio value is assumed to indicate an improvement of the search result quality. Secondly, the system logs the position of a click within the search result list. Therefore each result in the search result list is assigned a position number, starting with one at the top of the list. In consequence, a distribution of click positions can be derived that point towards the interest of the users. Again, an adjustment to lower click position is assumed to indicate an improvement in search result quality. Each of the different centrality measures together with different calculations of the document relevance is evaluate with a certain number of search queries in order to compare results.

Summary and outlook

In the study we presented an analysis and classification of information sources in the World Wide Web that can be used to extract social networks for the use in information retrieval systems. This was followed by an analysis of SNA methods that are considered to be useful to create a relevance ranking for documents based on the extracted social network. Both analysis' have then been integrated into a model of an information retrieval system based on a document relevance ranking that is derived from centrality measures. This model has been explained in detail on the implementation of ScientificCommons.org together with an explanation how the evaluation is conducted.

The suggested research approached showed that the co-author relationships are valuable information in the process of computing the importance of documents in an information retrieval system for scientific publications. The evaluation of the impact of the different centrality measures on the relevance ranking and therefore on the position of documents in

the results set is still ongoing, but first results show that the relevance of documents could be significantly enhanced.

Although the first results are quite promising there are numerous questions that still need to be answered. We only demonstrated the use of social networks extracted from scientific publication so far. It would be interesting to add data from other social network information sources to the social network graph and analyse what impact this has on the document relevance, thus following the argumentation of Simmel (1955) that different social networks of a person constitute its identity. Therefore it would be necessary to develop a model for the integration of different social networks into the centrality calculation. Furthermore the impact of personal filtering of the results based on the social network of the user is an interesting field to look at.

References

- Adamic, L.A., and Adar, E. "Friends and Neighbors on the Web," *Social Networks* (25:3) 2003, pp 211-230.
- Aingworth, D., Chekuriym, C., Indykz, P., and Motwanix, R. "Fast Estimation of Diameter and Shortest Paths (Without Matrix Multiplication)," *SIAM Journal on Computing* (28:4) 1999, pp 1167-1181.
- Barabási, A.-L., Jeong, H., Neda, Z., Ravasz, E., Schubert, A., and Vicsek, T. "Evolution of the social network of scientific collaborations," *Physica A* (311:3-4) 2002, pp 590-614.
- Borgatti, S.P., and Foster, P.C. "The network paradigm in organizational research: A review and typology," *Journal of Management* (29:6) 2003, pp 991-1013.
- Boyd, D.M. "Friendster and publicly articulated social networks," ACM Conference on Human Factors in Computing Systems ACM Press, Vienna, 2004, pp. 1279-1282.
- Boyd, D.M., and Ellison, N.B. "Social network sites: Definition, history, and scholarship," *Journal of Computer-Mediated Communication* (13:1) 2007.
- Brandes, U., and Erlebach, T. *Network Analysis: Methodological Foundations* Springer, Berlin / Heidelberg, 2005.
- Brin, S., and Page, L. "The Anatomy of a Large-Scale Hypertextual Web Search Engine," 7th WWW conference, Elsevier Science, Brisbane, Australia, 1998.
- Burt, R.S., and Schøtt, T. "Relation contents in multiple networks," *Social Science Research* (14:4), December 1985 1985, pp 287-308.
- Carpenter, T., Karakostas, G., and Shallcross, D. "Practical Issues and Algorithms for Analyzing Terrorist Networks," WMC 2002, 2002.
- Cross, R., Borgatti, S.P., and Parker, A. "Beyond answers: dimensions of the advice network," *Social Networks* (23:3), July 2001 2001, pp 215-235.
- Donath, J., and Boyd, D.M. "Public Displays of Connection," *BT Technology Journal* (22:4) 2004, pp 71 - 82

- Eppstein, D., and Wang, J.Y. "Fast Approximation of Centrality," *Journal of Graph Algorithms and Applications* (8:1) 2004, pp 27–38.
- Everett, M.G., and Borgatti, S.P. "Ego network betweenness," *Social Networks* (27:1), January 2005 2005, pp 31-38.
- Finin, T., Ding, L., Zhou, L., and Joshi, A. "Social Networking on the Semantic Web," *The Learning Organization* (5:12), December 31, 2005 2005, pp 418-435.
- Flake, G., Lawrence, S., and Giles, C.L. "Efficient Identification of Web Communities," 6th ACM Conference on Knowledge Discovery and Data Mining, Boston, MA, USA, 2000, pp. 150–160.
- Freeman, L.C. "The impact of computer based communication on the social structure of an emerging scientific specialty," *Social Networks* (6:3), September, 1984 1984, pp 201-221.
- Garton, L., Haythornthwaite, C., and Wellman, B. "Studying Online Social Networks," *Journal of Computer Mediated Communication* (3:1), June, 1997 1997.
- Gibson, D., Kleinberg, J.M., and Raghavan, P. "Inferring web communities from link topology," 9th ACM Conference on Hypertext and Hypermedia, 1998.
- Glänzel, W., and Schubert, A. "Analyzing scientific networks through co-authorship," in: *Handbook of Quantitative Science and Technology Research*, H.F. Moed, W. Glänzel, U. Schmoch and P.i.t. Netherlands. (eds.), Kluwer Academic Publishers, Dordrecht, 2004, pp. 257-276.
- Hangwoo, L. "Privacy, publicity, and accountability of self-presentation in an on-line discussion group," *Sociological Inquiry* (76:1) 2006, pp 1-22.
- Herring, S.C., Kouper, I., Paolillo, J.C., Scheidt, L.A., Tyworth, M., Welsch, P., Wright, E., and Yu, N. "Conversations in the Blogosphere: An Analysis "From the Bottom Up"," 38. Hawaii International Conference on System Sciences (HICSS-38), IEEE Press., Hawaii, 2005.
- Herring, S.C., Paolillo, J.C., Ramos-Vielba, I., Kouper, I., Wright, E., Stoerger, S., Scheidt, L.A., and Clark, B. "Language Networks on LiveJournal," Proceedings of the Fortieth Hawai'i International Conference on System Sciences (HICSS-40), IEEE Press, Los Alamitos, 2007.
- Hui, H., Giles, C.L., Hongyuan, Z., Cheng, L., and Kostas, T. "Two supervised learning approaches for name disambiguation in author citations," in: *Proceedings of the 4th ACM/IEEE-CS joint conference on Digital libraries*, ACM, Tuscon, AZ, USA, 2004.
- Jackson, M.H. "Assessing the Structure of Communication on the World Wide Web," *Journal of Computer-Mediated Communication* (3:1), 1997 1997.
- Järvinen, P.H. "Research Questions Guiding Selection of an Appropriate Research Method," Proceedings of the Eighth European Conference on Information Systems (ECIS2000), Vienna, 2000, pp. 124-131.
- Johnson, B., and Oppenheim, C. "How socially connected are citers to those that they cite?," *Journal of Documentation* (63:5) 2007, pp 609-637.
- Joinson, A.N. "Self-disclosure in computer-mediated communication: The role of self-awareness and visual anonymity," *European Journal of Social Psychology* (31:2) 2001, pp 177-192.
- Kautz, H., Selman, B., and Shah, M. "The Hidden Web," *The AI Magazine* (18:2) 1997, pp 27-36.
- Krebs, V. "Working in the connected world book network. ," *International Association for Human Resource Information Management Journal* (4:1), 87-90 2000.

- Kumar, R., Novak, J., Raghavan, P., and Tomkins, A. "Structure and evolution of Blogspace," *Communications of the ACM* (47:12) 2004, pp 35-39.
- Lawrence, J., Payne, T.R., and De Roure, D. "Co-Presence Communities: Using Pervasive Computing to Support Weak Social Networks," 17th IEEE International Workshop on Enabling Technologies: Infrastructures for Collaborative Enterprises (WETICE '06), IEEE Press, Rome, Italy 2006, pp. 149-156.
- Li-Chun, Y., Kretschmer, H., Hanneman, R.A., and Ze-Yuan, L. "The evolution of a citation network topology: The development of the journal *Scientometrics*," International Workshop on Webometrics, Informetrics and Scientometrics & Seventh COLLNET Meeting, Nancy, France, 2006.
- Marlow, C.A. "Linking without thinking: Weblogs, readership, and online social capital formation," International Communication Association Conference, Dresden, 2006.
- Martino, F., and Spoto, A. "Social Network Analysis: A brief theoretical review and further perspectives in the study of Information Technology," *PsychNology Journal* (4:1) 2006, pp 53-86.
- Mika, P. "Bootstrapping the FOAF-Web: An Experiment in Social Network Mining," 1st International Workshop on FOAF, Social Networks and the Semantic Web, Galway, Ireland, 2004.
- Mika, P. "Ontologies Are Us: A Unified Model of Social Networks and Semantics," in: *The Semantic Web – ISWC 2005*, Springer Link, 2005, pp. 522-536.
- Neuhaus, C., and Daniel, H.-D. "Data sources for performing citation analysis an overview," *Journal of Documentation*), 30 June 2006 2006.
- Newman, M.E.J. "Coauthorship networks and patterns of scientific collaboration," *PNAS* (101:Suppl. 1), April 6, 2004 2004, pp 5200-5205.
- Nunamaker, J.F.J., Chen, M., and Purdin, T.D.M. "System Development in Information Systems Research," *Journal of Management Information Systems* (7:3) 1991, pp 89-106.
- O'Reilly, T. "What is Web 2.0: Design Patterns and Business Models for the Next Generation of Software," 2005.
- Otte, E., and Rousseau, R. "Social network analysis: a powerful strategy, also for the information sciences," *Journal of Information Science* (28:6) 2002, pp 441-453.
- Palla, G., Derényi, I., Farkas, I., and Vicsek, T. "Uncovering the overlapping community structure of complex networks in nature and society," *Nature* (435), 9 June, 2005 2005, pp 814-818.
- Park, H.W. "Hyperlink Network Analysis: A New Method for the Study of Social Structure on the Web," *Connections* (25:1) 2003, pp 49-61.
- Park, H.W., and Thelwall, M. "Hyperlink Analyses of the World Wide Web: A Review," *Journal of Computer-Mediated Communication* (8:4) 2003, p 32.
- Petróczi, A., Nepusz, T., and Bazsó, F. "Measuring tie-strength in virtual social networks," *Connections* (27:2) 2006, pp 39-52
- Poulin, R., Boily, M.C., and Mâsse, B.R. "Dynamical systems to define centrality in social networks," *Social Networks* (22:3) 2000, pp 187-220.
- Ralph, G., Alessandro, A., and H. John Heinz, I. "Information revelation and privacy in online social networks," Proceedings of the 2005 ACM workshop on Privacy in the electronic society, ACM, Alexandria, VA, USA, 2005, pp. 71-80.
- Redner, S. "How popular is your paper? An empirical study of the citation distribution," *European Physical Journal B* (4:2) 1998, pp 131-134.

- Rice, R.E. "Network analysis and computer-mediated communication systems," in: *Advances in social network analysis*, S. Wasserman and J. Galaskiewicz (eds.), Sage Publications, Thousand Oaks, 1994, pp. 167-203.
- Scott, W., and Padhraic, S. "Algorithms for estimating relative importance in networks," in: *Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining*, ACM, Washington, D.C., 2003.
- Sifry, D. "The State of the Live Web, April 2007," 2007.
- Simmel, G. *Conflict And The Web Of Group Affiliations* The Free Press, Glencoe, Illinois, 1955, p. 195.
- Skounakis, M., Craven, M., and Ray, S. "Hierarchical hidden markov models for information extraction," 18th International Joint Conference on Artificial Intelligence, Acapulco, Mexico, 2003.
- Takasu, A. "Bibliographic attribute extraction from erroneous references based on a statistical model," *Proceedings of the 3rd ACM/IEEE-CS joint conference on Digital libraries*, IEEE Computer Society, Houston, Texas, 2003, pp. 49-60.
- Vázquez, A. "Statistics of citation networks," *E-print: arXiv: cond-mat/0105031*) 2001.
- Walther, J.B. "Anticipated Ongoing Interaction Versus Channel Effects on Relational Communication in Computer-Mediated Interaction," *Human Communication Research* (20:4), June, 1994 1994, pp 473-501.
- Walther, J.B. "Relational Aspects of Computer-Mediated Communication: Experimental Observations over Time," *Organization Science* (6:2), March - April, 1995 1995, pp 186-203.
- Wasserman, S., and Faust, K. *Social Network Analysis: Methods and Applications* Cambridge University Press, Cambridge, 1994.
- Wellman, B. "Physical Place and CyberPlace: The Rise of Personalized Networking," *International Journal of Urban and Regional Research* (25:2) 2001, pp 227-252.
- Wellman, B., and Gulia, M. "Virtual communities as communities," in: *Communities in cyberspace* P. Kollock and M.A. Smith (eds.), Routledge, London, 2003, pp. 167-195).
- White, H.D., Wellman, B., and Nazer, N. "Does citation reflect social structure?: Longitudinal evidence from the "Globenet" interdisciplinary research group," *Journal of the American Society for Information Science and Technology* (55:2) 2004, pp 111--126.