Social Network Models for Enhancing Reference Based Search Engine Rankings

Nikolaos Korfiatis[1], Miguel-Angel Sicilia[2], Claudia Hess[3], Klaus Stein[3], Christoph Schlieder[3]

[1] *Department of Informatics*
Copenhagen Business School (CBS)
Howitzvej 60, DK-2000 Frederiksberg (Copenhagen)
Denmark

[2] *Department of Computer Science*
University of Alcala
Ctra. Barcelona, km. 33.6 – 28871
Alcala de Henares (Madrid)
Spain

[3] *Laboratory for Semantic Information Technology*
University of Bamberg
Feldkirchenstraße 21, D-96045 Bamberg
Germany

Abstract

This chapter elaborates on a twofold approach on models of web search engine retrieval and more specifically the way information resources are ranked in the query results. In particular, current models of information retrieval are blind to the social context that surrounds information resources thus do not consider the trustworthiness of their authors when they present the query results to the users. Following this point we elaborate on the basic intuitions that highlight the contribution of the social context – as can be mined from social network positions for instance – into the improvement of the rankings provided in reference based search engines. A review on ranking models in web search engine retrieval along with social network metrics of importance such as prestige and centrality is provided as a background. Then a presentation of recent research models that utilize both contexts is provided along with a case study in the internet based encyclopedia Wikipedia based on the social network metrics.

Key-Words

Search Engines, Rankings, HITS , PageRank, Social Networks, Centrality

Chapter Contents

1 Introduction

Since the introduction of information technology, information retrieval (IR) has been an important branch of computer and information science mainly due to the ability to reduce the time required by a user to gather contextualized information and knowledge (Baeza-Yates & Ribeiro-Neto, 1999). With the introduction of hypertext (Conklin, 1987), information retrieval methods and technologies have been able to increase their accuracy because of the high amount of meta-information available for the IR system to exploit. That is not only information about the documents per se but information about their context and popularity. However the development of the World Wide Web has introduced another dimension to the IR domain by exposing the social aspect of information (Brown & Duguid, 2002) produced and consumed by humans in this information space.

Current ranking methods in information retrieval – which are used in web search engines as well - exploit the references between information resources such as the hypertextual (hyperlinked) context of web pages in order to determine the rank of a search result (Dhyani, Keong, & Bhowmick, 2002; Faloutsos, 1985). The well-known PageRank algorithm (Page, Brin, Motwani, & Winograd, 1998 ; Brin & Page, 1998) has proved to be a very effective paradigm for ranking the results of Web search algorithms. In the original PageRank algorithm, a single PageRank vector is computed, using the link structure of the Web, to capture the relative "importance" of Web pages, independent of any particular search query. Nonetheless, the assumptions of the original PageRank are biased towards measuring external characteristics. In fact, Page, Brin, Motwani, & Winograd (Page, Brin, Motwani, & Winograd, 1998) conclude their article with the sentence "*The intuition behind PageRank is that it uses information which is external to the Web pages themselves - their backlinks, which provide a kind of peer review"*.

That is to say that backlinks[a] (i.e. incoming links) are considered as a positive evaluation of the respective web site. The PageRank therefore does not distinguish whether the user setting the link agrees with the content of the other webpage or whether she or he disagrees. This fact underlines that current reference-based ranking algorithms often do not take into consideration that an information resource is a result of cognitive and social processes. In addition to its surrounding hyperlinks a social context (Brown & Duguid, 2002) underlies the referencing of those resources.

This suggests that a critical point in improving link-based metrics would be that of augmenting or weighting the pure backlink or reference model with social information, provided that linking is in many cases influenced by social ties, and trustworthiness critically depends on the social relevance or consideration of the authors of the pages. Recently, several independent research has provided different models for this kind of social network analysis as applied to ranking or assessing the quality of Web pages (Sicilia & Garcia, 2005; Hess, Stein & Schlieder, 2006; Stein & Hess, 2005) or activity spaces such as the usenet and wikis (Korfiatis & Naeve, 2005). Other approaches such as those presented by Borner, Maru and Goldstone (Borner, Maru, & Goldstone, 2004) have also dealt with integrating different networks by analyzing the simultaneous growth of coauthor and citation networks in time.

The main objective of this chapter is to provide a survey and a generic framework for such approaches as a roadmap for further research in this direction. Furthermore the chapter attempts to provide a bridge to the areas of Social Network Analysis (SNA) and Information Retrieval by highlighting the benefits of their integration in the case of a web search engine.

To this end this chapter is organized as follows: Section 2 describes the basic intuition behind the concepts and the definitions provided by this chapter. Section 3 provides an insight to the classic models of citation and link based ranking methods such as PageRank and HITS. Section 4 provides an insight to sociometric models that can be used in order to evaluate users as well as a reference to the FOAF vocabulary which is used to define social connections on the web. Section 5 provides an overview of two

[a] With the term backlink we refer to the backwards reference given to an object by another referring object. The most well known type of a backlink is the citation provided to scientific articles and scholarly work.

exemplar hybrid ranking models which integrate in an equal way both the social and hypertextual context. Then, Section 6 provides a case study of modeling authoritativeness on data obtained from the English-language version of the Wikipedia, by using some of the metrics discussed in Section 4. Finally section 7 presents the conclusions and provides an outlook of future research directions which can be exploited towards the combined consideration of hypertextual and social context in a web IR system such as a web search engine.

2 Background and Motivation

The original intuition behind the design of the World Wide Web and the Hypertext Markup (HTML) language is that authors can publish web documents which can provide pointers to other documents available online. This simplified aspect of a hypertext model adapted by Tim Berners-Lee in the original design of WWW (Berners-Lee & Fischetti, 1999) has given to the web the morphology of an open publication system which can evolve simultaneously by providing references and pointers to existing documents available. However, as in the original publication contexts where an author gains credibility due to the popularity of her or his productions/affiliations, the credibility/appropriateness of a page on the web is to some extent correlated with its popularity. What the first search engines on the web failed to capture was exactly this kind of popularity which attributes the appropriateness to the query results of the search engine.

Although advances in query processing have made the retrieval of the query results computationally efficient (Wolf, Squillante, Yu, Sethuraman, & Ozsen, 2002) the precision and the recall of the web document retrieval systems such as search engines is under the negative effects of issues such as the ambiguity inherent to natural language processing. Further, backlink models are blind to the authors of the pages, which entails that every link is equally important for the final ranking. This overlooks the fact that some links can be more relevant than others, depending on their authors or other parameters such as affiliation with organizations, and even in some cases, the links may have been created with a malicious purpose, such as spamming or the popular case of “Google bombing” (Mathes, 2005).

Probably the best example where the above phenomena are manifested is the field of scientometrics and in particular the scholar evaluation problem. Earlier from the introduction of the impact factor by Garfield (Garfield, 1972) several researchers have considered the development of metrics that can assist the evaluation of a scholar based on the reputation of his scientific works usually denoted as references to his/her work (Leydesdorff, 2001).

In particular the scholar evaluation problem can be formalized as follows:

Considering a scholar S with a collection of scientific output represented by a set of scholarly productions/publications as $A = \{a_1, a_2 \ldots, a_n\}$ then his/her research impact B is the aggregation if the impact of the individual scholarly works as:

$$B = \Omega\{P(a_1), P(a_2), \ldots, P(a_n)\}$$

Where, Ω is an aggregation operation (e.g. averaging) over the popularity $P(a_i)$ of his/her research production a_i. To the above formalization a set of open issues exist:

- Regarding the variability of his/her research work *how can we aggregate the impact of his production to a representative and generally accepted number?*
- How do we measure the popularity of the production? For example *do we count the same the reference provided to a research work by a technical report and by a generally accepted "prestigious" journal?*

Scientometrics provide an insight to open issues in the evaluation of the importance of documents that represent the scientific production whereas in our case we look into the evaluation of the importance of web pages. We follow the assumption that those web pages/documents are productions of social entities which transpose a degree of credibility from their social context.

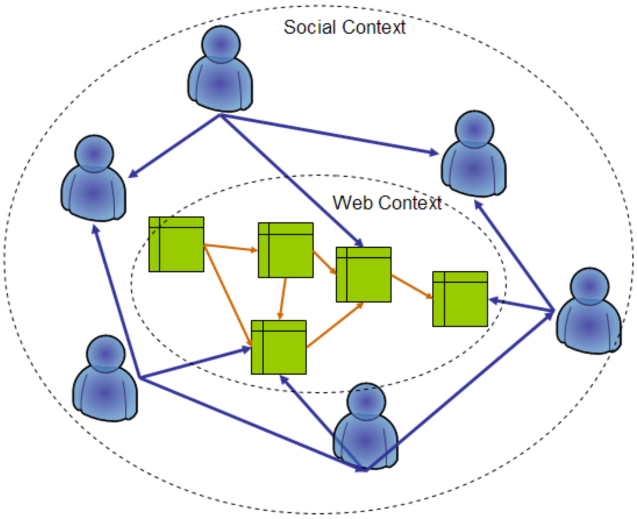


Figure 1: Layers of Context in the mixed mode network of authors or readers (blue) and web resources (green).

Based on this assumption we develop in the section that follow, a basic framework that underlines the core of the credibility models presented through the rest of this chapter both for the social and hypertextual context. As can be seen in Figure 1 our framework considers a kind of affiliation network (see definition bellow) where hypertext documents (web pages) are connected to their authors.

2.1 Formal Representations

Before continuing with the models representing credibility in both social and hypertextual contexts we should first have a look on the formalizations used in order to be able to comprehend the metrics and the models using them.

A social/hypertextual network is formalized as a graph $G := (V, E)$ which is an ordered pair of two sets. A set of vertices $V = (V_1, V_2, V_3, ..., V_n)$ which represents the social entities/pages and a set of edges $E = (E_{11}, E_{12}, E_{21}, ..., E_{ij})$ where E_{ij} represents the adjacent connection between the nodes i and j. In abstract form the network structure can be represented as a symmetric matrix (adjacency matrix) in which the nodes are listed in both axes and a Boolean value is assigned to E_{ij}, which depicts the existence of a relational tie between the entities, represented in E_{ij}.

Depending on the type of the network the adjacent connection may be:

- ***Edge:*** Indicates a connection between two nodes. In that case the graph is in the general form: $G := (V, E)$ When the connection is directed the graph has the general form $G := (V, E, E_a)$ denoting that the connection is a directional.
- ***Signed Edge:*** Indicates a connection between two nodes which is assigned with a value. In that case the graph is in the general form $G := (V, E, E_d)$ where the set E_d is the value set for the mapping $E \rightarrow E_d$. When the connection is directional the graph has the general form $G := (V, E, E_a, E_d)$ where E_a and E_d is the directed connection and value set respectively

An edge is often referenced in the literature as Arc. In a social network context, an edge and an arc differ on the fact that edge denotes a reciprocal connection whereas arc is a directed one. However when the network is a directed graph (digraph) those two terms become synonymous.

Furthermore depending on the group size a relation can be dichotomous, *trichotomous* or connecting subgroups.

- A *dichotomous* relation is a bivalence type of relational tie where the valued set E_d is mapped to [0; 1]. Dichotomous relationships form units called dyads that are used to study indirect properties between two individual entities. In the web for example the referring link from one page to another represents a dichotomous relation.
- A *trichotomous* or *triad* is a container of at most six dyads that can be either signed or unsigned. Usually in a triad a dyadic relation is referenced by a third entity which provides a point of common affiliation. For instance two pages are referenced by a directory page which can be studied in a triad.
- A *subgroup* relation is a more complex relation that acts as a container of triads where the members are interacting using a common flow or path.

2.2 Network Modes and Data

In principle network data consists of two types of variables: *Structural* and *Compositional*. *Compositional* variables represent the attributes of the entities that form apart the network. This kind of data can be for instance the theme of the webpage etc. *Structural* variables define the ties between the network entities and the mode under which the network is formed. Depending on the domain considered, different terminologies can be used.

Structural variables	Compositional variables
links	Web pages/documents
References/citations	Papers/articles

Table 1: Example structural and compositional variables for different applications/domains

Depending on the measurement of structural variables a network can have a mode. The term mode refers to the number of entities which the structural variables address in the network.

In particular we can have the two following types of networks:

- *One-mode networks:* This is the basic type of network where structural variables address the relational ties between entities belonging on the same set. The web is in an abstract form a one mode network linking web documents.
- *Two-mode networks or affiliation networks:* This type of networks contain two set of entities e.g. authors and articles. An affiliation network is a special kind of two mode network where at least one member of the set has a relational tie with the member of the other set. E.g. an author has written an article. The networks studied through the rest of this chapter are affiliation networks whereas authors produce documents with which are affiliated.

In social network theory several other kind of network exist such as for instance ego-centered networks or networks based on special dyadic relations. As aforementioned we intent to study only affiliation networks such as the author – article network or the reviewer-article network presented in section 5.

3 Reference Based Rankings in Web Context.

As noted in Sec.1, using the hyperlink structure of the web (or any linked document repository) to compute document rankings is a very successful approach (visible in the success of Google). The basic idea of link structure based ranking methods is that the prominence of a document p_d can be determined by looking at the documents citing

p_d (and the documents cited by p_d)[b]. A paper cited by many other papers must be somehow important otherwise it would not be cited so much. And a webpage linked by many other pages gains prominence (visibility).

Another view is the random surfer model: if an user starts on an arbitrary web page, follows some link to another page, follows a link from this page to a third one and so on, the likelihood for getting a certain page will depend on two aspects: from how many other pages it is linked and how visible the linking pages are.

3.1 PageRank

In 1976, Pinski and Narin (Pinski & Narin, 1976) computed the importance (rank) vis_d of a scientific journal p_d by using the weighted sum of the ranks vis_k of the journals p_k with papers citing p_d .[c] A slightly modified version of this algorithm (the PageRank algorithm) is used by the search engine Google[d] to calculate the visibility vis_d of a webpage p_d :

$$vis_d = (1-\alpha) + \alpha \sum_{p_k \in R_d} \frac{vis_k}{|C_k|} ,$$

where R_d is the set of pages citing p_d and C_k is the set of pages cited by p_k .
This resembles the idea of a random surfer: Starting at some webpage p_a with probability α she follows one of $|C_a|$ links while with $(1-\alpha)$ she stops following links and jumps randomly to some other page. Therefore $(1-\alpha)$ can be considered as a kind of basic visibility for every page. From the ranked document's perspective a page p_d can be reached by a direct (random) jump with probability $(1-\alpha)$ or by coming from one of the pages $p_k \in R_d$, where the probability to be on p_k is vis_k .

[b] Technically a document reference network is a graph with documents as nodes and references (links) as directed edges. The visibility of a node is determined by analyzing the link structure.

[c] The basic idea of this algorithm goes back to Seelay (1949) who used it in sociometrics (first published in L. Katz (1953), but without normalization by the number of outgoing edges and without damping term (which was introduced by Charles H. Hubbell (1965).

[d] We do not really know *how* google does its ranking. They claim that their ranking algorithm is based on PageRank with modifications.

The equation shown above gives a recursive definition of the visibility of a webpage because vis_a depends on the visibility of all pages p_i citing p_a , and vis_i depends on the visibilities of the pages citing p_i and so on. For n pages this is a system of n linear equations that has a solution. In praxis, solving this system of equations for some million pages would be much too expensive (takes to much time), therefore an iterative approach is used (for the details see (Page et al., 1998).

For even the iterative approach is rather time consuming, the rank (visibility) of all pages is not computed at query time (when users are waiting for their search results) but all documents are sorted by visibility offline, and this sorted list is used to fulfill search requests by selecting the first k documents matching the search term.[e]

3.2 HITS

The hypertext induced topic selection (HITS) is an alternative approach using the network structure for document ranking introduced by Kleinberg (Kleinberg, 1999). Kleinberg identifies two roles a page can fulfill: hub and authority. Hubs are pages referring to many authorities (e.g. linklists), authorities are pages that are linked by many hubs. Now for each page p_d its hub-value h_d and its authority value a_d can be computed:

$$h_d = \delta \sum_{p_k \in C_d} a_k \quad , \qquad a_d = \lambda \sum_{p_l \in R_d} h_l$$

Here a page p_d has a high h_d if it links many authorities (i.e. pages p_i with high a_i), in other words: it is a good hub, if it knows the important pages. The other way around p_d , has a high a_d , if it is linked by many good hubs, i.e. it is listed in the important link lists. While the authority a_d gives the visibility of the document ($vis_d = a_d$), h_d is only an auxiliary value.

But HITS not only differs from PageRank by the equations used to calculate the visibility. While the linear equation system is solved iteratively in both algorithms, they differ in the set of pages used. As PageRank does the calculation offline on all pages $p_i \in P$ and at query time selects the first k matching pages from this presorted list, HITS in a first step selects all pages matching the search term at query time (this gives a

[e] For the technical details see Brin and Page (1998).

subset $M \subset P$) and then adds all pages linked by pages from M ($M^C =_{p_i \in M} C_i$) and all pages linking pages from M ($M^R = ¿_{p_i \in M} R_i$). Then h_d and a_d are computed for all pages $p_d \in M' = \left(M^C \cup M \cup M^R \right)$). Only including pages somehow relevant to the search query (matching pages and neighbors) may improve the quality of the ranking, but produces high costs at query time for a_d and h_d are not computed offline because M' is dependent on the search term. This is a problem in praxis because a single search term can have up to some million matching webpages, even if there are strategies for pre-calculation.[f]

3.3 Other Approaches

There are other link structure based ranking algorithms, e.g. the hilltop algorithm[g] which in a first step identifies so called expert pages (pages linking to a number of unaffiliated pages[h]) and then each page is ranked by the rank of the expert pages it is linked by, or the TrustRank algorithm (Gyongyi, Garcia-Molina, & Pedersen, 2004), where link spam is semi-automatically identified by using a small seed of user-rated pages.

All these algorithms are best suited for networks with cyclic link structures. Repositories with mainly acyclic reference structure like citation networks of scientific papers (where each paper cites older ones) or online discussion groups where each thread consists of one initial posting and sequences of replies/follow-ups, need other visibility measures, such as those provided by (Malsch et al., 2006) where creation time of a document is much more important (as scientific papers or postings are not changed after publishing in contrast to webpages).

[f] There is no obvious reason not to use the hub and authority approach for offline calculation on the whole net. There are search engines claiming to use a HITS-based algorithm, but without giving detailed information.

[g] Also patented by Google.

[h] Unaffiliated means that they do not belong to the same organization. This is determined by comparing URLs, IP-addresses and network topology.

4 Linking and evaluating users

Ranking is not only a phenomenon that takes places in the web but derives from the archetype stratification of a society into degrees of power and dominance. In principle in social contexts two types of ranking exist: Explicit (Formal) and Implicit (Informal). In explicit ranking it is declared who is the one who has the authority to command due to insignia, symbols or status that he/she is being attributed from the beginning. On the other side, in implicit ranking it is the opinion and the behaviors of the surrounding entities that facilitates who has the authority to command and decide.

Both implicit and explicit forms of ranking may exist depending on the purpose of the social group and the kind of interactions between the members. However our study is focused only on implicit forms of ranking where status is a social attribute that emerges rather than being set. In sociology there are several kinds of studies that tend to explain how status emerges, therefore several research models exist. Those models are analyzed further in the sections bellow.

4.1 Sociometric Status and Prominence

In a social group one recognizes several strata that characterize their members with a kind of implicit rank which is known in the social science literature as “status” (Katz, 1953). Usually in a sociological interpretation status denotes power expressed in different contexts such as political or economical. The most basic theoretical implication of status is the availability of choices the entity receives in the network which gives the entity the advantage of negotiation over the others. Depending on the topology of the network status can be attributed to several nodes of the sociogram (see figure 3).

In SNA studies, status is depicted upon the generalization of the location of the actor in the network. In particular status is addressed by how strategic the position of that entity in a network is (e.g. how it affects the position of the others). Theoretical aspects of status were defined by Moreno and Jennings (Moreno and Jennings, 1945) as the instances of the sociometric “star” and “isolate”.

Quantifications of status employ techniques of graph theory such as the centrality index (Sabidussi, 1966; Freeman, 1979; Friedkin, 1991) which have been adjusted to the various representations of ties in a network. In our case we summarize the two most noteworthy measures of an actor in a directed network which are the “Prestige” and “Centrality”.

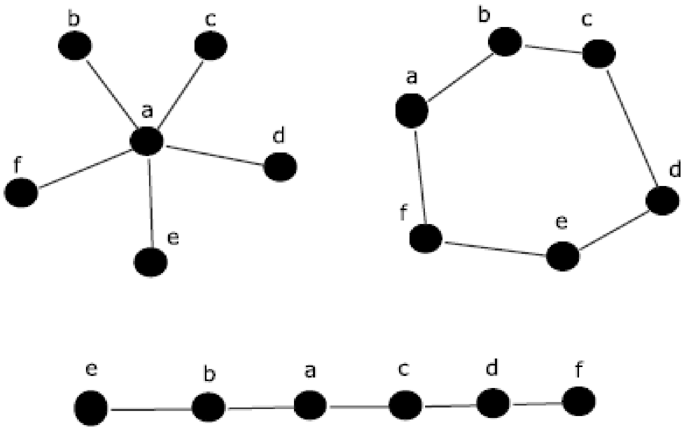


Figure 2: Types of connection degree in the network. It is obvious that the sociometric “star” (first diagram) is considered the one who has the highest inner degree thus is more popular. When there is reciprocity the degree of influence can be used to provide an indication of prestige in the network.

4.1.1 Prestige

A prestige measure is a direct representation of status which often employs the non-reciprocal connections/choices provided to that entity along with the influence that this entity might provide to the neighboring entities. In principle Prestige is a non-reciprocal characteristic of an entity. That means that if the entity is considered prestigious by another entity it doesn’t mean that this entity will consider the referring entity prestigious as well.

Inner degree

The simplest measure of prestige is the inner degree index or the popularity of an entity, which is defined as follows. Considering a graph $G := (V, E_A)$ and E_A the set of the directed connections $E = (e_{1,1}, e_{1,2}, ..., e_{i,j})$ between the members of the set V then the inner degree K_i^{in} of a vertex V_i is the sum of the incoming connections to that vertex.

$$K_i^{in} = \sum_{j=1} e_{j,i}$$

Degree of Influence

However the inner degree index of a vertex makes sense only in cases where a directional relationship is available (the connection is non-reciprocal) and this cannot be applied in the study of non directed networks. In that case the prestige is computed by the influence domain of the vertex. For a non directed graph $G := (V, E, \overline{E})$ the influence domain of a vertex V_i is the number or proportion of all other vertices which are connected by a path to that particular vertex.

$$\overline{d_i} = \frac{1}{N-1} \sum_{j=1} \overline{e}_{j,i}$$

On the above measure E represents the set of paths between the vertices V_i and V_j and $N-1$ is the number of all available nodes in the graph G (The total number of nodes $N = |V|$ minus the node that is subject to the metric).

Proximity Prestige

A combination of the above two metrics of prestige is known as the "Proximity Prestige" (PP_i) of vertex V_i encompasses the normalization of the inner degree of the vertex by its degree of influence such as:

$$PP_i = \frac{k_i^{in}}{\overline{d}_i}$$

4.1.2 Centrality

Unlike prestige measures which rely mainly on directional relations of the entities the centrality index of a graph can be calculated in various ways taking also into account the non-directional connections of the vertex which is examined. The most noteworthy measures of centrality are classified by the degree of analysis they employ in the graph.

Actor Degree Centrality

Actor degree centrality is the normalized index of the degree of an actor divided by the maximum number of vertices that exist in a network. Considering a graph $G = (V, E)$ with n vertices then the actor degree centrality $C_d(n_i)$ will be:

$$C_d(n_i) = \frac{d(n_i)}{n-1}$$

Where $n-1$ is the number of the remaining nodes in the graph G. Actor degree centrality is often interpreted in the literature as the “ego density” (Burt, 1982) of an actor since it evaluates the importance of the actor based on the ties that connect him/her to the other members of the network. The highest the actor degree centrality is, the most prominent this person is in a network, since an actor with a high degree can potentially directly influence the others.

Closeness or Distance Centrality

Another measure of centrality the closeness centrality, considers the “geodesic distance” of a node in a network. For two vertices a geodesic is defined as the length of the shortest path between them. For a graph $G=(V,E)$ the closeness centrality $C_c(V_i)$ of a vertex V_i is the sum of geodesic distances between that vertex and all the other vertices in the network.

$$C_c(V_i) = \frac{n-1}{\sum_{i=1} d(u_i,u_n)}$$

Where the function $d(u_i,u_n)$ calculates the length of the shortest path between the vertices i and j and n¡1 is the number of all other vertices in the network. The closeness centrality can be interpreted as a measurement of the influence of a vertex in a graph: the higher its value, the easiest it is for that vertex to spread information into that network. Distance Centrality can be valuable in networks where the actor possesses transitive properties that through transposition can be spread through direct connections such as the case of reputation.

Betweenness Centrality

Betweenness is the most celebrated measure of centrality since not only measures the prominence of a node based on the position or the activity but also the influence of the node in information or activity passed to other nodes Considering a graph $G=(V,E)$ with n vertices, the betweenness $C_B(u)$ for vertex $n \in V$ is:

$$C_B(u) = \frac{\sum_{s \neq u \neq t \in V} \sigma_{st}(u)}{(n-1)(n-2)}$$

Where $C_B(u) = 1$ if and only if the shortest path from s to t passes through v and 0 otherwise. Betweenness can be the basis to interpret roles such as the "gatekeeper" or the "broker" which are studies extensively in communication networks. A vertex is considered as a "gatekeeper" if its betweenness and inner degree is relatively high. The "broker" is a vertex which has relatively high outer degree and betweenness.

4.2 Cliques and Cohesive Subgroups

Although the metrics presented above provide an insight to the evaluation of dyadic relations it's often necessary to adapt the above measures to the study of larger subset of entities that share common characteristics such as for instance web pages regarding the same topic. In network analysis this kind of structure is a subgroup or a clique.

The graph theoretic definition of a clique has its qualitative interpretation in sociometric research as a discrete social structure (subgroup) shaped and contained within the structure of a social group (Scott, 2000). The number of cliques contained in a group and their diameter is subject to the *cohesion* that characterizes the social structure. Cohesion depicts how strong the ties between the members of a social group are and how homogeneous are their properties regarding the overall structure. The highest the cohesion of a group is then the minimal becomes the size of the cliques (subgroups) contained and formatted within the group. According to Friedklin (Friedkin, 1998) cohesion is a factor of influence of the individual by the group standards. A member of a group is highly connected with the group if and only if he/she accepts and/or possesses the characteristics of the groups.

The study of cliques can give indications regarding the qualitative aspects of connections between members of the same structure. For instance, authors of blogs that provides reference to other blogs that they read and backwards, form a clique of the general graph of bloggers. Studying how this clique evolves we can model for instance the spreading of news and blog posts on a perspective of social influence, which can provide as a way of evaluating the authoritativeness of a specific piece of information appearing on a website or a blog.

4.3 The FOAF vocabulary

Although sociometric models can provide several pathways on evaluating the importance of users, there is a crucial need for an accurate way on mining the relations between them in order to be able to apply those models. This can be done by using expressive vocabularies such as FOAF.

The Friend-of-a-Friend vocabulary (Brickley & Miller, 2005) is an expressive vocabulary set which syntax is based on RDF syntax (Klyne & Carroll, 2004) that is gaining popularity nowadays as it is used to express the connections between social entities in the web along with their hypertextual properties such as their homepages or the emails. In a best case scenario the author of a web page (information resource) will attach his FOAF profile in the resource in order to make it identifiable as an own production by the visitors of that page. This can be observed clearly in cases such as Blogs where the information resource represents the person that expresses his/her views through the blog. Furthermore connection between blogs represents also a kind of a directional dichotomous tie between the authors of those blogs.

Vocabulary Element	**Description**	**Type of Relational tie**
foaf:knows	links foaf:persons	Direct
foaf:member	Provides affiliation/membership(relates an entity with a social group)	indirect
foaf:maker	Indicates authorship (relates an information resource with it's creator)	indirect
foaf:based_near	Indicates a spatial affiliation of a social entity	indirect
foaf:currentproject	Indicates a temporal affiliation of a social entity with a project or an activity	indirect

Table 1: Some representative elements of the FOAF vocabulary and the representation of the relational tie

In FOAF standard RDF syntax is used to describe the relations between various acquaintances (relations) of the person described by the FOAF profile. This relation is depicted in the `<foaf:knows>` predicate which denotes that the person who has a description of <foaf:knows> in his profile for another person, has a social connection with that person as well. For example a FOAF profile for one of the authors of this chapter and the connection he has with his coauthors can be described by the following fragment of RDF code:

```
<?xml version="1.0" encoding="UTF-8" ?>
<rdf:RDF
   xmlns:xml="http://www.w3.org/XML/1998/namespace"
   xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
   xmlns:rdfs=”http://www.w3.org/2000/01/rdf-schema#”>
<foaf:Person rdf:nodeID="friend1">
   <foaf:name>Nikolaos Korfiatis</foaf:name>
   <foaf:mbox rdf:resource="mailto:nk.inf@cbs.dk"/>
   </foaf:Person>
<foaf:Person rdf:nodeID="friend2">
   <foaf:name>Claudia Hess</foaf:name>
   <foaf:mbox rdf:resource="mailto:claudia.hess@wiai.uni-bamberg.de"/>
  </foaf:Person>
<foaf:Person rdf:nodeID="friend3">
   <foaf:name>Klauss Stein</foaf:name>
   <foaf:mbox rdf:resource="mailto:klaus.stein@wiai.uni-bamberg.de"/>
   </foaf:Person>
<foaf:Person rdf:nodeID="me">
   <foaf:title>Dr. </foaf:title>
   <foaf:givenName>Miguel-Angel</foaf:givenName>
   <foaf:family_name>Sicilia</foaf:family_name>
   <foaf:mbox rdf:resource="mailto:msicilia@uah.es"/>
   <foaf:knows rdf:nodeID="friend1"/>
   <rel:collaboratesWith rdf:nodeID="friend2"/>
   <rel:collaboratesWith rdf:nodeID="friend3"/>
</foaf:Person>
</rdf:RDF>
```

Research on descriptions of social relations in the semantic web[i] is an undergoing effort which has been initiated lately to address the various concerns and sociological implications for the expressiveness of social connections in the web. One particular issue is that although the FOAF vocabulary (see the table 1) has a set of properties for the description of several kinds of relationships the <foaf:knows> property is the most common relation that is expressed in a publicly available FOAF profile. According to the general discussion in the FOAF project the reason for this is that many users prefer not to express their strength of social connections publicly than to do it with a general way which the <foaf:knows> property implies.

[i] See the proceedings on the 1st FOAF Workshop, Galway

Within well-defined application domains however, more expressive constructs are required for describing the relationships between persons. Using the social relationships for recommending sensitive information such as Avesani, Massa & Tiella (Avesani, Massa & Tiella, 2005) in Moleskiing, a platform for exchanging information on ski tours (e.g. snow conditions, risk of avalanches), it is highly critical to get these information only from persons that you consider as trustworthy, and not from everyone you know (or who is known by some of your friends). Golbeck, Parcia and Hendler (Golbeck, Garcia & Hendler, 2003) therefore provided an extension to the FOAF vocabulary so that users can explicitly specify their degree of trust in another person. Specifying the degree of trust in someone with respect to her or his recommendations of scientific papers gives according to the trust ontology[j] by Golbeck et al. the following:

```
<foaf:Person rdf:ID="Claudia">
      <trust:trustsRegarding>
            <trust:TopicalTrust>
                  <trust:trustSubject
rdf:resource="#RecommendationsForPapers"/>
                  <trust:trustedPerson rdf:resource="#Klaus"/>
                  <trust:trustValue>9</trust:trustValue>
            </trust:TopicalTrust>
      </trust:trustsRegarding>
</foaf:Person>
```

5 Evaluating web page importance through the analysis of social ties.

Based on the two previous chapters, we present different approaches to integrate the social information in link-based analysis and the respective algorithms for measuring the ranking of web pages.

5.1 Integrating imprecise expressions of the social context in PageRank

Garcia and Siclia (Sicilia & Garcia, 2005) have introduced an approach for integrating the social context of an information resource into the core of ranking algorithms such as PageRank. Although a flavor of PageRank (PageRank with priors) can be used to normalize existing rankings the core of the method relies on the expression of the

[j] The trust ontology is available at http://trust.mindswap.org/ont/trust.owl (last access date 30th of May)

relational ties of the social context (e.g. connections in a social network) using imprecise expressions.

The idea is that of using imprecise assessments of social relationships as a weighting factor for algorithms like PageRank. Whereas imprecise expressions map better to social relation due to the difficulty of making an accurate estimation of the strength of the relational ties and their influence to the rest of the structure members.

The model begins by computing a metric called PeopleRank (PPR). PPR is based on the declared relationships `<foaf:knows>` connecting pairs of `<foaf:Person>` specifications. As aforementioned in section 4.3 the FOAF vocabulary has deliberately avoided more specific forms of relation like friendship or endorsement, since social attitudes and conventions on this topic vary greatly between countries and cultures.

In consequence, the strength is provided explicitly as part of the link. In the case of absence of such value, an “indefinite” middle value is used. With the above, the PeopleRank can be defined by simply adapting the original PageRank definition:

We assume a social entity/person A that has persons $T_1 \dots T_n$ which declare they know him/her (i.e., provide FOAF social pointers to it). The parameter d is a damping factor which can be set between 0 and 1. [...]. Also C(A) is defined as the number of (interpreted) declarations of backlinks going out of A's FOAF profile.

Following the above definition the PeopleRank of a person A ($PpR(A)$) can be defined as follows:

$$PpR(A) = (1-d) + d\sum_{i=1}^{n}\frac{PpR(T_i)}{C(T_i)}$$

PageRank	**PeopleRank**
Intuitively, pages that are well cited from many places around the web are worth looking at.	Intuitively, the trust on the quality of pages is related to the degree of confidence we have on their authors.
Also, pages that have perhaps only one citation from something like the Yahoo! home-page are also generally worth	Pages authored or owned by people with a larger positive prestige should somewhat be considered more relevant.

looking at	

Table 2: Intuitions behind the development of the PeopleRank algorithm. Adapted from Garcia and Sicilia (Sicilia & Garcia, 2005)**.**

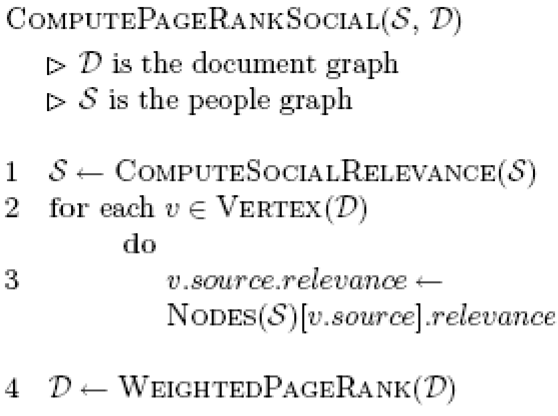
Even though the idea of ranking by peer's declarations seems intuitive and is coherent with the concepts explained in Section 4, the original intuitive justification provided for PageRank requires a re-formulation. Table 1 provides the original and the socially-oriented justifications.

An important parameter of the PpR computation is that declarations of social awareness should be interpreted. Here the management of vagueness plays a role, since there is no single model or framework that provides a metric for social distance.

The relevance r of a social tie in that model is provided by the following general expression, which provides a value for each edge (e_1, e_2) in the directed graph formed by the explicitly declared social relationships.

$r((p_1, p_2)) = S((p_1, p_2)) \cdot e((p_1, p_2))$ Asserting that $p_1 \neq p_2$

The relevance r of an edge is determined from a degree of strength S (in a scale of fuzzy numbers [~0; ~ 10]) weighted by a degree of evidence e about the relationship. These strengths could be provided by extending the current FOAF schema with an additional attribute.



```
ComputePageRankSocial($\mathcal{S}$, $\mathcal{D}$)
  ▷ $\mathcal{D}$ is the document graph
  ▷ $\mathcal{S}$ is the people graph

  $\mathcal{S} \leftarrow$ ComputeSocialRelevance($\mathcal{S}$)
  for each $v \in$ Vertex($\mathcal{D}$)
        do
            $v.source.relevance \leftarrow$
            Nodes($\mathcal{S}$)[$v.source$].$relevance$

  $\mathcal{D} \leftarrow$ WeightedPageRank($\mathcal{D}$)
```

Figure 3: The steps of the PpR algorithm.

Degrees of evidence support a notion of "external" evidence on the relation that completes the subjectively stated strength. The following expression provides a formalization for this evidence that is based both on the perceptions of "third parties" and/or affiliations.

$e((p_1, p_2)) = \max(\Phi_{i \in U} S^{P_i}(p_1, p_2), P(p_1, p_2))$ Asserting again that $p_i \neq p_1 \neq p_2$ and having $i \in U$

According to the above expression, the strengths of a social tie provided by third parties p_i in a group of users U are aggregated through simple fuzzy averaging (Φ), and the evidence provided by work in common projects in which two persons $P(p_1, p_2)$ (p1; p2) collaborate $P(p_1, p_2)$ is obtained from FOAF declarations.

Concretely if we follow the affiliation by the FOAF descriptor e.g. `foaf:Project`, people co-working in a project (as declared by `<foaf:pastProject>`) are credited an amount of 1, while people that were co-workers (as declared by `<foaf:pastProject>`) are credited an amount of 0.1 per common past project.

These somewhat arbitrary values should be complemented by more detailed measures in the future. Note that "projects" in FOAF are not only formal professional projects but the concept is open to any informal activity even non-profit or re-creative. Other usage oriented metrics could be used to complement this approach, such as data obtained from web usage mining.

Of course the above model can be modified and more parameters could be added. However but it provides an initial straight forward model for the integration of social ties in the PageRank from which further empirical analysis could be carried out.

5.2 Integrating trust networks and document reference networks

An alternative approach for integrating different types of networks has been presented by Hess, Stein & Schlieder (Hess et al., 2006; Stein & Hess, 2005). They propose to integrate information from a trust network between persons and a document reference network. A trust network has been chosen as social network because trust relationships represent a strong basis for recommendations. In contrast to social networks based only on a 'knows' or 'co-workers' relationship, trust relationships are more expressive, although more difficult to obtain. Trust networks cannot automatically be extracted from data on the web but users have to indicate their trust relationships as they are doing it already in an increasing number of trust-based social networks. Examples for such trust networks are the epinions web of trust in which users rate other users with respect to the quality of their product reviews, or social networking services such as Orkut and Linked-in, in which you can rate your friends with respect to their trustworthiness (among other criteria such as 'coolness'). In trust networks, trust values can be inferred for indirectly connected persons by using trust metrics such as those presented in Goldbeck, Parsia and Hendler (Golbeck, Parsia, & Hendler, 2003) or Ziegler and Lausen (Ziegler & Lausen, 2004). This reflects the fact that we trust in the real world the friends of our friends to a certain degree. Trust is therefore not transitive in the strict mathematical sense but decreases with each additional step in the trust chain. In contrast to the measures presented in section 4, most trust metrics calculate highly personalized trust values for the users in the network: a trust value for a user is always computed from the perspective of a specific user. The evaluations of one and the same user can therefore greatly vary between users depending on their personal trust relationships.

We can distinguish two roles that actors play in the trust networks: they can either be the authors or the reviewers of documents. The term 'reviewer' encompasses all persons having an opinion about the document, such as persons who read the document or editors who accepted the document for publication. Both cases are addressed in the following separately, leading to two different trust-enhanced visibility functions.

Case 1: Reviewers & Documents

In this first type of a two-layer-architecture, a trust network between reviewers is connected with a document reference network. In the document reference network, documents are linked via citations or hyperlinks. In the second layer, reviewers make statements about the degree of trust they have in other reviewers in making 'good' reviews, i.e. above all that they apply similar criteria to the evaluation as themselves. Depending on the trust metric that is used for inferring trust values between indirectly connected persons, trust statements are in [0, 1] with 0 for no trust and 1 for full trust, or even [-1, 1] ranging from distrust[k] to full trust. Both layers are connected via reviews, i.e. reviewers express their opinion on documents.

We have now two types of information, firstly the reviews of documents made by persons in which the requesting user has a certain degree of trust, i.e., *trust-weighted reviews*, and secondly the *visibility* of the documents calculated on the basis of the document network. We now integrate the trust-weighted reviews in the visibility function and hence personalize it. This new trust-review-weighted-visibility function (in the following called twr-visibility) has the property that reviews influence the rank of a document to the degree of trust in the reviewer, i.e. a review by someone deemed as highly trustworthy influences the rank to a considerable part whereas reviews of less trustworthy persons have few influence. Having only untrustworthy reviewers, the user likely prefers a recommendation merely based on the visibility of the documents.

This architecture permits to deal with two interpolation problems. Firstly, propagating trust values in the trust network makes it possible to consider reviews of users who are directly linked via a trust statement as well as indirectly via some “friends”, i.e. via chains of trust. Secondly, reviews influence not only the visibility of the reviewed papers but also indirectly the visibility of other papers, namely of those papers cited by the reviewed documents. Figure 4 illustrates these interpolations. Person 1 asks for a recommendation about document 4 that has not directly been reviewed but which visibility is influenced by the twr-visibility of document 2 that has been reviewed by a person deemed as trustworthy by person 1.

[k] Although distrust statements are difficult to handle in propagation: should I trust someone who is distrusted by someone I distrust? Or should I distrust this person even more?

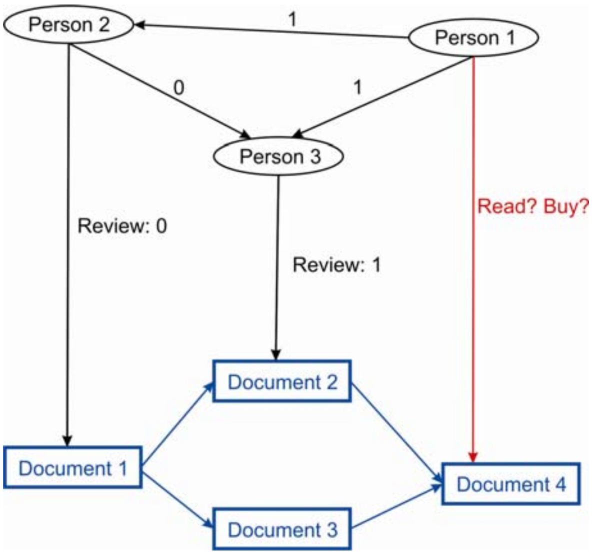


Figure 4- Interpolation in a Document and Reviewer Network
(from Hess, Stein and Schlieder, 2006)

The following function gives the twr-visibility for a document p_d in which r_d is the review of document p_d and t_{rd} is the trust in the reviewer. The trust in the reviewer is directly taken as the trust in the review.

$$vf^{TWR}(p_d)=t_{r_d} r_d+(1-t_{r_d})vf(p_d)$$

The trust in the reviewers of a document therefore determines to which degree the trust-weighted reviews influence the twr-visibility. In the case that the trust in the reviewer is not absolute (trust=1.0), the twr-visibility of the documents citing this document determines its twr-visiblity. Reviews by trustworthy persons therefore influence indirectly the twr-visibility of not directly evaluated documents. The vf^{TWR} function permits to use any visibility function such as the PageRank-formula for the propagation of the trust-enhanced visibilities. The indirect connections in the trust network are computed before calculating the twr-visibility. Any trust metric can be used for the propagation of the trust statements. The above presented function therefore represents a general framework for integrating trust and document reference networks. A detailed description of this approach and a simulation demonstrating the personalization provided by this function is in Hess, Stein and Schlieder (Hess, Stein & Schlieder, 2006).

Case 2: Authors & Documents

In the second case, the trust network is built up between authors of documents. The trust statements refer to the trust in the quality and accurateness of the evaluated author's documents. As in the case of the reviewer network, the range of the trust values depends on the trust metric chosen. Authors are connected via an 'is-author' relationship with the documents they have written. Documents can be written by several persons. 'Is-author'-edges are not weighted. The structure of the document network is identical as in the first case: references connect the documents.

This two-layered architecture permits to calculate a trust-weighted visibility for each document in the document reference network. The idea is that the semantics of links will differ based on whether there is a trust or a distrust relationship between the citing and the cited author. This reflects the fact that a link can be set in different contexts. On the one hand, a link can affirm the authority, the importance or the high quality of a document, for example the citing author confirms the results of the original work by his or her own experiments and therefore validates it. On the other, disagreement can be expressed in a link ranging from different opinions to suspicions that information is incorrect or even faked.

Several steps are required to compute the trust-weighted visibility:

1. Trust relationships between the authors of a citing and the cited paper are attributed to the reference between the documents in the document network. In the case of co-authorship, more than one trust value is available. Assuming that the opinions of coauthors do not vary extremely, the average of the trust values is attributed to the reference.
2. The trust values which are attributed to the references have to be transformed into edge weights, so an edge between two documents p_a and p_b has a weight $w_{a \to b}$ reflecting the relationship (trust) between the authors of the corresponding papers. This step is necessary as attributed trust values might be negative due to distrust between the authors. In a visibility function however, we need edge weights > 0. A mapping function defines how trust values are transferred into edge weights. Depending on the definition of the mapping function, different trust semantics can be realized. The mapping function can for instance be defined in a way that only such references are heavy weighted which express

high trust whereas another definition could give a high weight to distrust values in order to get an overview on papers which are not appreciated in a certain community.

3. The trust-enhanced visibility of the documents can now be calculated by a visibility function that considers the edge weights. We illustrate this general framework by using a concrete visibility function for calculating the trust-weighted visibility, namely the PageRank. The original PageRank formula is modified such that each page p_i contributes not any longer with[1] $\frac{vis_i}{|C_i|}$ to the visibility vis_d of another page p_d but distributes its visibility according to the edge weight, so we get

$$vis_d = (1-\alpha) + \alpha \sum_{p_i \in R_d} \frac{w_{i \to d}}{\sum_{p_j \in C_k} w_{i \to j}} vis_i$$

This function can further be personalized so that it calculates individual recommendations for each user. It can also be adapted such that documents are highly ranked which are controversially discussed. This approach again is a general framework. Other visibility functions can be applied instead of the PageRank.

6 Case study: Evaluating Contributions in the Wikipedia

Having presented metrics of social and hypertextual evaluation we choose as a case study to model the authoritativeness of a system rich of social interactions such as Wikipedia. Wikipedia is based on Wiki software (Leuf & Cunningham, 2001) and is considered as one of the most successful collaborative editing projects on the web since it currently contains over 1 million articlesm and it has an extensive community of contributors contributing content and improving the quality of the articles.

[1] with C_i being the set of pages cited by p_i

[m] Statistics and data for the English Language Wikipedia. For further information about the current size of the wikipedia the reader can visit : http://en.wikipedia.org/wiki/Wikipedia:Statistics (Last access date 30th of May, 2006)

Wikipedia is a voluntary project and since it facilitates a large amount of social connections over a common affiliation we consider it as an interesting example to discuss authoritativeness over evolving documents. In our case we consider the following social interactions:

- When a contributor edits content that has been submitted by someone else then it establishes a tie with him/her. This is depicted by an acceptance factor which represents the percentage of the content of the previous contributor that is visible after.
- Every contributor that has a single or more contribution to the article establishes a relational tie with the other content contributors of the article. Evidence of participation in common projects strengthens this tie.

As can be seen in Figure 5 we define two different networks the articles network and the contributors network.

- ***The Articles Network:*** Every article in the WIKIPEDIA contains reference to other articles as well as external references. A set of links used for classification purposes is also available in most of the active articles of the encyclopaedia. Every article represents a vertex in the article network and the internal connections between the articles the edges of the network.

- ***The Contributors Network:*** WIKIPEDIA is a collaborative writing effort which means that an article has multiple contributors. We assume that a contributor establishes relationship with another contributor if they work on the same article. In the resulted weighted network each contributor is represented by a vertex and their social ties (positive or negative) are represented by an edge denoting the sequence of their social interaction.

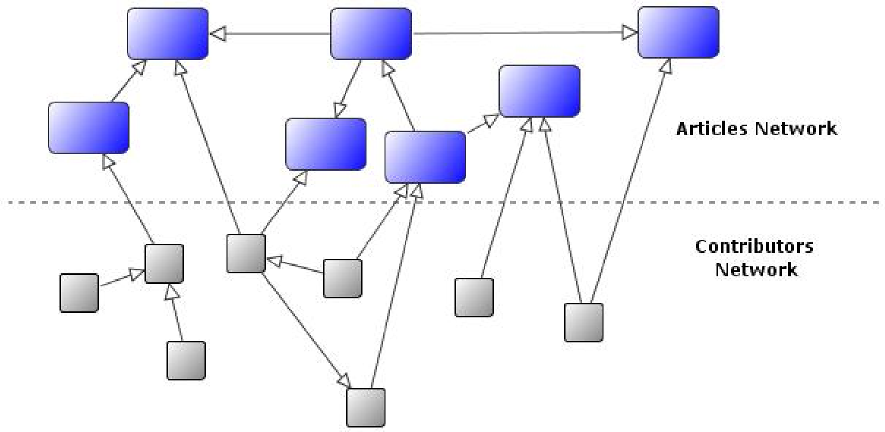


Figure 5: Network layers in the wiki publication model. Contributors are linked together by working on common projects (articles) in the same topic.

The development of those metrics is based on the following assumptions:

- The more decentralized the editing of an article, then the better this article represents a consensus about it.
- The contributors whose content has been most accepted (seen from the result of the diff operation in the wiki) are attributed a level of authority regarding the article.
- This level of authority remains only in the domain of the article. However, domains which belong in the same topic can retain the level of authority for a contributor.

The graph that we model is a signed directed network with arcs signed as a factor depicting the level of acceptance of the content submitted by contributor ***A*** and accepted by contributor ***B***. In order to model the authoritativeness of contributors, we selected the degree centrality. In our case study, we use the degree centrality index which is the simplest definition of centrality and as aforementioned is based on the incoming and outgoing adjacent connections to other contributors in an article graph. To measure the centrality at an individual level, we define the contributor degree centrality; and to an article level, the article degree centralization which represents the variability of Contributor Degree Centrality over the specific article.

6.1 Contributor Degree Centrality

As have been already discussed the inner degree (in a graph theoretic interpretation: the amount of edges coming into a node) represents the choices the actor has over a set of other actors. However, in our wiki network model the amount of incoming edges represents edits to the text; therefore the metric of inner degree is the opposite, meaning that the person with the biggest inner degree has the biggest amount of objection/rejection in the contributor community and thus receives a kind of negative evaluation from his/her fellow contributors. On the other hand, the outer-degree of the vertex represents edits/participation in several parts of the article and thus gives a clue to the activity of the person in relation to the article and the domain. Mathematically we can represent such formalism as follows: Considering a graph representing the network of contributors for an article contributed in the wiki, then the Contributor Degree Centrality - a contextualized expression of actor degree centrality - is a degree index of the adjacent connections between the contributor and others who edit the article. From graph theory, the outer degree of a vertex is the cumulative value of its adjacent connections.

$$C_D(n_i) = d(n_i) = \sum_j x_{ij}$$

The adjacent x_{ij} represents the relational tie between the contributors and their contribution over the domain of the article. This also is characterized by the visibility of the contribution in the final article and can be either 1 or 0. To provide the centrality, we divide the degree with the highest obtained degree from the graph which in graph theory is proved to be the number of remaining vertices (g) minus the ego (g-1). Therefore the contributor degree centrality can be calculated as an instance of the actor centrality index:

$$C'_D(n_i) = \frac{d(n_i)}{g-1}$$

6.2 Article Degree Centralization

We define an Article's degree centralization C_{DM} as the variability of the individual contributor centrality indices. The $C_D(n^*)$ represents the largest observed contributor degree centrality

$$C_{DM} = \frac{\sum_i^g [C_D(n^*) - C_D(n_i)]}{(g-1)(g-2)}$$

Again we divide the variability with the highest variability observed in the graph in order to get a normalized calculation.

Cluster (Outerdegree)	Freq	Freq%	CumFreq	CumFreq%	Representative
1	1	0.4329	1	0.4329	65.6.92.153
2	199	86.1472	200	86.5801	82.3.32.71
4	14	6.0606	214	92.6407	80.202.248.28
6	6	2.5974	220	95.2381	Snowspinner
8	3	1.2987	223	96.5368	Tim Ivorson
10	1	0.4329	224	96.9697	StirlingNewberry
12	2	0.8658	226	97.8355	24.162.198.123
16	2	0.8658	228	98.7013	JimWae
18	1	0.4329	229	99.1342	Jjshapiro
20	1	0.4329	230	99.5671	SlimVirgin
31	1	0.4329 s	231	100	Amerindianart

Table 3: Contributor Degree Centrality for the wikipedia article "Immanuel Kant".

6.3 Interpretation of the results

Having defined the metrics, we provide some indicative measurements for a set of articles from the category philosophy.

Article Name	Number of Contributors	Article Degree Centralization (max 1)
Adam Smith	276	0.039114
Aristotle	274	0.0232
Immanuel Kant	231	0.20484
Johann Wolfgang von Goethe	242	0.016682
John Locke	292	0.008581

Karl Marx	232	0.006601
Ludwig Wittgenstein	220	0.006328
Philosophy	280	0.00254
Plato	284	0.001207
Socrates	289	0.000405

Table 4: Articles used in the case study along with the number of contributors and their degree centralization

As can be observed from the table 3, the article degree centralization is relatively low because of the small collections of articles used in the case study and the inter-connections of the actors in the domain. However, it is enough to let us discuss some qualitative interpretations such as:

- The dispersion of the actor indices denotes how dependent this article is on individual contributors. For instance, if an article has a very low degree of centralization, then it means that the social process to shape it was highly distributed, thus resulting in an article which has been submitted by multiple authorities. In our case, the articles represent a low degree of centralization which means that contributions have been done by individuals with interests in other domains as well.
- The range of the group degree centralization reflects the heterogeneity of the authoring sources of the article. In our case, the article “Immanuel Kant” has a significantly higher degree of centralization which means that it has been contributed by authorities most concentrated in the domain of the article and thus have contributed to other articles.

Contributors with higher inter-relation over the same domain represent higher authorities based on the assumptions that their interest spans the domain to which the article belongs and therefore they have conducted background research regarding the material they have contributed. On the other hand contributors with lesser authority tend to have their content erased/objected by contributors with higher authority.

7 Conclusions & Outlook

Current information retrieval models that use links to compute rankings as measures of the popularity of the Web pages fail to consider the social context that is tacit in the authorship of the pages and their links. Social Network Analysis provide sound models for dealing with these kind of models, and can be combined with existing backlink models to come up with richer models.

This chapter has provided background on backlink models, social network concepts and their potential application to a combined model of Web retrieval that considers the links but also the relations between the creators of the documents and the links.

However research issues are open to the direction of applying such models in real context since the social context is something that is very difficult to extract. The semantic web can contribute to this direction by advancing the development and use of vocabularies that can express social relations in various ways and associate them with content. Nevertheless privacy issues should be also considered since the expression of social relations in a publicly accessible information space is something that exhibits vulnerabilities in cases such as “phising” attacks and social engineering (Levy, 2004).

8 Biographical Notes

Nikolaos Korfiatis is a PhD Student in the Department of Informatics at the Copenhagen Business School (CBS), Denmark. He is currently pursuing his PhD in the area of recommender systems with a research emphasis on the applications from behavioral economics and social network theory into the design of more effective recommender systems. He obtained a BSc in Information Systems from Athens University of Economics and Business (AUEB), Greece in 2004 and a Master of Science in Engineering of Interactive Systems from the Royal Institute of Technology (KTH), Stockholm Sweden in 2006.

Miguel-Ángel Sicilia obtained a Ms.C. degree in Computer Science from the Pontifical University of Salamanca, Madrid (Spain) in 1996 and a Ph.D. degree from the Carlos III

University in 2003. He worked as a software architect in e-commerce consulting firms, being part of the development team of a Web personalization framework at Intelligent Software Components (iSOCO). From 2002 to 2003 he worked as a full-time lecturer at the Carlos III University, after which he joined the University of Alcalá. His research interests are driven by the directions of the Information Engineering Research Unit which he directs.

Claudia Hess received her diploma in Information Systems from Bamberg University, Germany in 2004. Currently, she is research assistant at the Laboratory for Semantic Information Technology at Bamberg University. In her Ph.D. studies, which are jointly supervised at Bamberg University and University Paris 11, France, she is interested in the integration of social relationship data, above all trust relationships, into recommendation and ranking technology.

Klaus Stein obtained an Informatics (CS) diploma in 1998 and a doctoral (PhD) degree in 2003 from the Munich University of Technology, Germany in the field of spatial cognition. He currently works on computer-mediated communication processes (COM) at the Laboratory for Semantic Information Technology at Bamberg University, Germany.

Christoph Schlieder is Professor and Chair of Computing in the Cultural Sciences since 2002 at Bamberg University, Germany. He holds a Ph.D. and a Habilitation degree in computer science, both from the University of Hamburg. Before coming to Bamberg he was professor of computer science at the University of Bremen where he headed the Artificial Intelligence Research Group. His primary research interests lie in developing and applying methods from semantic information processing to problems from the cultural sciences. Applications areas of his research work include GIS and mobile assistance technologies as well as digital archives.

9 References

Avesani, P., Massa, P. & Tiella, R. (2005). A Trustenhanced Recommender System application: Moleskiing. *Proceedings of the 20th ACM Symposium on Applied Computing*, Santa Fe, New Mexico, USA.

Baeza-Yates, R., & Ribeiro-Neto, B. (1999). *Modern Information Retrieval.* Addison-Wesley Harlow, England.

Berners-Lee, T., & Fischetti, M. (1999). *Weaving the web: The original design and ultimate destiny of the world wide web by its inventor.* Harper, San Francisco.

Borner, K., Maru, J. T., & Goldstone, R. L. (2004). The simultaneous evolution of author and paper networks. *Proceedings of the National Academy of Sciences, 101*(5266 – 5273)

Brickley, D., & Miller, L. (2005). FOAF Vocabulary Specification.

Brin, S., & Page, L. (1998). The anatomy of a large-scale hypertextual web search engine. *Proceedings of the Seventh International Conference on World Wide Web,* Brisbane, Australia

Brown, J. S., & Duguid, P. (2002). *The Social Life of Information.* Harvard Business School Press.

Burt, R. S. (1980). Models of Network Structure. *Annual Review of Sociology, 6*, 79-141.

Conklin, J. (1987). Hypertext: An introduction and survey. *IEEE Computer, 20*(9), 17-41.

Dhyani, D., Keong, N. W., & Bhowmick, S. S. (2002). A survey of web metrics. *ACM Computing Surveys, 34*(4), 469-503.

Faloutsos, C. (1985). Access methods for text. *ACM Computing Surveys, 17*(1), 49-74.

Freeman, L. C. (1979). Centrality in social networks: Conceptual clarification. *Social Networks, 1*(3), 215-239.

Friedkin, N. E. (1991). Theoretical foundations for centrality measures. *The American Journal of Sociology, 96*(6), 1478-1504.

Friedkin, N. E. (1998). *A structural theory of social influence (Structural analysis in the Social Sciences).* Cambridge University Press.

Garfield, E. (1972). Citation analysis as a tool in journal evaluation. *Science, 178*(4060), 471-479.

Golbeck, J., Parsia, B., & Hendler, J. (2003). Trust networks on the semantic web. *Proceedings of Cooperative Intelligent Agents,* Helsinki, Finland.

Gyongyi, Z., Garcia-Molina, H., & Pedersen, J. (2004). Combating web spam with TrustRank. *Proceedings of VLDB, 4*

Hess, C., Stein, K., & Schlieder, C. (2006). Trust-enhanced visibility for personalized document recommendations. *Proceedings of the 21st ACM Symposium on Applied Computing,* Dijon, France.

Hubbell, C. H. (1965). An input-output approach to clique identification. *Sociometry, 28*(4), 377-399.

Katz, L. (1953). A new status index derived from sociometric analysis. *Psychometrika, 18*, 39-43.

Kleinberg, J. M. (1999). Authoritative sources in a hyperlinked environment. *Journal of the ACM (JACM)*, 46(5), 604-632.

Klyne, G., & Carroll, J. J. (2004). Resource description framework (RDF): Concepts and abstract syntax. *W3C Recommendation,W3C.*

Korfiatis, N., & Naeve, A. (2005). Evaluating wiki contributions using social networks: A case study on Wikipedia. Proceedings of the *First on-Line Conference on Metadata and Semantics Research (MTSR'05).* Rinton Press.

Leuf, B., & Cunningham, W. (2001). *The wiki way: Collaboration and sharing on the Internet.* Addison-Wesley Professional.

Levy, E. (2004). Criminals become tech savvy. *Security & Privacy Magazine, IEEE, 2*(2), 65-68.

Leydesdorff, L. (2001). *The challenge of scientometrics: The development, measurement, and self-organization of scientific communications.* Universal Publishers.

Malsch, T., Schlieder, C., Kiefer, P., Lubcke, M., Perschke, R., Schmitt, M. & Stein, K. (2006). Communication between process and structure: Modeling and simulating message-reference-networks with COM/TE. *Journal of Artificial Societies and Social Simulation* (accepted).

Mathes, A. (2005). Filler Friday: Google bombing. From *http://uber.nu/2001/04/06/, Accessed, 27/05/2006*

Page, L., Brin, S., Motwani, R., & Winograd, T. (1998). *The PageRank citation ranking: Bringing order to the web.* Technical Report. Stanford Digital Library Technologies Project.

Pinski, G., & Narin, F. (1976). Citation Influence for Journal Aggregates of Scientific Publications: Theory, With Application to the Literature of Physics. *Information Processing and Management, 12*(5), 297-312.

Sabidussi, G. (1966). The Centrality Index of a Graph. *Psychometrika, 31*, 581-603.

Scott, J. (2000). *Social network analysis: A handbook* (2nd Ed.). London; Thousands Oaks, Calif.: SAGE Publications.

Sicilia, M. A., & Garcia, E. (2005). Filtering Information with Imprecise Social Criteria: A FOAF-Based Backlink Model. *Proceedings of the Fourth Conference of the European Society for Fuzzy Logic and Technology (EUSLAT).*Barcelona, Spain

Stein, K., & Hess, C. (2005). Information retrieval in trust-enhanced document networks. *Proceedings of the European Web Mining Forum,* Porto, Portugal.

Wolf, J. L., Squillante, M. S., Yu, P. S., Sethuraman, J., & Ozsen, L. (2002). Optimal crawling strategies for web search engines. *Proceedings of the Eleventh International Conference on World Wide Web,* Honolulu, Hawaii, USA. 136-147.

Ziegler, C. N., & Lausen, G. (2004). Spreading activation models for trust propagation. *Proceedings of IEEE International Conference on e-Technology, e-Commerce and e-Service (EEE04).* Taipei, Taiwan, IEEE Computer Society Press.