

Filtering Information with Imprecise Social Criteria: A FOAF-based backlink model

Elena García-Barriocanal

Computer Science Dept.
University of Alcalá
elena.garciab@uah.es

Miguel–Angel Sicilia

Computer Science Dept.
University of Alcalá
msicilia@uah.es

Abstract

Several current approaches to information filtering in Web search engines implement models that use *backlinks* as a metric of subjective value that complements information retrieval techniques based on the content of the documents. Nonetheless, these models are somewhat “blind” to the reputation or trustworthiness of the creators of the links. The growing increase in interest in applications of social network analysis to the Web provides an opportunity to introduce such social metrics as a complement to existing backlink-based algorithms. In this paper, we describe an approach that uses a FOAF-like vocabulary to derive imprecise degrees of strength on explicitly declared social relations. A backlink model called *PeopleRank* — that uses the same scheme of the popular *PageRank* algorithm of the Google search engine — is used as a metric of social relevance. Then, this metric is used to weight the results of the *PageRank*, obtaining an straightforward “socially weighted” version of the algorithm. Illustrative examples of an implementation using the JUNG Java libraries are also provided.

Keywords: Semantic Web, *PageRank* algorithm, Social Networks, fuzzy sets.

1 Introduction

Many of the search engines use well-known information retrieval (IR) algorithms and techniques [3]. However, IR algorithms were developed for relatively small and coherent collections such as newspaper articles or book catalogs in a (phys-

ical) library. The Web, on the other hand, has been labeled as a “linked anarchy” [6] that poses different challenges that also involve page structure. Recent advances in Web structure analysis have resulted in several metrics for utility as *PageRank* or *hub–authority* models [2].

The celebrated *PageRank* algorithm [7] 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. To yield more accurate search results, Haveliwala [4] proposed computing a set of *PageRank* vectors, biased using a set of representative topics. Nonetheless, the assumptions of the original *PageRank* are biased towards measuring external characteristics. In fact, Page et al. [7] 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*”. This suggests that a critical point in improving link-based metrics would be that of improving the material information in which they are based, i.e. subjective value indicators. A relevant step in that direction is the growing consensus on open social metadata materialized in the *Friend-of-a-friend* (FOAF) project¹. The FOAF vocabulary² provides a rich language to describe relationships that could be interpreted as elements of subjective value between pairs of individuals or organizations. One of the FOAF utilities is to provide a

¹<http://www.foaf-project.org/>

²<http://xmlns.com/foaf/0.1/>

metadata set that enables the annotation in the future Semantic Web of Web resources like Web pages, documents, links and so on, providing a powerful way to identify detailed social networks. The resulting social metrics extracted from Semantic Web resources could then be used for the purpose of weighting or adjusting link-based metrics like the ones used by *PageRank*, integrating a form of implicit assessment of trustworthiness.

In this paper, an initial attempt to combine document link metrics with social relation metrics is described. Concretely, social relations described in the FOAF vocabulary are interpreted to have a given imprecise *strength* value [5], and metrics derived from those connections are combined with metrics of link structure. Since the approach is based on modified *PageRank* computations, the complexity of the algorithm remains $O(|E| \cdot I)$ where $|E|$ is the number of edges (be them social relationships or document links, as described below) and I is the number of iterations until convergence, and possible optimized computation mechanisms still remain valid. An example implementation using the JUNG³ Java libraries has been used to provide illustrative examples of the effect of such combined metrics in the final relevance of page ranks. The approach presented here could be further extended, modified or even replaced with other schemes that provide a numerical interpretation of social ties.

The rest of this paper is structured as follows. The combined social and link model for page ranks with imprecise weights is described in Section 2. Then, some illustrative data about the resulting scoring scheme are described in Section 3. Finally, Section 4 provides conclusions and future research directions.

2 A Model of Relevance based on Socially-Aware Linking

In this section we describe a core model for combined social and document relevance. The idea behind the model is that of using imprecise assessments of social relationships as a weighting factor for algorithms like *PageRank*.

Only positive “votes” are considered (thus neglecting the sign or valence of the relations [5]), following the *PageRank* approach. For the same reason, relations are not considered reciprocal even though their semantics indicate so, since it is the explicit declaration of relations that is valued in the approach. The degrees of strength in relations are modeled through fuzzy numbers that allow the specification of a relative uncertainty in the relation by the spread of the function.

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 (groups and organizations are left for future work). 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 [1]:

We assume *person A* 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* going out of *A's FOAF profile*. The *PeopleRank* of a *person A* is given as follows:

$$PpR(A) = (1 - d) + d * (PpR(T_1)/C(T_1) + \dots + PpR(T_n)/C(T_n))$$

Even though the idea of ranking by peer's declarations seem intuitive and is coherent with the discipline of Social Network Analysis, the original intuitive justification provided for *PageRank* in [1] 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

³<http://jung.sourceforge.net>

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! homepage are also generally worth looking at”.	Pages authored or owned by people with a larger positive prestige should somewhat be considered more relevant.

Table 1: Fundamental intuitive justifications for PR and PpR

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 of a social tie in our approach is modeled by the general expression (1), which provides a value for each edge (p_1, p_2) in the directed graph formed by the explicitly declared social relationships.

$$r((p_1, p_2)) = s((p_1, p_2)) \cdot e((p_1, p_2)); p_2 \neq p_1 \quad (1)$$

The relevance r of an edge is determined from a degree of strength s (in a scale of fuzzy numbers $[\tilde{0}, \tilde{10}]$) weighted by a degree of evidence e about the relationship. These strengths could be provided by extending recurrent FOAF schema with an additional attribute.

Degrees of evidence support a notion of “external” evidence on the relation that completes the subjectively stated strength. Expression (2) provides our current approach for this evidence that is based both on the perceptions of “third parties” and on the declaration of common projects.

$$e((p_1, p_2)) = \max(\Phi_{i \in U} s^{p_i}(p_1, p_2), P(p_1, p_2)); \quad (2)$$

$$p_i \neq p_2 \neq p_1, i \in U$$

The strengths of a social tie provided by third parties p_i in a group of user U are aggregated

through simple fuzzy averaging (Φ), and the evidence provided by work in common projects in which two persons (p_1, p_2) collaborate $-P(p_1, p_2)-$ is obtained from FOAF declarations. Concretely, people co-working in a project (as declared by `foaf:currentProject`) 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 recreative. Other usage-oriented metrics could be used to complement this approach, if available. An example of such metrics is described in [8].

Of course that the mathematical definition provided may be modified and more parameters could be added, but it provides an initial straightforward model from which further empirical analysis could be carried out.

3 Example Implementation

In this section, some illustrative cases are described, based on a prototype implementation of the model described above.

The example implementation has used the `PageRank` class provided by the JUNG libraries. The algorithm implemented is sketched in the code below. The proposed algorithm combines social with standard backlink consideration and has been labeled `PageRanks`. Inputs of the algorithm are the document graph and the people graph obtained from the FOAF annotations in Web resources.

COMPUTEPAGERANKSOCIAL(\mathcal{S}, \mathcal{D})

- ▷ \mathcal{D} is the document graph
- ▷ \mathcal{S} is the people graph

```

1  $\mathcal{S} \leftarrow \text{COMPUTESOCIALRELEVANCE}(\mathcal{S})$ 
2 for each  $v \in \text{VERTEX}(\mathcal{D})$ 
   do
3      $v.\text{source.relevance} \leftarrow$ 
        $\text{NODES}(\mathcal{S})[v.\text{source}].\text{relevance}$ 
4  $\mathcal{D} \leftarrow \text{WEIGHTEDPAGERANK}(\mathcal{D})$ 

```

The pseudocode above specifies the two step process, with a first social-specific computation COMPUTESOCIALRELEVANCE using expression (1) and a defuzzification method to compute fuzzy numbers resulting of that, and then an standard use of the weighted page rank provided in JUNG libraries, using the previously computed weights as source-related relevance indicators. Concretely, line (3) describes the actual modification introduced, which is traduced in JUNG-using code in the use of a user defined weighting for the computation of the transition probabilities that are used in the original *PageRank* algorithm.

Figure 1 depicts a possible case of social network, in which the connections were computed through the *relevance* procedure described above. To highlight the effects of social relevance, connections lines representing social relevance above 0.4 have been increased in thickness.

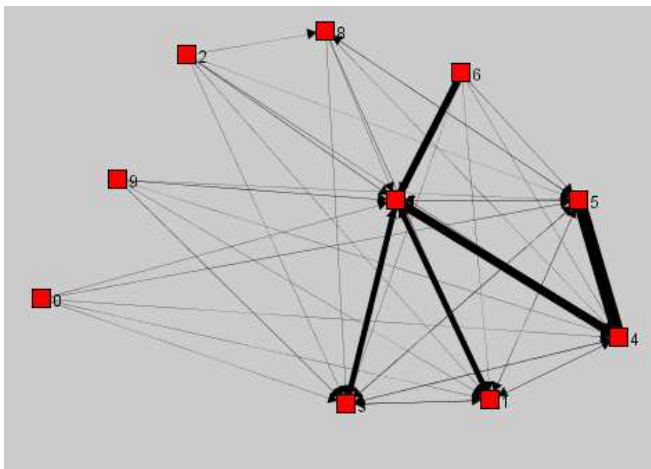


Figure 1: Social network view of *PeopleRank*

Rank	Node	value
1	10/16	0.084322/0.096884
2	12/12	0.083818/0.089171
3	2/14	0.083221/0.081886
4	9/5	0.082503/0.080030
5	6/9	0.073127/0.078538

Table 2: Example top ranks without and with consideration of social relevance

Figure 2 provides the document relevance obtained by the application of the weighting in the previous step. This can be used to provide an example of the effect of considering social relevance. Table 2 provides the values for the top ranks for the with/without social consideration executions of the algorithm.

The numbers in Table 2 provide a clear example of the effect of social relevance. Concretely, node 16 goes up in the ranking and differentiates it significantly. This is the effect of the weighting of incoming links from documents that belong to highly relevant persons, concretely document 14.

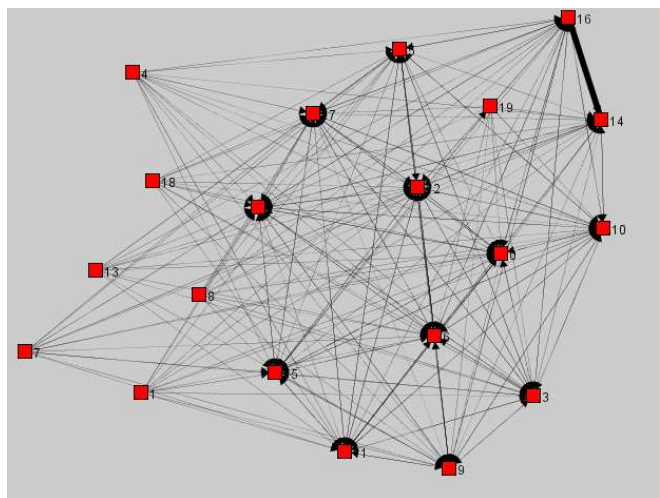


Figure 2: Resulting document back-link relevance network with social weighting

Random simulations with the kind of network presented yielded significant ranking differences, even though the ranking algorithm does not experiment significant variations in the lower ranking levels. For example, following the above example, the worst seven rankings of the 20 documents used in the example were the same in

both versions of the ranking computation (with or without social consideration).

4 Conclusions and Future Work

A model of Web page ranking combining link-based metrics with a basic account of social relations has been described. The model provides a tentative account for evidence and strength of relationships considering both the opinion of the individuals in the relationship and also the external view of others. In addition, it integrates a view on co-working as an indicator of evidence about the relationship. The approach taken for information filtering proceeds in two stages. First, a measure of social relevance is computed, and this measure is then used as a weighting factor for computing the relevance of documents. In both steps, the original *PageRank* algorithm is used.

Many other elements specified in the FOAF vocabulary could be subject to a specific and differentiated consideration in ranks that consider social relations. Examples are `<foaf:Group>`, `<foaf:Organization>` or `<foaf:Project>`, each of which entail a different consideration of relationship that would deserve a separate attention. In addition, the tentative scheme provided here requires much empirical study to come up with the better computational scheme with regards to reflecting the effects of “prestige” in Web documents. This could possibly require the introduction of other well-known social measures as “interconnectedness” that can be easily computed from graphs as those depicted above, but reflect social structural facts that are common to sub-graphs.

References

- [1] S. Brin, L. Page (1998) The Anatomy of a Large-Scale Hypertextual Web Search Engine. *WWW7 / Computer Networks*, 30(1-7): 107-117.
- [2] D. Dhyan, W. Keong Ng, S.S. Bhowmick (2002). A survey of Web metrics. *ACM Computing Surveys*, 34(4), 469 - 503.
- [3] C. Faloutsos (1985) Access methods for text. *ACM Computer Surveys* 17(1) (Mar.), 497-514.
- [4] T. Haveliwala. (2002) Topic-sensitive PageRank. In Proceedings of the eleventh international conference on World Wide Web, 517 - 526.
- [5] P. Mika, A. Gangemi (2004) Descriptions of Social Relation. *First International Workshop on FOAF, Social Networks and the Semantic Web*, Galway, Ireland.
- [6] A. Naevé (2001) The Concept Browser - a new form of Knowledge Management Tool. *Proceedings of the 2nd European Web-based Learning Environments Conference (WBLE 2001)*, Lund, Sweden, 24-26.
- [7] L. Page, S. Brin, R. Motwani, T. Winograd (1998) The PageRank citation ranking: Bringing order to the web. *Technical report, Stanford Digital Library Technologies Project*, paper SIDL-WP-1999-0120 (version of 11/11/1999).
- [8] M.A. Sicilia, E. García-Barriocanal (2004) Fuzzy Group Models for Adaptation in Cooperative Information Retrieval Contexts. *Lecture Notes in Computer Science* 2932 Springer, 324-334.