Benchmarking link analysis ranking methods in assistive environments

Konstantinos Georgatzis
Department of Information and Computer Science
Aalto University, Finland
konstantinos.georgatzis@aalto.fi

Panagiotis Papapetrou
Department of Information and Computer Science
Aalto University, Finland
panagiotis.papapetrou@aalto.fi

ABSTRACT

Several assistive applications exhibit a network structure. Characterizing the structure of such networks is critical in many assistive applications. Existing methods in the areas of social network analysis aim to detect, analyze, and summarize interesting or surprising components and trends in the network. In this paper, we provide a benchmark of two graph ranking methods: pagerank and HITS. The methods are tested on real social network data from three different domains: citation graphs, road networks, and a subgraph of Google. Our findings suggest that the quality of the ranking as well as the speed of convergence of both algorithms highly depends on the underlying network structure.

Categories and Subject Descriptors

H.4 [**Information Systems Applications**]: Miscellaneous

General Terms

Algorithms, experimentation

Keywords

Social networks, link analysis ranking, network analysis, american sign language.

1. INTRODUCTION

Several assistive applications exhibit a network structure. Characterizing the structure of such networks is critical in many assistive applications. Existing methods in the areas of social network analysis aim to detect, analyze, and summarize interesting or surprising components and trends in the network.

A large family of social network analysis methods focus on link analysis ranking. Given a graph that depicts the connection structure between the nodes in the network, the main task in link analysis ranking is to extract a ranking of the nodes in the underlying graph based on their linkage.

PETRA'12 June 6 - 8, 2012, Crete Island, Greece.

An example of such a graph is shown in Figure 1. In this example the graph is directed.

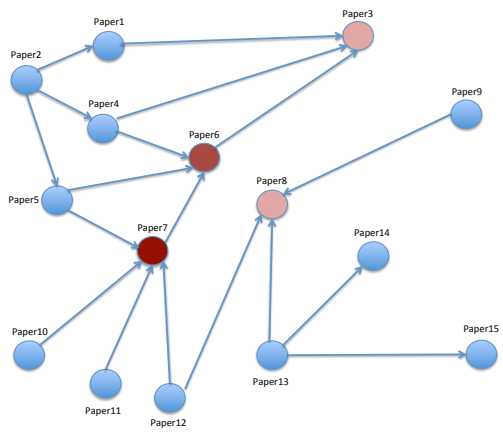

Figure 1: Example of a social network represented by a directed graph.

Several ranking algorithms have been developed to analyze sets of hyperlinked documents, such as web pages in the Internet. In this paper, we benchmark two of these methods PageRank[3] and HITS[10]. The methods are tested on real social network data from three different domains: citation networks, road networks, and a Google subgraph. Our experimental findings show the suitability of such methods in social network data related to assistive environments.

The main contributions of this paper include: (1) the presentation of two common ranking algorithms, pagerank and HITS, and (2) an experimental evaluation of the two methods on three datasets.

The remainder of this paper is organized as follows: in Section 2 we summarize the related work, in Section 3 we present the two benchmarked methods, in Section 4 we show our experimental findings, and in Section 5 we conclude the paper and discuss directions for future work.

2. RELATED WORK

Many methods have been developed in social-network analysis to assess the "authoritativeness" or "importance" of individuals in implicitly- or explicitly-defined social networks. The network, in its basic model, is represented by a directed graph $G = (V, E)$, where the nodes represent individuals,

and the edges model "endorsement" from one individual to another. For such a directed graph, the concept of *in-degree* (the number of incoming edges at a node), or refinements [9, 7] , is the simplest measure of importance of an individual. Other notions of importance in social networks include degree centrality, closeness centrality, betweenness centrality, eigenvector centrality Newman [11].

A prominent application domain of importance measures is in the area of web search and hypertext ranking [3, 10]. The goal is to assign importance scores to web documents in order to assist users locate the most relevant results for their searches. Not surprisingly many of the importance measures discussed above can also by used in the case of web search ranking, however, the two most well-known techniques are PageRank [3] and HITS [10]. Many variants of those methods have been proposed, as well as adaptations of those basic methods for different objectives. Several methods for ranking semantic Web resources have been proposed, including SemRank [1], which is based on information theory, and OntoRank[4] which is an adaptation of PageRank for Semantic Web resources. Finally, ReconRank [6] is a link analysis ranking method applied at query time for computing the popularity of resources and documents.

When the network is characterized by heterogeneous links, standard ranking methods may not provide accurate results due to the fact that different types of links may have different impact on the resulting ranking; hence, a weighted version of PageRank has been proposed [14] that can handle different types of relations between the nodes in the network. PopRank [12], employs a machine learning approach where each type of connection (edge) in the graph is assigned with a "popularity propagation factor". In ObjectRank [2], an authority-based ranking is applied to keyword search.

The hierarchical structure of the Web has also been studied quite extensively in the literature. A hierarchical model of the Web is described in Kamvar et al. [8] along with the dersirable computational properties, while it is shown [15] that hierarchical ranking algorithms outperform qualitatively other well-known algorithms, such as PageRank.

3. METHODS

A social network can be represented as a graph. Each node in the network corresponds to a node in the graph and a conenction between two nodes in the network is represented by an edge in the graph. The graph can be either directed or undirected.

More formally, let $G = (V, E)$ be a graph, where $V = \{v_1, \ldots, v_{|V|}\}$ is the set of vertices and $E = \{e_1, \ldots, e_{|E|}\}$ the set of edges in G.

3.1 Hubs and Authorities

The hubs-and-authorities [10] algorithm, also known as HITS, is a link analysis ranking algorithm precursor to PageRank. The intuition behind this algorithm is based on the way web pages are formed. That is, some web pages act as information hubs, in the sense of links to highly authoritative web pages. Hence, they are not authoritative themselves but they point to many authoritative pages. At the same time, a web page is considered authoritative if it is pointed to by many "strong" hubs.

Each node is ranked with respect to its "importance" as a *hub* or *authority*. A strong hub is a node that is connected to many strong authorities; a strong authority is a node that is connected to many strong hubs.

Given G, each node v_i is associated with a hub weight h_i and an authority weight a_i. The magnitude of the weight corresponds to the strength or importance of each node. During the application of the iterative algorithm, the sum of the squared hub and authority weights is enforced to be invariant and equal to unity. This is achieved by successive normalizations of the weights.

At the intialization step, each node is assigned with hub and authority weights of $\frac{1}{\sqrt{(|V|)}}$. Then at each iteration, authority and hub weights are updated as a function of the hub and authority weights, respectively, or

$$a_j \leftarrow \sum_{i=1}^{|V|} h_i \text{ and } h_i \leftarrow \sum_{j=1}^{|V|} a_j$$

Authority weights are thus a weighted sum of the ingoing hub weights, and correspondingly, hub weights are maintained as a weighted sum of the authority weights of the pages linked to by a hub. To guarantee invariance of the weights at each iteration it should hold that

$$\sum_{i}^{|V|} a_i^2 = 1 \text{ and } \sum_{i}^{|V|} h_i^2 = 1$$

Hence, after each iteration the authority values are divided by the square root of the sum of the squares of all authority values, and the hub values are divided by the square root of the sum of the squares of all hub values. An illustration of the above iterative process is given in Figure 2.

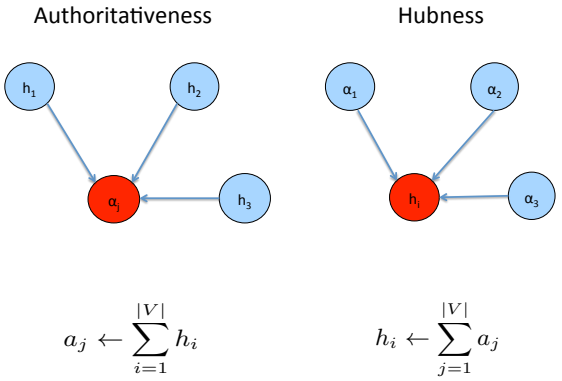

Figure 2: Illustration of HITS.

In linear algebra terminology, if A denotes the adjacency matrix of graph G, HITS computes the dominant eigenvector of $A^T A$ using the Power Method [5]. This suggests that the speed of convergence highly depends on the connectivity of G and it is highly tied to the separation of the first and second eigenvalues of that matrix. Nonetheless, no bounds on this separation are known, since for arbitrary matrices of any fixed size the separation can be arbitrarily small and the convergence arbitrarily slow.

3.2 PageRank

PageRank [3, 13] has been developed for link analysis ranking. The key intuition behind this method is the assumption of a user browsing web pages randomly. The user starts from a random web page and then follows an outgoing link with probability d and resets to some other random web page with probability $1-d$. Overall, Pagerank expresses a probability distribution that represents the likelihood that a person who is randomly following edges in a graph will arrive at any particular node. For initialization, it is assumed that the distribution is evenly divided among all nodes in the graph. Pagerank is an iterative algorithm and its computation requires several passes through the whole graph to ensure that the approximate Pagerank values more closely reflect the theoretical true value.

Let PR be the vector of PageRank values. Using the recursive definition of PageRank [3, 13], PR is computed as follows:

$$PR(v_i) = \frac{1-d}{|V|} + d \sum_{\forall v_j \to v_i} \frac{PR(v_j)}{|F(v_j)|}, \quad (1)$$

where $F(v_j)$ corresponds to the out-degree of node v_j.

Note that in order to avoid getting trapped to loops of sink states (i.e., nodes that have no outgoing edges) PageRank assumes that nodes with no outbound edges should link to all other nodes in V. For this purpose the regulating residual probability $1-d$ is used, also called *damping factor*. Assigning an appropriate value to $1-d$ depends on the graph we are studying. In the case of WWW, a random user surfing the web will typically follow the order of 6 hyperlinks after becoming bored and choosing some other random web-page to surf. Thus in this case, $1-d = 1/6 \approx 0.15$. An illustration of the above iterative process is given in Figure 3.

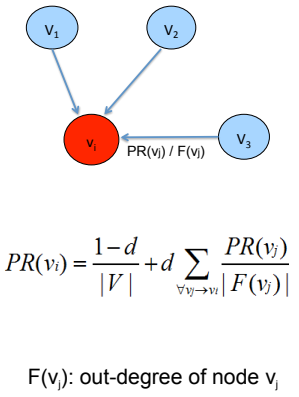

Figure 3: Illustration of Pagerank.

4. EXPERIMENTS

4.1 Setup

The two methods have been benchmarked on three real datasets. All datasets can be found at the Stanford Large Dataset Collection repository [1].

- CitationNet: this is the *Arxiv HEP-TH* high energy physics theory citation graph. It is taken from the e-print arXiv and covers all the citations within a dataset of 27,770 nodes with 352,807 edges. Each node is a paper and a directed edge indicates a citation from the source node to the destination node. The papers covered by this dataset have been published in the period from January 1993 to April 2003.

- RoadNet: This is part of a road network of Pennsylvania. Intersections and endpoints are represented by nodes and roads connecting these intersections or endpoints are represented by undirected edges. The dataset contains 1,088,092 nodes and 3,083,796 edges.

- GoogleNet: This data was released in 2002 by Google as a part of Google Programming Contest. The dataset consists of 875,713 nodes and 5,105,039 edges. Nodes represent web pages and directed edges represent hyperlinks between them.

In Table 1 we can see a summary of the main statistics of each dataset.

Dataset	number of nodes	number of edges
CitationNet	27,770	352,807
RoadNet	1,088,092	3,083,796
GoogleNet	875,713	5,105,039

Table 1: Dataset Statistics.

We have implemented HITS and Pagerank in Python. The source code can be found online[2].

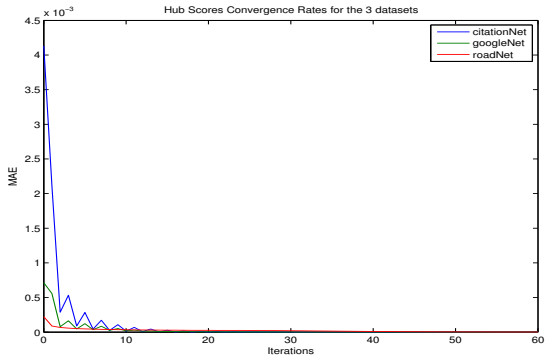

Figure 4: Convergence of hubness scores for the three datasets.

4.2 Evaluation

We compared the performance of HITS and Pagerank on the three datasets.

First, we studied their convergence. In Figures 4, 5, and 6 we see the convergence of the hubness, authoritativeness, and pagerank scores, respectively, for the three datasets. As it can be seen from the results, the citation network is the hardest to rank, while the road network is the easiest.

[1] `http://snap.stanford.edu/data/.`

[2] `http://users.ics.tkk.fi/panagpap/ranking/.`

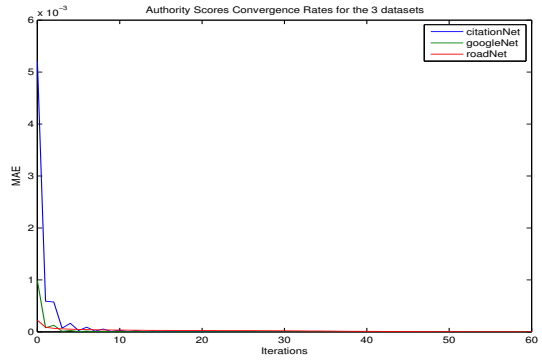

Figure 5: Convergence of authoritativeness for the three datasets.

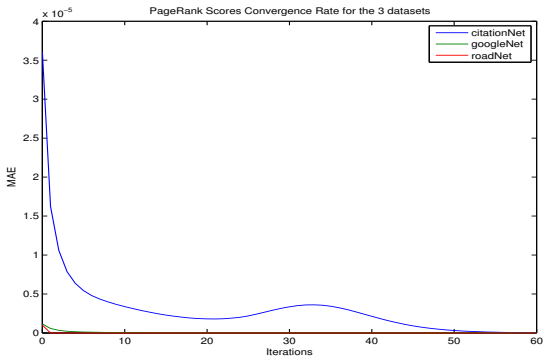

Figure 6: Convergence of pagerank for the three datasets.

Specifically, for the road dataset all three scores converge after the third iteration, while for the citation dataset Pagerank needs approximately 50 iterations. In overall, Pagerank shows a faster convergence rate than HITS with the exception of the citation dataset. This is due to the high sparseness of this dataset as well as the heavy tail indegree distribution that is more extreme in this dataset than in the other two.

Furthermore, we observed a very high positive correlation of the hubness and authoritativeness scores of the nodes in the road network. This can be explained by the high connectivity in this specific network, which results in a high "agreement" between hubness and authoritativeness. This behavior was not observed in the other two networks.

5. SUMMARY

We performed a qualitative and quantitative analysis of two link analysis ranking methods that are highly used in social networks. Our findings suggest that the quality of the ranking as well as the speed of convergence of both algorithms highly depends on the underlying network structure. The main goal of this paper is to introduce these ranking methods to the domain of assistive environments which can highly benefit by employing or adapting the high variety of existing link anaylsis ranking methods.

6. ACKNOWLEDGMENTS

This work has been supported in part by the Academy of Finland Centre of Excellence grant 118653 (ALGODAN)

7. REFERENCES

[1] K. Anyanwu, A. Maduko, and A. Sheth. Semrank: ranking complex relationship search results on the semantic web. In *Proceedings of the 14th international conference on World Wide Web*, WWW '05, pages 117–127, 2005.

[2] A. Balmin, V. Hristidis, and Y. Papakonstantinou. Objectrank: Authority-based keyword search in databases. In *Proceedings of the International Conference on Very Large DataBases*, pages 564–575, 2004.

[3] S. Brin and L. Page. The anatomy of large-scale hypertextual web search engine. *Computer Networks and ISDN Systems*, 30:107–117, 1998.

[4] L. Ding, R. Pan, T. Finin, A. Joshi, Y. Peng, and P. Kolari. Finding and ranking knowledge on the semantic web. In *In Proceedings of the 4th International Semantic Web Conference*, pages 156–170, 2005.

[5] G. H. Golub and C. F. V. Loan. *Matrix Computations, 3rd Edition*, 30:107–117, 1996.

[6] A. Hogan, A. Harth, and S. Decker. Reconrank: A scalable ranking method for semantic web data with context. In *In 2nd Workshop on Scalable Semantic Web Knowledge Base Systems*, 2006.

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

[8] S. Kamvar, T. Haveliwala, C. Manning, and G. Golub. Exploiting the block structure of the web for computing pagerank. Technical Report 2003-17, Stanford InfoLab, 2003.

[9] L. Katz. A new status index derived from sociometric analysis. *Psychometrika*, 18:39–43, 1953.

[10] J. M. Kleinberg. Authoritative sources in a hyperlinked environment. *Journal of the ACM (JACM)*, 46(5):604–632, September 1999.

[11] M. E. J. Newman. The mathematics of networks. *The New Palgrave Encyclopedia of Economics*, 2007.

[12] Z. Nie, Y. Zhang, J.-R. Wen, and W.-Y. Ma. Object-level ranking: bringing order to web objects. In *Proceedings of the 14th international conference on World Wide Web*, WWW '05, pages 567–574, 2005.

[13] L. Page, S. Brin, R. Motwani, and T. Winograd. The pagerank citation ranking: Bringing order to the web. Technical Report 1999-66, Stanford InfoLab, November 1999.

[14] W. Xing and A. Ghorbani. Weighted pagerank algorithm. In *Communication Networks and Services Research, 2004. Proceedings. Second Annual Conference on*, pages 305 – 314, 2004.

[15] G.-R. Xue, Q. Yang, H.-J. Zeng, Y. Yu, and Z. Chen. Exploiting the hierarchical structure for link analysis. In *Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '05, pages 186–193, 2005.