1. Summary

The focus of this paper is the distillation of *broad* search topics down to a small size: from *global* (not restricted to a focused set of pages or a single website) WWW to "authoritative" representations. Meanwhile, for scalability, it is desired not to directly maintain an index of WWW or its link structure. In addition, broad-topic queries face complexities such as overwhelming relevant resources and search terms not explicitly included in authorities.

In Section 2 and 3, the author discusses algorithms of constructing a focused subgraph of WWW based on the *link* structure and iteratively computing *authorities* and *hubs* with convergence.

The Subgraph algorithm in Section 2 basically starts from a root set *R* (subscript omitted) containing top pages returned by available search engines such as AltaVista, then grows *R* to base set *S* by making use of the *link* structure to include any page pointed to by a page in *R* and some (not all) pages point to a page in *R*. *S* satisfies properties including 1) small in size, 2) rich in relevant pages, and 3) containing most (or many) strongest authorities. In addition, only transverse links with more information (compared to intrinsic links) are kept and *S* becomes *G*.

The Iterative algorithm in Section 3 takes *G* in Section 2 and extracts authorities through an analysis of its link structure. In addition to the simple heuristic that authorities are of large in-degrees, an important observation is the *mutually reinforcing* relationship between *authorities* (pages that have links from multiple relevant hubs) and *hubs* (pages that have links to multiple relevant authorities), based on which "I" and "O" operations (with normalization operation) are applied alternately on authority and hub weight vectors of pages respectively to reach the fixed points of weight vectors ("equilibrium"). The paper also discusses the convergence and the fixed points of authority and hub as principle eigenvectors of A^TA and AA^T respectively (with an assumption about principle eigenvectors, *A* is the adjacency matrix of *G*). As a result, the largest coordinates in weight vectors are reported as authorities and/or hubs.

Section 4 shows that, with change on content of the root set *R*, similar-page queries could be dealt with the same methods without essential modification.

Section 5 discusses connections with related work (link structure), which covers 1) the use of a link structure for defining standing, impact and influence, 2) other ways in which links have been integrated into hypertext and WWW search techniques, and 3) how a link structure is utilized for explicit clustering of data.

The topic of Section 6 is *multiple* collections authorities and hubs. Well-separated collections exist naturally due to multiple meanings of the query string in different communities or it being a polarized issue. An extension of the algorithms shown in Section 2 and 3, which makes use of non-principle eigenvectors of matrixes A^TA and AA^T, is able to produce multiple collections of authorities and hubs.

In Section 7, diffusion and generation of "*too-specific*" queries on the methods is investigated. To mitigate this issue, combination with textual content might be able to generate more relevant results.

In Section 8, a system name CLEVER based on the proposed methods are evaluated with Yahoo! by a group of human users. Within a threshold of statistical significance, CLEVER's performance is competitive with Yahoo's on about 81% of 1369 user responses.

2. Comments

The methods/algorithms look elegant and powerful. Linear algebra concepts such as vector and matrix correspond with elements in the focused subgraph such as pages and links. Moreover, the mutual enhancing relationship between authorities and hubs implies that authority and hub vectors will converge to the principle eigenvector of matrixes A^TA and AA^T (A is the adjacency matrix of the focused subgraph) with the assumption on the value of principle eigenvalues. More interestingly, non-principle eigenvectors also play important roles in obtaining information about the clustering of authorities.

Moreover, the paper is strict with concept and connections. In Section 5, some similar notions (also using a link structure) to authority in the area of scientific literature are mentioned: standing, impact and influence. Relevant definitions, meanings and mutual connections are discussed. It also compares scientific journals with WWW, arguing that they are governed by different principles and should be considered with different models. The former is homogeneous and thus one-level model where authorities directly endorse each other work well. In contrast, the latter is much more heterogeneous, which make it is necessary to introduce a two-level model with hub pages as an intermediate layer between different authorities.

3. Questions

My problem is about the proof of Theorem 3.1. To be specific, the idea of power iteration (power method) for calculating the eigenvalues/eigenvectors of a diagonalizable matrix looks abstract. I guess I need to review on some basic knowledge of linear algebra. Also, I am curious about the comparison between PageRank and the proposed methods here. Given differences on the passage of authority (without or with hubs), usage of random jumps or not, direct or indirect access to WWW, what can we expect on their complexity, scalability and maintainability?