Toward a Unification of Text and Link Analysis

Brian D. Davison
Computer Science & Engineering
Lehigh University
Bethlehem, PA 18015 USA
davison@lehigh.edu

ABSTRACT

This paper presents a simple yet profound idea. By thinking about the relationships between and within terms and documents, we can generate a richer representation that encompasses aspects of Web link analysis as well as text analysis techniques from information retrieval. This paper shows one path to this unified representation, and demonstrates the use of eigenvector calculations from Web link analysis by stepping through a simple example.

Categories and Subject Descriptors

H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval—*clustering, retrieval models*; H.3.1 [**Information Storage and Retrieval**]: Content Analysis and Indexing

General Terms

Algorithms

Keywords

knowledge representation, eigenvectors, link analysis, search engines, RageRank, hubs, authorities, HITS

1. INTRODUCTION

The ubiquity of the World-Wide Web has placed information retrieval systems at the fingertips of millions of people, in the form of Web search engines. While those search engines initially used textual analysis to match documents with queries, the use of link analysis techniques have become common.

By link analysis, we refer to the study or use of algorithms operating over the Web's link graph. This graph defines the relationships between pages, based on the hyperlinks from page to page. Such algorithms might help to find relevant documents for a query, or find similar or duplicate documents. Link analysis has similarities to, and has benefited from, social network analysis and bibliographic citation analysis. Thus, Web link analysis is one form of the more generic problem of what we term *relationship analysis* which includes any algorithm operating over a network graph or matrix built from relationships between entities in the graph.

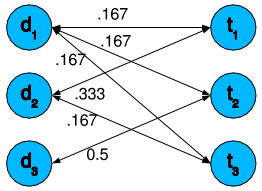
However, in the Web, most link analysis has only limited textual components. In systems based on Kleinberg's HITS [3], information retrieval based on text is used to select the initial core set of documents. PageRank [4], in contrast, doesn't use text at all to determine a document's authority score. Likewise, in traditional information retrieval research, the use of eigenvectors — which underlies most Web link analysis — is commonly limited to the dimension-reduction approach found in LSI [1].

SIGIR'03, July 28–August 1, 2003, Toronto, Canada.
ACM 1-58113-646-3/03/0007.

	d_1	d_2	d_3
t_1	1	2	0
t_2	1	0	3
t_3	1	1	0

$$M := \begin{bmatrix} .167 & .333 & 0.0 \\ .167 & 0.0 & 0.5 \\ .167 & .167 & 0.0 \end{bmatrix}$$

(a) Sample terms and documents, and their matrix view.



(b) Graph view.

Figure 1: A simple term-document matrix M and equivalent graph with three documents and three terms, using length-normalized TF-IDF weighting.

2. REPRESENTATIONS

Typical information retrieval approaches represent documents as vectors of term weights. When placed together, these vectors form a term-document matrix, in which one axis enumerates each document, and the other enumerates each term found in the collection. Figure 1(a) displays this matrix, using a simple length-normalized form of TF-IDF weighting. This matrix corresponds to a bipartite graph in which terms and documents are nodes, with undirected links between them (Figure 1(b)) wherever there are non-zero entries in the matrix. In our thinking, these links are really relationships. That is, a document is linked to a term when that term is found in the document and vice versa. However, this model is limited to undirected links. To use directed links, we need to expand the matrix. We propose the use of four submatrices (as shown in Figure 2), including the same term-document submatrix (M), but now also includes a document-term submatrix (M^T), plus new term-term and document-document submatrices. This particular matrix still represents the same graph, but now there are directed links that happen to have the same weights in both directions. However, this matrix provides a richer representation. If desired, we can

	Terms	**Docs**
Terms	term-term	term-doc
Docs	doc-term	doc-doc

Figure 2: A generic augmented matrix with sub-matrices.

	EV_1	EV_2	EV_3	EV_4	EV_5	EV_6
E.value	0.541	-0.541	0.420	-0.420	0.061	-0.061
t_1	0.200	0.200	0.570	0.570	0.367	-0.367
t_2	0.662	0.662	-0.248	-0.248	0.024	-0.024
t_3	0.148	0.148	0.337	0.337	-0.604	0.604
d_1	0.311	-0.311	0.262	-0.262	-0.578	-0.578
d_2	0.169	-0.169	0.587	-0.587	0.357	0.357
d_3	0.612	-0.612	-0.295	0.295	0.196	0.196

Figure 3: Eigenvalues and eigenvectors of the running example.

now have different weights for links in different directions, as we will demonstrate shortly. We can also have non-zero weights between nodes of the same type — that is, we can have links between terms, or between documents. The links between documents easily correspond to citations (whether hypertext or bibliographic), and the links between terms might be similarity values.

The doc-doc submatrix is exactly the matrix used in Web link analysis algorithms (typically a weighted adjacency matrix). By varying the mechanisms to determine weights, one can specify different algorithms. For example, the simplified form of the PageRank algorithm [4] doesn't use the adjacency matrix directly — it uses the inverse of the number of outgoing links. In text analysis we often do something similar — one way to normalize the term frequencies is to divide by the document length. In our augmented matrix, we can use different weights for different link directions. Therefore, we can still use a normalized term frequency, just as in PageRank, for the links from docs to terms. We can even use the same principle (the inverse of the number of outgoing links) to weight the links from terms to documents. This, in a sense, corresponds to the inverse document frequency of TF-IDF. However, other variations are possible.

3. ALGORITHMS

Most Web link analysis algorithms revolve around the use of eigenvector calculations, and differ in their matrix weights and generation. For example, the random surfer model in PageRank, while described as a probabilistic jump from any document to any other, can be implemented as a simple operation over the existing doc-doc matrix (adding low-weight links between all pages). This factor in PageRank is described as helping to prevent rank sinks.

Web link analysis generates eigenvectors primarily for two reasons. The first is to generate a total ordering of documents, typically from the principal eigenvector. This total ordering may then be combined with other factors, such as textual relevance, to generate a final ordering for presentation to the user. The second reason for generating eigenvectors is to investigate communities within some topic, as suggested by Kleinberg [3].

We propose using these approaches on text. We can directly calculate eigenvectors of our augmented matrix, and consider some fraction of documents with high values in the principal eigenvector and at both ends of non-principal eigenvectors as clusters. A nicety of our model is that the clusters generated will be self-describing because they will include terms as well as documents.

Figure 3 shows the results of calculating the eigenvectors from our augmented sample matrix. In the principal eigenvector, we see that term t_2 scores highly, followed closely by document d_3. This matches the original distribution of terms to documents (since d_3 contained only instances of t_2). Likewise, in EV_3, we see that document d_2 is ranked highest, followed closely by term t_1.

Calculating the principal eigenvector of the standard augmented matrix is not suggested, as while it will generate an order for the terms and documents in the matrix, that order has no underlying meaning for our purposes (unlike the calculation in the Web matrix, which corresponds to authoritativeness, assuming links convey authority upon the target of the link). Additionally, it doesn't take the query into consideration, and so separate text analysis would be needed.

$$M_2 := \begin{bmatrix} 0.200 & 0.0 & 0.0 & 0.133 & 0.267 & 0.0 \\ 0.200 & 0.0 & 0.0 & 0.133 & 0.0 & 0.400 \\ 0.200 & 0.0 & 0.0 & 0.133 & 0.133 & 0.0 \\ 0.333 & 0.133 & 0.133 & 0.0 & 0.0 & 0.0 \\ 0.467 & 0.0 & 0.133 & 0.0 & 0.0 & 0.0 \\ 0.200 & 0.400 & 0.0 & 0.0 & 0.0 & 0.0 \end{bmatrix}$$

Figure 4: The skewed matrix M_2 emphasizing t_1.

Our approach is based on the idea that PageRank can be personalized [4, 2]. Instead of using a uniform probability of jumping randomly from one page to any other, an emphasis can be made to a certain page or set of pages. This has the effect of ranking other pages in relation to the emphasized ones. We can do the same with our augmented matrix. A given query, either in the form of a page in the corpus or a set of terms can be emphasized by modifying the augmented matrix so that all objects have a link to the emphasized object(s). With enough emphasis, the query objects will rise to the top of the principal eigenvector, and remaining objects will be ordered in terms of their relevance to the initial query objects.

For example, we saw above that the principal eigenvector placed terms t_1 and t_3 well below the value assigned to t_2. The documents were ordered d_3, d_1, d_2. We can skew the eigenvector by emphasizing t_1, as if it were the given query. To do so, we arbitrarily modify the network to incorporate links from all objects to t_1 with weight .2, and decrease all existing weights by 20%. This modified matrix M_2 is shown in Figure 4. The principal eigenvector of M_2 contains the values [0.854, 0.100, 0.148, 0.246, 0.415, 0.067], resulting in a rank ordering of $t_1, d_2, d_1, t_3, t_2, d_3$, which is in fact the desired ordering.

4. DISCUSSION

The ideas presented here are only exploratory. Additional effort will be needed to experimentally verify (on a large scale) the claims made here. Moreover, we make no claims of optimality in the details of our approach.

We envision many extensions to this work, the most obvious being the incorporation of term-term and doc-doc links, as mentioned in Section 2. However, some care may be needed to balance the relative influence of various kinds of relationships, and prevent domination by a single submatrix.

5. REFERENCES

[1] S. C. Deerwester, S. T. Dumais, T. K. Landauer, G. W. Furnas, and R. A. Harshman. Indexing by latent semantic analysis. *Journal of the American Society of Information Science*, 41(6):391–407, 1990.

[2] T. H. Haveliwala. Topic-sensitive PageRank. In *Proceedings of the Eleventh International World Wide Web Conference*, Honolulu, Hawaii, May 2002.

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

[4] L. Page, S. Brin, R. Motwani, and T. Winograd. The PageRank citation ranking: Bringing order to the Web. Unpublished draft, 1998.