Beyond Relevance Ranking: Hyperlink Vector Voting

Yanhong Li, Larry Rafsky
GARI Software/IDD Information Services
293 Eisenhower Pkwy, Suite 250
Livingston, NJ 07039 USA
{yli,lcr}@iddis.com

Abstract

A new method for hypertext indexing and retrieval called Hyperlink Vector Voting (HVV) is proposed. It combines relevance ranking and quality ranking for hypertext retrieval systems. Ranking of search results no longer depends on the content of the document being ranked but rather on the number of hyperlinks pointing to the document and the descriptions of these links. An experimental World Wide Web search engine is also described, and the results it produces appear quite satisfactory.

1 Introduction

As the popularity of the Internet and World Wide Web grows, people begin to experience the pressure of information explosion. Hunting for information on the Web become more important than ever before. Many current Internet search engines, such as Excite or Infoseek serve as agents to help people find the information they want. Casual users and simple queries comprise the majority of Internet search activities [3]. A simple query could result in tens of thousands of hits; ranking of the search results therefore becomes so important that one Excite ad says "You realize that when you find too much, all you actually found is that you have to keep searching". However, current Internet search engines are not good at ranking results. For example, when you search for "netscape", "http://home.netscape.com" is *not* ranked 1st by most search engines.

Traditional relevance ranking models such as the Vector Space Model, Probabilistic Models, Fuzzy Logic Models, etc.[1] make the assumption that all the documents being ranked are almost equally good in terms of quality, and authors of these documents are not trying to manipulate search engines. Therefore, almost all the ranking techniques in use today depend on the frequency of query terms in a given document. But as is now well known, many web site authors pack repeated word occurances in a single page in order to be ranked higher by search engines. On the Web, where everyone can be a publisher, there should be other measurements beyond word counts, quality and popularity matter.

2 Hyperlink Vector Voting Method

We introduce a new representation for hypertext documents and algorithms for indexing and retrieving hypertext systems. Our technique integrates relevance ranking and quality ranking; it also solves vocabulary problems such as synonyms and foreign language terms under certain conditions.

Without loss of generality, we use the Vector Space Model to describe the Hyperlink Vector Voting (HVV) method. In the Vector Space Model, a document is represented by a vector, and each dimension of the vector is a term or concept extracted from the content of the document. In HVV, a document is represented by zero or more link vectors. Each link vector represent a hyperlink pointing to the document. The description of the hyperlink – the anchor-text - is treated like the content of the hyperlink, and the vector representation of the link is formed by the term weighting of each term in the anchor-text. In a hypertext system, there can be zero or more hyperlinks whose destination anchor (head anchor) is a given document, so a given document is represented by zero or more link vectors as shown in equation 1.

$$D_j = \begin{bmatrix} \vec{L}_1 \\ \vec{L}_2 \\ \cdots \\ \vec{L}_i \end{bmatrix} \qquad (1)$$

For example, if there are 189 links pointing to "http://www.javasoft.com/tutorial", then this Java tutorial document from Sun is represented by 189 link vectors.

Each dimension of a link vector is represented by the weight of the term extracted from the link description (anchor text), typically each distinct word in the anchor text can be considered as a separate dimension. Term weighting is given by the well known equation 2.

$$W_{x,t} = f_{x,t} \cdot log(N/f_t) \qquad (2)$$

where $f_{x,t}$ is the number of occurrences of word t in x; N is the number of documents in the collection; and f_t is the number of documents whose representation contains term t. Note that f_t is usually referred to as Document Frequency, it is different in HVV than in standard IR systems. Explicitly,

Definition The HVV Document Frequency for a given term t is the number of documents that are referred as the head anchor in any hyperlink whose link description (anchor-text) contains t.

During a typical retrieval process, query words are first matched against an inverted file to locate which documents should be retrieved. The ranking is then based on the document representations. The ranking process here is called Hyperlink Vector Voting. Similar to link descriptions, queries are also represented as query vectors. The ranking score is defined as the summation of all the dot products between the query vector and each hyperlink vector for a given document. The summation process is like a voting - more links usually result in higher scores. But it is a weighted voting: the weights depend on how similar the link vectors are to the query vector. Finally, our ranking formula is expressed in equation 3.

$$R = \sum_{i=1}^{n} (\vec{Q} \cdot \vec{L}_i) \qquad (3)$$

where R is the ranking score, $\vec{Q}$ is the query vector, and $\vec{L}_i$ is the link vector.

3 Discussion and Experimental Results

Because the hyperlink vector representation and link based inverted file only have link information, the ranking of retrieval results does not depend on the words appeared in the documents themselves, but only on the descriptions of those hyperlinks pointing to them, or on how other people describe the documents and how many people cited the document. Thus "search hostile" documents with "keyword spamming" will not get unfairly high scores, moreover, size of a document is no longer a factor in relevance ranking and thus problems associated with document size can be avoided. Thesaurus or knowledge bases may not be crucial because even if the word "lawyer" never appears in a document titled "California Immigration Attorneys", someone may have a hyperlink pointing to this document and the anchor-text might read "California Immigration Lawyers". The key idea is : When a hypertext document database is large enough, such as the World Wide Web, search results should become a kind of "voting" result: the suitability of a document is determined by how other documents describe it, not only how the document "describes" itself.

There are other advantages: Images, graphics, sound file, etc. are not searchable by conventional information retrieval methods, but they are searchable by the description of the hyperlinks pointing to them. The same applies to documents in foreign languages. There may be hyperlinks pointing to them and the description of those hyperlinks are in the user's native language. (It is true that anchor-text could also be in the form of images, graphics etc. but the index engine can substitute it with head anchor's document title if applicable.)

An experimental Web search engine called "Rankdex" based on the Hyperlink Vector Voting (HVV) is available at "http://rankdex.gari.com". Our spider has collected about 5.3 million hypertext documents on the Web, and they are indexed by the HVV index structure. The system ranks "http://home.netscape.com" first when a user searches for "netscape" and rank "weber.u.washington.edu/~louie/sandra.html" (this site has won a lot of awards) first when searching for "Sandra Bullock". A detailed report and analysis of the search results and the comparison with other search engines can be found in [2], a more comprehensive paper by the first author (the inventor of HVV on which a patent is pending).

4 Conclusions

This paper presents a new representation for hypertext documents and indexing and retrieval algorithms based on the link vector representation. It is more efficient than traditional search engines, because it only indexes hyperlinks and hyperlink descriptions. It is much more effective for the simple queries used by most Web users. In addition, it partialy avoided many vocabulary and language problems and makes some non-textual information searchable by their textual descriptions.

The authors thank Doran Howitt for his effort in deploying a realworld search service based on HVV, Ron Katriel for his constructive suggestions. Many stuff members of GARI Software also provided important feedback after using Rankdex. This research is funded by IDD Information Services.

References

[1] Donna Harman. Ranking algorithms. In W.B Frakes and R. Baeza-Yates, editors, *Information Retrieval Data Structures and Algorithms*, pages 363-392. Prentice Hall, 1992.

[2] Yanhong Li. Towards a qualitative search engine: Hyperlink vector voting. *IEEE Internet Computing*, (submitted).

[3] Erik Selberg and Oren Etzioni. Multi-service search and comparison using the metacrawler. In *Proceedings of the 4th International World Wide Web Conference*, 1995.