A Unified Model of Literal Mining and Link Analysis for Ranking Web Resources

Yinghui Xu Kyoji Umemura
Toyohashi Unversity of Technology Dept. of Information and Computer Sciences
1-1, Hibarigaoka, Toyohashi, Aichi
+81-532-44-6762
xyh@ss.ics.tut.ac.jp umemura@tutics.tut.ac.jp

ABSTRACT

Web link analysis has been proved to provide significant enhancement to the precision of Web search in practice. The PageRank algorithm, which is used in Google Search Engine, plays an important role on improving the quality of its resuls by employing the explicit hyperlink structure among the Web pages. The prestige of Web pages defined by PageRank is purely derived from surfer random walk on the Web graph without textual content content consideration. However, in the practical sense, user surfing behavior is far from random jumping. In this paper, we present a unified model for a more accurate page rank. User's surfing is guided by a probabilistic model that is based on literal matching between connected pages. The result shows that our proposed ranking algorithms do perform better than the original PageRank.

Categories and Subject Descriptors

H.3.3 [**Information Storage and Retrieval**]: Information search and retrieval- Retrieval models. Search process

General Terms

Algorithm

Keywords

PageRank, Language Modeling, Virtual document (VD)

1. INTRODUCTION

The PageRank algorithm [1] uses the hyperlink structure of the Web to build a stochastic irreducible Markov chain with transition matrix P. The transition matrix was built on the assumption that a "random surfer" who is given a web page at random and keeps clicking on links, never hitting "back" but eventually gets bored and starts on another random page. The irreducible of the chain guarantees that the long-run stationary vector π^T, known as the PageRank vector, exists. The calculated PageRank vector will be used to measure the importance of web resources. From the view of PageRank algorithm, we note that the PageRank model is just based on the hyperlink structure without considering the literal information that are carried by the edge between connected pages and the transition probability from a given page to its outgoing links are weighted with equal chance. However, in practical sense, user surfing behavior is far from random jumping. We may describe it as the following procedure: user goes somewhere according to his interests. They arrive there and begin thinking what they are looking at. They may find something interesting which are somewhat similar to their intensions but do still not satisfy their requirements based on some literal indications in the page. Therefore, user will continue searching. Where will user go? The indications obtained through reading page and the literal information of outgoing links will help user weighting those candidates because searching provides a literal match between the words in the user's mind and the words on the page. That is, the transition probability will be influenced by those factors. The intuition of us is that inbound links from pages with similar topic to our own have larger influence on PageRank then links from unrelated pages. In this paper, we propose a unified model to modify the underlying Markov process by giving different weights to different outgoing links from a page based on not only hyperlink structure but also some important literal clues.

SIGIR'04, July 25–29, 2004, Sheffield, South Yorkshire, UK.
ACM 1-58113-881-4/04/0007.

2. OUR APPROACH

2.1 Virtual Document

What is the virtual document? The concept of virtual document is introduced by Glover [4]. There may several possible ways to define VD. In this paper, the virtual document of a given page is comprised of the expanded anchor text from pages that point to him and some important words on the page itself. The definition equation shows as follows:

$AnchorText\,(i,j)$: set of terms that appear in and around the anchor text from page i to j.
$TitleText\,(j)$: set of terms in the title tag of page j.
$VD\,(j)$: set of terms in the virtual document j.
$VD\,(j) = \bigcup_i (AnchorText\,(i,j)\,, TitleText\,(j))\,.$

What is functionality of virtual document? The anchor text, which is typically a short summary of the target pages within the context of the source page being viewed, can be regarded as objective impression of Web pages and the text information in the title tag, which are usually succinct and representative, can be looked at as subjective presentation of author's motivation. Therefore, VD, the collective wisdom of what the page is about does a useful implicit resource for representing the user's mind of Web pages.

2.2 Algorithms

How does the surfer decide his target after viewing the current page? We can look at user selecting behavior as the process of virtual document generation. Therefore, language modeling algorithms can be put into use for calculating the probability of a selected outgoing link. Here, the Okapi model [2] was used for calculating the similarity of the page with its target documents, denoted by $SIM(i,j)$. In addition, we noted that the virtual document of a page play an important role on attracting the user's attention to it. The literal information in the virtual document shows the user's intensions to some extent. We assume that user intensions are somewhat consistent in the searching path. Therefore, in our approach, summation of inverse document frequency of terms in the intersection between two connected virtual documents is used as an informative weighting factor, denoted by $VI(i,j)$, to indicate the dependent degree of the two connected pages. $TranOdds(i,j)$ is the value for indicating the likelihood that a user will surf from i to j. The equation is:

$$
\begin{aligned}
&Content(j) \ : \text{ set of terms in the page j.}\\
&df(w) = \#\{j|w \in Content(j)\}\\
&TF(w,j) = \#\{w|w \in Content(j)\}\\
&VDDF(w) = \#\{j|w \in VD(j)\}\\
&IVDDF(w) = \log {}^{N}/_{VDDF(w)} + 1\\
&SIM(i,j) = \sum_{w \in VD(j)} \frac{TF(w,j)}{TF(w,j)+0.5+1.5\frac{dl}{avg.dl}} \times \frac{\log\left(0.5+{}^{N}/_{df(w)}\right)}{\log(1.0+\log N)}\\
&VI(i,j) = \sum_{w \in (VD(i) \cap VD(j))} IVDDF(w)\\
&Lit : Link(i,j) \rightarrow \begin{array}{l} 1, \ if : A \wedge B \wedge C \\ 0, \ otherwise \end{array}
\end{aligned}
$$

where:

$$
\begin{aligned}
&A : VD(j) \neq \emptyset; \ B : Content(i) \neq \emptyset\\
&C : \{w|w \in (VD(j) \cap Content(i))\} \neq \emptyset\\
&if : Lit(Link(i,j)) = 1\\
&TranOdds(i,j) = \begin{cases} VI(i,j) \times SIM(i,j), \\ \ if: \ \{w|w \in (VD(i) \cap VD(j))\} \neq \emptyset \\ SIM(i,j), \ otherwise \end{cases}
\end{aligned}
$$

Our unified model is also based on the Markov chain. The calculating method shows in the following equations. The parameter γ is used for adjusting the probability that surfer tends to follow those links with literal matching information. In this paper, the γ is set 0.7.

$$
\begin{aligned}
&PR(j) = (1-\lambda)\,{}^{1}/_{N} + \lambda \sum_{i \in B_j} PR(i)\,prob(i \rightarrow j)\\
&prob(i \rightarrow j) = \begin{cases} \frac{\gamma \times TranOdds(i,j)}{\sum_{k \in F_i} TranOdds(i,j)}, \ Lit(Link(i,k)) = 1 \\ \frac{(1-\gamma)}{\#F(i) - litLink(i)}, \ otherwise \end{cases}
\end{aligned}
$$

$$
\begin{aligned}
&Where :\\
&B(i) : \ set\ of\ pages\ which\ link\ to\ page\ i;\\
&F(i) : \ set\ of\ pages\ which\ page\ i\ links\ to;\\
&r \ : \ the\ transition\ probability\ follow\ literal\ link.\\
&litLink(i) = \#\{k|k \in F(i) \wedge Literal(Link(k,i)) = 1\}
\end{aligned}
$$

3. EXPERIMENT RESULTS

We ran experiments on NTCIR 100G Web data [3]. To evaluate the effectiveness of our approach, the simple comparison was performed based on only rank value of the right answer (relevant files) in the result sets between PageRank and our approach. Among 350 relevant documents of 38 queries, we got 206 win, 128 fail and 16 equal. The rough metric is defined as:

$$
\begin{aligned}
&R : return\ doc.\ sets\ for\ a\ given\ query.\\
&\tau_2 : \ doc.\ in\ R\ sort\ by\ original\ pagerank\ value.\\
&\tau_3 : doc.\ in\ R\ sort\ by\ pagerank\ value\ of\ our\ approach\\
&\tau_k(i) : rank\ of\ i\ in\ \tau_k\\
&\begin{cases} win \ : \ \tau_2(i) > \tau_3(i) \\ fail \ : \ \tau_2(i) < \tau_3(i) \quad , i \in R \\ equal \ : \ \tau_2(i) = \tau_3(i) \end{cases}
\end{aligned}
$$

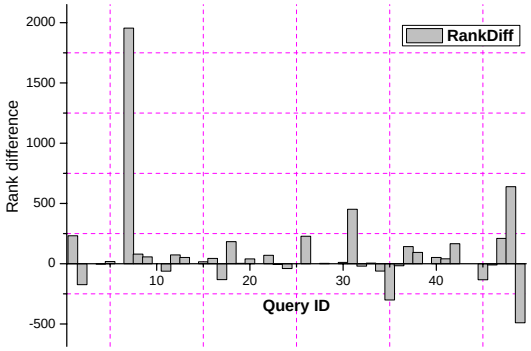

Figure 1: Rank summation difference of relevant files for each query

To observe the difference degree clearly, plot based on the summation of rank difference, $\tau_2(i) - \tau_3(i)$, of right answers for each query shows in Fig 1. It shows clearly that the rank values based on our approach are less than that of original PageRank in most cases. Therefore, our proposed model does make sense for getting more accurate page importance.

4. CONCLUSIONS

In this paper, we proposed a unified model, which combines literal matching with link structure analysis. We provide a good mechanism for unifying the literal mining and link analysis for Web information retrieval. Experiment results show that our approach will potentially do a better job for IR task than original PageRank.

5. ACKNOWLEDGMENTS

This work was supported by the 21st Century COE program "Intelligent Human Sensing", from the Ministry of Education, Culture, Sports, Science and Technology.

6. REFERENCES

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

[2] W. B. Croft. *Language modeling for information retrieval*. Kluwer Academic Publishers, 2003.

[3] K. Eguchi, K. Oyama, E. Ishida, N. Kando, and K. Kuriyama. System evaluation methods for web retrieval tasks considering hyperlink structure. In *the 12th international world wide wed conference*, page 344. WWW, 2003.

[4] E. J. Glover, K. Tsioutsiouliklis, S. Lawrence, D. M. Pennock, and G. W. Flake. Using web structure for classifying and describing web pages. In *Proc. 11th WWW*, pages 562–569. WWW, 2002.