A Link-Based Ranking Scheme for Focused Search

Tony Abou-Assaleh Tapajyoti Das Weizheng Gao
Yingbo Miao Philip O'Brien Zhen Zhen

GenieKnows.com
Halifax, Nova Scotia, Canada
research@genieknows.com

ABSTRACT

This paper introduces a novel link-based ranking algorithm based on a model of focused Web surfers. FocusedRank is described and compared to implementations of PageRank and Topic-Sensitive PageRank. We report a user study that measures the relevance and precision of each approach. FocusedRank gives superior relevancy over PageRank, while significantly reducing the computational complexity compared to the Topic-Senstivice PageRank.

Categories and Subject Descriptors

H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval—*Retrieval models*

General Terms

Algorithms, Experimentation

Keywords

search, link analysis, ranking, surfer model, focused search

1. INTRODUCTION

As the wealth of searchable data grows, the need for specialized tools focused exclusively on a particular topic of interest grows, prompting the inception of *focused search engines*. Focused search addresses the multi-dimensionality of Web content and the relationship between pages of similar content. It limits the scope of the documents indexed to those whose content falls under a particular topic. Recent topic-oriented ranking schemes have provided some solutions to this end [2, 3, 4].

Traditional strategies for computing topic-sensitive link-based ranks of Web documents have stored multiple ranks for each document; one for each topic. While some are entirely query-dependent [4], others compute query-document scores online by first classifying a query according to the same topics used for classifying the document collection [2]. We show that a single rank can be computed off-line for each document based on the documents which link to it and the topics of these documents. Additionally, we show that the model introduced in this paper can achieve accuracy and relevance equivalent to that of multiple rank-vector approaches, with a substantial improvement in on-line query processing performance and reduced storaged requirements.

In our model, a link from page u to page v exists only if there is a hyperlink between these and they belong to at least one common topic. For example, if many small commercial Web sites were to provide PDFs for download and link to a site where a PDF reader could be downloaded, no degree of rank is carried between these since they will not share any topics. The premise of this decision is that popular pages will only convey a vote of importance to other pages on the same or similar topic. This notion is especially useful in the implementation of link-based ranking for focused search engines where the results of queries are required to be accurate and on the same topic or set of topics.

2. METHODOLOGY

For a collection of documents D and a list of topics τ we obtain a set of probabilities $P(d, t_i)$, where $d \in D$ and $t_i \in \tau$, of the document d belonging to a particular topic t_i. The vector C_d contains the topic membership probabilities for d.

The topical overlap between two documents is computed by calculating the normalized sum-of-products of the topic-membership probability vectors of two documents. When two documents have few topics in common, the topical overlap between them is low. Moreover, when two pages are on the same topic(s), but the probability of their inclusion in those topics $P(d, t_i)$ is low, the topical overlap remains low. In this way, the topical overlap score strengthens propagation of rank to pages that belong to the same topics and have a high probability of belonging to these topics. Where a link exists from page u to page v, the topical overlap score is computed as

$$T(u, v) = \sum_{j \in \tau} C_u(j) C_v(j).$$

Given a focused surfer on page u, our model dictates that the probability of the surfer navigating to v is a function of the two pages' topic overlap. Computing T for each pair of hyperlinked documents forms a matrix M where $M_{uv} = T(u, v)$. The probability of a focused surfer following a link from u to v is a function of the topical overlap scores between u and each of the pages to which it links,

$$P_T(u \to v) = \frac{T(u, v)}{\sum_{d \in D} T(u, d)}$$

where $T(u, d)$ is 0 if u does not link to d. $P_T(u \to v)$ is the portion of page u's rank conferred to page v. Since T is a function of the topics shared between two pages, P_T is a

WWW 2007, May 8–12, 2007, Banff, Alberta, Canada.
ACM 978-1-59593-654-7/07/0005.

Table 1: Average relevance scores

Query	PR	TSPR	FR
calorie diet	0.22	0.52	0.82
cardinal fitness	0.17	0.1	0.17
act health mental	0.12	0.26	0.54
avocado	0.78	0.97	1.47
beauty and fitness	0.2	0.33	0.43
berkeley optometry	0.75	0.95	0.34
nursing board	-0.7	0.9	1.0
allergy medicine	0.77	0.9	1.25
cancer chemotherapy	0.5	1.07	1.33
cosmetic dentistry	-0.57	0.27	-0.33
child nutrition	0.35	0.78	0.9
pharmacy schools	-0.53	-0.33	0.27
online drug store	-0.07	0.97	0.97
occupational therapy	0.57	1.2	1.13
clinical health insurance	-0.55	0.2	0.15
appetite zantac	-0.21	-0.55	-1.1

Table 2: Performance comparison and analysis

Metric	PR	TSPR	FR
MAS	0.112	0.532	0.583
MAP	0.315	0.563	0.575

(a) Comparison of mean average score and precision

Metric	PR	TSPR
MAS	0.0058	0.5878
MAP	0.0179	0.9071

(b) P-values when compared to FocusedRank

model of focused surfing behaviour as a function of the set of topics shared, the probability of inclusion in those topics, and the topology of the link graph.

At each iteration of the link-based ranking algorithm, the importance conferred converges until a stable distribution is achieved. The rank at iteration i is

$$Rank_i(u) = \frac{\alpha}{N} + (1-\alpha)\sum_{d\in D} Rank_{i-1}(d)P_T(d \rightarrow u).$$

where α is a scaling factor, N is the number of documents in D, and $P_T(d \rightarrow u)$ is equal to 0 if d does not link to u.

3. EXPERIMENT

Using a collection of support vector machine (SVM) classifiers trained on 60,000 health documents taken from the Open Directory Project (ODP) health category listing, a collection of 42 million health-related documents were classified. This classification produced a list of probabilities for each document, indicating the likelihood that a document belongs to one of the 45 ODP health categories.

Using this classification output, together with the link graph produced during crawling, we computed M, as described in the previous section. Link-based ranking was computed using M and a uniform initialization vector over 30 iterations to produce a final ranking for the documents.

We compared our ranking to the traditional PageRank [1] and Topic-Sensitive PageRank [2] schemes computed on the same 42 million Web documents. To test the performance of each algorithm, 16 health-related queries were selected and executed on the data set using each ranking scheme. For each query, the top 10 search results generated by all ranking schemes were randomly mixed and duplicates were removed. To inspect the results manually, 17 participants were recruited. Queries were distributed to participants such that each received 3 query result pages and no pair of participants shared more than one query. Furthermore, no indication of which ranking scheme generated a result was given.

Participants were asked to rate a result as very relevant, somewhat relevant, not sure, irrelevant, or totally irrelevant. These ratings were internally given the values 2, 1, 0, -1, and -2 respectively. A result is deemed "relevant" if its average score across all participants is greater than or equal to 0.5.

4. RESULTS

The scores received by each search result, for each query, and for each ranking scheme were recorded. The average scores for each query are given in Table 1. Table 2(a) shows performance comparisons of the mean average score (MAS) and mean average precision (MAP). Results of the two-tailed t-tests are listed in Table 2(b).

Topic-Sensitive PageRank (TSPR) and our FocusedRank (FR) out-perform traditional PageRank (PR) with almost a 5-fold increase in mean average score and a 2-fold increase in mean average precision. While FR out-performed TSPR in this experiment, two-tailed t-tests show that the improvement is not statistically significant (p-value > 0.5). The improvement over PR is shown to be significant ($p < 0.02$).

Although the accuracy improvements over TSPR are not significant, FR consistently achieves at least equivalent accuracy with a significant reduction in online processing overhead. As with the approaches discussed in [3] and [4], TSPR is query-dependent; it combines multiple document-topic rank vectors with categorizations of query terms to compute a final rank. FocusedRank achieves the same relevance accuracy and precision without the online expense by computing, offline, a single rank vector.

5. CONCLUSIONS AND FUTURE WORK

We have introduced a focused surfer model and ranking method for improving search results for focused search paradigms. Our model is shown to achieve significant performance and accuracy benefits over vanilla PageRank and approximately equivalent accuracy as Topic-Sensitive PageRank without the online processing overhead.

Our future work will explore classification within topic hierarchies. When topics have subtopics and documents belong to these with varying probabilities, comparison of topic membership and propagation of rank is a nontrivial task. Also, we look to optimize FocusedRank by inspecting the storage and computational performance benefits.

6. REFERENCES

[1] S. Brin and L. Page. The anatomy of a large-scale hypertextual web search engine. In *7th International World Wide Web Conference*, pages 107–117, 1998.

[2] T. H. Haveliwala. Topic-sensitive pagerank. In *11th International World Wide Web Conference*, pages 517–526, New York, NY, USA, May 2002.

[3] L. Nie, B. D. Davison, and X. Qi. Topical link analysis for web search. In *29th Annual International ACM SIGIR Conference*, August 2006.

[4] M. Richardson and P. Domingos. The intelligent surfer: Probabilistic combination of link and content information in pagerank. *Advances in Neural Information Processing Systems*, 14, 2002.