Affinity Rank: A New Scheme for Efficient Web Search

Yi Liu[1], Benyu Zhang[2], Zheng Chen[2], Michael R. Lyu[1], Wei-Ying Ma[2]

[1]Department of Computer Science and Engineering
The Chinese University of Hong Kong
Shatin, NT, Hong Kong
{yliu,lyu}@cse.cuhk.edu.hk

[2]Microsoft Research Asia
49 Zhichun Road, Haidian District
Beijing 100080, P.R. China
{byzhang, zhengc, wyma}@microsoft.com

ABSTRACT

Maximizing only the relevance between queries and documents will not satisfy users if they want the top search results to present a wide coverage of topics by a few representative documents. In this paper, we propose two new metrics to evaluate the performance of information retrieval: diversity, which measures the topic coverage of a group of documents, and information richness, which measures the amount of information contained in a document. Then we present a novel ranking scheme, Affinity Rank, which utilizes these two metrics to improve search results. We demonstrate how Affinity Rank works by a toy data set, and verify our method by experiments on real-world data sets.

Categories & Subject Descriptors: H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval – *retrieval models*, *search process*; H.2.8 [**Database Management**]: Database Applications – *Data Mining*

General Terms: Algorithms, Performance

Keywords: Affinity Rank, Link Analysis, Diversity, Information Richness

1. INTRODUCTION

The top few search results play an important role in the user satisfaction. However, when the user query is short or ambiguous, the top search results are always dominated by very few topics with most popularity or authority. Such topic concentration can hardly meet the needs of diversified information from various users. A possible solution to this problem is to include more topics in the top search results. Furthermore, since fewer results per topic could appear in the top positions, we also hope them to be representative in their topic locality. To satisfy these purposes, we propose two new metrics, *diversity* and *information richness*, to evaluate the retrieval performance. We also introduce algorithms to calculate information richness for each document by analyzing the link graph constructed from the similarity relationship between documents. Then a penalty is imposed on the score of each document to measure its influence on the topic diversity. The combination of information richness and diversity penalty constitute our new ranking scheme: Affinity Rank.

Our new ranking scheme is highly related to many research efforts on link analysis for retrieval performance improvement, including the well-known Google's PageRank algorithm [1] and Kleinberg's HITS algorithm [2]. Actually the computation of information richness in our method is very similar to that of PageRank. However, the link structure we exploit is not based on explicit hyperlinks on the web pages, but the similarity between document pairs. Mining similarity data as link graphs has been discussed at the theoretical level in some research work in the field of statistics [3] and in some applications such as image retrieval [4]. These efforts have also motivated us to apply a similar concept to the area of web information retrieval.

2. AFFINITY RANK SCHEME

2.1 Definitions

First we give formal definitions on the two new metrics:

Definition of *Diversity*: *Given a set of documents* $R=\{d_1,d_2,\cdots d_m\}$ *, we use diversity* $Div(R)$ *to denote the number of different topics to measure the topic diversity contained in R.*

Definition of *Information Richness*: *Given a document collection* $D=\{d_i\,|\,1\le i\le n\}$ *, we use information richness* $InfoRich(d_i)$ *to denote the informative degree of the document* d_i*, i.e., the richness of information contained in the document* d_i *with respect to the entire collection D. Without loss of generality, we let* $InfoRich(d_i)\in[0,1]$.

Specifically, *information richness* measures how much information a single document contains in its topic locality. A document with high information richness should be inclusive of other similar ones so that it can well represent the topic. *Diversity*, on the other hand, measures how many different topics are covered by a group of documents.

2.2 Algorithms

Let $D=\{d_i\,|\,1\le i\le n\}$ denote a document collection. According to vector space model, each document d_i can be represented as a vector $\vec{d}_i$. The *similarity* between any pair of documents can be calculated as $sim(d_i,d_j)=\cos(\vec{d}_i,\vec{d}_j)$. Using a threshold S_t, we construct a link for each pair of documents d_i and d_j if $sim(d_i,d_j)>S_t$ is satisfied. At the same time $sim(d_i,d_j)$ is also assigned as the weight to the link. Thus a link graph is built and it depicts the similarity relationship in the whole document collection. The link graph can be represented by an adjacency matrix $\mathbf{M}$, each of whose entry represents the weight of a link.

The computation of information richness is based on two intuitions: 1) the more neighbors a document has, the more informative it is; and 2) the more informative a document's neighbors are, the more informative it is as well. Formulating the above notions in a matrix form, we can compute the principal eigenvector λ of $c\widetilde{\mathbf{M}}^T+(1-c)\mathbf{U}$, where $\widetilde{\mathbf{M}}$ is a matrix by normalizing the sum of each row of $\mathbf{M}$ to 1, $\mathbf{U}=[\frac{1}{n}]_{n\times n}$, and c

WWW 2004, May 17–22, 2004, New York, New York, USA.
ACM 1-58113-912-8/04/0005.

is a dumping factor whose value is always set to 0.85 (similar to the formulation of PageRank [1]). Each entry of the eigenvector is then the value of the document's information richness, i.e., $\lambda = [InfoRich(d_i)]_{n\times 1}$.

Computing information richness can help us choose more informative documents for each topic, but not to preclude the possibility of excessively selecting similar ones from the same topic. Furthermore, we impose different penalties to the score of information richness of each document in terms of its influences to the topic diversity. The combination of information richness and the diversity penalty leads to a new score, called *Affinity Rank* score. The following greedy algorithm is used to iteratively impose penalty to documents topic by topic and update the Affinity Rank scores.

The Greedy Algorithm for Diversity Penalty

```
(1) Initialize two sets $A = \Phi, B = \{d_i \mid i = 1,2,\cdots n\}$, and set the
    initial Affinity Rank scores to the value of information richness, i.e.,
    $AR_i = InfoRich(d_i), i = 1,2,\cdots n$
(2) Sort the documents in $B$ by $AR_i$ in descending order.
(3) Suppose the document ranked highest in $B$ is $d_i$. Move document $d_i$
    from $B$ to $A$, and then decrease the Affinity Rank scores of less
    informative documents by the part conveyed from the most informative one.
    E.g., for each document $d_j (j \neq i)$, let $AR_j = AR_j - \widetilde{M}_{j,i} \cdot InfoRich(d_i)$.
(4) Re-sort the documents in B by the updated rank scores $AR_i$ in
    descending order.
(5) Go to (3) until $B = \Phi$ or until a predefined maximum loop
    count is reached.
```

Using Affinity Rank we can re-rank the preliminary search results which are ordered by full-text search. The most straightforward re-ranking mechanism is a linear combination of each result's ranks in full-text search and in Affinity Rank.

3. EXPERIMENTS

3.1 Toy Data

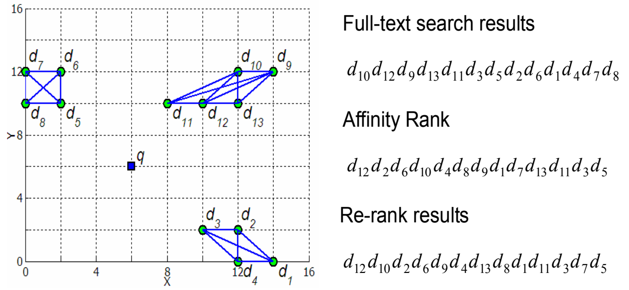
16
12
Y 8
4
0
0
4
8
X
12
16
d_7
d_6
d_8
d_5
d_{10}
d_9
d_{11}
d_{12}
d_{13}
q
d_3
d_2
d_4
d_1
Full-text search results
$d_{10}d_{12}d_9d_{13}d_{11}d_3d_5d_2d_6d_1d_4d_7d_8$
Affinity Rank
$d_{12}d_2d_6d_{10}d_4d_8d_9d_1d_7d_{13}d_{11}d_3d_5$
Re-rank results
$d_{12}d_{10}d_2d_6d_9d_4d_{13}d_8d_1d_{11}d_3d_7d_5$

Figure 1. Toy Data Demonstration of Affinity Rank

Figure 1 demonstrates a toy dataset to show how Affinity Rank works. Suppose that the circles represent documents and the square represents a query, their positions on the 2-dimension grid corresponding to their vector representation. Those circles form three clusters, indicating three different topics of documents. Links are also labeled as connections between circles. By threshold selection we can keep most links within each topic (in our toy data we show an ideal case that no link is constructed across different topics by setting the threshold to 0.9).

Figure 1 also shows the results by using the query in the toy data for retrieval. We can see that the top five positions by full-text search are occupied by $d_9 \sim d_{13}$, which are the most relevant five documents to the query, but all of them come from the same topic. However, the highest three in Affinity Rank, d_{12}, d_2 and d_6, not only come from three different topics but also are central in each topic respectively. Re-ranking by combining the above two ranks with a 1:2 weighting ratio, the top four results become d_{12} and d_{10}, which are the two most relevant documents, followed by d_2 and d_6, which are two central documents from the other two different topics. The toy data demonstrates that our new ranking scheme gives attentions to all the three metrics: diversity, information richness, as well as relevance.

3.2 Real-world Data

We also conduct experiments on two sets of real-world data. One experiment is retrieval on web pages crawling from the domain of cs.berkeley.edu, which consists of over 73,000 pages. Another experiment is a newsgroup search, in which we collect 256,449 posts from 117 Microsoft newsgroups. Re-ranking is performed on the top 50 results from full-text search. For the top 10 search results, we achieve improvements shown as Table 1. The results suggest that by Affinity Rank we efficiently improve the diversity and information richness in the top search results without a significant change in relevance.

Table 1. Improvement in top 10 search results

	Diversity	Information Richness	Relevance
Berkeley Data	+22.29%	+19.17%	-2.50%
Newsgroup Data	+31.02%	+11.97%	+0.72%

4. CONCLUSION

In this paper we introduce a novel ranking scheme, Affinity Rank, to improve information retrieval performance based on two proposed evaluation metrics, diversity and information richness. We demonstrate the effectiveness of our method by toy data and also verify it with real-world data experiments.

5. ACKNOWLEDGMENTS

The work was done in Microsoft Research Asia, and was partially supported by grants from the Research Grants Council of the Hong Kong Special Administrative Region, China (Project No. CUHK4182/03E and CUHK4351/02E).

6. REFERENCES

[1] Page, L., Brin, S., Motwani, R. and Windograd, T. *The pagerank citation ranking: Bring order to the web*, Stanford Digital Library Technologies Project, 1998.

[2] Kleinberg, J.M. Authoritative sources in a hyperlinked environment. *Journal of the ACM (JACM)*, 46 (5). 604-632.

[3] Meila, M. and Shi, J., A random walks view of spectral segmentation. In *Proceedings of the International Workshop on AI and Statistics(AISTATS)*, (Florida, 2001), 177-182.

[4] He, X., Ma, W.-Y. and Zhang, H.-J., Spectral Techniques for Structural Analysis of Image Database. In Proceedings of the 2003 International Conference on Multimedia and Expo, (Baltimore, 2003), 25-28.