Diversification of Search Results using Webgraphs

Praveen Chandar and Ben Carterette
{pcr,carteret}@udel.edu
Department of Computer and Information Sciences
University of Delaware
Newark, DE, USA 19716

ABSTRACT

A set of words is often insufficient to express a user's information need. In order to account for various information needs associated with a query, diversification seems to be a reasonable strategy. By diversifying the result set, we increase the probability of results being relevant to the user's information needs when the given query is ambiguous. A diverse result set must contain a set of documents that cover various subtopics for a given query. We propose a graph based method which exploits the link structure of the web to return a ranked list that provides complete coverage for a query. Our method not only provides diversity to the results set, but also avoids excessive redundancy. Moreover, the probability of relevance of a document is conditioned on the documents that appear before it in the result list. We show the effectiveness of our method by comparing it with a query-likelihood model as the baseline.

Categories and Subject Descriptors: H.3.3 [Information Storage and Retrieval]

General Terms: Algorithms

Keywords: information retrieval, diversity, webgraphs.

1. INTRODUCTION

Users express information needs using a set of keywords. Current retrieval systems fail to capture the different information needs that could be expressed by users using the same set of keywords. Clearly this leads to multiple interpretations for a given query. For example, consider the query *kcs*. There are multiple interpretations for this query, one being the Kansas City Southern railroad; another, being Kanawha County Schools in West Virginia; one more interpretation is information on KCS Energy, Inc.

In order to maximize the user experience it appears reasonable to *diversify* the result set. *Diversify* means to examine the query with a broader perspective and account for the multiple information needs for the query. This diversification would provide complete coverage of subtopics for a given query to the user. Ranking with diversity requires moving away from the assumption that documents are independently relevant to the query. Each document must be ranked based not just on its similarity to the query but also based on the documents retrieved before it.

SIGIR'10, July 19–23, 2010, Geneva, Switzerland.
ACM 978-1-60558-896-4/10/07.

Our task is same as the "diversity" task of the TREC Web Track [4]; the goal of the system is to return a ranked list of documents that provides complete coverage for a given topic, while avoiding excessive redundancy in the result set. We have used the topics created for this task. These topics consists of a query, a description of an information need, and one or more subtopics or alternative interpretations of the query. These topics were developed from information extracted from the logs of a commercial Web search engine, thereby ensuring a good mix of user needs for a given query.

Most previous work, including the MMR approach of Carbonell & Goldstein [2] and the language modeling framework proposed by Zhai et al. [7], involve a greedy approach to finding subtopics. In this work, we propose a method using the link structure of the web to maximize the subtopics covered for a given query. Our method identifies authoritative documents in a set and assumes that these authoritative documents represent a subtopic. We evaluate our proposed method using the α-nDGC measure proposed by Clarke [5] and intent aware precision (P-IA) proposed by Agarwal et al [1] and compare it to a query-likelihood baseline.

2. THE WEBGRAPH METHOD

The link structure has often provided a rich source of information about the content of the environment. Our method uses the information provided by the link structure to find several densely linked collections of hubs and authorities within a subset of the results. Each densely linked collection could potentially cover different subtopics for a given query.

In our approach, we re-rank an initial ranking of documents (query-likelihood results) to provide a diverse ranking of documents. The documents in this initial ranking consisting of hyperlinked pages are represented as a directed graph $G = (V, E)$: nodes corresponds to pages and a directed edge $(p, q) \in E$ correspond to the presence of link from page p to q. We expand the subgraph to include all the *in-links* to the subgraph and *out-links* from the subgraph. The *hubs* and *authorities* scores are calculated for each document using the iterative procedure described by Kleinberg [6].

Kleinberg's procedure begins by representing the directed graph as an adjacency matrix. The principal and non-principal eigenvectors are calculated from this matrix multiplied by its transpose. Each value in an eigenvector represents a document score. The values in the principal eigenvector correspond to the Kleinberg's *hub* score for a document. The non-principal eigenvectors represent other densely-linked clusters in the graph; they have both positive and negative entries, but we consider only the positive entries.

No. of Eigenvectors	α-nDCG10	P-IA10
5	0.100	0.038
10	0.154	0.050
25	0.143	0.051
50	0.169	0.057
100	0.142	0.047
0 (baseline)	0.124	0.061

Table 1: Diversity results for varied number of eigenvectors and 50 terms.

For each eigenvector we construct a language model using the documents corresponding to the k greatest values. Therefore, the m language models constructed from the documents correspond to the k greatest values in each of the first m eigenvectors. The intuition is that the link structure clusters the documents into subtopics, therefore these language models provide a hypothetical set of subtopic models. The language model corresponding to each subtopic is evaluated against the query and then we take the document with the greatest score. This produces a set of documents (possibly fewer than m) which are the highest scoring for the hypothesized subtopics that are then ranked in decreasing order of the original query-likelihood scores. We iterate in this way, taking the highest-scoring set of documents remaining, until we rank the top 200 documents in the original ranking. This method of iterating to obtain the final ranking is similar to the one described by Cartertte et al [3].

3. IMPLEMENTATION AND RESULTS

In our experiments, we used the ClueWeb09 dataset consisting of one billion web pages (5 TB compressed, 25 TB uncompressed), in ten languages, crawled in January and February 2009. We indexed the smaller set of "Category B" which consists of 50 million web pages in English. We used the *webgraphs* in the dataset which has about 428,136,613 unique URLs and 454,075,638 outlinks. This test collection was used for the diversity task at TREC'09. A total of 50 queries were evaluated and the subtopics for each query ranged from 3 to 8. We used the Lemur Toolkit and the Indri search engine in our experiment. The query-likelihood result set with Dirichlet smoothing ($\mu = 2000$) was used as our baseline results for reranking.

Our method was evaluated using the two measures which reward novelty and diversity, namely α-normalized discounted cumulative gain (α-nDCG) and intent-aware precision (P-IA). All our methods were evaluated at rank 10 with $\alpha = 0.5$ in α-nDCG. To see whether the setting of parameters such as m (the number of eigenvectors) and n (number of terms) may affect the performance, we compare the results for a range of values.

By comparing the results of the two parameters in Figure 1 we see that in general the performance increases and reaches a maximum at 50 eigenvectors and starts to decrease again. The number of terms in the model has less effect on the results. We report the diversity results by varying the *number of eigenvectors* along the Indri baseline model in Table 1. This table shows that our method did considerably well in diversifying the results set for all parameter values according to the α-nDCG measure although for the P-IA measure the results were below the baseline.

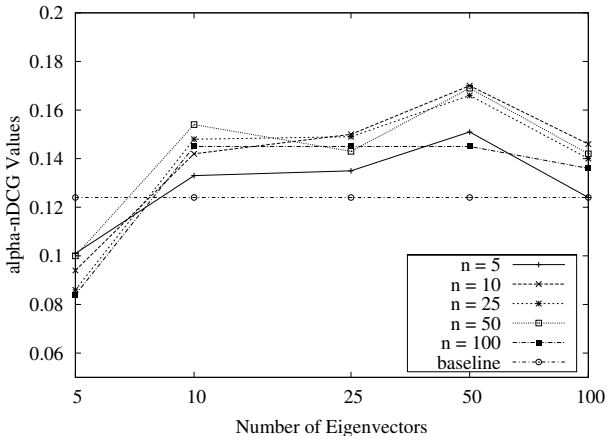

Figure 1: α-nDCG averaged over 50 queries with increasing numbers of eigenvectors (subtopic models) and terms in each model.

4. CONCLUSIONS AND FUTURE WORK

In this work, we have proposed a novel method for diversifying search results. The webgraph method produces a diverse ranking from an initial set of documents for a given query by considering the underlying link structure of the retrieved documents. We believe more information can be harnessed from the hyperlink structure of retrieved documents; our work provides enough evidence for future work along these lines.

5. REFERENCES

[1] R. Agrawal, S. Gollapudi, A. Halverson, and S. Ieong. Diversifying search results. In *Proceedings of WSDM '09*, pages 5–14, 2009.

[2] J. Carbonell and J. Goldstein. The use of mmr, diversity-based reranking for reordering documents and producing summaries. In *Proceedings of SIGIR '98*, pages 335–336, 1998.

[3] B. Carterette and P. Chandar. Probabilistic models of ranking novel documents for faceted topic retrieval. In *Proceeding of CIKM '09*, pages 1287–1296, 2009.

[4] C. L. Clarke, N. Craswell, and I. Soboroff. Overview of the trec 2009 web track. In *Proceedings of TREC*, 2009.

[5] C. L. Clarke, M. Kolla, G. V. Cormack, O. Vechtomova, A. Ashkan, S. Büttcher, and I. MacKinnon. Novelty and diversity in information retrieval evaluation. In *Proceedings of SIGIR '08*, pages 659–666, 2008.

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

[7] C. X. Zhai, W. W. Cohen, and J. Lafferty. Beyond independent relevance: methods and evaluation metrics for subtopic retrieval. In *Proceedings of SIGIR '03*, pages 10–17, 2003.