Distribution of Relevant Documents in Domain-level Aggregates for Topic Distillation

Vassilis Plachouras
University of Glasgow
Glasgow, G12 8QQ, U.K.
vassilis@dcs.gla.ac.uk

Iadh Ounis
University of Glasgow
Glasgow, G12 8QQ, U.K.
ounis@dcs.gla.ac.uk

ABSTRACT

In this paper, we study the distribution of relevant documents in aggregates, formed by grouping the retrieved documents according to their domain. For each aggregate, we take into account its size, and a measure of the correlation between its incoming and outgoing hyperlinks. We report on a preliminary experiment with two TREC topic distillation tasks, where we find that larger aggregates, or those aggregates with correlated hyperlinks, are more likely to contain relevant documents. This result shows that the distribution of domain-level aggregates is potentially useful for finding relevant documents.

Categories and Subject Descriptors

H.3.3 [**Information Storage and Retrieval**]

General Terms: Experimentation

Keywords: Web IR, aggregates, distribution of relevant documents

1. INTRODUCTION

Web documents are organised in various ways. For example, documents are connected by hyperlinks, providing the means for navigation. In Web Information Retrieval (IR), hyperlink analysis approaches have been used to detect and retrieve the most authoritative documents. In addition, closely related Web documents are found together in aggregates, organised hierarchically in directories, sites and domains. There have been efforts to automatically identify aggregates of hypertext or Web documents, using either graph theory [2], or heuristics based on observations of the structure of sites [3, 5]. In the context of TREC experiments, grouping documents according to their domain has been employed in order to limit the redundancy of retrieving many documents from a given site [4]. However, the usefulness of evidence from the distribution of domain-level aggregates has not been fully studied.

In this paper, we look at the distribution of relevant documents in domain-level aggregates, and how it is related to two features of the aggregates: the size of an aggregate, and the correlation between the incoming and outgoing hyperlinks within the aggregate. Our data set is the standard .GOV TREC collection, and its associated topic distillation queries from TREC11 and TREC12[1].

[1]http://trec.nist.gov/pubs.html

WWW2004, May 17–22, 2004, New York, New York, USA.
ACM 1-58113-912-8/04/0005.

2. FEATURES OF AGGREGATES

In this paper, the aggregates are formed at query time, by grouping documents from the same domain. Let $D = \{d_i\}$ be the set of retrieved documents for a query. For each unique domain that appears in D, we create one aggregate a_j. We look at two basic features of these domain-level aggregates.

The first feature is related to the size $s(a_j)$ of an aggregate a_j, formed for a given query. We assume that if a_j is relatively large, then a_j contains more documents related to the query topic. In this case, a retrieval approach for finding the entry points of a_j may improve retrieval effectiveness.

The second feature we consider is the correlation between the outgoing and incoming links within an aggregate a_j. The existence of such a correlation suggests that there is a pattern in the distribution of the outgoing and incoming hyperlinks within a_j. Therefore, employing additional evidence from hyperlink analysis may be appropriate for enhancing retrieval results. For each aggregate a_j, we compute the percolation threshold $q_c(a_j)$ [6], as follows:

$$q_c(a_j) = \frac{\langle outdegree_i \rangle}{\langle outdegree_i \cdot indegree_i \rangle} \quad (1)$$

where $outdegree_i$ and $indegree_i$ stand for the number of outgoing and incoming links of document d_i within a_j respectively, and $\langle x \rangle$ stands for the average of x. We consider only the hyperlinks within a_j, since we want to measure the cohesiveness of the specific aggregate. If $q_c(a_j)$ is undefined (for $\langle outdegree_i \rangle = 0$), or $q_c(a_j) \to +\infty$, then there is no correlation between the outgoing and the incoming links within the aggregate a_j, and therefore, the hyperlinks are distributed without any apparent pattern. When $q_c(a_j) \in (0, +\infty)$, we assume that there is some pattern in the distribution of hyperlinks within a_j; the lower the value of $q_c(a_j)$, the stronger the correlation is.

We assume that if an aggregate a_j is large, or the outgoing and incoming hyperlinks within a_j are correlated, then a_j is more likely to contain relevant documents. We will test the validity of this assumption in the following section.

3. EXPERIMENT AND RESULTS

In order to test the assumption that relevant documents are more likely to be found in large aggregates, or in aggregates with correlated hyperlinks, we experiment with a standard TREC collection, the .GOV, and the topic distillation queries from both TREC11 and TREC12. More specifically, the .GOV is a recent crawl of approximately 1.25 million documents from the .gov domain. Both topic distillation tasks involve finding useful entry points for the query topics. However, for the TREC12 queries, the relevant documents were restricted to be homepages of relevant sites, thus resulting in a lower number of relevant documents per query.

In order to form the aggregates of documents for each query, we perform a content-only retrieval using the weighting scheme $PL2$ from Amati and van Rijsbergen's Divergence From Randomness (DFR) framework [1]. In order to reduce the overhead of forming the aggregates, we select the set of the top 20000 documents per query. Within this set, we identify all the distinct domains and form the corresponding aggregates a_j. Next, we compute the sizes $s(a_j)$, as well as the average size $\langle s(a_j)\rangle$, and the percolation coefficient $q_c(a_j)$ for each aggregate.

We test the assumption of Section 2 by comparing the number $n(a_j \text{ rel})$ of aggregates with at least one relevant document to: (a) the number of aggregates a_j with at least one relevant document and $s(a_j) > \langle s(a_j)\rangle$, and (b) the number of aggregates a_j with at least one relevant document and $q_c(a_j) \in (0, +\infty)$. Both conditions are based on the characteristics of the distributions of aggregates.

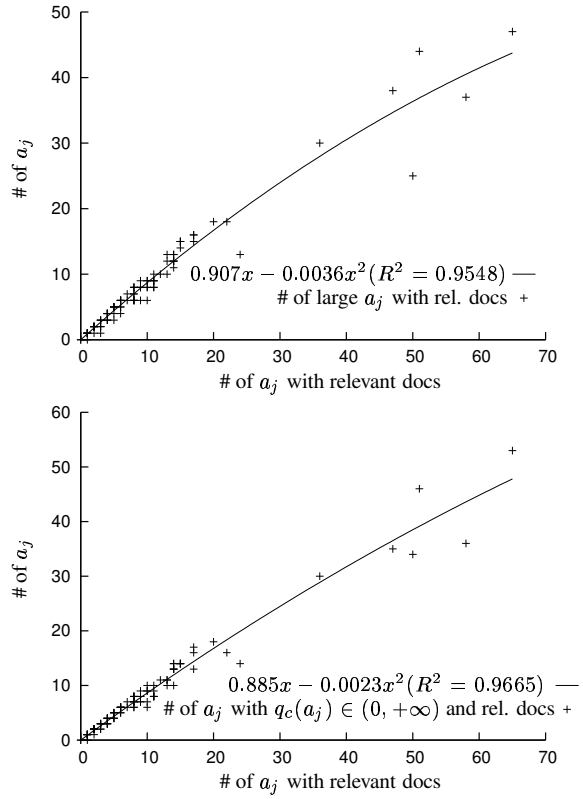

Figure 1: Scatter plots of the number of a_j with at least one relevant document versus the number of a_j with at least one relevant document and $s(a_j) > \langle s(a_j)\rangle$ (top), or $q_c(a_j) \in (0, +\infty)$ (bottom), for both TREC11 and TREC12.

From Figure 1, we can see that the distribution of relevant documents in aggregates and the two features of the aggregates are related. For both cases, a quadratic curve fits the data accurately (linear fitting matched the data nearly as well, returning $R^2 = 0.9373$ and $R^2 = 0.9598$). This shows that if $n(a_j \text{ rel})$ is low for a query, then the sizes of the corresponding aggregates are more likely to be higher than $\langle s(a_j)\rangle$ for the same query. Moreover, we find that the outgoing and incoming links within these aggregates are correlated for most of the queries. As the number of aggregates with relevant documents increases, we expect that some of the relevant documents will appear in either smaller aggregates, or in aggregates without correlated hyperlinks, justifying a quadratic curve fitting.

Moreover, we compare the probability $P(a_j \text{ rel})$ of finding an aggregate with at least one relevant document, to the conditional probabilities that consider the size and the percolation coefficient of the aggregates. We find that the probability $P(a_j \text{ rel} | s(a_j) > \langle s(a_j)\rangle)$ of finding a large aggregate with at least one relevant document is on average 5.21 times higher than $P(a_j \text{ rel})$. In addition, we find that the probability $P(a_j \text{ rel} | q_c(a_j) \in (0, +\infty))$ of finding an aggregate with at least one relevant document and correlated hyperlinks is on average 4.93 times higher that $P(a_j \text{ rel})$. Therefore, it is more likely to find relevant documents in aggregates satisfying either of the two conditions.

4. DISCUSSION AND CONCLUDING REMARKS

In this paper, we have presented results from an ongoing study of the distribution of relevant documents in domain-level aggregates, for two TREC topic distillation tasks. Since Web documents are often organised in aggregates of closely related documents, we use this information, and introduce a new source of evidence that can be used for documents ranking. Our results show that both the size of aggregates, as well as the correlation between their incoming and outgoing hyperlinks, are effective for identifying the aggregates, which are more likely to contain relevant documents.

From a Web IR perspective, these findings underline the fact that apart from using the textual content of Web documents, along with hyperlink analysis, evidence from the distribution of aggregates can also be exploited. Indeed, in a refined model for topic distillation, we could focus on the aggregates that are more likely to contain relevant documents. For example, given a new query, we could select the large aggregates, which are more likely to contain relevant documents, and apply hyperlink analysis to retrieve their best entry points.

In future experiments, we will employ alternative features and more refined approaches for detecting aggregates, using different collections, as they become available. Moreover, an interesting direction of further research is the incorporation of these features in ranking the documents.

5. ACKNOWLEDGEMENTS

This work is funded by a UK EPSRC project grant, number GR/R90543/01. The project funds the development of the Terrier IR framework (http://ir.dcs.gla.ac.uk/terrier).

6. REFERENCES

[1] G. Amati and C. J. van Rijsbergen. Probabilistic models of information retrieval based on measuring divergence from randomness. *ACM TOIS*, 20(4):357–389, 2002.

[2] R. A. Botafogo and B. Shneiderman. Identifying aggregates in hypertext structures. In *Proceedings of the 3rd ACM conference on Hypertext*, pp. 63–74, 1991.

[3] N. Eiron and K. McCurley. Untangling compound documents on the web. In *Proceedings of the 14th ACM conference on Hypertext and hypermedia*, pp. 85–94, 2003.

[4] K. L. Kwok, P. Deng, N. Dinstl, and M. Chan. TREC2002 Web, Novelty and Filtering Track Experiments using PIRCS. In *Proceedings of TREC11*, pp. 520–528, 2002.

[5] W.-S. Li, O. Kolak, Q. Vu, and H. Takano. Defining logical domains in a web site. In *Proceedings of the 11th ACM conference on Hypertext and Hypermedia*, pp. 123–132, 2000.

[6] N. Schwartz, R. Cohen, D. ben Avraham, A.-L. Barabási, and S. Havlin. Percolation in directed scale-free networks. *Physical Review E*, 66, 2002.