Information Processing and Management 000 (2016) 1–22

Contents lists available at ScienceDirect

Information Processing and Management

journal homepage: www.elsevier.com/locate/ipm

Reducing hardware hit by queries in web search engines

Marcelo Mendoza[a,*], Mauricio Marín[b], Verónica Gil-Costa[c], Flavio Ferrarotti[d]

[a] Universidad Técnica Federico Santa María, Santiago, Chile
[b] Universidad de Santiago de Chile, Santiago, Chile
[c] CONICET, Universidad Nacional de San Luis, Argentina
[d] Software Competence Center Hagenberg, Austria

ARTICLE INFO

Article history:
Received 5 December 2015
Revised 19 April 2016
Accepted 23 April 2016
Available online xxx

Keywords:
Query routing
Distributed information retrieval
Incremental learning

ABSTRACT

In this paper, we introduce a new collection selection strategy to be operated in search engines with document partitioned indexes. Our method involves the selection of those document partitions that are most likely to deliver the best results to the formulated queries, reducing the number of queries that are submitted to each partition. This method employs learning algorithms that are capable of ranking the partitions, maximizing the probability of recovering documents with high gain. The method operates by building vector representations of each partition on the term space that is spanned by the queries. The proposed method is able to generalize to new queries and elaborate document lists with high precision for queries not considered during the training phase. To update the representations of each partition, our method employs incremental learning strategies. Beginning with an inversion test of the partition lists, we identify queries that contribute with new information and add them to the training phase. The experimental results show that our collection selection method favorably compares with state-of-the-art methods. In addition our method achieves a suitable performance with low parameter sensitivity making it applicable to search engines with hundreds of partitions.

1. Introduction

The Web has significantly expanded forcing search engines to handle document collections with millions of websites and web pages. Search engines use collection partition strategies enabling operations on small subcollections and render their processing feasible for both indexing and retrieving operations. A prevalent trend is to consolidate thematic collections consisting of documents that address specific topics. This trend is known as vertical search and requires adequate collection partitioning strategies.

A search engine can receive an enormous amount of queries on a daily basis. For this reason, one or more machines are exclusively devoted to the task of receiving and processing these queries. These machines are referred to as brokers. Brokers route queries to machines that have documents and local inverted indexes. Two routing methods are employed: the first routing method, known as a broadcast, sends a query to all partitions; the second method, known as a multicast, only sends a query to some partitions. In the latter case, a selective partition routing method is required (query routing). These methods generally employ descriptions of the contents of each partition in order to determine their similarity to the

* Corresponding author. Tel.: +56223037213.
E-mail addresses: marcelo.mendoza@usm.cl, mmendoza@inf.utfsm.cl (M. Mendoza), mauricio.marin@usach.cl (M. Marín), gvcosta@unsl.edu.ar (V. Gil-Costa), flavio.ferraroti@scch.at (F. Ferrarotti).

http://dx.doi.org/10.1016/j.ipm.2016.04.008
0306-4573/

query. Depending on this information, they select the most promising partitions for processing. This query routing method is referred to as collection selection.

Two methods for collection partitioning are available. The first method, known as term or horizontal partitioning, consists of spliting the vocabulary. This strategy enables routing the inverted index of an entire collection and indexing each partition of the vocabulary in different machines. The second strategy is known as document or vertical partitioning and consists of dividing a document collection in disjoint subsets. This strategy by placing each document subset in a machine along a local inverted index. These strategies enable to process queries in multicast mode, which reduces the number of partitions that are required to process each query. This approach results in lower computational costs at the partition level and improvements in a complete system, in terms of query throughput. In this paper, we employ this scenario, assuming a vertical partition of the collection, and applying a collection selection method that is capable of identifying a subset of promising partitions in the broker to process a given query.

We propose a new collection selection method. Our method builds vector representations of each document partition over the query term space. The construction of these representations is performed using a learning algorithm based on logistic regression, which is capable of finding regions in the query space where an information recovery reward function is maximized. As a reward function, we employ the recall of each partition over the exact top-k document list and the normalized discount cumulative gain (NDCG), which considers not only the fraction of relevant and recovered documents but also their ranking. The learning algorithm operates over all collection responses, which are calculated using a broadcast and correspond to the gold standard of the learning process. Beginning with this information, we can generalize to new queries that share terms with the set of queries initially employed to build each partition representation. When detecting new queries, our method is capable of determining whether they should be included in the learning phase. The decision is simply attained by comparing the ranking inversions in the processing list of broadcast versus multicast. Using a statistical criterion, we can determine if both lists were generated using the same conditions or were randomly jointly generated.

Our proposal's main strength is the possibility of providing control to precision versus workload tradeoff. By implementing our method in search engines with vertical partitioning, the method can be employed in high-query traffic conditions without losing response quality.

The main contributions of this study are as follows:

- The construction of representations of each partition's content, without document content: We employ the query terms, which enable us to model the aspect of each partition from the user's point of view. This approach enables us to conduct a learning process over a term space with lower dimensionality and higher representation, reducing the computational costs associated with the training phase of the collection selection method;
- The modeling of the processor lists in the learning process using recall or the NDCG as a reward function for the process: This step enables the incorporation of a reward function, whose optimization is required within the collection selection method; and
- The incorporation of new queries into the learning process using information novelty favoring the elimination of redundancy avoiding overfitting.

This paper is organized as follows. In Section 2, we perform a review of the literature, identify a previous study whose proposal is the closest to ours, and compare our own results with results of the selected study. In Section 3, we describe the necessary background for constructing a vertically partitioned search engine, on which our proposal is constructed. In Section 4, we introduce our learning-based collection selection method. In Section 5, we present our incremental learning strategy. In Section 6, we present experiments using real data. We conclude in Section 7, where we highlight this study's main achievements.

2. Related work

2.1. Data distribution strategies

When document's distribution is done following an architecture partitioned by terms, each processor stores a portion of the global index representing a fraction of the vocabulary. The problem of identifying an appropriate partition for the global index allowing a balanced workload is complex. Moffat, Webber, and Zobel (2006) modelled the term partitioning as a *bin packing* problem. There, each bin represents a partition where each term is an object to be placed into the bin. By using the frequency of each query term registered in a log file, they associate each term with a weight proportional to its frequency and with the length of its respective posting list. Through the use of these weights, a function of the query processing costs is constructed. Then, the bin packing problem is solved by minimizing the cost function. Using a cluster of eight processors, the authors show that it is possible to improve query throughput by up to 30% when compared to a centralized system.

Lucchese, Orlando, Perego, and Silvestri (2007) also modelled the term partitioning as a bin packing problem. Unlike the work of Moffat et al., they seek to model queries instead of isolated terms in each of the bins. To do this, they analyzed a query log determining the term co-occurrence in queries adding this variable to the cost function. After minimizing the cost function, they can determine good solutions for partitioning the global index, improving results based on isolated terms.

Cambazoglu, Kayaaslan, Jonassen, and Aykanat (2013) proposed a term-based index partitioned scheme by using hypergraph partitioning. The goal is to minimize the communication overhead while achieving good computational load balance

Please cite this article as: M. Mendoza et al., Reducing hardware hit by queries in web search engines, Information Processing and Management (2016), http://dx.doi.org/10.1016/j.ipm.2016.04.008