Chapter 19
Development of a Novel Compressed Index–Query Web Search Engine Model

Hussein Al-Bahadili
Petra University, Jordan

Saif Al-Saab
University of Banking & Financial Sciences, Jordan

ABSTRACT

In this paper, the authors present a description of a new Web search engine model, the compressed index-query (CIQ) Web search engine model. This model incorporates two bit-level compression layers implemented at the back-end processor (server) side, one layer resides after the indexer acting as a second compression layer to generate a double compressed index (index compressor), and the second layer resides after the query parser for query compression (query compressor) to enable bit-level compressed index-query search. The data compression algorithm used in this model is the Hamming codes-based data compression (HCDC) algorithm, which is an asymmetric, lossless, bit-level algorithm permits CIQ search. The different components of the new Web model are implemented in a prototype CIQ test tool (CIQTT), which is used as a test bench to validate the accuracy and integrity of the retrieved data and evaluate the performance of the proposed model. The test results demonstrate that the proposed CIQ model reduces disk space requirements and searching time by more than 24%, and attains a 100% agreement when compared with an uncompressed model.

INTRODUCTION

A Web search engine is an information retrieval system designed to help finding information stored on the Web (Levene, 2005). It allows us to search the Web storage media for a certain content in a form of text meeting specific criteria (typically those containing a given word or phrase) and retrieving a list of files that match those criteria (Brin & Page, 1998). Web search engine consists of three main components: Web crawler, document analyzer and indexer, and search processor (Calishain, 2004).

DOI: 10.4018/978-1-4666-2157-2.ch019

Due to the rapid growth in the size of the Web, Web search engines face enormous performance challenges, in terms of: storage requirement, data retrieval rate, query processing time, and communication overhead. Large search engines, in particular, have to be able to process tens of thousands of queries per second on tens of billions of documents, making query throughput a critical issue (Fagni, Perego, & Silvestri, 2006). To satisfy this heavy workload, Web search engines use a variety of performance optimizations including succinct data structure (Ferragina et al., 2005; Gonzalez & Navarro, 2006), compressed text indexing (Ferragina & Manzini, 2006; Ferragina et al., 2009), query optimization (Chen, Gehrke, & Korn, 2001; Long & Suel, 2003), high-speed processing and communication systems (Badue et al., 2002), and efficient search engine architectural design (Zobel & Moffat, 2006).

Compressed text indexing has become a popular alternative to cope with the problem of giving indexed access to large text collections without using up too much space. Reducing space is important because it gives the chance of maintaining the whole collection of data in main memory. The current trend in compressed indexing is full-text compressed self-indexes (Ferragina et al., 2007). Such a self-index replaces the text by providing fast access to arbitrary text substrings, and, in addition, gives indexed access to the text by supporting fast search for the occurrences of arbitrary patterns. It is believed that the performance of current search engine models that base on compressed text indexing techniques only, is still short from meeting users and applications needs.

In this work, we present a description of a novel Web search engine model that utilizes the concept of compressed index-query search; therefore, it is referred to as the CIQ Web search engine model. The new model incorporates two bit-level compression layers implemented at the server side, one after the indexer acting as a second compression layer to generate a double compressed index (index compressor), and the other one after the query parser for query compression (query compressor) to enable bit-level compressed index-query search. The main features of the new model are it requires less index storage requirement and I/O overheads, which result in cost reduction and higher data retrieval rate or performance. Furthermore, the compression layers can be used to compress the any index regardless of indexing technique.

The data compression algorithm that will be used in this model is the novel Hamming codes-based data compression (HCDC) algorithm (Al-Bahadili, 2008), which is a lossless bit-level data compression algorithm. The main reason for using this algorithm is that its internal structure allows compressed data search. Moreover, recent investigations on using this algorithm for text compression showed that the algorithm can provide an excellent performance in comparison with many widely-used data compression algorithms and state-of-the-art tools (Al-Bahadili & Rababa'a, 2010).

The different components of the new Web model are implemented and integrated to build a prototype CIQ Web search engine, which also used as a test bench to validate the accuracy and evaluate and compare the performance of the CIQ model, namely, the CIQ test tool (CIQTT) (Al-Saab, 2011). The CIQTT was used to collect a test corpus of 104000 documents from 30 well-known Websites; process and analyze the test corpus; generate five inverted indexes of different sizes (1000, 10000, 25000, 50000, and 75000 documents), compress the indexes and measure the compression ratio and the storage reduction factor; search the indexes for 29 different keywords in both compressed and uncompressed forms; and finally, compare the outcomes of the different search processes and estimate the speedup factor and the time reduction factor.

The test results demonstrate that the HCDC algorithm achieves a compression ratio of more than 1.3, which reduces the storage requirement by more than 24%. The searching processes can

Studying and Analysis of a Vertical Web Page Classifier Based on Continuous Learning Naïve Bayes (CLNB) Algorithm
H. A. Ali, Ali I.El Desoukyand Ahmed I. Saleh (2007). *International Journal of Information Technology and Web Engineering (pp. 1-44).*
www.irma-international.org/article/studying-analysis-vertical-web-page/2625