Simulation Modelling Practice and Theory 22 (2012) 74–91

Contents lists available at SciVerse ScienceDirect

Simulation Modelling Practice and Theory

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

Improved retrieval effectiveness by efficient combination of term proximity and zone scoring: A simulation-based evaluation

Leonidas Akritidis, Dimitrios Katsaros *, Panayiotis Bozanis

Department of Computer and Communication Engineering, University of Thessaly, Greece

ARTICLE INFO

Article history:
Received 30 August 2011
Received in revised form 2 December 2011
Accepted 4 December 2011
Available online 30 December 2011

Keywords:
Web
Search engines
Inverted index
Simulation
Evaluation

ABSTRACT

During the past few years, the commercial Web search engines have augmented their underlying index structures by significantly enriching the information which describes the appearance of a word within a document Dean (2009) [7]. This enriched information is now used in complex and effective functions which rank documents by taking into consideration hundreds of features, with respect to a user query. Despite the evolution of the search engines, the past research has mainly concentrated on improving plain Web indexes storing typical data only. In this work we study the problem of organizing an inverted index storing additional information. In particular, we examine how the physical locations of a document, called zones, can be efficiently integrated with such an index structure. We introduce TZP, an encoder which compresses these zones in combination to the positions of a word in a document, by employing a fixed number of bits for each portion of a word's inverted list. We demonstrate that our method allows direct access to the compressed zones and positions without expensive look-ups, avoids decoding any unnecessary information, while its overall index size is analogous or even better when compared against state-of-the art schemes. Moreover, we examine how the word positions can be combined to the zones to improve retrieval effectiveness. We introduce BM25TOPF, a scheme which incorporates term proximity and zone weighting into a single ranking formula. Unlike other term proximity approaches, BM25TOPF also takes into account the ordering of the query terms by rewarding the documents containing them in the correct order. Our experiments with the Web Adhoc Task of TREC 2009 and a set of own queries show that BM25TOPF outperforms the current state-of-the-art approaches by a margin between 6% and 11%.

© 2011 Elsevier B.V. All rights reserved.

1. Introduction

Nowadays, the repositories of the major search engines consist of tens of billions of documents Dean [7] and as the Web becomes larger and the crawling technology evolves, these repositories are expected to grow further. Furthermore, search engines accept and answer thousands of queries per second attempting to quickly retrieve the most suitable documents for each submitted query. In such a dynamic environment where the available information, the workload and the user expectations expand, search engines have to constantly scale up in terms of both efficiency (query throughput) and effectiveness (quality of query results).

The inverted index is the primary data structure used by search engines for storing document-related data and metadata. According to [27,31], an appropriately constructed inverted index can improve the performance of query processing

* Corresponding author. Tel.: +30 2421074975.
E-mail address: dkatsar@inf.uth.gr (D. Katsaros).

1569-190X/$ - see front matter © 2011 Elsevier B.V. All rights reserved.
doi:10.1016/j.simpat.2011.12.002

dramatically. Due to the importance of the inverted index in the overall efficiency of a search system, there has been a lot of research conducted towards its optimization. Optimization primarily regards two critical issues: *compression* and *organization*. The former is a key issue for reducing the overall index size and minimizing the transfer costs from either disk or main memory. The latter enables partial access of the index structure, that is, a query can be answered without having to traverse all the available information stored in it.

Several engineers (see for instance Dean [7]) have revealed that the information stored in the inverted index search engines has tripled during the past few years. However, in the literature we mainly encounter strategies and algorithms concerning typical inverted indexes, which almost always store very limited data: document identifiers, word-document frequencies and word positions in a document. Compared to the hundreds of the parameters employed by the major search engines for ranking their documents [28,24], this data is apparently inadequate.

In this work we study the potential of including additional information within the inverted index. In particular, we adopt the idea of partitioning a Web document into locations of special interest, namely *zones*. The document zones were introduced in Manning et al. [12], but to the best of our knowledge, issues regarding the compression and organization of such indexes have never been studied before. In this paper we investigate the meaning of a word's appearance within a document; we replace the plain positional data by the occurrences, a piece of information which contains both the position and the zone of the document where this specific word appears.

In the sequel, we propose a method which allows compact storage of zones along with the corresponding word positions. Our approach, namely TZP, operates in combination with the *block-based* inverted list organization, a strategy introduced by Moffat and Zobel [14] which splits an inverted list into blocks. Block-based schemes allow us to skip large, unneeded portions of the index during query processing. TZP is designed to support all the partitioning strategies that have been proposed so far (refer to [15,3,1,30,6]), and operates in two steps: In the first step, the compressor packs each position-zone pair of a block into a 32-bit space and in the next phase, these packets are encoded together by employing a fixed number of bits. This scheme enables the direct access of the occurrence data for each posting, by using a limited number of pointers (one pointer per block).

Finally, one of our main motives was to examine whether the usage of the additional information can really lead to search results of higher quality. Discovering a ranking function which combines many different parameters (i.e. frequencies, term proximity, zone weighting, document lengths etc) is a challenging task. In this paper we initially examine some state-of-the-art probabilistic retrieval functions, such as BM25 (firstly introduced Robertson and Jones [18]), a variant which takes into consideration the zone of the document where a term appears (BM25F, Lu et al. [11]), and another variant which takes into consideration term proximity, namely BM25TP Buttcher et al. [5]. In the sequel, we propose an enhancement to the BM25TP, *BM25TOP*, which takes into consideration both term proximity and correct term ordering (that is, whether the terms in a document appear in the same order as in the query). Finally, we inject the concepts of BM25TOP to BM25F to produce *BM25TOPF*, a ranking function which is sensitive to term proximity, correct term ordering and zone weighting.

As a summary, the contributions of this paper are:

- We introduce a two-level encoder for positions and zones, namely TZP. TZP compresses the data by using a fixed number of bits for each position-zone pair of the same block of an inverted list.
- We show that the fixed-bit policy adopted by TZP allows very fast decompression, whereas it also enables us to access directly only the data actually required for processing a query. In other words, with TZP we are not obliged to look-up for the positional and zone data of a particular posting since we are allowed to compute their location, and moreover, we do not have to decode any unnecessary information.
- We investigate the usefulness of combining term proximity and zone weighting for document ranking. In particular, we first propose BM25TOP, an enhancement to the original BM25TP function which is sensitive to the correct term ordering. We also introduce BM25TOPF, a ranking method that allows term proximity, correct document ordering and zone weighting to be combined into a single scoring formula.
- All our contributions are experimentally evaluated by using the Clueweb09-T09B document collection consisting of roughly 50 million English documents.

The rest of the paper is organized as follows: In Section 2 we examine the state-of-the-art methodologies for organizing inverted indexes and we cite the relevant work. Section 3 contains the description of zones and consists of three Subsections: SubSection 3.1 discusses the new form of postings after the inclusion of zones, SubSection 3.2 introduces the TZP compression algorithm for zones and positions, and SubSection 3.3 demonstrates how the data encoded by TZP can be efficiently accessed and decompressed. In Section 4 we provide descriptions of some popular ranking functions and we propose our own scoring approaches. Finally, Section 5 contains the experimental evaluation of our methods and propositions, whereas in Section 6 we finalize this work by stating our conclusions.

2. Preliminaries and related work

The inverted index is the primary data structure constructed and maintained by the search engines to serve user queries. There is a significant amount of research regarding the efficient organization of these indexes and in this Section we briefly describe some basic elements deriving from the related theory.

Download English Version:

https://daneshyari.com/en/article/492252

Download Persian Version:

https://daneshyari.com/article/492252

Daneshyari.com