Chapter 6
New FastPFOR for Inverted File Compression

V. Glory
Periyar University, India

S. Domnic
National Institute of Technology Tiruchirappalli, India

ABSTRACT

Inverted index is used in most Information Retrieval Systems (IRS) to achieve the fast query response time. In inverted index, compression schemes are used to improve the efficiency of IRS. In this chapter, the authors study and analyze various compression techniques that are used for indexing. They also present a new compression technique that is based on FastPFOR called New FastPFOR. The storage structure and the integers' representation of the proposed method can improve its performances both in compression and decompression. The study on existing works shows that the recent research works provide good results either in compression or in decoding, but not in both. Hence, their decompression performance is not fair. To achieve better performance in decompression, the authors propose New FastPFOR in this chapter. To evaluate the performance of the proposed method, they experiment with TREC collections. The results show that the proposed method could achieve better decompression performance than the existing techniques.

INTRODUCTION

Information Retrieval System (IRS) is receiving substantial attention due to the exponential increase of the quantity of information available in recent years. Digital Library, Search engines, E-commerce and Electronic news are the some of the applications of the information retrieval system (Kobayashi & Takeda, 2000; Williams & Zobel, 2002). The main objective of IRS is to provide the maximum efficiency (speed) and effectiveness (relevance) with proper balance between them. Particularly, IR effectiveness deals with retrieving the most relevant information to a user's need, while IR efficiency deals with providing fast and ordered access to huge amounts of information. Indexing is one of the efficient ways to improve the fast retrieval in IRS. Compared to the Signature file (Faloutsos,1985), Bitmaps (Chan &

DOI: 10.4018/978-1-5225-3004-6.ch006

Ioannidis, 1998) and Pat Tree (Morrison, 1968), the inverted index is the most suitable indexing structure to locate the data quickly, offers quick response time and supports the various searching techniques (Zobel & Moffat, 2006).

An inverted index contains two parts: lexicon file (dictionary) which stores a distinct list of terms found in the collection and document frequency (total number of documents in which term appears). For each term, an inverted list (posting list) is maintained and it contains a sequence of document identifiers (id), term frequency (tf) (number times the particular term appears) and positions. In each inverted list, the increasing order of document identifiers is replaced by D-Gap (difference between the document identifiers except the first one to enable efficient compression). The compression of inverted index is essential because it is potentially taking the less storage space and gives faster query performance to improve the efficiency of IRS. Compression techniques are classified into two categories such as integer compression and integer list compression techniques. Each integer is processed individually in integer compression whereas the group of integers are processed in integer list compression. Unary code (UC) (Salomon, 2004), Golomb code (GC) (Golomb, 1966), Rice code (RC) (Rice, 1979), Elias Gamma code (EC) (Elias, 1975), Elias Delta code (DC) (Elias, 1975), Variable Byte code (VBC) (Salomon, 2007), Fast Extended Golomb code (FEGC) (Domnic & Glory, 2012) and Re-ordered Fast Extended Golomb code (RFEGC) (Glory & Domnic, 2013) are the some of the integer compression techniques. VBC is faster than GC, RC, EC and DC. Compared to VBC, RFEGC gives better compression and decompression when the occurrence of middle and large range of integers are more in the data (Glory & Domnic, 2013). Some of the integer list compression techniques are Interpolative Code (Moffat & Stuiver, 2000), Simple Family (Anh & Moffat, 2005), Frame-Of-Reference (FOR) (Goldstein, Ramakrishnan & Shaft, 1998; Ng, & Ravishankar, 1997) and Patched coding techniques (PFORDelta, NewPFD, OptPFD and FastPFOR) (Zukowski et al., 2006; Yan, Ding & Suel, 2009; Lemire & Boytsov, 2015). Interpolative code is slower than GC (Anh & Moffat, 2005; Yan et al., 2009). Simple Family coding is generally slower and it is slightly better in compression. FOR and Patched coding techniques are the faster decoding coder in the recent years (Zukowski et al., 2006; Yan et al., 2009; Lemire & Boytsov, 2015). Depending on the range of integers, sometimes FOR gives the poor compression. FastPFOR is one of the recent patched coding techniques, which gives the better compression rate and fast decoding performance compared to other patched techniques. But the decompression performance (disk access time + decoding time) of FastPFOR is not fair.

In this paper, we propose a new patched coding technique called New FastPFOR to achieve the better compression and decompression performances. New FastPFOR is based on FastPFOR technique, but it uses new cost formula to determine optimal b value. In the proposed scheme, the positional values of the exceptions, the number of exceptions and maximum bit width are together represented by unary code, which leads the better result than FastPFOR. In our work, we have used TREC dataset to evaluate the performance of proposed method and other existing methods.

The rest of the paper is organized as the review of some of the compression techniques, the proposed method, the experimental results and conclusions are derived.

COMPRESSION TECHNIQUES

Some of the compression techniques which have been used for inverted list compression are discussed in this section