Information Systems 49 (2015) 25–39

Contents lists available at ScienceDirect

Information Systems

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

Inverted index maintenance strategy for flashSSDs: Revitalization of in-place index update strategy

Wonmook Jung [a,1], Hongchan Roh [a,b,1], Mincheol Shin [a], Sanghyun Park [a,*]

[a] Department of Computer Science, Yonsei University, South Korea
[b] ICT R&D Division, SK Telecom, South Korea

ARTICLE INFO

Article history:
Received 8 May 2013
Received in revised form
10 July 2014
Accepted 7 November 2014
Available online 15 November 2014

Keywords:
FlashSSDs
Inverted index
Index maintenance
Information retrieval
Search engines

ABSTRACT

An inverted index is a core data structure of Information Retrieval systems, especially in search engines. Since the search environments have become more dynamic, many on-line index maintenance strategies have been proposed. Previous strategies were designed for HDDs. Consequently, in order to avoid expensive random access cost, Merge-based strategies have been preferred to In-place index update strategies on HDDs. However, flashSSDs have become solid alternatives to HDDs. FlashSSDs currently are adopted in a wide range of areas due to their superior features such as the short access latency, energy efficiency, and high bandwidth. In this article, we first reexamined potentials of In-place index update strategies on flashSSDs. Thanks to the insignificant access latency of flashSSDs, we discovered that In-place index update strategies outperform Merge-based strategies, since In-place index update strategies generate much less amount of I/O than Merge-based strategies despite inducing frequent random accesses. Based on this discovery, we suggest a new inverted index maintenance strategy based on an In-place index update strategy for flashSSDs, called Multipath Flash In-place Strategy (MFIS). To enhance the index maintenance performance, MFIS stores the posting list of each term non-contiguously and exploits the internal parallelism of flashSSDs. Thus, MFIS not only induces the minimum amount of I/O but also utilizes the maximum bandwidth of flashSSDs. Furthermore, MFIS is designed to show high query processing performance by utilizing the internal parallelism of flashSSDs even though the posting list of each term is stored non-contiguously. In our experiments, the index maintenance performance of MFIS was considerably better than other previous maintenance strategies. The index maintenance performance was up to 14.93, 4.04, 5.12, and 2.33 times higher than Merge-based strategies such as Immediate Merge, Geometric Partitioning, Hybrid, and SSD-aware Hybrid, respectively. The query processing performance of MFIS was up to 1.62 times higher than non-contiguous In-place. In addition, MFIS showed almost the best query processing performance as Merge-based strategies did. In conclusion, MFIS is the best on-line inverted index maintenance strategy on flashSSDs in terms of both index maintenance and query processing performance.

1. Introduction

In Information Retrieval (IR) systems, an inverted index is used as a core data structure. It consists of a list of terms (vocabulary) and a list of postings for each term (posting list). Each posting contains term's information such as document ID and term frequency in the document. Many studies have focused on maintaining the inverted index. It is

* Corresponding author. Tel.: +82 2 2123 5714; fax: +82 2 365 2579.
E-mail addresses: jngwnmk@cs.yonsei.ac.kr (W. Jung), hongchan.roh@sk.com (H. Roh), smanioso@cs.yonsei.ac.kr (M. Shin), sanghyun@cs.yonsei.ac.kr (S. Park).
[1] These authors have contributed equally to this paper.

http://dx.doi.org/10.1016/j.is.2014.11.004
0306-4379/