Document Retrieval using Efficient Indexing Techniques: A Review

Shweta Gupta, Department of Computer Science and Engineering, Ajay Kumar Garg Engineering College, Ghaziabad, India & Dr. A.P.J. Abdul Kalam Technical University, Uttar Pradesh, India
Sunita Yadav, Department of Computer Science and Engineering, Ajay Kumar Garg Engineering College, Ghaziabad, India & Dr. A.P.J. Abdul Kalam Technical University, Uttar Pradesh, India
Rajesh Prasad, Department of Computer Science, Yobe State University, Damaturu, Nigeria

ABSTRACT

Document retrieval plays a crucial role in retrieving relevant documents. Relevancy depends upon the occurrences of query keywords in a document. Several documents include a similar key terms and hence they need to be indexed. Most of the indexing techniques are either based on inverted index or full-text index. Inverted index create lists and support word-based pattern queries. While full-text index handle queries comprise of any sequence of characters rather than just words. Problems arise when text cannot be separated as words in some western languages. Also, there are difficulties in space used by compressed versions of full-text indexes. Recently, one of the unique data structure called wavelet tree has been popular in the text compression and indexing. It indexes words or characters of the text documents and help in retrieving top ranked documents more efficiently. This paper presents a review on most recent efficient indexing techniques used in document retrieval.

KEYWORDS

Document Retrieval, Information Retrieval, Inverted Index and Full Text Index, Text Indexing

1. INTRODUCTION

With the day-to-day advancement in every domain, the volume of textual information is expanding rapidly. Internet has become an important source of information. User perform web search to obtain solution for their queries. Web searching is becoming more complex day-by-day because of availability of billions of web pages. This is the reason that why finding relevant information for a user query is still a big problem (Croft et al., 2010). Since search engine first extract relevant web pages and then return the Search Engine Result Pages (SERP’s), therefore it plays an important role between User and World Wide Web (WWW). Initially, Crawler browses and stores the relevant web pages for the corresponding query. It further performs specific tasks such as indexing on the web. Also, a collection of seed URL’s is the basis for the searching (Olston & Najork, 2010). Therefore, pages are fetched and accessed from the seed URL’s and become SERP’s. Here crawling, indexing, query and ranking are the web entities as shown in Figure 1. These web entities act as key components for text retrieval. It is difficult to analyze whether user is satisfied by the top result or not? Due to excess amount of data availability, it is really tough for a search engine to retrieve relevant information via web entities that satisfy user and also loyal to a relevant top result (Gurumurthy et al., 2010).

DOI: 10.4018/IJBAN.2016100104

Copyright © 2016, IGI Global. Copying or distributing in print or electronic forms without written permission of IGI Global is prohibited.
Initially a query pattern is given by user for which results are retrieved and ranked (Jansen et al., 2010). Query text can be in the terms of English words or some biological sequences like DNA sequence or a protein code words etc. It is assumed that retrieved web pages contain content of similar keywords of user query. Here, the relevancy is achieved on the basis of most occurrences of query terms (Jansen et al., 2011). Crawling plays a decisive role in the web indexing as it view websites and download web pages. Then, these web pages are indexed further based on their relevancy. After indexing, resultant web pages are displayed to the user via SERP’s on web browser. For word-based query, inverted index has been most popular indexing technique while for alphabet-based queries; full text indexing (or self indexing) is efficient for web pages. Inverted index creates lists and support word-based pattern queries effectively. In order to reduce space requirement of lists, word-based compression techniques have been applied further on natural language texts. Text results may include web pages, articles of journals or newspapers clippings, advertisements, images, videos, Wikipedia’s, ancient historical data, and biological data (Berry & Browne, 2005). Further, search engine perform ranking for their SERP’s.

Therefore, if user finds a result relevant, user click on the corresponding link and retrieve the required information (Shin et al., 2012). Here, search engine acts as an intermediate system that satisfies user requirement using information system. It requires at least two or three words to understand the subject of query which is used to find accurate and valid information quickly. Although information can be available through an audio/video file/animated images but the major part of searching is still through text. To approximate relevancy of results retrieved by a search engine, is the biggest challenge faced by users (Fatima et al., 2014). Therefore, online searching becomes a key focus to retrieve desired text in the field of information retrieval, where these web pages are referred as documents.

Generally, information required by user expressed as sequence of words via search engine. In order to evaluate user query fast, inverted index prepare lists to match word-based query from the documents given in database, where, inverted lists are represented as (d, fd, t) where d is document and f is frequency of words (t) in d. Then, ranking the documents on the basis of relevancy score, assigned to the words of inverted list. This classic approach has several good characteristics over word-based search. In case of repetition of words in a document, it maintains the list by keeping repetitive word as one record. While another data structure i.e. full text index lists the positions of all suffixes of the text which unfortunately increase the size of index as compared to inverted index. Here, suffix array can be used as full text index to allow substring search. Though inverted index cannot handle other language query in which words cannot be separated as in English language. In that case, full text

Figure 1. Text retrieval via search engine
How Business Intelligence Creates Value: An Empirical Investigation
[www.igi-global.com/article/how-business-intelligence-creates-value/83476?camid=4v1a](www.igi-global.com/article/how-business-intelligence-creates-value/83476?camid=4v1a)