ISSN: 2393-8528

IJICSE
Open Access

Contents lists available at www.ijicse.in

International Journal of Innovative Computer Science & Engineering

Volume 4 Issue 1; January-February-2017; Page No. 05-09

PDF Crawler using Inverted Index and Interval lists

Snehal S. Kadwe[1], Prof.Shrikant Ardhapurkar[2]

[1]Student, CSE Department, YCCE Nagpur, Maharashtra, India

snehalkadwe@gmail.com

[2]Professor of Computer Science & Engineering, YCCE Nagpur,

Maharashtra, India

shrikant.999@gmail.com

Received 27 Dec. 2016; Accepted 22 Jan. 2017

ABSTRACT

The search operation in PDF document has become very indispensable now a days and loads of research have being organized to store and process the index required for search operation in a very simple and effective manner. Whenever indexes are stored, its access time is large and it requires large amount of storage space. The above techniques have some limitation like it can be done only for small number of PDF documents. To increase the access time and to reduce the storage space we are using the concept of inverted index and interval list. With the help of inverted index of a keyword available in PDF it can easily retrieve the PDF document. It can assign unique id to each and every document (docID) available in repository. Interval list is used for lower bound and upper bound of document present in repository. The inverted index and interval list make it easy to retrieve information of PDF document with the help of keyword. The combination of both can improve the information retrieval system (IR) and it allows us to search millions of PDF document.

Index terms: keyword search, key-phrase search, inverted index, interval list.

INTRODUCTION

The today's web contains a lot of information and it is keeps on increasing every-day. There are availability of abundant data on web, to search for particular data and information in this collection it has become very difficult task [16]. A focused web crawler traverses the web for selecting relevant pages of predefined information, data which is available in the form of text file, links, PDF or on blogs over the World Wide Web. Although only relevant pages are to be considered for any search query but still huge data needs to be explored for example if we want to search the information or data which present only in pdf document, but the web crawler will show all other links related to that particular information or data for that we need a crawler such that it will crawl the web only for PDF documents. So our survey is to build a PDF crawler such that it will search only for PDF document which is very useful in future. Web crawler has some drawback like its traffic may blocked. It has limits in processing of pages and storage. While survey on PDF crawler we focused on the drawback of web crawler and try to overcome it with the help of inverted index and interval list as well as try to improve the quality of search result by optimizing its search result. The main aim of this research project is to determine to what amount the visibility of these PDF documents can be improved.

USE OF INVERTED INDEX

Inverted index have long been the standard data structure underlying the implementation of information system (IR). Other data structures have been proposed, generally with the intention of providing more efficient support for specific retrieval operations. However, none of these data structures provide the flexibility and generality of inverted index, and they have mostly fallen out of use. Pre-processing is used extract the keyword from document and it extract the keyword with the help of term frequency (TF).

OVERVIEW

Inverted index and interval list for keyword search: Databases or repository can store

*Corresponding author: Snehal S. Kadwe | E-mail: snehalkadwe@gmail.com

numerous numbers of documents and pdf with a unique index id for user reference and those pdf's and documents are arranged in specific manner so that they can be easily available and accessible. When user has to access those pdf documents then it provides efficient search and data retrieval operations. Search system in inverted index usually supports union and intersection operations. UNION operation provides functionalities for OR query whereas interception provides the functionalities for AND query. This process of searching involves high processing and storage overhead. The processing of search operation on inverted index an algorithm requires searching entire list of indices while algorithm on interval list requires only lower and upper bound of intervals. The keyword which has highest term frequency count can be placed on the top of the search result and the remaining keywords are arranged in descending [2] order. The lists in inverted indexes can be very long for large datasets, and many existing approaches paying more attention to how to compress them to provide efficient search result. The observation it provides that there are many consecutive IDs on the inverted lists and the size of the whole inverted index can be reduced by merging these groups of consecutive IDs into intervals, hence each interval is denoted by (r), can be represented by only two numbers (the lower and upper bounds, denoted by lb.(r)/ and ub.(r)/). For example, the ID list of databases in the sample dataset can be converted into an interval list.

Search performance:

Performance of keyword search algorithm is compared by synthetic quires. For example: each dataset of document had 9 query workloads, each containing 1000 k-word queries, where k D 2, 3, _ _ _, 10. The keyword in each query are drawn according to the word's term frequencies occurred in each document, in other words if keywords occurs more frequently in document then it is drawn as query keyword. The synthetic query uses memory-based algorithm to draw a search result. The memory based algorithm has their indexes in main memory and it requires large amount of memory for calculation of keyword occurrences in each document and huge amount of storage to store the keyword search result according to term frequency of keyword present in each document within a dataset. The InvIndex algorithm and memory based algorithm requires much more time for keyword search operation in each documents.

LITERATURE SURVEY

This section presents the detailed review about how keyword are extracted using inverted index and how it is used in search engine, various algorithms used for keyword and key-phrases search.

Ian H. Witten et al [1] in 2016 It defines that the Key-phrase provide semantic metadata which summarize and characterize the documents. Kea, is an algorithm which automatically extracts the key-phrases from text. Kea identifies candidate key-phrases using lexical methods and calculates feature values for each candidate, which uses a machine-learning algorithm to predict which candidates are good key-phrases. The machine learning scheme first builds a prediction model using training documents with known key-phrases, after that it will use the model to find key-phrases in new documents. We use a large test set of documents to evaluate. The results show that Kea has an average match between one and two of the five key-phrases chosen by the client. Hence it is considered as good performance. Although Kea find less than half the client's phrases, it must choose from many thousands of candidates, also, it is highly unlikely that even another human like select the set same of phrases as the as author. Therefore, it leaves the author's phrases behind and evaluates Kea's phrases with a more robust measure. We will use human judges to rate how well a set of extracted key-phrases summarize a particular document. Although this experiment will provide a more realistic assessment, it is clear that some of the Kea's phrases are very poor regardless to the measure. The poor phrases are not easy to weed out: the reason that *garbage* is a poor keyword (see Table 1) is subtle from a computational view. Therefore, it will also investigate techniques to determine what makes a phrase reasonable from a human perspective. At present, Kea's performance is sufficient for the applications it is designed for provide support for summarizing, searching, browsing and clustering in cases where manual key-phrase assignment is infeasible. It can and will greatly assist designers and users of large document collections. Kea's effectiveness in terms of how many author-assigned key-phrases are correctly identified. The system is simple, robust, and publicly available.

Giridharan J et al [2] Search operations have become quite indispensable in recent days and loads of research are being organized to store and

process the indices required for search operations in a simple and effective in manner. It introduces a new and effective way to store and a process index, namely using interval lists which drastically reduces the storage space and improves the access time. It will then introduce basic search algorithms that use indices stored as interval lists. It uses basic search operations like Union and Intersection operations on inverted indices and interval lists. This algorithm tries to improve the efficiency of interval lists while it is fast scalable method to enhance the search speed of interval lists by reordering documents in the datasets.

Qiuying Bai et al [3] Indexes are the kernels of search engines. This paper presents a Combination Inverted Index (Cn), which is a new inverted index. Cn contains three components: appendix inverted index prime inverted index and deleted file list. The addition of appendix inverted index and deleted file list construct a new methodology of creating and updating the index. Compared to traditional inverted index, cn updates indexes promptly and is appropriate for subject-oriented search engines. The performance of the search engine is proved by experiments. CII inverted index shows the superiority in retrieval time, while the retrieval documents of two indexes are the same. Hence CII inverted index provides the same recall and higher retrieval efficiency.

Milos Ilic et al [4] in 2014 Data mining has its origins in various disciplines. Two most important data mining disciplines are statistics and machine this is techniques provide faster and better search for large amounts of data. Inverted index is structure that can be used in data mining process. That is it is used for sorting a list of keywords, with the list of corresponding documents attached to each keyword. Authors create an application that use inverted index structure. Application uses open source library named Lucene. Inverted index and search that use inverted index structure are very useful in the field of data mining and information retrieval. Fast search in text or web documents is the key in massive database. This application present use of inverted index for text documents and gives practical knowledge. This application can be practically used in search process in any kind of documents stored on user's computer. Logical continuation of this application is web documents indexing in which web crawlers must be used, and comparing the performance for text and web documents.

Hao Wu et al [5] in 2013 Keyword search has become a ubiquitous method for users to access text data in the face of information explosion. Inverted lists are usually used to index underlying documents to retrieve documents according to a set of keywords efficiently. Inverted lists are usually large, many compression techniques is used to reduce the storage space and disk I/O time. These techniques usually perform decompression operations which increase the CPU time. It presents a more efficient index structure like Generalized Inverted Index (Ginix), it merge consecutive IDs in inverted lists into intervals to save storage space. With this index structure, more efficient algorithms can be devised to perform basic keyword search operations, i.e., the union operation and the intersection operations take the advantage of intervals. Specifically, these algorithms do not require conversions from interval lists back to ID lists. The Ginix performance is also improved by reordering the documents in datasets using two scalable algorithms. Ginix requires less amount of storage space, as well as it improves the keyword search performance, compared with traditional inverted indexes.

Adriana Szekeres et al [6] in Peer-to-Peer (P2P) networks are largely used for file-sharing and hence must provide efficient mechanisms for searching the files stored at various nodes. The existing structured of Peer-to-Peer support only "exact-match" look-up which is hardly sufficient in a file-sharing network. It addresses the problem of keyword-based search in structured P2P networks. A new keyword-based searching algorithm which can be implemented on top of any structured Peer-to-Peer overlay. The proposed algorithm used to achieve very good searching results as it requires the minimum number of messages are sent to find all the references to the files containing at-least the given set of keywords.

Yong Zhang et al [7] Keyword search for smallest lowest common ancestors (SLCA) is important to identify interesting data nodes in XML documents. It uses parallelization while searching the keyword from XML documents and split it into several parts to be processed parallel.

MeliusWeideman [8] in 2010 Digital library users might not enter a digital library with help of homepage menus. Thus digital library owners should consider the visibility of search engines of stored PDF documents. The aim of this research

project was to determine to what extent the visibility of these PDF documents can be improved. In a series of experiments, the 100'of PDF documents are stored on the digital libraries were it is inspected and identified. The current visibility of these documents can be calculated. After submission to the Google, a waiting period was then allowed for crawler visitation and the searches are repeated. The visibility can be only improved marginally and this can be proved with the help of these experiments. It provides text extracts of PDF documents, to enhance the overall visibility of content. The 100 PDF document's ranking improved very little over time but the PDF format is not enough to guarantee visibility; providing a clear menu structure from the homepage does not necessarily add to crawler visibility; even manual submission of each document to a crawler (as the only method of visibility improvement) is not acceptable. Possible actions could include the creation of text-based webpages containing the titles and abstracts of the PDF documents. These pages should be stored on the same website. Secondly, many in-links from the home- and other pages of the website should be set up to point to these new pages. Finally, these "new" webpages should be manually submitted to search engines which enhance the visibility of some of these PDFs, as indicated in the results

B.M. ThosiniKumarika et al [9] in 2014 It introduce two practices for evaluating the effectiveness of key phrase extraction system. 1. Human judgement. 2. Another method is less costly which will measure how well system generated key phrases are. Key phrases are the phrases consisting of one or more than one significant words. It incorporated the search results as subject of metadata to facilitate the information searches on the web. The final outcome of this is a smart content bookmarking tool which bookmarks only specific textual contents of web page which is relevant to each web user's interest.

Haitao Wu et al [11] The XML becomes a de-factorized standard for presenting and exchanging information. The keyword search has become the focus of information retrieval and effective in identifying return information and it uses efficient algorithm for returning meaningful node, based on the specific data indexes. Inverted index also used to search the keyword from XML documents.

Aliya Nugumanova et al [12] Automatic Keywords Extraction Using Chi-square Test. The aim of automatic keyword extraction is to develop the algorithm for method based on distributed computing model. So it describes the implementation of the algorithm based on the MapReduce model of distributed computing and present the results of experiments showing the benefits of distributed computing. MapReduce computation model is an effective tool for parallel computing in large data arrays. The model is intuitive and allows us to effectively use computing resources and also reduce the complexity (and hence time) of algorithm design.

Aviral Nigam [13] "Web Crawler" uses various kinds of algorithms to search the document on web. The algorithm used to search the document on web it uses various kinds of heuristic functions which increase efficiency of the web crawlers. A* and Adaptive A* are the search algorithm used in web crawler these are best at path finding of document where it is stored. Best First Search and A* search show nearly equal search time and can be improved using heuristic functions. A numbers of users try to search same type of content repeatedly using Adaptive A* search which will prove efficient as it stores the history of previous searches and with every search, the efficiency of search will increase.

Ammar Al-Dallal et al [14] It defines Genetic-based algorithm which uses inverted index model for pre-processing step which is called as GAWS. It is used as a method for finding the best set of the documents which is related to the keyword entered by users. These keywords are divided into three types: main keywords, should exist keywords and should not exist keywords. There are different types of sets of data that are used to evaluate GAWS each of which is double of the initial space size. Experimental results show that GAWS demonstrate high quality and also found to be competitive with the standard search engines. It uses WTindex which is a powerful tool for pre-processing of documents. GAWS proved its ability to retrieve documents and the experimental results show that GAWS is a promising method.

There is problem associated with GAWS which is solved by using a modifier operator. However more work is needed for improving the probability of words found rather than just counting the number of occurrences of the keywords provided by user. Improvement can also be achieved by applying this method to a large set

of documents for obtaining better and more accurate results.

CONCLUSION

The main objective of this survey is to build a PDF crawler with the help of inverted index and interval list by overcoming the drawbacks of web crawler. We also discussed the use of inverted index and interval list to search the keyword present in the PDF document. With the help of inverted index the storage capacity to store the keyword in PDF document can be increased and access time is optimized.

REFERENCES

1. Ian H. Witten, Gordon W. Paynter, Eibe Frank, Carl Gutwin and Craig G. Nevill-Manning "*KEA: Practical Automatic Keyphrase Extraction*," *Proceedings of the fourth ACM conference on Digital libraries*, 2013.
2. Gridharan J1, Vairavan S.V2, "*Inverted Index and Interval Lists for Keyword Search*," *Green Computing Communication and Electrical Engineering (ICGCCEE), International Conference*, March 2014.
3. Qiuying Bai, ChiMa, Xuechang Chen "A New Index Model Based on Inverted Index," *International Conference on Software Engineering and Service Science*, 2016.
4. Milos Ilic, PetarSpalevic, MladenVeinovic "Inverted Index Search in Data Mining," *22nd Telecommunications forum TELFOR*, 2014.
5. Hao Wu, Guoliang Li, and Lizhu Zhou, "Ginix: Generalized Inverted Index for Keyword Search," *TSINGHUA SCIENCE AND TECHNOLOGY ISSNll1 Vol 18,* February 2013.
6. Adriana Szekeres, SilviuHoriaBaranga, CiprianDobre, Valentin Cristea, "A Keyword Search Algorithm for Structured Peer-to-Peer Networks," *12th International Symposium on Symbolic and Numeric Algorithms for Scientific Computing*, 2010.
7. Yong Zhang, Quanlin Li, Bo Liu, "MapReduce Implementation of XML Keyword Search Algorithm," *IEEE International Conference on Smart City/SocialCom/SustainCom together with DataCom* and *SC2,* 2015
8. MeliusWeideman, "Empirical Study on Crawler Visibility of PDF Documents in Digital Libraries," *IEEE*, 2010.
9. B.M. ThosiniKumarika, N. G. J. Dias, "Smart Web Content Bookmarking with ANN based Key Phrase Extraction Algorithm," *International Conferences on Advances in ICT for Emerging Regions*, 2014.
10. Haitao Wu, Zhenmin Tang, "Effective XML Keyword Search Algorithm for Meaningful Return Information," *1st International Conference,* 2009.
11. Aliya Nugumanova, ArtemNovosselov, YerzhanBaiburin, Alexey Karimov, "Automatic keywords extraction from the domain texts: Implementation of the algorithm based on the MapReduce model" *International Conference on Current Trends in Information Technology (CTIT)*, 2013.
12. Aviral Nigam, "Web Crawling Algorithms", *International Journal of Computer Science and Artificial Intelligence Vol. 4 Iss. 3, PP. 63-67*, September2014**.**
13. Ammar Al-Dallal, Rasha Shaker, "Genetic Algorithm in Web Search Using Inverted Index Representation," *5th IEEE GCC Conference & Exhibition,*2009.
14. Anshika Pal, Deepak Singh Tomar, S.C. Shrivastava, "Effective Focused Crawling Based on Content and Link Structure Analysis," *International Journal of Computer Science and Information Security, Vol. 2, No. 1,* June 2009.
15. ApoorvVikramSingh ,Vikas , Achyut Mishra "A Review of Web Crawler Algorithms" *(IJCSIT) International Journal of Computer Science and Information Technologies, Vol. 5 (5) ,* 2014.