AN INVESTIGATIVE SCHEME FOR KEYWORD SEARCH USING INVERTED KEY TACTIC

Dnyaneshwar K. Bhujbal [1], Preeti Sharma[2]

[1]Computer Engineering, SKN Sinhgad Institute Technology & Science, Lonavla, Maharashtra, India
[3]Assistant Professor, Computer Engineering, SKN Sinhgad Institute Technology & Science, Lonavla, Maharashtra, India

Abstract

Unverified classification of outlines that is data items, observations or feature vectors into groups is called as Clustering. To retrieve a document from a group of documents according to a set of keywords efficiently, Inverted lists are mostly used instead of Keyword Search. Keyword Search is one of the most important activities in Information Retrieval System. As study shows in latest years, Keyword search is widely used by users to access text data. Keyword search is appropriate for document gatherings as well as for accessing structured or semi-structured data, XML documents relational databases and relational tables which can also be regarded as sets of documents. This is done by a keyword as query user retrieve documents. To proficiently retrieve documents, a data structure is used, that maps each word in the dataset, to a list of IDs of documents in which the word appears. The inverted index for a document collection consists of a set of so-called inverted lists, known as posting lists.

Keywords: *Keyword Search, Index Compression, Document Reordering, Rendering*

---***---

1. INTRODUCTION

Keyword Search is basis of Information Retrieval System. As new information is continuously added, the data size keeps on increasing and in these cases Keyword search is helpful for user to access text datasets. This textual data consists of web pages, XML documents and relational tables which can also be regarded as sets of documents. But even Keyword Search leads to a Datasets containing large amount of textual data. To overcome this problem & to retrieve data efficiently Inverted Keys are used which is also known as posting lists.

Inverted Keys are used to assess queries in all practical search engines. Keys or Indexes have three major benefits with respect to Compression for performance.
1) Requires less storage space
2) Better use of the available communication Bandwidth.
3) Avoiding a disk access compression [1].

Cluster of Inverted Indexes of a document Datasets is called as Posting Lists.

Information retrieval system (IRS) are widely developed for to finding useful information from a large-scale information datasets in which specialized indexing mechanism is required for efficient retrieval. Inverted files and signature files are two indexing tools in which inverted file is more popular because of better performance [2,3].

The new compression technique, that is, Inverted Index Compression Using Word-Aligned Binary Codes, provides compression rates, that are inferior to the best compression rates that have been previously achieved. But has an advantage of requiring exceptionally low computational effort at decoding time [4].

There is huge amount of data stored in different form like Unstructured, Semi-structured and Structured. An Effective 3-in-1 Keyword Search Method known as EASE for Unstructured, Structured and Semi-structured Data for indexing and querying large collections of heterogeneous data. EASE achieves both high accuracy and high search efficiency, and outperforms the existing approaches significantly [5].

The main concept *Inverted Index* is also used in current web search engines to perform some tasks efficiently. Good answers must be provided by engine within a fraction of a second, while simultaneously serving multiple requests, which gives interactive applications with respect to latency demands [6].

The section [II] describing Literature Survey of An Investigative Scheme for Keyword Search Using Inverted Key Tactic.

The Section [III] Explaining Propose System & Architecture of An Investigative Scheme for Keyword Search Using Inverted Key Tactic.

The Section [IV] gives the Conclusion of An Investigative Scheme for Keyword Search Using Inverted Key Tactic.

2. LITERATURE SURVEY

The use of clustering has been reported by studies in *Inverted Indexing* for keyword search in Information Retrieval. Basically, The use of classic algorithm for clustering data is described by most of the research such as Expectation-Maximization (EM) for unsupervised learning of Gaussian Mixture Models, K-means, Fuzzy C-means (FCM), and Self-Organizing Maps (SOM). These algorithms have well-known properties and are widely used in practice [9].

The main task is Keyword search through Indexing technique which is one of the most important activities in Information Retrieval System. Keyword search is critical for users to access text datasets, as huge amount of new information is added over the time. In that textual document web pages XML documents, and relational tables which can also be regarded as sets of documents is also includes a keyword as query user can retrieve documents. To efficiently retrieve documents, a data structure is used that maps each word in the dataset to a list of IDs of documents in which the word appears. The inverted index for a document collection consists of a set of so-called inverted lists, known as posting lists.

The Concept described here uses Inverted indexes to evaluate queries in all practical search engines. Indexes has three major benefits with respect to Compression for performance, so,
1) Requires less storage space
2) Better use of the available communication Bandwidth so that more information can be transferred per second than, when the data is uncompressed. The total time of transferring compressed data and subsequently decompressing is potentially much less than the cost of transferring uncompressed data for fast decompression schemes
3) Avoiding a disk access compression which increases the part of an index required to evaluate a query, is already cached in memory, Because of the above reasons index compression can reduce cost in retrieval systems [1].

Information Retrieval System (IRS) are widely developed for finding useful information from a large-scale information. So, inverted file compression through document identifier reassignment has tried. Compression of inverted lists of document postings that store the position and frequency of indexed terms. Better implementation and better choice of integer compression schemes are two approaches for improving retrieval efficiency [2,3].

New compression technique that is Inverted Index Compression Using Word-Aligned Binary Codes provides fast access into the compressed inverted lists, and overall provides faster query processing than previous techniques by allowing some inefficiency in the compressed representation. In support of these claims, This gives compression results for several large text collections as well as retrieval throughput results that show the overall benefit of the new approach compared to previous index coding mechanisms [4].

Keyword search is not only convenient for document collections as well as for accessing semi-structured or structured data, like relational databases and XML documents Data is present in different formats like Unstructured, Semi-structured and Structured. An Effective 3-in-1 Keyword Search Method (EASE) for Unstructured, Structured and Semi-structured Data for indexing and querying large collections of heterogeneous data, first model unstructured, structured and semi-structured data as graphs to achieve high efficiency in processing keyword queries, and then instead of using traditional inverted indices summarize the graphs and construct graph indices. For enhancing search effectiveness propose an extended inverted index to facilitate keyword-based search, and present a different ranking mechanism. An extensive experimental study by real datasets and the output show that EASE achieves both high accuracy and high search efficiency, and outperforms the existing approaches significantly [5].

Inverted Index is also used in current web search engines to perform some tasks efficiently. The good answers must be provided by engine within a fraction of a second, while simultaneously serving multiple requests, which gives interactive applications w.r.t. latency demands [6].

To reduce the space cost of storing inverted indexes various compression techniques are used. IDs in inverted lists are sorted in ascending order, to store the differences between IDs many existing techniques are used called d-gaps, and then to encode these d-gaps using shorter binary representations use various techniques. Because of extra computational cost during query processing, An Investigative Scheme for Keyword Search Using Inverted Key Tactic which is an extension of the traditional inverted index (denoted by InvIndex), to support keyword search [7].

An Investigative Scheme for Keyword Search Using Inverted Key Tactic uses no. of Search algorithms for keyword search. Scan-line algorithm, Improved scan-line algorithm, Twin-heap algorithm, Probe-Based Algorithm An Algorithm is given for an Efficient support to basic operations on interval lists, like union and intersection without decompression. The performance of An Investigative Scheme for Keyword Search Using Inverted Key Tactic is enhanced by document reordering, and two scalable and effective algorithms based on signature sorting and greedy heuristic of Traveling Salesman Problem (TSP). A Heuristic Inverted Index Approach of Keyword Search reduces the index size as well as improves the search performance on real datasets [7].

Keyword search is widely used by users to access text data with many studies in recent years. Keyword search is convenient for document collections as well as for accessing semi-structured or structured data, like relational databases and XML documents [8].

Most techniques first convert each ID in an inverted list to the difference between it and the preceding ID, known as d-gaps, and then encode the list using integer compression algorithms [3,4,8]. Simple and fast decoding provided by Variable-Byte Encoding [3].

Other studies have focused on how to improve the compression ratio of inverted index using document reordering[6]. The overall compression ratio is improved by reassigning document IDs so that similar documents are close to each other, & then there are more small d-gaps in the converted lists.

Figure.1 shows the index sizes using different compression techniques. The widely-adopted VBE is used to evaluate the present technique of converting consecutive IDs to intervals in An Investigative Scheme for Keyword Search Using Inverted Key Tactic. Figure 1 compares the original inverted index (denoted by InvIndex), the inverted index compressed by VBE (denoted by InvIndex+VBE), the present inverted index (denoted by Heuristic Inverted Index Approach of Keyword Search), and the present inverted index compressed by VBE (denoted by Heuristic Inverted Index Approach of Keyword Search +VBE) for PubMed datasets. The results show that the Architecture of an Investigative Scheme for Keyword Search Using Inverted Key Tactic Fig. 1 Comparison of sizes of indexes using different compression techniques. Compression is much better than that of VBE. An Investigative Scheme for Keyword Search Using Inverted Key Tactic + VBE result has the smallest index size. [7]

The Figure 1 is given below,

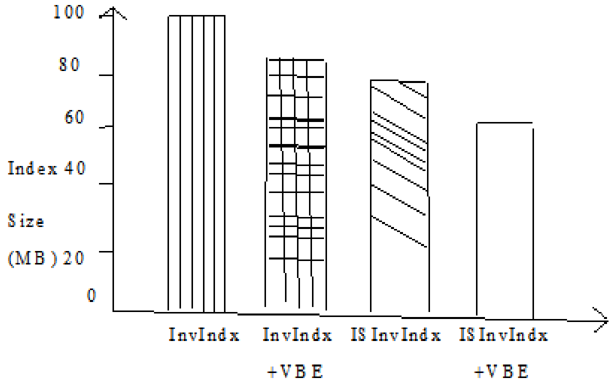


Fig 1 Architecture of An Investigative Scheme for Keyword Search Using Inverted Key Tactic

In our propose system basically there are three important steps which are as follows

1) Labelling
2) Twin Heap Selection
3) Document Retrieving

Labelling: In this 1st step there are three sub parts,
a) Visit to an individual documents
b) Give them label
c) Sort in Ascending order.

Twin Heap Selection: This is the main Algorithm method used for selection. The steps are as follows
a) Find out the Maximum entity in Lower Heap as well as Minimum entity in upper heap.
b) After that apply twin heap selection algorithm which gives smallest d-gap that most similar documents in one cluster.
c) Set the bound which is a limit.

Document Retrieval: This is last step of An Investigative Scheme for Keyword Search Using Inverted Key Tactic in which initially binary search is applied for better performance. Then again sort in ascending order for lower bound [7]. After sorting decide limits which will lead to retrieve documents & it gives Required clustered Documents.

To reduce the space cost of storing inverted indexes various compression techniques are used. IDs in inverted lists are sorted in ascending order. To store the differences between IDs many existing techniques are used called as d-gaps, and then to encode these d-gaps using shorter binary representations using various techniques. Because of extra computational cost during query processing, An Investigative Scheme for Keyword Search Using Inverted Key Tactic which is an extension of the traditional inverted index denoted by InvIndex is used to support keyword search [7].

The input document will be in any of the format like text, word, pdf & so on.

3. MATHEMETICAL MODULE

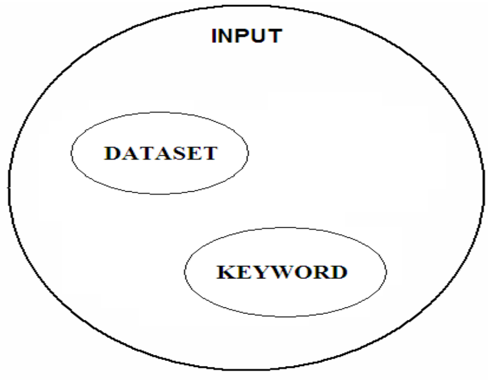
Document Retrieval
Set D
D0= Get all Interval List
D1=Remove Empty Lists
D2=For Each interval list
D3= Get upper limit
D4= Get Lower Limit
D5=If interval size is greater than 1
D6=Then add in retrieval vector
D7= Return Document retrieval vector

```
Algorithm:    PROBISECT+(R)
Input:   R    A set of interval lists
Output:  G    The resulting interval list.
    Sort R in ascending order of list lengths
    for all r ϵ R1 do Add CASPROBE(r,R-R1) to G
     return G
    procedure CASPROBE(r,R)
Input    : r0   A non-empty interval.
R          An set of interval lists.
Output   : G    The resulting interval list.
          R1*← CASPROBE(r0,R1)
           for all r ϵ R1* do
              r* ← [max(lb(r0),lb(r)),min(ub(r0),ub(r))]
    if n > 1 then Add CASPROBE(r*,R-R1) to G
           else Add r* to G
    return G
end procedure
```

4. RESULTS

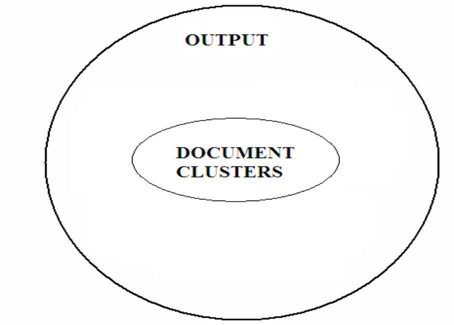
4.1 Data Set

Here, input is Data Set in which no. of files are there. The input document will be in any of the format like text, word, pdf & so on. In future work we can extend this with audio video input. We will any give keyword as an input to search, then it will search that required keyword in given input dataset.



4.2 Result Set

An Investigative Scheme for Keyword Search Using Inverted Key Tactic has an effective index structure and efficient algorithms to support keyword search, This Fast scalable methods enhance the search speed of An Investigative Scheme for Keyword Search Using Inverted Key Tactic by reordering documents in the datasets. It also requires smaller storage size than the traditional inverted index. It has a higher keyword search speed and using compression techniques, performance can be improved.
This is used in Document Searching, Search Engine, Bio-Informatics & So on.



5. CONCLUSIONS

By doing the survey on An Investigative Scheme for Keyword Search Using Inverted Key Tactic it was determined that clustering on data is not an easy task. There is huge data to be cluster in this Approach, So, to overcome this problem, an approach is presented that applies document clustering methods to An Investigative Scheme for Keyword Search Using Inverted Key Tactic.

An Investigative Scheme for Keyword Search Using Inverted Key Tactic has an effective index structure and efficient algorithms to support keyword search, Fast scalable methods of An Investigative Scheme for Keyword Search Using Inverted Key Tactic enhance the search speed by reordering documents in the datasets. Experiments show that This Approach of Keyword Search not only requires smaller storage size than the traditional inverted index, but also has a higher keyword search speed and using compression techniques, performance can be improved.

In future work we can extend this project with audio video input.

ACKNOWLEDGMENTS

I express my sense of gratitude towards my Project guide **PROF.R.V.PUJERI & PROF. PREETI SHARMA**, who has been very concerned and has aided for all the material essential for the preparation of this paper, he helped me to explore this vast topic in an organized manner and provided me with all the ideas on how to work towards a research oriented venture.

I am also thankful to **PROF. SUPRIYA B. SARKAR**, P.G Coordinator for the motivation and Inspiration that triggered me for the thesis work.

REFERENCES

[1]. I. Witten, A. Mo_at, and T. Bell. *'Managing Gigabytes: Compressing and Indexing Documents and Images.'* Morgan Kaufmann Publishers, Los Altos, CA 94022, USA, second edition, 1999.
[2]. S. Ding, J. Attenberg, and T. Suel, *"Scalable techniques for document identifier assignment in inverted indexes"* in Proc. of the 19th International Conference on World Wide Web, Raleigh, North Carolina, USA, 2010, pp. 311-320
[3]. F. Scholer, H. E. Williams, J. Yiannis, and J. Zobel, *"Compression of inverted indexes for fast query evaluation"* in Proc. of the 25th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Tammpere, Finland, 2002, pp. 222-229,Georgia, USA, 2006, pp. 59.
[4]. V. N. Anh and A. Moffat, *"Inverted index compression using word-aligned binary codes, Information Retrieval"*, vol. 8,no. 1, pp. 151-166, 2005.
[5]. G. Li, B. C. Ooi, J. Feng, J.Wang, and L. Zhou, *"EASE: An effective 3-in-1 keyword search method for unstructured, semi-structured and structured data"*, in Proc. of the ACM SIGMOD International Conference on Management of Data, Vancouver, BC, Canada, 2008, pp. 903-914.
[6]. S. Ding, J. Attenberg, and T. Suel, *"Scalable techniques for document identifier assignment in inverted indexes"* in Proc. of the 19th International Conference on World Wide Web, Raleigh, North Carolina, USA, 2010, pp. 311-320
[7]. Hao Wu_, Guoliang Li, and Lizhu Zhou*,"Ginix: Generalized Inverted Index for Keyword Search",* IEEE transactions on knowledge and data mining vol:8 no:1 year 2013.
[8]. S. Agrawal, S. Chaudhuri, and G. Das, *"DBXplorer: A system for keyword-basedsearch over relational databases"*,

in Proc. of the 18th International Conference on Data Engineering, San Jose, California, USA, 2002, pp. 5-16.

[9]. L. Liu, J. Kang, J. Yu, and Z. Wang, "A comparative study on unsupervised feature selection methods for text clustering," in *Proc. IEEE Int. Conf. Natural Language Processing and Knowledge Engineering*, 2005, pp. 597–601.

[10]. V. Hristidis and Y. Papakonstantinou, DISCOVER: Keyword search in relational databases, in Proc. of the 28th International Conference on Very Large Databases, Hong Kong, China, 2002, pp. 670-681.

[11]. V. Hristidis, L. Gravano, and Y. Papakonstantinou, Efficient IR-style keyword search over relational databases, in Proc. of the 29th International Conference on Very large Databases, Berlin, Germany, 2003, pp. 850-861.

BIOGRAPHIES

Mr.Dnyaneshwar Kisan Bhujbal is pursuing his Masters of Computer Engineering in SKN Sinhgad Institute of Technology & Science, Lonavla, University of PUNE. He received his BEng. Info.Tech. degree in 2012 from Sinhgad Institute of Technology & Science, Narhe, PUNE, He has published 2 paper in International Conference & Journal.

Prof. Preeti Shrama is an Asst. Prof. at SKN Sinhgad Institute of Technology & Science Lonavala. She received her Master of Computer Engineering degree from Sinhgad Institute of Technology Lonavla, University of PUNE and BCSEng. Degree from MIT Aurangbad. She has published around 10 paper in International Conference & Journals