ISSN (Online) 2394-2320

International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)
Vol 4, Issue 10, October 2017

An Analysis of Keyword Search Retrieval Techniques on Multi-Dimensional Datasets

[1] Mrs. R.Ramya Devi, [2] T.Shanmuga Priya
[1] Assistant Professor , [2] PG Scholar
[1][2] VELAMMAL Engineering College, Chennai.

***Abstract* - Information retrieval (IR) is the activity of obtaining information resources relevant to an information need from a collection of information resources. Searches can be based on full-text or other content-based indexing. Information retrieval is the most concentrated research field in the real world environment which enables users to retrieve the contents stored in the web. Keyword based search retrieval make ease of this task which can find the most similar information present web source accurate by computing similarity distance between the keywords submitted and the information present on web source. However keyword based similarity finding on multi-dimensional data would be more difficult task which required more computation to be performed. In this analysis work, comparison evaluation of different methodologies that are implemented to perform keyword based search retrieval on multi-dimensional data set has been discussed. Those research methodologies have been discussed based on their working procedures in detail. The different existing research methods are compared in detail based on their merits and demerits. The comparison analysis of research methods are conducted by comparing them with each other in terms of their merits and demerits. From this analysis, efficient research methodology that can perform key word based search retrieval in the flexible and accurate way can be identified in terms of performance improvement.**

Keywords: Information retrieval, multi dimension data, Keyword based search, spatial properties.

1. INTRODUCTION

Nearest Keyword set inquiries on content rich different types of data sets [1]. The NKS inquiry is an arrangement of catchphrases in view of theme. Also, the arrangement of the question consolidates ‘’K’’ type of catchphrases as a group and concentrates each and every set which posses data based bunches along with structures in which bunches of multi-dimensional region is created. Each point is labeled with an arrangement of clusters. When all is said in done, PromishA is more time and space effective compared to PromishE which can get close ideal outcomes practically speaking.

The file structure and the hunt technique for PromishA are like promishE, along these lines, we just depict the contrasts in the procedures. Here list design of PROMISH- A varies with promishE by the method for apportioning projectional region of irregular bits of vector space. Promish an allotments projection region into canisters of equivalent width which are not covered, not at all like promishE parceling projections of words into covering receptacles. Accordingly, every information sets get one receptacle point from an irregular vector z in promishA algorithm. Only a solitary check is created due to association of own compartment focuses produced by every single m discretionary vectors. Every single id is made using it's stamp in the vector space.

Keyword search is the most popular information discovery method because the user does not need to know either a query language or the underlying structure of the data [2]. The search engines available today provide keyword search on top of sets of documents. When a set of query keywords is provided by the user, the search engine returns all documents that are associated with these query keywords. Solution to such queries is based on the IR2-tree, but IR2-tree having some drawbacks. Efficiency of IR2-tree badly is impacted because of some drawbacks in it. The solution for overcoming this problem should be searched. Spatial inverted index is the technique which will be the solution for this problem. Spatial database manages multidimensional data that is points, rectangles

In this analysis paper, comparison evaluation of the different keyword based search retrieval techniques on multi-dimensional dataset has been discussed. It is done with the concern of different search retrieval techniques which varies with their working procedure and the metrics evaluated for differing their working improvement. This comparison evaluation is conducted in terms of various performance metrics and merits and demerits involved in it.

The overall organization of the research method is given as follows: In this section introduction about the information retrieval and their importance has been given. In section 2, different related research methodologies which are implemented by different authors. In section 3, comparison analysis of research methods in terms of merits and demerits has been given. Finally section 4, overall analysis work is concluded in terms of merits and demerits.

II. ANALYSIS OF KEYWORD BASED SEARCH RETRIEVAL TECHNIQUES

Nearest neighbour search (NNS), also known as closest point search, similarity search. It is an optimization problem for finding closest (or most similar) points. We can search closest point by giving keywords as input; it can be spatial or textual.

In [3] Aggregate nearest keyword search in spatial databases,‖ in Asia-Pacific Web Conference, 2010. Keyword search on relational databases is useful and popular for many users without technical background. Recently, aggregate keyword search on relational databases was proposed and has attracted interest. However, two important problems still remain. First, aggregate keyword search can be very costly on large relational databases, partly due to the lack of efficient indexes. Second, the top-k answers to an aggregate keyword query have not been addressed systematically, including both the ranking model and the efficient evaluation methods. We also report a systematic performance evaluation using real data sets.

Many applications require finding objects closest to a specified location that contains a set of keywords [4]. For example, online yellow pages allow users to specify an address and a set of keywords. In return, the user obtains a list of businesses whose description contains these keywords, ordered by their distance from the specified address. The problems of nearest neighbor search on spatial data and keyword search on text data have been extensively studied separately. However, to the best of our knowledge there is no efficient method to answer spatial keyword queries, that is, queries that specify both a location and a set of keywords.

Locality-sensitive hashing scheme based on p-stable distributions,‖ in SCG, 2004 [5]. We present a novel Locality-Sensitive Hashing scheme for the Approximate Nearest Neighbor Problem under lp norm, based on pstable distributions. Our scheme improves the running time of the earlier algorithm for the case of the l2 norm. It also yields the first known provably efficient approximate NN algorithm for the case p < 1. We also show that the algorithm finds the exact near neigbhor in O (log n) time for data satisfying certain ―bounded growth‖ condition. Unlike earlier schemes, our LSH scheme works directly on points in the Euclidean space without embedding's. Consequently, the resulting query time bound is free of large factors and is simple and easy to implement. Our experiments (on synthetic data sets) show that the our data structure is up to 40 times faster than kd-tree. Our algorithm also inherits two very convenient properties of LSH schemes. The first one is that it works well on data that is extremely high dimensional but sparse. Specifically, the running time bound remains unchanged if d denotes the maximum number of non-zero elements in vectors. To our knowledge, this property is not shared by other known spatial data structures.

Keyword search on spatial databases [6]. This work, mainly focus on finding top-k Nearest Neighbors, in this method each node has to match the whole querying keywords. As this method match the whole query to each node, it does not consider the density of data objects in the spatial space. When number of queries increases then it leads to lower the efficiency and speed. They present an efficient method to answer top-k spatial keyword queries. This work has the following contributions: 1) the problem of top-k spatial keyword search is defined. 2) The IR2-Tree is proposed as an efficient indexing structure to store spatial and textual information for a set of objects. There are efficient algorithms are used to maintain the IR2-tree, that is, insert and delete objects. 3) An efficient incremental algorithm is presented to answer top-k spatial keyword queries using the IR2-Tree. Its performance is estimated and compared to the current approaches. Real datasets are used in our experiments that show the significant improvement in execution times.

Location based information stored in GIS database [7]. These information entities of such databases have both spatial and textual descriptions. This paper proposes a framework for GIR system and focus on indexing strategies that can process spatial keyword query. The following contributions in this paper: 1) It gives framework for query processing in Geo- graphic Information Retrieval (GIR) Systems. 2) Develop a novel indexing structure called KR*-tree that captures the joint distribution of keywords in space and significantly improves performance over existing index structures. 3) This method have conducted experiments on real GIS datasets showing the effectiveness of our techniques compared to the existing solutions. It introduces two index structures to store spatial and textual information
There is more and more research interest in location-based web search, i.e. searching web content whose topic is related to a particular place or region [8]. This type of search contains location information; it should be indexed as well as text information. Text search engine is set-oriented

where as location information is two-dimensional and in Euclidean space. In previous paper we see same two indexes for spatial as well as text information. This creates new problem, i.e. how to combine two types of indexes. This paper uses hybrid index structure, to handle textual and location based queries, with help of inverted files and R*-trees. It considered three strategies to combine these indexes namely: 1) inverted file and R*-tree double index.2) first inverted file then R*-tree.3) first R*-tree then inverted file. It implements search engine to check performance of hybrid structure, that contains four parts:(1) an extractor which detects geographical scopes of web pages and represents geographical scopes as multiple MBRs based on geographical coordinates. (2) The work of indexer is use to build hybrid index structures integrate text and location information. (3) The work of ranker is to ranks the results by geographical relevance as well as non-geographical relevance. (4) An interface which is friendly for users to input location-based search queries and to obtain geographical and textual relevant results [9].

C.R. Barde, Pooja Katkade, Deepali Shewale, Rohit Khatale et al describes the problem of Secured Multiple-keyword Search (SMS) on encrypted cloud data. It constructs a group of confidentiality policies for safe cloud data consumption method [10]. This reference paper has picked a standard called coordinate matching, which is utilized to distinguish the similitude between search inquiry and data documents.

C. Rajeshkumar, Dr.K.RubaSoundar et al portrays the idea of Encrypted Cloud data recovery utilizing multi-keyword. As and when the users send their own data onto the cloud, the administration supplier ought to be equipped for dealing with the data and the link between the users and cloud. The issue of data confidentiality is one of the significant issues in the cloud. With the end goal of data confidentiality, delicate data must be encrypted some time ago sending it to cloud [11].

K. Manoj Kumar, E. Purushotham et al describes the cloud data privacy and elimination of information leakage in cloud using Two Round Searchable Encryption (TRSE) [12]. In cloud computing environment data owner keeps up an arrangement of documents to send its encrypted structure into the cloud server. Towards this, the data owner needs to make a searchable index utilizing the keywords and afterward it sends both the encrypted index and encrypted records into the cloud server.

Amol D. Sawant, Prof. M.D Ingle et al depicts the issue of saving total of multiple keywords search records ranking score [13]. This additionally enhances the ranking score of the documents which expands the framework ease of use by giving critical requested ranking as opposed to giving indistinguishable yield.

Wenhai Sun, Bing Wang,Ning Cao, Ming Li,Wenjing Lou, Y. Thomas, Hui Liet al depicts the issue of client data confidentiality and secure cloud search capacities over encrypted cloud data [14]. They tended to the issue utilizing security defensive Multi-Keyword Text/Keyword search (MTS) technique alongside equivalent word ranking. It proposes how search index will be established by Term Frequency (TF) and the Vector Space Model (VSM) alongside level of cosine comparability. It can be utilized as a part of request to achieve propelled search yield precision. With a specific end goal to expand search fitness, recommends index structure and Multi-Dimensional (MD) algorithm.

Cao et al. [15] presented paper on the new problem for retrieving a spatial object‟s groups that are nearest to the query point location and each associated with a set of keywords. They collective known as spatial keyword query.
G. Cong, C.S. Jensen, and D. Wu [16] proposed an approach that computes the relevancy of a query result by means of language models and a probabilistic ranking function. This relevance is then incorporated with the Euclidean distance between object and query to calculate an overall similarity of object to query.

In [17] focused on location based information retrieval. A framework is proposed for Geographical Information Retrieval (GIR) system and focus on indexing strategies can process Spatial Keywords (SK) queries effectively. Two type of indexing mechanism is used. First method is separate index for spatial and text attribute. Second method is hybrid indices techniques combine the spatial and inverted file indices.

This hybrid index structure is used to search m-closest keywords [18]. This technique finds the closest tuples that matches the keywords provided by the user. For processing the m-closest keyword query BR*-tree structure combines the features of R*-tree and bitmap indexing and returns the spatially closest objects matching m keywords as a results. A priori based search strategy is used to reduce the search space. To facilitate efficient pruning, two monotone constraints are used as a priori properties. But this method has some disadvantages while handling ranking queries and in this number of false hits is large.

IR2-tree is a combination of R-tree and signature files [19]. R-trees and the best-first algorithm for NN search are well known techniques in spatial databases. In general, Signature file refers to a hashing-based framework which is known as superimposed coding (SC), which is more effective than other instantiations. It is designed to perform membership tests: determine whether a query word„ t‟ exists in a set„ T‟ of words. SC is conservative, if it says "no", then w is definitely not in W. On the other hand, if SC returns "yes", the true answer can be either way, in which case the whole W must be scanned to avoid a false hit.

III. COMPARISON EVALUATION OF RESEARCH METHODS

In this section, analytical evaluation of the research methodologies that are discussed in the previous section has been given. Here the comparison is made based on merits and demerits of research methods which are given in tabular format.

Table 1. Comparison analysis of research methods

S.No	Authors	Method	Merits	Demerits
1	Z. Li et al 2010	Aggregate nearest keyword search	useful and popular for many users without technical background	Aggregate keyword search is costly on large relational databases top-k retrieval not addressed
2	I. De Felipe et al 2008	Keyword search on spatial databases	Obtains a list of businesses whose description contains these keywords	there is no efficient method to answer spatial keyword queries
3	M. Datar et al 2004	Locality-sensitive hashing scheme based on p-stable distributions	Bound remains unchanged if d denotes the maximum number of non-zero elements in vectors	property is not shared by other known spatial data structures
4	D. Zhang et al 2009	Keyword search on spatial databases	significant improvement in execution times	Each node has to match with querying keyword. So it affects on performance also it becomes time consuming and maximizing searching space
5	R. Hariharan et al 2007	Location based information stored in GIS database	Easy of maintaining two separate indices Performance bottleneck lies in the number of candidate object generated during the filtering stage	If spatial filtering is done first, many objects may lie within a query is spatial extent, but very few of them are relevant to query keywords. This increases the disk access cost by generating a large number of candidate objects. The subsequent stage of keyword filtering becomes expensive

6	Y.-Y. Chen et al 2006	Hybrid Index Structures for Location-based Web Search	Instead of using two indexes for textual and spatial information. this paper gives hybrid index structures that integrate text indexes and spatial indexes for location based web search	In ranking phase, it combine geographical ranking and non-geographical ranking, combination of two rankings and the computation of geographical relevance may affects on performance of ranking
8	C. R. Barde et al 2014	Secured Multiple-keyword Search (SMS)	Accurately distinguish the similitude between search inquiry and data documents	Search enquiry complexity is high
9	C.Rajeshkumar and Dr.K.Rubasoundar 2014	Encrypted Cloud data recovery	Ensured security level	data confidentiality is one of the significant issues
10	K. Manoj Kumar, E. Purushotham 2014	Two Round Searchable Encryption	Ensured security level Accurate retrieval	More computation overhead
11	Amol D et al 2014	Multiple keywords search records ranking	Enhances the ranking score	Highly correlated data items cannot be distinguished accurately
12	Wenhai Sun et al 2013	Multi-Keyword Text/Keyword search (MTS) technique	Increased precision rate More accuracy	Required more computation time to accurately retrieve contents
13	X. Cao et al 2011	retrieving a spatial object‟s groups	Accurate retrieval of contents	Need more historic knowledge for accurate retrieval of contents
14	G. Cong 2009	probabilistic ranking function	Can accurately retrieve the data contents stored in the database	More computation overhead
15	R. Hariharan et al 2007	Geographical Information Retrieval (GIR) system	can process Spatial Keywords (SK) queries effectively	Need to learn more volume of data to retrieve the accurate information
16	Yufei Tao and Cheng Sheng 2014	hybrid index structure priori based search strategy	efficient pruning	While handling ranking queries and in this number of false hits is large
17	Cao 2010	IR2-tree	more effective than other instantiations	Whole W must be scanned to avoid a false hit.

IV. CONCLUSION

Keyword based search retrieval is the most complex task in the real world, which attempts to retrieve the most similar contents from the web source. In this analysis work, various research methodologies that are conducted towards achieving efficient keyword search retrieval has been discussed in detail. However keyword based similarity finding on multi dimensional data would be more difficult task which required more computation to be performed. In this analysis work, comparison evaluation of different methodologies that are implemented to perform keyword based search retrieval on multi-dimensional data set has been discussed. Those research methodologies have been discussed based on their working procedures in detail. The different existing research methods are compared in detail based on their merits and demerits. The comparison analysis of research methods are conducted by comparing them with each other in terms of their merits and demerits. From this analysis, efficient research methodology that can perform key word based search retrieval in the flexible and accurate way can be identified in terms of performance improvement.

REFERENCE

[1] Lakshmi, L., Reddy, P. B., Bindu, C. S., & Sri, S. B. (2017). An Effective nearest Keyword Search in Multifaceted Datasets. International Journal, 8(5).

[2] Hristidis, V., & Papakonstantinou, Y. (2002, August). Discover: Keyword search in relational databases. In Proceedings of the 28th international conference on Very Large Data Bases (pp. 670-681). VLDB Endowment.

[3] Z. Li, H. Xu, Y. Lu, and A. Qian, ―Aggregate nearest keyword search in spatial databases,‖ in Asia-Pacific Web Conference, 2010.

[4] I. De Felipe, V. Hristidis, and N. Rishe, ―Keyword search on spatial databases,‖ in ICDE, 2008, pp. 656–665.

[5] M. Datar, N. Immorlica, P. Indyk, and V. S. Mirrokni, ―Locality-sensitive hashing scheme based on p-stable distributions,‖ in SCG, 2004

[6] D. Zhang, Y. M. Chee, A. Mondal, A. K. H. Tung, and M. Kitsuregawa. Keyword search in spatial databases: Towards searching by document. In ICDE, pp. 688– 699, 2009

[7] R. Hariharan, B. Hore, C. Li, and S. Mehrotra, "Processing Spatial- Keyword (SK) Queries in Geographic Information Retrieval (GIR) Systems," Proc. Scientific and Statistical Database Management (SSDBM), 2007.

[8] Y.-Y. Chen, T. Suel, and A. Markowetz. Efficient query processing in geographic web search engines. In SIGMOD, pp. 277–288, 2006.

[9] X. Cao, G. Cong, and C. S. Jensen. Retrieving top-k prestige-based relevant spatial web objects. PVLDB, 3(1):373–384, 2010.

[10] C. R. Barde, Pooja Katkade, Deepali Shewale and Rohit Khatale, "Secured Multiple-keyword Search over Encrypted Cloud Data ",International Journal of Emerging Technology and Advanced Engineering , Volume 4, Issue 2, February 2014

[11] C.Rajeshkumar and Dr.K.Rubasoundar, "Retrieval of Encrypted cloud data using multi-keyword", International Journal of Innovative Research in Computer and Communication Engineering, Vol.2, Special Issue 1, March 2014.

[12] K. Manoj Kumar, E. Purushotham, "Top-K Retrieval of Encrypted Cloud Data by Using Secure Multi-Keyword", International Journal of Software and Hardware Research in Engineering, Vol 2,Issue 8, August 2014

[13] Amol D. Sawant and Prof. M.D Ingle," Indexing and Advanced Relevance Ranking Score Preserving for Multi Keyword Search over Encrypted Cloud Data", International Journal of Computer Science and Information Technologies (IJCSIT), Vol. 5 (3),3165 – 3169,2014

[14] Wenhai Sun, Bing Wang, Ning Cao, Ming Li, Wenjing Lou, Y. Thomas Hou and Hui Li, "Privacy-preserving Multi keyword Text Search in the Cloud Supporting Similarity based Ranking", ASIA CCS, 2013

[15] X. Cao, G. Cong, C.S. Jensen, and B.C. Ooi, "Collective Spatial Keyword Querying", in Proc. ACM SIGMOD Int"l Conf. Management of Data, pp. 373- 384, 2011.

[16] G. Cong, C.S. Jensen, and D. Wu, "Efficient Retrieval of the Top-k Most Relevant Spatial Web Objects", PVLDB, vol. 2, no. 1, pp. 337- 348, 2009.

[17] R. Hariharan, B. Hore, C. Li, and S. Mehrotra, "Processing spatial-keyword (SK) queries in geographic information retrieval (GIR) systems", in Proc. of Scientific and Statistical Database Management (SSDBM), 2007.

[18] Yufei Tao and Cheng Sheng, "Fast Nearest Neighbor Search with Keywords", IEEE transactions on knowledge and data engineering, VOL. 26, NO. 4, APRIL 2014.

[19] Cao, Cong, G., and Jensen, C. S., "Retrieving top-k prestige based relevant spatial web objects", PVLDB, 3(1), pp. 373–384, 2010