



ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijircce.com

Vol. 5, Issue 1, January 2017

A Review on Fast Query Processing Techniques and Algorithms

Pallavi D. Nikam, Prof. M. S. Burange

Dept. of Computer Science & Engineering, P.R.Pote (Patil) College of Engineering and Management, Amravati, India

Assistant Professor, Dept. of Computer Science & Engineering, P.R.Pote (Patil) College of Engineering and

Management, Amravati, India

ABSTRACT: Keyword search is very popular method for searching from large datasets nowadays. There are objects associated with different domains that are having their significances in variety of applications. Today, several trendy applications call for novel varieties of queries that aim to seek out objects satisfying both a spatial predicate, and a predicate on their associated texts. The presence of keywords in feature space allows for the development of new tools to query and explore from these multi-dimensional datasets. In normal web based applications search box is at the top of any browser or document. This will carried out the needs for the research to develop the Nearest Keyword Search methods. There are lots of application of nearest keyword based searching, and as the amount of data is growing on increasing. Lots of Research in different areas of keywords based searching in multi-dimensional environment are performed. This paper presents study related to the different terms in the nearest keyword searching (NKS) and multidimensional datasets. As lots of research work is being perform in the field of keyword based searching, this paper performs review on some of the important techniques.

KEYWORDS: Nearest Keyword Set (NKS), Multi-Dimensional data, Indexing, Hashing, Querying.

I. INTRODUCTION

The World-Wide Web has reached a size where it is becoming increasingly challenging to satisfy certain information needs. While search engines are still able to index a reasonable subset of the web. The pages a user is really looking for are often buried under hundreds of thousands of less interesting results. Thus, search engine users are in danger of drowning in information. Adding additional terms to standard keyword searches often fails to narrow down results in the desired direction. A natural approach is to add advanced features that allow users to express other constraints or preferences in an intuitive manner, resulting in the desired documents to be returned among the first results. In fact, search engines have added a variety of such features, often under a special advanced search interface, but mostly limited to fairly simple conditions on domain, link structure, or modification date.

Most of the searching is carried out with the help of objects that are having their significances in variety of domains. Objects in images, chemical compounds, documents, or experts in collaborative networks are often characterized by a collection of relevant features, and are commonly represented as points in a multi-dimensional feature space [1]. For example, images are represented using color feature vectors, and usually have descriptive text information e.g., tags or keywords associated with them. In this paper, the study is performed on multi-dimensional datasets where each data point has a set of keywords. The presence of keywords in feature space allows for the development of new tools to query and explore these multi-dimensional datasets. Keyword-based search in text-rich multi-dimensional datasets facilitates many novel applications and tools. In this paper, we consider objects that are tagged with keywords and are embedded in a vector space. Some algorithms are design to work on two dimensional data sets, when applied on multi-dimensional data sets it take more time to retrieve data. Retrieving data for multi-dimensional data set for these algorithms is crucial task. There is a need for an efficient algorithm that scales with dataset dimension and yields practical query efficiency on large datasets. When search for keyword set the threading is



International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijircce.com

Vol. 5, Issue 1, January 2017

necessary to increase the retrieval time of the query [2]. The one by one search of the word take more time to retrieve the data. Thus multi-threading is necessary to minimize the time for large set of keywords.

In this paper, the study of fast query processing techniques based on keywords is performed. For this the nearest keyword searching seems effective technique. This nearest keyword set referred to as NKS queries on text-rich multi-dimensional datasets are performed. An NKS (nearest keyword set) query is a set of user-provided keywords, and the result of the query may include k sets of data points each of which contains all the query keywords and forms one of the top-k tightest cluster in the multi-dimensional space. NKS queries are useful for many applications, such as photo-sharing in social networks, graph pattern search, geo location search in GIS systems and so on.

In cloud computing environment as now-a-days data owners are motivated to delegate complex data managements to the commercial cloud for economic savings [3]. Sensitive data is usually encrypted before being uploaded to the cloud, which unfortunately makes the frequently-used search function a challenging problem. In this paper, a new multi-keyword dynamic search scheme is used with result ranking to make search over encrypted data more secure and practical. In the scheme, a powerful function-hiding inner product encryption to enhance the security by preventing the leakage of search pattern is performed. For the concern of efficiency, we adopt a tree-based index structure to facilitate the searching process and updating operations.

New keyword suggestions will be determined in keeping with their linguistics connection to the initial keyword question. One of the novel method called ProMiSH (Projection and Multi Scalez Hashing) that uses random projection and hash-based index structures, and achieves high scalability and speedup [1]. This methods is useful to work on real and synthetic datasets with more faster than traditional methods that are based on tree-based techniques. However, it is necessary to increase the ways and techniques to search based on the keywords that will efficiently give the outcome to the user, in which he is interested in. And this techniques should be compatible with multi-dimensional data models. The remaining paper is organized as, Section II study about some the important terms and applications related to the NSK and multidimensional data. Section III studies some techniques and algorithms that are related to develop our fast query processing and searching model. Section IV gives some Literature Survey which gives brief information of the study done by different researchers in this field. Finally, Section IV concludes the paper.

II. IMPORTANT TERMS

A. KEYWORDS:

Keywords are terms extracted from document or single sentence generating sense when clustered in context. Same Keywords might be present in different document but structure and position of keywords build different meaning, this highlights keyword importance. Keywords are daily used atomic terms which when used in two or more group to represent information form phrases or short sentences and require concept generated by keywords. Success of keyword technology is simple match of documents consisting Question Keyword.

B. MULTIDIMENSIONAL DATA MODEL:

The multidimensional data model is an integral part of On-Line Analytical Processing, or OLAP. Because OLAP is on-line, it must provide answers quickly; analysts pose iterative queries during interactive sessions, not in batch jobs that run overnight. And because OLAP is also analytic, the queries are complex. The multidimensional data model is designed to solve complex queries in real time. As the data is come from online description this data is of vast amount, and growing on increasing continuously. The different types of multidimensional data models having their different representation and storage are:

- The Logical Multidimensional Data Model
- The Relational Implementation of the Model
- The Analytic Workspace Implementation of the Model

C. DATA MINING:

Data mining is a new powerful technology with great potential to help companies focus on the most important information in their large amount of data contained in data warehouses. It has been defined as the fast analysis of large



International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijircce.com

Vol. 5, Issue 1, January 2017

or complex data sets in order to discover significant patterns or trends that would otherwise go unrecognized. Here, the Data mining techniques used that automates the process of shifting through historical data in order to discover new and required information. Here a model is usually devised by a statistician to deal with a specific analysis problem. It also differentiates data mining from expert systems and the model is built by a knowledge engineer from rules extracted from the experience of an expert.

D. APPLICATIONS OF NEAREST KEYWORD SEARCH (NSK):

There are lots of application of NSK some of which are explained below:

- In a photo-sharing social network like Facebook, where photos are tagged with people names and locations. These photos can be embedded in a high-dimensional feature space of texture, color, or shape [4]. Here an NKS query can find a group of similar photos which contains a set of people.
- A drug company will analyze its recent sales department activity and their results to boost targeting of high-value physicians and verify that selling activities can have the best impact within the next few months. The information has to embrace challenger market activity still as information regarding the native health care systems.
- NKS queries are useful for graph pattern search, where labeled graphs are embedded in a high dimensional space for scalability. In this case, a search for a sub graph with a set of specified labels can be answered by an NKS query in the embedded space [5].
- A heterogeneous company with an oversized direct sales department will apply data processing to spot the most effective prospects for its services. Mistreatment data processing to investigate its own client expertise, this company will build a singular segmentation characteristic the attributes of high- value prospects.
- NKS queries can also reveal geographic patterns. GIS can characterize a region by a high-dimensional set of attributes, such as pressure, humidity, and soil types. Meanwhile, these regions can also be tagged with information such as diseases.

III. RELATED TECHNIQUES AND ALGORITHMS

For performing fastest query processing in large amount of multidimensional datasets data mining operations and algorithms are required. There are variety of techniques available that are useful for performing this task. Every technique will itself be enforced in numerous ways in which, employing a kind of algorithms. Some of the useful algorithms related to fastest keyword searching are as follows [6].

A. CLASSIFICATION AND PREDICTION:

Classification is most ordinarily supported operation by industrial data processing tools. This operation allows organizations to get patterns in massive or complicated information sets so as to resolve specific business issues. Classification is that the method of sub dividing knowledge set with relation to variety of specific outcomes. For an example, we would need to classify our customers into 'high' and 'low' classes with relation to credit risk. The class or 'class' into that every client is placed is that the outcome of our classification. The foremost common techniques for classification square measure call trees and neural networks. If a call tree is employed, it will offer a group of branching conditions that in turn split the purchasers into teams outlined by the values within the freelance variables.

B. CLUSTERING ALGORITHMS:

Cluster analysis is that the method of distinctive relationships that exist between things on the idea of their similarity and difference. In contrast to classification, cluster doesn't need a target variable to be known beforehand. A cluster algorithmic rule takes Associate in Nursing unbiased investigate the potential groupings at intervals an information set Associate in Nursing makes an attempt to derive an optimum delineation of things on the idea of these teams. To spot things that belong to a cluster, some live should be used that gauges the similarity between things at intervals a cluster and their difference to things in alternative clusters.



International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijirccce.com

Vol. 5, Issue 1, January 2017

C. NEAREST NEIGHBORS:

Nearest Neighbors could be a technique appropriate for classification models. Once a new case or instance is conferred to the model, the algorithmic rule appearance in the slightest points the information to search out a set of cases that are most almost like it and uses them to predict the result. There are two principal drivers within the k-NN algorithm: the quantity of nearest cases to be used (k) and a metric to live what's meant by nearest. Every use of the k-NN algorithmic rule needs that we tend to specify a positive number worth for k. This determines what number existing cases are checked out once predicting a new case. K-NN refers to a family of algorithms that we tend to may denote as 1-NN, 2-NN, and 3- NN, so forth. For instance, 4-NN indicates that the algorithmic rule can use the four nearest cases to predict the result of a new case. K-NN is predicated on a thought of distance, and this needs a metric to work out distances that helps in searching keywords as we required.

D. NEURAL NETWORKS:

A Neural Network could be a set of connected input/output units wherever every association includes a weight associated with it. Throughout the learning section the network learns by adjusting the weights thus on are able to predict the proper category label of the input samples. Neural network learning is additionally observed as connectionless learning because of the connections between units. A key distinction between neural networks and lots of different techniques is that neural nets solely operate directly on numbers. As a result, any non-numeric information in either the freelance or output columns should be reborn to numbers before we are able to use the info with a Neural Network.

E. NAIVE-BAYES:

Naive-Bayes could be a classification technique that is each prophetic and descriptive. It analysis the connection between every experimental variable and also the variable to derive a chance for every relationship. Nave-Bayes needs just one experience the coaching set to come up with a classification model. This makes it the foremost economic data processing technique. However, Naive-Bayes doesn't handle continuous information, therefore any freelance or dependent variables that contain continuous values should be binned or bracketed.

F. DECISION TREES:

Decision trees are one among the foremost common data processing technique and therefore the best liked in tools geared toward the business user. They are simple to line up, their results are graspable by an end-user, they will address a large vary of classification issues, they are strong within the face of various knowledge distributions and formats, and that they are effective in analyzing giant numbers of fields. The foremost common forms of call tree rule are CHAID, CART and C4.5. CHAID (Chi- square automatic interaction detection) and CART (Classification and Regression Trees) were developed by statisticians. CHAID will turn out tree with multiple sub-nodes for every split. CART needs less knowledge preparation than CHAID, however produces solely two-way splits. C4.5 comes from the globe of machine learning, and relies on scientific theory.

G. R-TREE:

R-Tree makes use of solely Associate in Nursing R-Tree organization [7]. Given a distance-first top-k spatial keyword query, the algorithmic rule initial finds the top-1 nearest neighbour object to the query purpose. Then it retrieves that object and compares that object's matter description with the keywords of the query. If the comparison fails then that object is discarded, and therefore the next nearest object is retrieved. The progressive NN algorithmic rule is employed. This method continues till Associate in nursing object is found whose matter description contains the query keywords. Once a satisfying object is found its came back and therefore the method repeats till k objects are came back. The drawback of this algorithmic rule is that it's to retrieve each object came back by the NN algorithmic rule till the top-k result objects are found. This doubtless will result in the retrieval of the many "useless" objects. Within the worst case the whole tree must be traversed and each object must be inspected.



International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijirccce.com

Vol. 5, Issue 1, January 2017

IV. LITERATURE REVIEW

The authors in [1] suggested that Keyword-based search in text-rich multidimensional datasets facilitates many novel applications and tools. The author in this paper consider objects that are tagged with keywords and are embedded in a vector space. For these datasets, we study queries that ask for the tightest groups of points satisfying a given set of keywords. Author proposed a novel method called ProMiSH (Projection and Multi Scale Hashing) that uses random projection and hash-based index structures, and achieves high scalability and speedup. Authors present an exact and an approximate version of the algorithm. The studies, both on real and synthetic datasets, show that ProMiSH has a speedup of more than four orders over state-of-the-art tree-based techniques. Our scalability tests on datasets of sizes up to 10 million and dimensions up to 100 for queries having up to 9 keywords show that ProMiSH scales linearly with the dataset size, the dataset dimension, the query size, and the result size. But in future, it is necessary to explore other scoring schemes for ranking the result sets. And each group of points can be scored based both on the distance between the points and weights of the keywords. Further, the criteria of a result containing all the keywords needs to be relaxed to generate results having only a subset of the query keywords.

The authors in [8] proposed an efficient a novel search implementation on spatial databases with simple implementation than the complex tree constructions like R trees, in both cache based and non-cache based with geocodings. The proposed algorithms shows an optimal results than the traditional approaches and for the Multi-dimensional databases with key word searches. Here the searching is performed for nearest neighbor locations and keywords. In this paper, solution was proposed to have remedied the situation by developing an access method called *the spatial inverted index i.e.* (SI-index). The SI-index is not only fairly space economical, but also it has the ability to perform keyword-augmented nearest neighbor search in time that is at the order of dozens of milliseconds. Furthermore, as the SI-index is based on the conventional technology of inverted index, it is readily incorporable in a commercial search engine that applies massive parallelism, implying its immediate industrial merits.

The authors in [9] proposed the keyword search over relational databases has been extensively studied because it promises to allow users lacking knowledge of structured query languages or unaware of the database schema to query the database in an intuitive way. The existing works about keyword search on databases proposed many approaches and have gain remarkable results. As the databases are becoming more and more, the existing methods to keyword search over relational databases with the centralized setting hardly can process the keyword queries effectively. To solve this problem, authors propose DKS, a distributed data-graph based approach to keyword search over relational databases with MapReduce. In this approach, authors partition the data graph into multiple sub graphs, and then employ a cluster of servers to search the *CS-trees* within each sub graph in the *map* operation. After combining the *CS-trees* produced by mappers, the combined results will be processed by reducers in a parallel way to construct the Steiner trees. Finally, the top-*k* results can be found by merging Steiner trees sent by reducers. Further authors plan to study some more efficient pruning rules to beforehand cut down the *CS-trees* that cannot be connected into a Steiner tree in the *reduce* operation. Furthermore, they continue to optimize the combine operation of DKS to reduce the cost of transmitting the *CS-trees*. Finally they make the DKS approach be more general to distributed keyword search over various very large databases.

The authors in [10] performs work for the presence of keywords in feature space that allows for the development of new tools to query and explore these multi-dimensional datasets. They propose a method called Projection and Multi Scale Hashing that uses random projection and hash-based index structures, and achieves high scalability and speedup. Here this proposed system provides accurate results in multiple keyword search. This is how user data can be used to enhance search list and to find interest of the user. In this project, they also proposed how social annotations will be useful in the field of complex word search, which gives optimization as day by day large size of data available for searching by interest will be the future for search engines. The main advantage of this system will save lacks of processor cycles used in multidimensional data sets for finding image. But at the same time, the develop system fail to provide real time answers on difficult inputs. The real nearest neighbor lies quite far away from the query point, while all the closer neighbors are missing at least one of the query keywords.

The authors in this paper [11], propose a new secure multi-keyword search scheme supporting both result ranking and dynamic document updating. As cloud computing is becoming prevalent, data owners are motivated to delegate complex data managements to the commercial cloud for economic savings. Sensitive data is usually encrypted before being uploaded to the cloud, which unfortunately makes the frequently-used search function a challenging problem. In



International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijirccce.com

Vol. 5, Issue 1, January 2017

this paper, new multi-keyword dynamic search scheme with result ranking to make search over encrypted data more secure and practical. In this scheme, a powerful function-hiding inner product encryption to enhance the security by preventing the leakage of search pattern. For the concern of efficiency, a tree-based index structure to facilitate the searching process and updating operations. Experiments over real-world data demonstrate the good performance of this scheme.

Another track of related works deal with m-closest keyword queries. In bR*-Tree is developed based on R*-tree[12] that stores bitmaps and minimum bounding rectangles(MBRs) of keywords in every node along with points MBRs. bR*-Tree also suffers from a high storage cost; therefore author Zang et al Modified bR*-Tree to create virtual bR*-tree in memory at run time. Virtual bR*-tree is created from a pre-stored r*-Tree, which indexes all the points, and an inverted index which stored keyword information and path from the root node in R*-Tree for each point. Both bR*-Tree and virtual bR*-Tree shares similar performance weaknesses as bR*-Tree.

The authors in this paper [13], investigate the problem of multi-keyword fuzzy ranked search over encrypted cloud data. They propose a multi-keyword fuzzy ranked search scheme based on Wang et al.'s scheme. Concretely, they have develop a novel method of keyword transformation and introduce the stemming algorithm. With these two techniques, the proposed scheme is able to efficiently handle more misspelling mistake. Moreover, our proposed scheme takes the keyword weight into consideration during ranking. Like Wang et al.'s scheme, our proposed scheme does not require a predefined keyword set and hence enables efficient file update too. They also give thorough security analyses and conduct experiments on real world data set, which indicates the proposed scheme's potential of practical usage. But at the same time Fuzzy ranked search supporting dynamic update: Though here proposed scheme can support update, it is failed to achieve the ideal state because of the keyword weight. So, it develop a way to reflect the keyword weight and enable update. Along with this as multi-data owner scheme has more realistic significance. Many of the work done were mainly focusing on the cases of single data owner and hence not effective for multi-data owner.

In [14], the author gathered the data by applying data mining techniques on instant messaging service on MSN, the search has been made in two ways, first the data which is gathered is based on user to user chats because the people who chat with each other mostly share their interests with each other, second the data is gathered on the basis of keywords which were entered on MSN search engine for searching different things. The author further applied Bayes' rule on the gathered data for calculating the probabilities.

The authors in [15] perform works on locations based query processing. The queries are made by the user manually, which are more time consuming and finding the route is difficult. So the android users makes a query to the cloud server so the data can be retrieved on the bases of geo tagged query and checking the privacy profile. So the user location in which query is requested and the wireless network is focused with the relation data and spatial database. Based on the spatiotemporal data the processing and retrieving can be done by hadoop framework. The result shows that the framework can give helpful suggestions. Within the future, we have to decide to more study the effectiveness of the LKS framework by collection additional information and planning a benchmark. Additionally, it is subject to the provision of knowledge, needs to be adapt and take a look at location-aware keyword question suggestion (LKS).

V. CONCLUSION

This paper performs the review on Keyword-based search in text-rich multidimensional datasets that facilitates many applications and tools. In this paper, we ponder objects that are tagged with keywords and are embedded in a vector space. For working with multidimensional datasets, the queries are perusal that ask for the tightest groups of points satisfying a given set of keywords. There are lots of application available and also are growing on increasing in the various field that require fast query processing. As the data is growing on increasing and all the activities are performed online, so to working in this text reach environment it is necessary to develop the method for fastest keyword searching. For this to understand all the important terms related to this, this paper perform a review. This paper also perform study of the different research work performed by different authors in the fields of nearest keyword set (NKS) for query processing. By understanding all this study, one can able to develop system for fastest query searching and processing in multi-dimensional data environment.



ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijircce.com

Vol. 5, Issue 1, January 2017

REFERENCES

- [1] Vishwakarma Singh, Ambuj K. Singh, "Nearest Keyword Set Search in Multi-dimensional Datasets", *Department of Computer Science, University of California, Santa Barbara, USA*. 12 Sep 2014.
- [2] Zhiguo Wan and Robert H. Deng, "VPSearch: Achieving Verifiability for Privacy-Preserving Multi-Keyword Search over Encrypted Cloud Data" *Transactions on Dependable and Secure Computing*, Citation information: DOI 10.1109/TDSC.2016.2635128, IEEE.
- [3] Sanket S.Pawar, Abhijeet Manepatil, Aniket Kadam, Prajakta Jagtap, "Keyword Search in Information Retrieval and Relational Database System: Two Class View", *International Conference on Electrical, Electronics, and Optimization Techniques (ICEEOT) – 2016*.
- [4] V. Singh, A. Bhattacharya, and A. K. Singh, "Querying spatial patterns," in *Proc. 13th Int. Conf. Extending Database Technol.: Adv. Database Technol.*, 2010, pp. 418–429.
- [5] H. He and A. K. Singh, "GraphRank: Statistical modeling and mining of significant subgraphs in the feature space," in *Proc. 6th Int. Conf. Data Mining*, 2006, pp. 885–890.
- [6] Chandrashekhar, "Fast Searching With Keywords Using Data Mining", *International Journal of Computer Science and Information Technology Research* Vol. 2, Issue 2, pp: (82-99), Month: April-June 2014, Available at: www.researchpublish.com.
- [7] I. Kamel and C. Faloutsos. Hilbert R-tree: An improved r-tree using fractals. In *Proc. of Very Large Data Bases (VLDB)*, pages 500–509, 1994.
- [8] C. Usha Rani, N.Munisankar, "Spatial Index Keyword Search in Multi-dimensional Database", C. Usha Rani et al, / (IJCSIT) *International Journal of Computer Science and Information Technologies*, Vol. 5 (5), 2014, 6468-6471.
- [9] Ziqiang Yu, Xiaohui Yu *IEEE Member*, Yuehui Chen, Kun Ma, "Distributed Top-k Keyword Search over Very Large Databases with MapReduce", *IEEE International Congress on Big Data*, 2016.
- [10] Ruksar I. Attar, Shraddha S. Hon, Ruchita M. Agrawal, Deepali R. Borse, Prof. R. B. Bhosale, "An Efficient Nearest Keyword Set Search in Multidimensional Dataset", *International Research Journal of Engineering and Technology (IRJET)*, Volume: 03 Issue: 12 | Dec -2016 www.irjet.net
- [11] Jingbo Yan, Yuqing Zhang, Xuefeng Liu, "Secure Multi-keyword Search Supporting Dynamic update and ranked retrieval", *China Communications* • October 2016.
- [12] 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*, 2009, pp. 688–699.
- [13] Zhangjie Fu, Xinle Wu, Chaowen Guan, Xingming Sun, and Kui Ren, "Toward Efficient Multi-Keyword Fuzzy Search Over Encrypted Outsourced Data With Accuracy Improvement", *IEEE TRANSACTIONS ON INFORMATION FORENSICS AND SECURITY*, VOL. 11, NO. 12, DECEMBER 2016.
- [14] Parag, Singla, Matthew, Richardson. Yes, there is a Correlation – From Social Networks to Personal Behavior on the Web. WWW-08 (pp. 1 -7).
- [15] S.Dhamodaran, IV.S.Mahesh, M. SaiSwaroop, "OPTIMISED KEYWORD SEARCH WITH PROXIMITY LOCATION BASED SERVICES", *International Conference on Computation of Power, Energy Information and Communication (ICCPEIC)*, 2016.