International Journal of Engineering & Scientific Research
Vol.5 Issue 9, September 2017,
ISSN: 2347-6532 Impact Factor: 6.660
Journal Home page: http://www.ijmra.us, Email:editorijmie@gmail.com
Double-Blind Peer Reviewed Refereed Open Access International Journal - Included in the International Serial Directories Indexed & Listed at: Ulrich's Periodicals Directory ©, U.S.A., Open J-Gage as well as in Cabell's Directories of Publishing Opportunities, U.S.A

AUTOMATIC ASSIGNED KEY BASED QUERY SEARCH ON TEXTUAL DATA SETS

Shilpa B Kodli*

Huda Fatima*

Rohini Mishra*

ABSTRACT

In multi dimension data sets key based search enables many new applications. Here we tagged the objects with a keyword and embed into the data base afterwards we examine the results. the datasets which are available in the database on that we make a query and it fetch those fields which are most satisfy the given set of query keywords. We introduce a new method which uses random projection and hash based index structure search so our technique is more faster and high scalable with compare to earlier tree based techniques.

1. INTRODUCTION

OBJECTS (e.g., 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. 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, we consider 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. In this paper, we study nearest keyword set (referred to as NKS) queries on text-rich multi-dimensional datasets. An NKS 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. Fig. 1 illustrates an NKS query over a set of two-dimensional data points. Each point is tagged with a set of keywords. For a query Q ¼ fa; b; cg, the set of points f7; 8; 9g contains all the query keywords fa; b; cg and forms the tightest cluster compared with any other set of points covering all the query keywords. Therefore, the set f7; 8; 9g is the top-1 result for the query Q.

Project Description

Key-based examination in text-rich multi-dimensional datasets enables many innovative applications and tools. Here we study the objects that are labelled with specific keywords and are fixed in a vector space. For these input datasets, we make queries that may fitted to the closest groups of points satisfy a given set of keywords. Here we implementing a new method called ProMiSH which stand for Projection and Multi Scale Hashing which uses random prediction and hash-based index structures. So that it can achieve high scalability and speed. We are here to describe an exact and an approximate of algorithm.

PROBLEM STATEMENT

Location-specific keyword queries on the web and in the GIS systems were earlier answered using a combination of R-Tree and inverted index. These existing works mainly focus on the type of queries where the coordinates of query points are known . Even though it is possible to make their cost functions same to the cost function in NKS queries, such tuning does not change their techniques. in multi-dimensional spaces, it is difficult for users to provide meaningful coordinates, and our work deals with another type of queries where users can only provide keywords as input.

OBJECTIVE OF THE STUDY

Here we are achieve more faster and reliable data retrieval technique with the help of automatically keyword comparison. Our technique will display results inmore faster and more accurate way.

SCOPE OF THE STUDY

Proposed technique can be applicable in any where we need a lot of data should be retrieved and stored for example railways, hospitals, companies etc.

METHODOLOGY

Work ofan our algorithm is divided into two sections namely Hash table inverted index and inverted index. Here we are considering keyword which are given by an end user to the system as an input. The point belongs to the keywords and respected hash bucket IDs can be find out all hash numbers in keyword bucket. The keyword bucket contains all query keywords. The index may not work properly if dimension of the data will increases so we came across the new method called nearest keyword search (NKS)in multi dimension datasets.

We will search the points associated with the hash bucket IDS i.e. there will be search for all the keywords, if there is no exact match for the keyword, then it will search for 2 keywords i.e. the multiple combination of the keywords then for the single keyword. Thus all the keywords are searched efficiently with fewer time and more correctness in multidimensional datasets And we proposed solution reimplementing multiple rounds in the top-k NKS in multidimensional data.

LITERATURE SURVEY

Litrature is the study of existing system here we have studied the different technique for ________________published by many authors by using different method and techniques.

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 effcient indexes. Second, the top-k answers to an aggregate keyword query has not been addressed systematically, including both the ranking model and the effcient evaluation methods. We also report a systematic performance evaluation using real data sets.

Many applications require finding objects closest to aspecified location that contains a set of keywords. 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.

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. 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

Rank based data inquiry in content rich variety datasets encourages numerous novel applications and devices. In this paper, we consider objects that are labeled with data and are inserted in a vector space. For these datasets, we ponder inquiries that request the most impenetrable gatherings of focuses fulfilling a given arrangement of data. We propose a novel technique called Best Data Matching Ranker Result (BDMRR) that utilizations irregular projection and Ranking Based hash file structures, and accomplish high adaptability and speedup. We exhibit a careful and a rough form of the calculation. Our trial results on genuine and manufactured datasets demonstrate that BDMRR has up to 70 times of optimized our data retrieval scheme

Near neighbor search in high dimensional spaces is useful in many applications. Existing techniques solve this problem efficiently only for the approximate cases. These solutions are designed to solve r-near neighbor queries for a fixed query range or for a set of query ranges with probabilistic guarantees, and then extended for nearest neighbor queries. Solutions supporting a set of query ranges suffer from prohibitive space cost. There are many applications which are quality sensitive and need to efficiently and accurately support near neighbor queries for all query ranges. In this paper, we propose a novel indexing and querying scheme called Spatial Intersection and Metric Pruning (SIMP). It efficiently supports r-near neighbor queries in very high dimensional spaces for all query ranges with 100% quality guarantee and with practical storage costs. Our empirical studies on three real datasets having dimensions between 32 and 256 and sizes up to 10 million show a superior performance of SIMP over LSH, Multi-Probe LSH, LSB tree, and iDistance. Our scalability tests on real datasets having as many as 100 million points of dimensions up to 256 establish that SIMP scales linearly with query range, dataset dimension, and dataset size

EXISTING SYSTEM

In Existing system, the Nearest Keyword Search (NKS) queries is applied on high dimension text datasets. Where an NKS query is a group of user given keywords to the graph pattern. After passing the NKS query we observe the K set of results which includes all the query keywords and forms in a cluster. Existing technique can work efficiently if the data is available in single dimension.If dimension of the data increases then it reduces its efficiency and gives possible solutions which are not an accurate.

Disadvantages:

1. Existing system technique can be used only on graph based pattern search. Where user provide the labels to the pattern and embed into the database.

2. Existing system technique always require the accurate information of the coordinates of a pattern. So that it became very difficult to remember exact coordinate of the graph based pattern.

PROPOSED SYSTEM

In proposed system we are introducing a novel method for NKS query search.Here we are proposing an efficient search algorithm which work with multi-scale indexes for good query processing result.The proposed system uses the data comparison method got the extraction of the required accurate results. The system stores the sub keys of the words from the file and considers these as the keyword which get compared with the data during the retrieval of the data of the keyword stored in the database.

Advantages of Proposed System:

1. Our Search algorithm will save time and space by direct index searching.

2. Proposed system can be applicable if the dimension of the data is more and it does not effect on the performance of the result.

SYSTEM DESIGN

The motive is to design the system as per the requirement. To examine the system we use Data flow Diagrams and the system architecture. System design describes the system by analyzing the existing system and what we are establishing in proposed system.

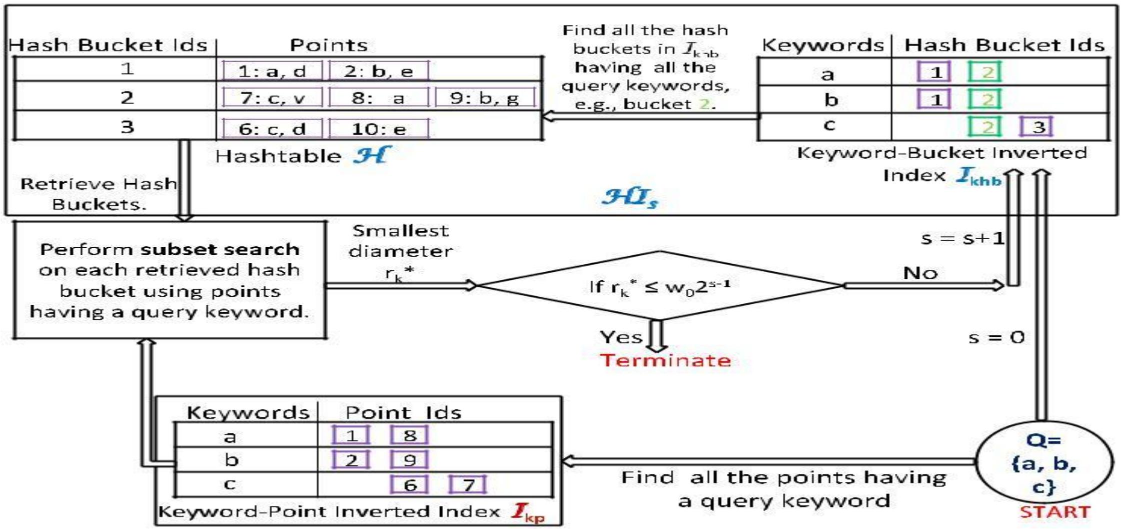
Hash Bucket Ids
Points
1
1: a, d
2: b, e
2
7: c, v
8: a
9: b, g
3
6: c, d
10: e
Hashtable $\mathcal{H}$
Find all the hash buckets in I_{khb} having all the query keywords, e.g., bucket 2.
Keywords
Hash Bucket Ids
a
1
2
b
1
2
c
2
3
Keyword-Bucket Inverted Index I_{khb}
$\mathcal{HI}_s$
Retrieve Hash Buckets.
Perform subset search on each retrieved hash bucket using points having a query keyword.
Smallest diameter r_k^*
If $r_k^* \leq w_0 2^{s-1}$
No
s = s+1
Yes
Terminate
s = 0
Keywords
Point Ids
a
1
8
b
2
9
c
6
7
Keyword-Point Inverted Index I_{kp}
Find all the points having a query keyword
Q= {a, b, c}
START

Fig1: System Architecture

Multi Dimension Data:

To search a meaningful record in multi dimension data set is very difficult for the user. And to search on a pattern its very difficult to remember an exact coordinates. To scale a data set in high dimension we require an efficient algorithm which provide the keyword based searching in multi dimension datasets.

Nearest Keyword Search:

Nearest keyword search (NKS) query can be user given keywords and the result gives k sets of record those are contain all the NKS query keyword and forms the top k cluster.

Indexing:

Indexing is the time taken to search a meaningful record from the database of a given NKS query. If the number of dimension increases our algorithm will take more time to execute the given query and gives the near optimal result.

Hashing:

Hashing sorts the indexing outcome based on the probability threshold. And selects only those records which is crossing the given threshold and gives the accurate result which are arrange in top-K cluster fashion.

CONCLUSION

In this paper, we are propose solutions to the problem of searching a record from multidimension data set which is the great deal for the researchers. And gives result in top-k fashion where individual datapoint contain all the NKS query keywords. We propose an efficient search technique based on word comparison and which extracts the words and use them as the keywords. Our technique suits well with earlier tree based techniques.

FUTURE ENHANCEMENT

In future enhancement, We explore the ranking to the result sets by using the different scoring schemes. Among those in one scheme, keywords may have weights which are assign by the user then, individual selected group of points can be valued based on distance between the other points and weights of the keywords. The result may having all the keywords which are satisfy by the given query.

BIBILOGRAPHY

[1] X. Cao, C. S. Jensen, and B. C. Ooi, "*Collective spatial keyword querying,*" in Proc. ACM SIGMOD Int. Conf. Manage. Data, (2011).

[2] G. Cong and D. Wu, "*Efficient retrieval of the top-k most relevant spatial web objects*" Proc. VLDB Endowment, vol. 2, (2009).

[3] Z. Li, H. Xu, Y. Lu, and A. Qian, "*Aggregate nearest keyword search in spatial data,*" in Proc. 12th Int. Asia-Pacific Web Conf., (2010).

[4] M. L. Yiu, X. Dai, N. Mamoulis, and M. Vaitis, "*Top-k spatial preference,*" in Proc. IEEE 23rd Int. Conf. Data Eng., (2007).

[5] V. Singh and A. K. Singh, "*SIMP: Accurate and efficient nearest neighbor search in high dimensional spaces,*" in Proc. 15th Int. Conf. Extending Database Technol., (2012).

[6] Y. Tao, K. Yi, C. Sheng, and P. Kalnis, "*Quality and efficiency in high dimensional near neighbor search,*" in Proc. ACM SIGMOD Int. Conf. Manage, (2009).