International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395 -0056
Volume: 03 Issue: 12 | Dec -2016 www.irjet.net p-ISSN: 2395-0072

An Efficient Nearest Keyword Set Search in Multidimensional Dataset

Ruksar I. Attar[1], Shraddha S. Hon[2], Ruchita M. Agrawal[3], Deepali R. Borse[4], Prof. R. B. Bhosale[5]

[1234] *B.E (Computer), S.V.I.T, Chincholi, Nashik, Maharashtra – 422102*
[5] *ME (Computer), S.V.I.T, Chincholi, Nashik, Maharashtra – 422102.*

Abstract - *Consider objects that are tagged with keywords and are embedded in a vector space. The presence of keywords in feature space allows for the development of new tools to query and explore these multi-dimensional datasets. We propose a method called Projection and Multi Scale Hashing that uses random projection and hash-based index structures, and achieves high scalability and speedup. 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. Images are represented using color feature vectors, and usually have descriptive text information (e.g., tags or keywords) associated with them. Our system is based on real datasets shows that we can show the efficient searching of keywords in multidimensional datasets.*

Key Words: Querying, Multi-dimensional Data, Indexing, Hashing.

1. INTRODUCTION

In today's digital world the amount of data which is developed is increasing day by day. There is different multimedia in which data is saved. It's very difficult to search the large dataset for a given query as well to archive more accuracy on user query. In the same time query will search on dataset for exact keyword match and it will not find the nearest keyword for accuracy. Ex: Flickr.

The amount of data which is developed is increasing day by day, thus it is very difficult to search large dataset for a given query as well to achieve more accuracy on user query.so we have implemented a method of efficient search in multidimensional dataset. This is associated with images as an input. Images 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 colour feature vectors, and usually have descriptive text information (e.g., tags or keywords) associated with them. 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.

Our main contributions are summarized as follows.
(1) We propose a novel multi-scale index for exact and approximate NKS query processing.
(2) We develop efficient search algorithms that work with the multi-scale indexes for fast query processing.
(3) We conduct extensive experimental studies to demonstrate the performance of the proposed techniques.

1. Filename: It is based on image filename.

2. CBIR (Content based image search): Content-based image retrieval (CBIR), also known as query by image content (QBIC) and content-based visual information retrieval (CBVIR) is the application of computer vision techniques to the image retrieval problem, that is, the problem of searching for digital images in large databases. Content-based image retrieval is opposed to traditional concept-based approaches (see Concept-based image indexing).

3. TBIR (Text based image search): Concept-based image indexing, also variably named as "description-based" or "text-based" image indexing/retrieval, refers to retrieval from text-based indexing of images that may employ keywords, subject headings, captions, or natural language text. It is opposed to Content-based image retrieval. Indexing is a technique used in CBIR.

Table -1: Comparison Table

	Filename	CBIR	TBIR	NKS (Extended TBIR)
No. of Result	Highest	Low	High	Low
Accuracy	Low	High	Medium	High
Performance	Highest	Low	High	High
User Satisfaction	<50%	90-100%	60-80%	90-100%

2. LITERATURE SURVEY

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 clusters in the multi-dimensional space. Illustrates an NKS query over a set of two-dimensional data points. Each point is tagged with a set of keywords. For a query the set of points contains all the query keywords and forms the tightest cluster compared with any other set of points covering all the query keywords. Therefore, the set is the

top-1 result for the query Q.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.

We present an exact and an approximate version of the algorithm. Our experimental results on real and synthetic datasets show that the method has more speedup over state-of-the-art tree-based techniques.

Other related queries include aggregate nearest keyword search in spatial databases, top-k preferential query, top-k sites in a spatial data based on their influence on feature points, and optimal location queries. Our work is different from these techniques. First, 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. The proposed techniques use location information as an integral part to perform a best first search on the IR-Tree, and query coordinates play a fundamental role in almost every step of the algorithms to prune the search space. Moreover, these techniques do not provide concrete guidelines on how to enable efficient processing for the type of queries where query coordinates are missing. Second, 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. Without query coordinates, it is difficult to adapt existing techniques to our problem.

Finding nearest neighbors in large multi-dimensional data has always been one of the research interests in data mining field. In this paper, we present our continuous research on similarity search problems. Previous work on exploring the meaning of K nearest neighbors from a new perspective in Pan KNN. It redefines the distances between data points and a given query point Q, efficiently and effectively selecting data points which are closest to Q. It can be applied in various data mining fields. A large amount of real data sets have irrelevant or obstacle information which greatly affects the effectiveness and efficiency of finding nearest neighbors for a given query data point. In this paper, we present our approach to solving the similarity search problem in the presence of obstacles. We apply the concept of obstacle points and process the similarity search problems in a different way. This approach can assist to improve the performance of existing data analysis approaches. The similarity between two data points used to be based on a similarity function such as Euclidean distance which aggregates the difference between each dimension of the two data points in traditional nearest neighbor problems.

In those applications, the nearest neighbor problems are solved based on the distance between the data point and the query point over a fixed set of dimensions (features). However, such approaches only focus on full similarities, i.e., the similarity in full data space of the data set. Also early methods suffer from the "curse of dimensionality". In a high dimensional space the data are usually sparse, and widely used distance metric such as Euclidean distance may not work well as dimensionality goes higher. Recent research [8] shows that in high dimensions nearest neighbor queries become unstable: the difference of the distances of farthest and nearest points to some query point does not increase as fast as the minimum of the two, thus the distance between two data points in high dimensionality is less meaningful. Some approaches are proposed targeting partial similarities. However, they have limitations such as the requirement of the fixed subset of dimensions, or fixed number of dimensions as the input parameter(s) for the algorithms. Keyword-based search in text-rich multi-dimensional datasets facilitates many novel applications and tools. We 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. We propose a method that uses random projection and hash-based index structures, and achieves high scalability and speedup. However, none of these algorithms considers detecting outliers simultaneously with clustering process. In many cases, outliers are as important as clusters, such as credit card fraud detection, discovery of criminal activities, discovery of computer intrusion, and etc. Analyzing the data distribution with the consideration of obstacles is critical for many data sets.

In recent years, various general techniques for analysis of movement data and human activities in particular were proposed. Different techniques for 3D geo-visualization of space-time patterns of people's travel experience and mobility is presented in .Two types of algorithms for mining interesting patterns from trajectories acquired by GPS-enabled devices are proposed. In the first type, the trajectories are converted into a sequence of stops or important parts (regions in which an object stayed more than a predefined time interval) before the algorithm for mining interesting patterns is applied. In the second type, the identification of important parts in a trajectory is part of the algorithm for mining patterns. Progressive clustering of trajectories of moving objects is presented. The authors combined clustering with visual interaction to let the analyst apply different distance functions based on the particular characteristics of trajectories under investigation. Visualization techniques (aggregations, ringmaps) of daily repeating activities like travel, work, shopping are presented. An algorithm for finding interesting places and mining travel sequences from GPS trajectories is proposed. The algorithm detects frequent sequences on different scales, taking into account the interestingness of the visited place and the experience of a user. Research on movement data is usually done on trajectories acquired by GPS-enabled devices. However, large scale GPS datasets, which would allow us to perform qualitative analysis on the level of a city or country, are still not available. On the other hand, geo tagged photo collections could be obtained on the world scale, which makes them a valuable resource for the analysis of people's activities. Concentration and movement of tourists at the scale of a city is analyzed using Flickr geo tagged photos. For this, the identified tourists in the city of

Rome using user profiles and built heat maps to visualize regions of high tourist concentration. The heat maps were created by dividing a region into cells, counting then number of people who took photos in every cell and smoothing the visualization by interpolating between values of every cell. However, no detailed analysis of the method, its advantages and disadvantages was provided. In addition, flow maps were used to visualize tourist movement between visited places. These places were connected by lines whose widths were proportional to the number of tourists. Mean-shift, a non-parametric clustering algorithm, was used in to find the most attractive places on Earth on a local and city scale using Flickr photos. The represented examples of maps with movements of people. However, no detailed analysis of the movement was presented.

Photo-sharing websites such as Flickr and Panoramio contain millions of geo tagged images contributed by people from all over the world. Characteristics of these data pose new challenges in the domain of spatio-temporal analysis. In this paper, we define several different tasks related to analysis of attractive places, points of interest and comparison of behavioral patterns of different user communities on geo tagged photo data. We perform analysis and comparison of temporal events, rankings of sightseeing places in a city, and study mobility of people using geo-tagged photos. We take a systematic approach to accomplish these tasks by applying scalable computational techniques, using statistical and data mining algorithms, combined with interactive geo-visualization. We provide exploratory visual analysis environment, which allows the analyst to detect spatial and temporal patterns and extract additional knowledge from large geo-tagged photo collections. We demonstrate our approach by applying the methods to several regions in the world.

Huge amount of data have been generated in many disciplines nowadays. The similarity search problem has been studied in the last decade, and many algorithms haves been proposed to solve the K nearest neighbor search. Previously proposed Pan KNN which is a novel technique that explores the meaning of K nearest neighbors from a new perspective. It redefines the distances between data points and a given query point Q, and selects data points which are closest to Q efficiently and effectively. In this paper, first a brief introduction about previous work on Pan KNN and discuss the Fuzzy concept; then, we propose to use the Fuzzy concept to design OPan KNN algorithm that targets solving the nearest neighbors problems in the presence of obstacles.

3. EXISTING SYSTEM

Using the combination of R-tree and inverted index, the location specific keyword queries were answered on web and GIS system. Ranking to the objects were done in existing system i.e. IR2 tree to rank objects based on the combination of their distances to the query location and the relevance of text description to the query keywords. Flickr uses keyword density and exact matching for keyword parsing technique. We can search one or two images related to the keyword not more than that. Cao et al and long et al proposed algorithm to retrieve a group of spatial web objects such that the group keywords cover the query keywords and the objects have lowest inter-object distances. Other related queries include aggregate nearest keyword search in spatial databases, top-k preferential query, top-k sites in spatial data base, and optimal location queries.

4. PROPOSED SYSTEM

In our proposed system the real data set is collected from photo sharing websites. In which we collect images from descriptive tags from Flickr and the images are transformed into grayscale and associate each data point, with a set of keyword that are derived from tags. We can collect number of datasets, suppose we collect five datasets (R1, R2, R3, R4, R5) with up to million data points, we can create multiple dataset to investigate performance. The query co-ordinates play a fundamental role in every step of algorithm to prune search space. Our work deals with providing keyword as an input. . We propose a novel multi-scale index for exact and approximate NKS query processing. We develop efficient search algorithms that work with the multi-scale indexes for fast query processing. Distance browsing is easy with R-trees. In fact, the best-first algorithm is exactly designed to output data points in ascending order of their distances.

In order to run the application efficiently the user must have following characteristics.

USER Module: User provides the input keyword as an image.

SYSTEM Module:

1) The system module retrieves all images from the database, and then it analyzes keywords.
2) The positive point relation is undertaken by the system.
3) It analyzes image keyword relation between points.
4) It filters the image based on the relations.
5) Applying nearest neighbor method retrieved images.
6) Displays nearest image as an output.

5. SYSTEM ARCHITECTURE

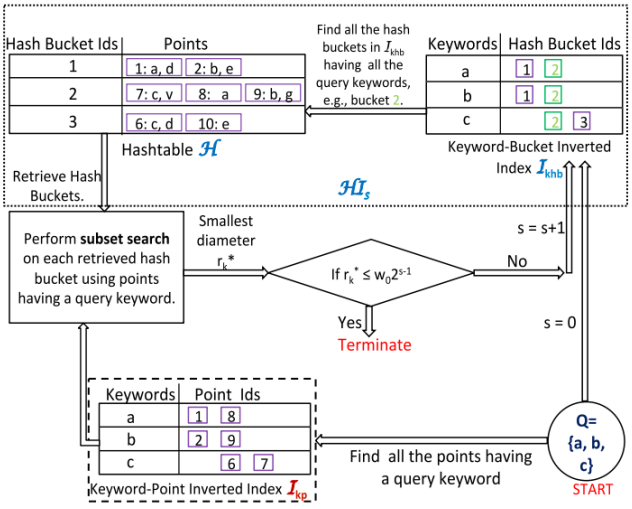

Fig -1: System Architecture

We start with the index for exact search. There are 2 main component included i.e. Inverted index ikp and Hashtable inverted index pairs (HI). We treat keyword as keys and provide it as an input to our system. There are hash bucket IDs and respective points associated with the keywords, it will find all the hash buckets in Ikhb (keyword bucket inverted index), having all query keywords.

In our system we are performing subset search on each retrieved hash bucket using points having query keywords. These indexes fail to scale dimension greater than 10 because of its dimensionality thus random projection with hashing and indexing has come up in the method of nearest keyword search in multidimensional datasets.

For e.g. Consider there are 3 keywords a, b, c. We will be searching 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, and then for the single keyword. Thus all the keywords are searched efficiently with less time and more accuracy in multidimensional datasets, and we proposed solution re-implementing multiple rounds in the top k nearest set in multidimensional datasets.

6. ADVANTAGES

- Distance browsing is easy with R-trees. In fact, the best-first algorithm is exactly designed to output data points in ascending order of their distances.
- It is straight forward to extend our compression scheme to any dimensional space.

7. DISADVANTAGES

- 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.

8. APPLICATIONS

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.

9. CONCLUSIONS

We have concluded that 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 our project we 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.

ACKNOWLEDGEMENT

We take this opportunity to express our hearty thanks to all those who helped us in the completion of the Paper. We express our deep sense of gratitude to our Project Guide Prof. R. B. Bhosale, Asst. Prof., Computer Engineering Department, Sir Visvesvaraya Institute of Technology, Chincholi for his guidance and continuous motivation. We gratefully acknowledge the help provided by him on many occasions, for improvement of this project report with great interest. We would be failing in our duties, if we do not express our deep sense of gratitude to Prof. S. M. Rokade, Head, Computer Engineering Department for permitting us to avail the facility and constant encouragement. Lastly we would like to thank all the staff members, colleagues, and all our friends for their help and support from time to time.

REFERENCES

[1] W. Li and C. X. Chen, "Efficient data modeling and querying system for multi-dimensional spatial data," in Proc. 16th ACM SIGSPATIAL Int. Conf. Adv. Geographic Inf. Syst., 2008, pp. 58:1 58:4.

[2] D. Zhang, B. C. Ooi, and A. K. H. Tung, "Locating mapped resources in web 2.0," in Proc. IEEE 26th Int. Conf. Data Eng., 2010, pp. 521-532.

[3] V. Singh, S. Venkatesha, and A. K. Singh, "Geo-clustering of images with missing geotags," in Proc. IEEE Int. Conf. Granular Comput., 2010, pp. 420-425.

[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] J. Bourgain, "On lipschitz embedding of finite metric spaces in hilbert space," Israel J. Math., vol. 52, pp. 46-52, 1985.

[6] 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.

[7] X. Cao, G. Cong, C. S. Jensen, and B. C. Ooi, "Collective spatial keyword querying," in Proc. ACM SIGMOD Int. Conf. Manage. Data, 2011, pp. 373-384.

[8] C. Long, R. C.-W. Wong, K. Wang, and A. W.-C. Fu, "Collective spatial keyword queries: A distance owner-driven approach," in Proc. ACM SIGMOD Int. Conf. Manage. Data, 2013, pp. 689-700.