Nearest Keyword Set Search In Multi-Dimensional Data Sets

[1] U. Venkatateja, [2]J. Krishna,[3]Dr. M. Rudra Kumar

[1] M.Tech.,(PG Scholar), [2]Assistant Professor,[2]Professor & Head of the Department

[1]Dept of CSE, [2]Dept of CSE, [3]Dept of CSE,

[1] Dept of CSE,Annamacharya Institute Of Technology & Sciences, Rajampet, Kadapa,

[2] Dept of CSE,Annamacharya Institute Of Technology & Sciences, Rajampet, Kadapa,

[3] Dept of CSE,Annamacharya Institute Of Technology & Sciences, Rajampet, Kadapa,

Abstract : Nearest neighbor search in multimedia databases needs more support from similarity search in query processing. Range search and nearest neighbor search depends mostly on the geometric properties of the objects satisfying both spatial predicate and a predicate on their associated texts. We do have many mobile applications that can locate desired objects by conventional spatial queries. Current best solution for the nearest neighbor search are IR2 trees which have many performance bottlenecks and deficiencies. So, a novel method is introduced in this paper in order to increase the efficiency of the search called as Spatial Inverter Index. This new SI index method enhances the conventional inverted index scheme to cope up with high multidimensional data [7] and along with algorithms that‟s compatible with the real time keyword search [2].

***IndexTerms* - Spatial Inverted Index, Nearest Neighbor Search, IR2 Trees, similarity search, Spatial Index**

1. Introduction

Many search engines are used to search anything from anywhere; this system is used to fast nearest neighbor search using keyword. Existing works mainly focus on finding top-k Nearest Neighbors, where each node has to match the whole querying keywords .It does not consider the density of data objects in the spatial space. Also these methods are low efficient for incremental query. But in intended system, for example when there is search for nearest restaurant, instead of considering all the restaurants, a nearest neighbor query would ask for the restaurant that is, closest among those whose menus contain spicy, brandy all at the same time, 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. The spatial inverted index is the technique which will be the solution for this problem.

Multidimensional objects such as points, rectangles managed by spatial databases provides fast access to those objects based on different selection criteria. For example, location of hospitals, hotels and theatres are represented as points whereas parks, lakes and shopping malls are represented as rectangles [1]. For instance, GIS range search gives all the cafes in certain area and nearest neighbour gives location of café near to our geometrical location.

Today, the search engine optimisation has made a realistic approach to write a spatial query in a brand new style. Some of may have few applications which finds the objects in a huge multidimensional data along with its geometrical locations and associated texts. There are easy ways to support queries that combine spatial and text features. For example, if we want to search a café whose menu contains keywords {Mocha, Espresso, Cappuccino} it would fetch all the restaurants with the keywords and from that list gives the nearest one. This approach can also be in another way but this straight forward approach has a drawback, which they will fail to provide real time answers on difficult inputs. A typical example, while all the closer neighbours are missing at least one of the query keywords, that the real nearest neighbour lies quite far away from the query point. The introduction of internet has given rise to an ever increasing amount of text data associated with multiple

dimensions (attributes), for example customer feedbacks in online shopping website like flipkart as they are always associated with the price, specifications and product model. Keyword query, one of the most popular and easy-to-use ways retrieves useful data from plain text documents. Given a set of keywords, existing methods aim to find joins or all the relevant items that contains a few or all the keywords. Spatial queries with keywords has not been explored. Recently, attention was diverted to multimedia databases [8]. The integration of two well-known concepts: R-tree [2], a popular spatial index, and signature file [4], an effective method for keyword-based document retrieval. This makes to develop a structure called IR2 trees, which has strengths of both signature files and R-Trees. Like R-Trees, IR2-Tree has object spatial proximity that solves spatial queries

efficiently. On the other side, the IR2-tree is able to filter a considerable portion of the objects that do not contain all the

query keywords, like signature files. Nearest neighbor search (NNS), also known as closest point search, similarity search. It is an optimization problem for finding closest (or most similar) points. Nearest neighbor search which returns the nearest neighbor of a query point in a set of points, is an important and widely studied problem in many fields, and it has wide range of applications. We can search closest point by giving keywords as input; it can be spatial or textual. A spatial database use to manage multidimensional objects i.e. points, rectangles, etc.

Some spatial databases handle more complex structures such as 3D objects, topological coverage's, linear networks. While typical databases are designed to manage various NUMERIC'S and character types of data, additional functionality needs to be added for databases to process spatial data type's efficiently and it provides fast access to those objects based on different selection criteria. 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. 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.

This paper gives importance spatial queries with keywords [5] [6] [9] [10]. Spatial queries with keywords take arguments like location and specified keywords and provide web objects that are arranged depending upon spatial proximity and text relevancy.

Some other approaches take keywords as Boolean predicates [1] [2], searching out web objects that contain keywords and rearranging objects based on their spatial proximity. Some approaches use a linear ranking function [7] [8] to combine spatial proximity and textual relevance. Earlier study of keyword search in relational databases is gaining importance. Recently this attention is diverted to multidimensional data [3] [4] [11]. N. Rishe, V. Hristidis and D. Felipe [12] has proposed best method to develop neighbor search with keywords. For keyword-based retrieval, they have integrated R-tree with spatial index and signature file [12]. By combining R-tree and signature they have developed a structure called the IR2-tree [12]. IR2-tree has merits of both R-trees and signature files.

The IR2-tree preserves object's spatial proximity which important for solving spatial queries.

II Literature Survey

We came across several methods like spatial index, inverted index, nearest neighbor index. The first method „Spatial Index" is used to create indices in order to store the huge amount of data to be search in the form of XML documents. In this technique space required for search and the time will be greatly reduced. Second technique is Inverted Index", which acts as a brain of typical search engine indexing algorithm [6]. This optimizes the speed of the query and find the documents where the query occurs. The inverted index data structure is introduced in order to list the documents per word instead of listing the words per article. Third technique is „Nearest Neighbor Search (NNS)", also identified as closeness search, parallel search is an optimization problem for finding closest points in metric spaces. Inverted Index methods provides index instead of providing whole data which is space consuming 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 stateof- 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 GPSenabled 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 geotagged 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.

III. PROPOSED WORK

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 multiscale 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.

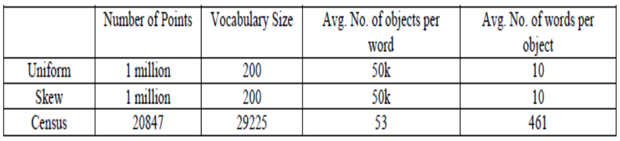
We start with the index for exact search. There are 2 main component included i.e. Inverted index ikp and Hashtable inverted index pairs (HI). We treat keyword as keys and provide it as an input to our system. There are hash bucket IDs and respective points associated with the keywords, it will find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

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.

IV. RESULTS AND DISCUSSION

The experimental evaluation of practical efficiency of our solutions with proposed and existing methods which are based on synthetic and real data. The synthetic category which consist of two sets, uniform and skew, that differ in distribution of data points and in defining a correlation between the spatial distribution and objects text documents. For the datasets, the vocabulary has 200 words, each word appears in 50k data points. In Uniform, the difference in association of words with points is completely random, whereas in skew it will be “word-locality”: points that are spatially close have almost identical text documents. Our real dataset, Census is a combination of US Census Bureau and web pages from Wikipedia.



	Number of Points	Vocabulary Size	Avg. No. of objects per word	Avg. No. of words per object
Uniform	1 million	200	50k	10
Skew	1 million	200	50k	10
Census	20847	29225	53	461

Dataset Statistics

The deficiency of IR2-tree is mainly caused by the need to verify a vast number of falsehits.To illustrate this, the figure below plots the average false hit number per query. We see an exponential escalation of the number on Uniform and Census, which explains the drastic explosion of the query cost on those datasets. Interesting is that the number of false hits fluctuates a little on Skew, which explains the fluctuation in the cost of IR2-tree. The space consumption of IR2 –tree, SI-Index on the datasets of uniform, skew, Census are explained in the figure below. IR2 Tree has much more space efficiency than any other technique but doesn‟t compensate with the expensive query time. The SI-Index accompanied by the proposed query algorithms, has presented itself as an excellenttradeoff between space and query efficiency. Compared to IR2 Tree, its superiority is very high as the factors of order magnitude is typically high than its query time.

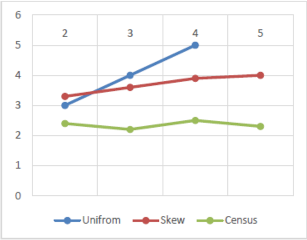


Fig 1.No. of False hits of IR2-Tree

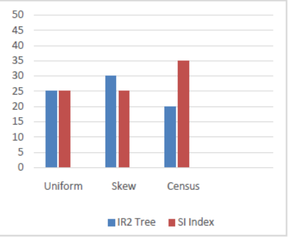


Fig. 2. Comparison of Space Efficiency

V. CONCLUSION

They are numerous number of applications of with a search engine which efficiently support novel forms of spatial queries integrated with keyword search. By all the above methods, the main goal is searching a relevant keyword with appropriate info with minimum time and with valid results. In this paper, we come to a conclusion by developing an access method called Spatial Inverted Index (SI-Index). SI Index has high space efficiency and also has the ability to perform keyword augmented NN search in time. The performance bottlenecks of SI index would be how to differentiate the keyword with searched one, if two nodes have same keyword. If the cluster is dynamically growing, the index of the cluster also keep grows 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.

REFERENCE

[1] I. De Felipe, V. Hristidis, and N. Rishe. Keyword search on spatial databases. In ICDE, pp. 656–665, 2008.

[2] 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

[3] 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.

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

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

[6] G. Cong, C. S. Jensen, and D. Wu. Efficient retrieval of the top-k most relevant spatial web objects. PVLDB, 2(1):337–348, 2009.

[7] I. De Felipe, V. Hristidis, and N. Rishe. Keyword search on spatial databases. In ICDE, pp. 656–665, 2008.

[8] Y. Zhou, X. Xie, C. Wang, Y. Gong, and W.-Y. Ma, “Hybrid Index Structures for Location-Based Web Search,” Proc. Conf. Information and Knowledge Management (CIKM), pp. 155-162, 2005.

[9] I.D. Felipe, V. Hristidis, and N. Rishe, “Keyword Search on Spatial Databases,” Proc. Int’l Conf. Data Eng. (ICDE), pp. 656-665, 2008.

[10] C. Faloutsos and S. Christodoulakis, “Signature Files: An Access Method for Documents and Its Analytical Performance Evaluation,” ACM Trans. Information Systems, vol. 2, no. 4, pp. 267-288, 1984.

[11] N. Beckmann, H. Kriegel, R. Schneider, and B. Seeger, “The R- tree: An Efficient and Robust Access Method for Points and Rectangles,” Proc. ACM SIGMOD Int’l Conf. Management of Data, pp. 322-331, 1990.

[12] G.R. Hjaltason and H. Samet, “Distance Browsing in Spatial Databases,” ACM Trans. Database Systems, vol. 24, no. 2, pp. 265-318, 1999