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

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 11, November 2015

Efficient and Flashing Nearest Neighbour Search with keywords

K. Vanajakshi Devi, P. V. Nagarjun, N. Praveen Kumar

Associate Professor, Dept. of C.S.E, Yogananda Institute of Technology &Science, JNTUA, AP, India

Dept. of C.S.E, Yogananda Institute of Technology & Science, JNTUA, Tirupati, AP, India

Dept. of C.S.E, Yogananda Institute of Technology & Science, JNTUA, Tirupati, AP, India

ABSTRACT: Nearest neighbor search in multimedia databases [7] 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 IR^2 trees [3] 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].

KEYWORDS: Spatial Inverted Index, Nearest Neighbor Search, IR^2 Trees, similarity search, Spatial Index

I. INTRODUCTION

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 IR^2 trees, which has strengths of both signature files and R-Trees. Like R-Trees, IR^2-Tree has object spatial proximity that solves spatial queries efficiently. On the other side, the IR^2-tree is able to filter a considerable portion of the objects that do not contain all the query keywords, like signature files.

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

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 11, November 2015

II. RELATED WORK

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.

TABLE 1: Comparison of DFS and Inverted Index

Parameters/Methods	Depth First Search	Inverted Index
Space	More	Less
Time Complexity	O(n+m)	O(n)
Working	Slow	Efficient

III. EXISTING SYSTEM

Present system gives the real nearest neighbour that lies quite far away from the query location, while all the closer objects missing one or any of the keywords. This system mainly focuses on finding the nearest neighbour where each node satisfies all the query keywords. This leads to low efficiency for incremental query. The problem is Implement k nearest neighbour search algorithm using for given data set and to find out closest point from give query also analyse the result fetch time and accuracy. Implement the Inverted index algorithm by extending point the k nearest neighbour and forming R tree to find closest point from given set of query and also analyse the result against time and result accuracy.

A. *IR^2-Trees:*

As mentioned earlier, IR^2 Trees are a combination of R-Trees and Signature files, which are well known technologies in spatial databases. Signature file in general refers to a hashing-based framework, whose instantiation in [1] is known as superimposed coding (SC), which is shown to be more effective than other instantiations [2]. It is designed to perform membership tests: determine whether a query word w exists in a set W of words. The whole set of words W must be scanned even for a false hit. If SC says "no", then w is not in W. It is "Yes" if it finds in it. In some context SC works as same as the classic technique called bloom filters.

The IR^2-tree is an R-tree where each (leaf or nonleaf) entry E is augmented with a signature that summarizes the union of the texts of the objects in the subtree. On conventional R-trees, Nearest Neighbour search is resolved by breadth- first algorithm.

B. *Other Relevant Trends:*

Spatial keyword search also comes under the nearest neighbour search treatment, which also gives rise to several alternative problems. As in an application it returns the result which matches all the spatial queries or not at all any of it.

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

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 11, November 2015

In Geometrical web search, each web page is associated with geographical location that is relevant to webpage content. In web search, that higher rankings are given to the pages in the same area as the location of the computer issuing the query.

The other so-called conventional work to find NN search would be 'm-closest key words problem', 'collective spatial keyword querying', 'prestige-based spatial keyword search.', 'reverse nearest neighbour queries'.

Keyword search [6] is already thoroughly studied in relational databases where the objective is to enable a querying interface that is similar to that of search engines, and can be easily used by naive users without knowledge about SQL. It also receives huge attention in spatial databases also

C. *Drawbacks:*

The IR^2-tree is the first access method for answering NN queries with keywords. Although IR^2 Trees gives pioneering solutions, it also has few drawbacks that affects its efficiency. The most important drawback is the result set may be empty or the number of false hits can be very large when the object of the final result is far away from query point. The query algorithm would need to load the documents of many objects, incurring expensive overhead.

The R-trees allow us to remedy an awkwardness in the way NN queries are processed with an I-index. Recall that, to answer a query, currently we have to first get all the points carrying all the query words. Although the distance browsing is easy with R-Trees, the best-first algorithm is exactly designed to output data points in ascending order of their distances. A serious drawback of R-Trees are its not space efficient, as the point needed to be duplicated once for every word in its description.

IV. PROPOSED SYSTEM

The drawbacks of R-Trees and inverted index can be overcome by designing a variant of inverted index that supports compressed coordinate embedding. This system deals with searching and nearer location issues and database manage multidimensional objects which resulted in failure of previous systems. To deal with spatial index as searching the entered keyword and from that find the nearest location having that keyword available and showing the location of restaurant having menus available in map. So easier to find the location of nearer restaurant in map having the available keyword.

Spatial databases [8] manages multidimensional objects and provide quick access to those objects. The importance of spatial databases is mirrored by the convenience of modeling entities of reality in an exceedingly geometric manner. The Inverted Index is compressed by coordinate encoding which makes Spatial Inverted Index (SI-Index). Query processing with an SI-index can be done either by together or merging with R-Trees in distance browsing manner. The inverted index compress eliminates the defect of a conventional index such that an SI-index consumes much less space.

Compression is already wide used technology to reduce the space of an inverted index where each inverted list contains only ids. So, the effective approach is used to record gaps with consecutive ids. Compressing an SI index is less straightforward than other approaches. For example, if we decide to sort the list by ids, gap-keeping on ids may lead to good space saving, but its application on the x- and y-coordinates would not have much effect.

V. SIMULATION RESULTS

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.

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

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 11, November 2015

TABLE 2: Dataset Statistics

	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

The deficiency of IR^2-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 IR^2-tree. The space consumption of IR^2 –tree, SI-Index on the datasets of uniform, skew, Census are explained in the figure below. IR^2 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 IR^2 Tree, its superiority is very high as the factors of order magnitude is typically high than its query time.

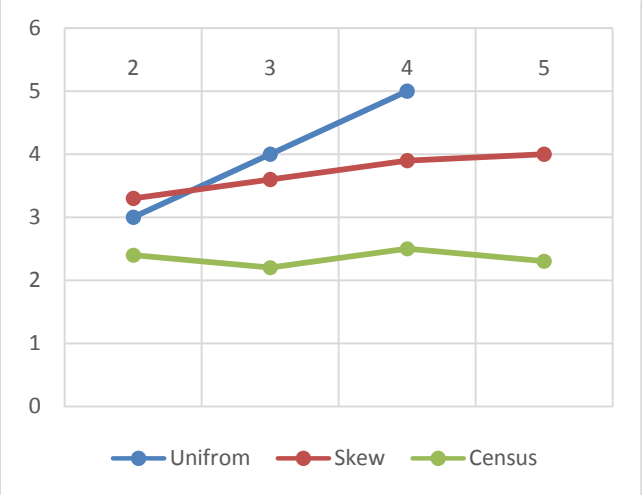

Fig.1.No. of False hits of IR2-Tree

Fig. 2. Comparison of Space Efficiency

VI. CONCLUSION AND FUTURE WORK

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.

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

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 11, November 2015

REFERENCES

1. S. Agrawal S. Chaudhuri, and G. Das. Dbxplorer: A system for keyword-based search over relational databases. In Proc. Of International Conference on Data Engineering (ICDE), pages 5–16, 2002.
2. Yufei Tao, Cheng Sheng. "Fast Nearest Neighbor Search with Keywords" IEEE transactions on Knowledge and Data Engineering.
3. Sonal S. Kasare, AnupBongale, "Efficiently Searching Nearest Neighbor In Documents Using Keywords," IJRET, Vol 2, Issue 1. 2013.
4. Vidya L. Tikone, Snehal M. Tembare, Manisha L. Narad, Bharat V. Supekar, Nilesh T. Pawar "Quick Retrieval of Nearest Neighbor by using Keywords" , IJSRET, Volume 4, Issue 4, 2015.
5. Hariharan, B. Hore, C. Li, and S. Mehrotra. Processing spatialkeyword (SK) queries in geographic information retrieval (GIR) systems. In Proc. of Scientific and Statistical Database Management (SSDM), 2007.
6. G. Bhalotia, A. Hulgeri, C. Nakhe, S. Chakrabarti, and S. Sudarshan,."Keyword searching and browsing in databases using banks.In Proc. Of International Conference on Data Engineering" (ICDE), Pages 431–440, 2002.
7. Rosslin John Robles, "Fast Nearest-Neighbor Search Algorithms Based on High-Multidimensional Data," Asia-pacific Journal of Multimedia Services Convergence with Art, Humanities and Sociology Vol.3, pp. 17-24, 20123
8. I. D. Felipe, V. Hristidis, and N. Rishe, "Keyword Search on Spatial Databases", Proc. of International Conference on Data Engineering (ICDE), pp. 656– 665.s., 2008

BIOGRAPHY

K. Vanajakshi Devi is an Associate Professor in the Computer Science & Engineering Department, Yogananda Institute of Technology & Science, JNTUA. She received Master's degree and published many journals. She has membership in ISTM, IAENG,ICST and MISTE.

P. V. Nagarjun received his Bachleor's Degree in the Computer Science & Engineering Department from Yogananda Institute of Technology & Science, JNTUA. Active member of IAENG and Secretary of CSE department.

N. Praveen Kumar received his Bachelor's Degree in the Computer Science & Engineering Department fromYogananda Institute of Technology & Science, JNTUA. He also has attended several National & International conferences. He got first place in National Symposium conducted by SVNE, Tirupati.