Closest Keyword Retrieval with Data Mining Approach

Ms. Sonali B. Gosavi[1], Dr.Shyamrao.V.Gumaste[2]

Abstract— As ***the use of internet is increasing nowadays there is need to write queries that can answer users request instantly which will be helpful for user to find the particular places, restaurants, book stores etc.. So as data is an asset, increasing size of data becomes obstacle to provide respected information to user is very difficult job. To process this huge amount of data, data mining approach is helpful. By making use of this information peoples can search the objects like places, restaurants, book stores and so on as per users request. but right now required information is not sufficient to users ,they also demand geographical information to reach to the exact location. So data mining applications are also required to mold such that which can process both queries textual as well as spatial information which need to satisfy users requirement.. For example if one considers book stores data, instead of searching all the books, one can search the book depend on its types etc. To satisfy the requirement of such queries the solution that is been used is the IR2-tree. By considering the disadvantages of IR2-tree an alternative to this is to use WIBR- Tree which is capable to handle multidimensional data and process closest item retrieval queries that can search the user's requests efficiently.***

***Index Terms*—Closest item search, nearest neighbor search, IR2-tree, WIBR-tree, R-tree, spatial inverted index.**

I. INTRODUCTION

The web is acquiring a spatial dimension with the proliferation of online objects with both an associated geo-location and a text description. Specifically, web users and content are increasingly being geo-positioned and geo-coded. At the same time, textual descriptions of points of interest, e.g., cafes and tourist attractions are increasingly becoming available on the web. This development calls for techniques that enable the indexing of data that contains both text descriptions and geo-locations in order to support the efficient processing of spatial keyword queries that take a geo-location and a set of keywords as arguments and return relevant content that matches the arguments [2].

Spatial keyword queries are being supported in real-life applications, such as Google Maps where points of interest can be retrieved, Foursquare where geo-tagged documents can be retrieved, and Twitter where tweets can be retrieved. Spatial keyword querying is also receiving increasing interest in the research community where a range of techniques have been proposed for efficiently processing spatial keyword queries.

Spatial database contains multidimensional objects like geographical data (such as points, rectangles, etc.). But because of the wide spread use of search engine, it is essential to search spatial data as well as information of that spatial data. let us focus on the basic notation of the spatial database. In the spatial database the locations of small entities are represented in the form of points in map e.g. book stores, restaurants, hospitals, hostels etc. and to represent larger entities such as gardens, grounds, etc. rectangle is used. GIS (Geographical Information System) gives all the restaurant in the address given in query. But the closest item methodology will give nearest restaurants of the given address. Now a days many search engines tries to answer spatial queries. But traditional spatial queries focus on the geometric properties of objects, such as whether point is in rectangle, or how close two points are from each other, etc. In modern era, it is quite essential to process queries that need both spatial properties as well as information about the spatial data. For example, it would be more useful if a search engine can be used to find the nearest restaurant that offers "veg afgani, kashmiri pulav and momos" all at the same time. The nearest neighbor algorithm gives the nearest restaurant among only those providing all the demanded drinks and foods. There are simple ways to process queries that contains both spatial as well as text features. For example to find restaurants for above query one can first retrieve the text features i.e. restaurant those menu is having given foods and drinks and then from resultant restaurants sort with spatial data i.e. nearest first. Alternative way to do it reversely by first finding spatial data of the query and then find the keywords. The major drawback of the above typical approach is that they are unable to process both spatial data and keywords at same time, so it is difficult to handle real time query using these approaches. This constraint may yields in missing of nearest neighbor. Previous study reflects that, authors used different approaches for retrieving data from server.[1]. Those concepts are R-tree for the spatial index and signature file for keyword based document retrieval. For that purpose they developed a structure named it as IR2 - tree which has advantages of both signature files and R-trees [3]. IR2-tree not only preserves the object of the spatial data like R-tree but also filter the exact portion of the query like signature file. Along with this IR2-tree also inherits pitfalls of signature files: false hits. In case of signature files, due to its conservative nature, sometimes it may still direct the search to some objects, even though they do not have all the keywords. Thus it needs cost to check an object which satisfying a query or not cannot be resolved using only its signature, but requires loading its whole text. It is very costly method for the Information retrieval. But the false hit problem is not only concern with signature file but also it may carry in other methodologies. Previous study reflects that, authors used different approaches for retrieving data from server. This method is very useful for the accessing

the point co-ordinates in to an inverted bitmaps with small space. WIBR stores the spatial locality of points in R-tree and textual part in inverted bitmaps at very small space.

II. RELATED WORK

A. Signature file

Signature files were introduced by Faloutsos and Christodoulakis [11] as a technique to with efficiency search a group of text documents. Signature files appear to be a promising access technique for text and attributes. Consistent
with this technique, the documents (or records) are keep consecutive in one file ("text file"), whereas abstractions of the documents ("signatures") are keep consecutive in another file ("signature file"). So as to resolve a query, the signature file is scanned 1st, and plenty of no qualifying documents are at once rejected. In general signature file refers to a hashing-based framework, whose internal representation in keyword search on spatial information is thought as superimposed committal to writing (SC). It's designed to perform membership tests that confirm whether or not a query word w exists in a very set W of words. SC is conservative, within the sense that if it says "no", then w is certainly not in W. on the opposite hand, if SC returns "yes", truth answer will be either manner, during which case the total W should be scanned to avoid a false hit.

Table I: Hash Format of string computation with L=5, M=2

word	Hashed bit string
A	00101
B	01001
C	00011
D	00110
E	10010

Table II: Spatial points with text

P	**W**p
P1	a, b
P2	b, d
P3	D
P4	a ,e
P5	c, e
P6	c, d, e
P7	b, e
P8	c,d

Consider a string of length L bits. Repeated bit M is 2. SC works in the same manner as of Bloom Filter. Let us consider a word / string of length L = 5 and repeated bit M = 2 as shown in Table I. For example: In the above table one have to assume l=5,m=2 means if one have h(a) of a word, third and fifth bit set to 1.from left. In the hashing function one will OR the bit string of all the bit. Suppose I want to calculate signature of a,b so 01101.For finding the query word wq from W the SC perform the membership test in that test it will check whether all 1's of W appear at same location. If not it is sure that wq is not in the W. But sometimes false hit may occurs like assume that one want to check 'c' is member of the a,b using the set of signature h(c)=00011.and h(a,b)=01101 in the h(c) the fourth bit is 1,and h(a,b) fourth bit is 0 so the 'c' is not the member of a,b.False hit example: Consider the membership test of SC in which 'c' will be test in (b,d) in h(b,d) the fourth bit is 0 and h(c) fourth bit is 0 and SC report that 'c' is member of (b.d) and that is false hit.

B. IR-Tree

IR2-tree is based on the R-tree. Each leaf, non-leaf entry is E which is summary of the text object. In the fig.1will illustrate the example based on data set of fig.1and hash value which is represented in the Table 1. The string value 01111 in Fig 1 is the leap entry, which is P2.The signature of the wP2 = b, d which is the document of the P2. The string 11111 is the non-leaf entry E3 is the signature of wp2+ wp6 means the signature of the non-leaf entry is the combination of the signature of leaf node. Normally R tree, best first search algorithm is the better option of the NN nearest neighbor search.[1]For IR2-tree one have to fire the query point 'q' with associated text wq. The IR2-tree generate the ascending order of the distance of MBR to 'q' (MBR is the leaf entry).Pruning the entry whose signature which is absent the any one word of Wq.so in the Fig.1 for the verification the algorithm read all node of the tree and fetch the entry of p1,p4,p6 for the word c,d because the Wq is c,d and the final answer is p6 while p2,p4 are the false hit. So in the IR2-tree avoid the false hit which was occurred in signature file.

Table III: Example of an inverted Index

word	Inverted list
a	p1,p4
b	p1,p2,p7
c	p5,p6,p8
d	p2,p3,p6,p8
e	p4,p5,p6,p7

1) Review study of Spatial invert index: Spatial invert index is the best method for accessing the keyword based retrieval .in the following list one will see the how to arrange the inverted index of points and the associated text of that point[5]. According to above list one have to create the list of inverted index which is having query word and associated point which having the same word [1]. One more point is that the list of the word is sorted order with regards point ID. So at the time of the query processing merge step will be performed on list. For example suppose one want find the point which is having words c, d because of that one will compute insertion of the inverted list. In the NN algorithm NN processing is with the IR2-tree. In that the points are retrieved from the index .Specifically NN query q with keyword set Wq the query method of I-index first determine the set of pq of all the points that will contains all the query word and then do —pq— randomly for finding the distance of pq from q.
2)Overview of Dbxplorer: The above paper is related to Keyword-Based Search over Relational Databases [2]. In day today internet is very user friendly for accessing the data. In this paper they have to give us the powerful question

language. It will find the keyword from the server and retrieve the related web pages for the user.

3) Query processing on geographic data: In the geographic search the search engine allow the user to fire the query or find the result based geographic region [7]. It's also called local search, it is also useful for the extracting the knowledge of any location. It is also useful in the GIS. For the geographical search engine one need association of text as well as spatial data.

C. Basic Technology for Geographic search

1) Geo coding: In the geo coding technique three steps are necessary that are geo extraction, geo matching, geo propagation. Geo extraction: All the elements from a page which indicate query location. That is city name, contact number distance and generate the footprint. For the second step that is geo matching that foot print of same page will be considered and in the third step that is geo propagation increase the quality scope of the geocoding by analyzing the link structure and the web pages topology [9],and from that site map they will generate tree result.

2) Geographic query processing: Each query is having text term and query footprint means geographic ranking regarding
user request. Thus in the above technique geographic ranking assign the score to each document footprint.

D. Concept of Bloom filter

Bloom filter [8] is one of the data structures that are useful for the membership queries to a set. The bloom filter needs very less space. Bloom filter avoids the false hit. It is normally used in the network. It is also used in the distributed
database. As per above section in the signature file also use the bloom filter for the membership testing. Also it used in the password data structure and spell checking.

E. View of Spatial Inverted Index

Above topic name suggest the retrieving concept that is one will retrieve the geographic location as well as associated text with the query. For the spatial ,keyword retrieval one need to first of all collective answer of the spatial query that full fill the user requirement for that one have to assume the database of spatial multidimensional object and after that one will find the set of keyword[6].

III. METHODOLOGY

This paper proposes a solution to search book stores according to keywords. Here Working of system is divided into different parts. The purpose of the system design is to plan the solution of the problem specified by the requirements document. This phase is the first step in moving from problem to the solution domain. The design of the system is perhaps the most critical factor affecting the quality of the software and has a major impact on the later phases, particularly testing and maintenance. System design describes all the major data structure, file format, output as well as major modules in the system and their Specification is decided.

A. Datasets:

For this study authors are considering Book store dataset which is prepared containing Book's details of location and types of Books available in book store.

B. Working Details:

The architecture of proposed system is shown in Fig.1. It works in following steps:

1. Merging and distance browsing: In the spatial retrieving the basic problem is the bottleneck so need to avoid it. But in the I-index is having the simple way to recover it. In the I-index one have to preserve co-ordinates of the point in one group in the inverted list and that co-ordinates of the list are used to generate the R-tree. Now discuss how to perform keyword based NN. In this technique NN queries are processed with I index. For answering the query first of all one will access all the points which is having all the query keywords in Wq. It is very useful if one find the p very early in all the relevant inverted list. In that case one can access the list of element which are having less distance with q. so the p will be discovered all points of the list.

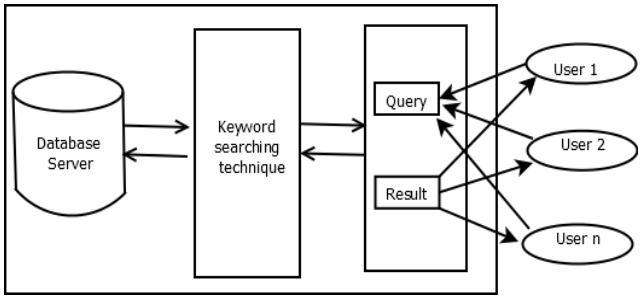

Fig. 2: System Interface

By using that one can count the number of copies of same point that will be relevant data. Consider an example if one want to NN search whose query point q and associated text is (c,d). for that search autnors have to use the list of word(c) and (d).from list 1.And now the new access order is depend on the distance of the given q. If one use the kNN then it will reported k nearest neighbor point and finish. Distance browsing is simple in the R-trees because R-tree uses best first algorithm which will give the exactly point with ascending order of the distance to q and R-trees are also the global access of the tree. For example at each step taking the point with the next point and return it. This algorithm is normally work in the condition when the Wq small. But if the Wq is large then the out performance of sequential algorithm will be merged.

2. Weighted Independent Binary Representation (WIBR)tree: WIBR-tree [10] is a variant of IR-tree. It aims at partitioning objects into multiple groups such that each group shares as few keywords as possible. To achieve this goal, the objects in D is partitioned first into two groups using the most frequent word w1: one group whose objects contain w1 and the other group whose objects do not. Then it partition each of these two groups by the next frequent word w2. This process is repeated iteratively until each partition contains a certain number of objects. After partitioning, each group of objects becomes the leaf node of the WIBR-tree. Afterwards the tree

is constructed following the structure of the IR-tree. When used for processing Boolean queries, the WIBR-tree [10] uses the inverted bitmap to replace the inverted file, which is denoted as the WIBR-tree, where a bitmap position corresponds to the relative position of an entry in its WIBR-tree node. The length of a bitmap is equal to the fan out of a node.

IV. RESULTS

	bookstoreid	bookstore	boc	address	latitude	longitude
	2	Pragati E	BT001	Shukrawa	18.5074	73.8567
	4	Anant Boc	BT001	Shaniwar	18.5186	73.8503
	1	Arihant E	BT001	ABC, Pur	18.5186	73.8503
	3	Vaibhav E	BT002	Narayan	18.5146	73.8484
	5	Agarwal E	BT002	Guruwar	18.507	73.8605
	6	Sanket Bc	BT003	Shivaji	18.5211	73.8566
*		(NULL)	(NULL	(NULL)	(NULL)	(NULL)

Fig.2: Spatial data of Book store

	booktypeid	booktypename	description
	BT001	Inspirational	(NULL)
	BT002	Historical	(NULL)
	BT003	Technical	(NULL)
	BT004	SciFi	(NULL)
	BT005	Political	(NULL)
	BT006	Fiction	(NULL)
	BT007	NonFiction	(NULL)
	BT008	Comics	(NULL)
	BT009	Spiritual	(NULL)
*		(NULL)	(NULL)

Fig.3: Book types table

Fig.4: Output window of book store searching

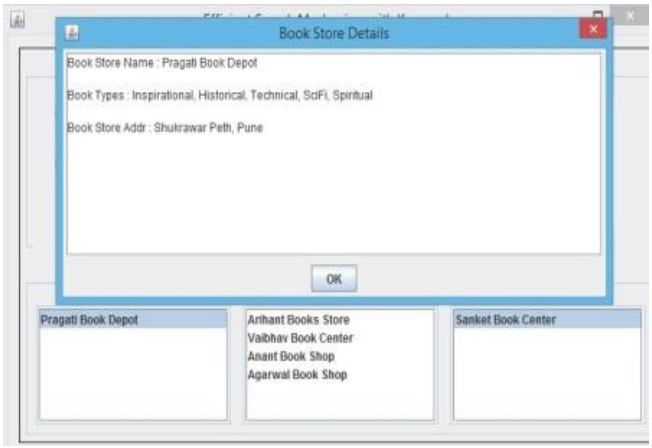

Fig.5: Output window of book store details

V. CONCLUSION

This paper provides solution to problem of spatial keyword search and improves the performance limitations of current approaches. This system proposes a solution which is dramatically faster than current approaches and is based on a WIBR- tree. In particular authors used the WIBR-Tree and showed how it is better than prior approaches. An efficient incremental algorithm was presented that uses the WIBR Tree to answer spatial keyword queries.

ACKNOWLEDGMENT

The authors would like to thank the researchers as well as publishers for making their resources available and teachers for their guidance. We also thank the college authority for providing required infrastructure and support.

REFERENCES

[1] Yufei Tao and Cheng Sheng, "Fast Nearest Neighbor Search with Keywords," National Research Foundation of Korea, GRF 4166/10,4165/11, and 4164/12 from HKRGC .

[2] 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 a 16, 2002.

[3] N. Beckmann, H. Kriegel, R. Schneider, and B. Seeger. The R*-tree, "An efficient and robust access method for points and rectangles", In Proc. Of ACM Management of Data (SIGMOD), pages 322 a 331, 1990.

[4] 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 a 440, 2002.

[5] X. Cao, L. Chen, G. Cong, C. S. Jensen, Q. Qu, A. Skovsgaard, D. Wu, and M. L. Yiu, "Spatial keyword querying," In ER, pages 16 a 29, 2012.

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

[7] X. Cao, G. Cong, C. S. Jensen, and B. C. Ooi, "Collective spatial keyword querying". In Proc. of ACM Management of Data (SIGMOD), pages 373 a 384, 2011.

[8] B. Chazelle, J. Kilian, R. Rubinfeld, and A. Tal. "The bloomier filter: an efficient data structure for static support lookup tables". In Proc. of the Annual ACM-SIAM Symposium on Discrete Algorithms (SODA), pages 30 a 39, 2004.

[9] Y.-Y. Chen, T. Suel, and A. Markowetz, "Efficient query processing in geographic web search engines". In Proc. of ACM Management of Data (SIGMOD), pages 277 a 288, 2006.

[10] D. Wu, M. L. Yiu, G. Cong, and C. S. Jensen, "Joint top-k spatial keyword query processing". IEEE TKDE, 24(10):18891903, 2012.

[11] C. Faloutsos and S. Christodoulakis. Signature files: An access method for documents and its analytical performance evaluation. ACM Transactions on Information Systems (TOIS), 2(4):267–288, 1984.

Ms. Sonali B. Gosavi received Bachelor of engineering degree in Computer Engineering from AISSMS in 2008. Now pursuing Master of Engineering in computer Engineering from SPCOE, Dumbarwadi, pune university.

Dr. S.V.Gumaste, currently working as Professor and Head, Department of Computer Engineering, SPCOE-Dumberwadi, Otur. Graduated from BLDE Association's College of Engineering, Bijapur, Karnataka University, Dharwar in 1992 and completed Post- graduation in CSE from SGBAU, Amravati in 2007. Completed Ph.D (CSE) in Engineering & Faculty at SGBAU, Amravati. Has around 22 years of Teaching Experience.

