Available Online at www.ijcsmc.com

International Journal of Computer Science and Mobile Computing

A Monthly Journal of Computer Science and Information Technology

ISSN 2320–088X

IJCSMC, Vol. 3, Issue. 12, December 2014, pg.37 – 43

RESEARCH ARTICLE

A Tool for Processing Spatial Queries with Multiple Keyword Support

Premkumar Reddy K[1], V.Santosh Kumar[2], Sirisha K L[3]

Student, Department of CSE, Sreyas Institute of Engineering and Technology, Hyderabad, India [1]

Associate Professor, Department of CSE, Sreyas Institute of Engineering and Technology, Hyderabad, India [2]

Asst.Professor, Dept. of CSE, Keshav Memorial Institute of Technology, Hyderabad, India [3]

[1] kattapremkumarreddy@gmail.com, [2] Veenu.santosh@gmail.com, [3] klssirisha@gmail.com

---

Abstract-- Spatial Data Mining (SDM) is the process of mining spatial databases. Spatial databases contain details of geographical objects. Spatial data is generally associated with non-spatial data as well. Queries on spatial databases can also predicates to obtain the required results. However, the predicates are pertaining to the geographical features of spatial objects. The applications that use spatial data mining need to use both spatial and non-spatial predicates in order to achieve high quality results. For speeding up the process of spatial queries, IR2-tree is used. There are some drawbacks of this data structure. In order to overcome this spatial inverted index came into existence which improves query processing significantly better. However, it supports search with a single keyword. In this paper we proposed a solution with multi-keyword search. Our prototype application demonstrates the proof of concept.

Index Terms – Spatial data mining, keyword search, nearest neighbor search

---

I. INTRODUCTION

Spatial database contain objects in the space. These objects are represented using special data types such as lines, points and rectangles. The spatial databases are built in such a way that they provide faster response to queries. Spatial databases contain spatial and non-spatial information so as to support comprehensive queries with spatial and non-spatial predicates. Nearest neighbor search is one of the commonly used search mechanism on spatial databases. For instance a customer wants to have a house for rent with other qualities. The other qualities might include a school within 1 KM distance, a hospital in 1.5 KM distance and a play ground in half KM. The results that meet this nearest neighbor query will surface so as to help the customer to make a well informed decision. This kind of search

can be made on various spatial objects such as hotels, lakes, cities, rivers etc. Geographical database has data that can be understood as different layers.

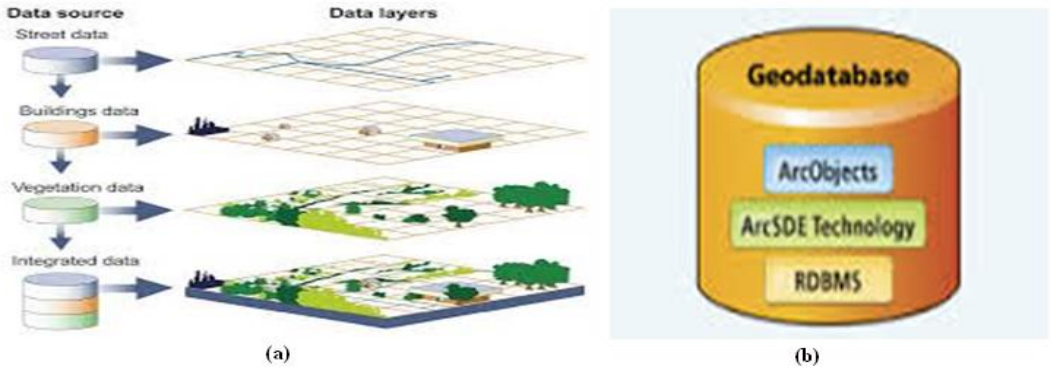

Figure 1 – Illustrates data layers in geodatabase

As can be seen in Figure 1, there are many layers in database (a) which are stored in a geographical database (b). The data which is in the form of layers such as streets data buildings data, vegetation data and integration data is subjected to frequent queries.

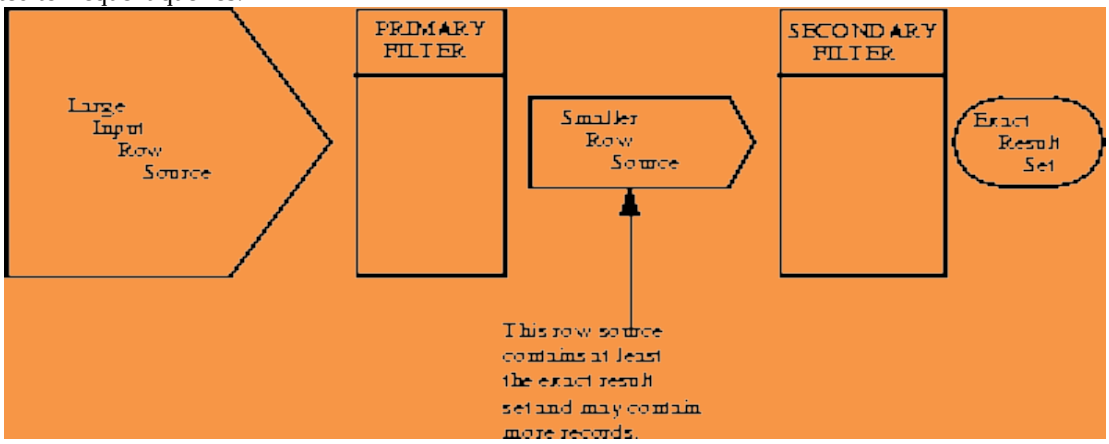

Figure 2 – General query processing on spatial databases

As shown in Figure 2, the spatial query processing involves primary filtering and secondary filtering. The primary filtering reduces search space while the secondary filtering gives rise to exact results. Spatial data has many types of geographical data. Queries can be made on such data so as to get results that satisfy given criteria. The queries generally have a mix of spatial and non-spatial features for better granularity. Quality of results cannot be obtained using only spatial predicates as the non-spatial predicates can qualify spatial objects to bring about useful results. Of late researchers focused on multi-dimensional data as explored in [1], [2] and [3]. Other important researches are spatial index and signature file [4] and R-tree [5]. These two features are combined well in [3] for best method to have nearest neighbor search. This solution was named IR2-tree [3]. However, this solution cannot have multiple keywords to examine all possible spatial objects. Therefore it resulted in false hits as well.

To overcome the drawbacks of IR2-tree, recently, Tao and Sheng [6] proposed a spatial inverted index based on conventional inverted index. This is a new access method that is successful in incorporating point coordinates well. However, it too does not support multiple keywords in query processing. In this paper we proposed a solution that supports multiple keywords. We have built a prototype application that demonstrates the usefulness of the application for fast nearest neighbor search. Our contributions in this paper include building a prototype web application based on the concepts proposed by Tao and Sheng [6] which demonstrates the spatial

nearest neighbor queries with multiple keywords. The remainder of the paper is structured as follows. Section II reviews literature pertaining to spatial nearest neighbor search. Section III provides information related to the proposed system. Section IV presents the prototype application and results while section V concludes the paper.

II. RELATED WORKS

A nice survey on spatial nearest neighbor queries can be found in [7]. Keyword based nearest neighbor queires was considered in [8] which is similar to the work done in [6] and this paper. The approach in [8] is to make use of IR for computing the relevance among the objects and query. It makes use of similarity measure to return object with highest similarity. Geographical web search concept was explored in [1], [2] and [11]. M-closest key-word problem was explored in [9]. It considered a set of points carrying only a single keyword each. This is very different from the work done in this paper. Prestige based spatial keyword search was proposed by Cong et al. [12] for evaluating similarity in spatial query processing. In [13]nearest neighbor queries and keyword search are combined in order to provide Search Engine kind of user interface to end users.

IR2-tree [3] is the most related work which is similar to our work. The IR2-tree combines the features of R-tree with signature files. Therefore it has the good features of both. Best-first algorithm for nearest neighbor search was proposed in [14] that makes use of well known techniques on spatial databases. Signature file is nothing but a hashing-based framework that is based on the concept of superimposed coding as explored in [3]. The superimposed code (SC) was first demonstrated in [4]. SC is conservative and specific in functionality. However, in IR2-tree false hits are possible as SC forced while set of words is to be scanned. As explored in [3] SC works in similar fashion to that of bloom filter. This process is better illustrated in Table 1.

WORD	HASHED BIT STRING
a	00101
b	01001
c	00011
d	00110
e	10010

Table 1 – Bit string computation with m=2 and l=5

The bit length is considered 5 while the 2 mutually independent choices are considered (m). The bit string computation process can be found in [14]. The problem with [3] is that as the number of words grows in size, scanning the entire list become tedious. When the list is not scanned it may result in false hits. As a solution to this problem inverted indexes were introduced in [6]. As inverted indexes became effective access methods, an example for inverted index appears as shown in Table 2.

WORD	INVERTED LIST
a	p1 p4
b	p1 p2 p1
c	P5 p6 p8
d	P2 p3 p6 p8
e	P4 p5 p6 p7

Table 2 – Illustrates inverted list

As scanning an inverted index is cheaper this solution is preferred in [6]. More details can be found on inverted index in [6].

III. PROPOSED SOLUTION

The proposed solution is based on the spatial inverted index (SI-Index) concept proposed in [6]. In fact it is an improved form of I-index with compressed coordinates. The query process in done by using either merging or in a distance browsing fashion. The convention index tree has defects which can be overcome with compression. Moreover the SI-index [6] takes very less space besides providing fast nearest neighbor results. Compression techniques have been around for many years in order to reduce size of index. However, it is very effective when gaps are between consecutive ids when compared with precise ids. For instance, consider a set of integers {2,3,6,8}. The gap – keeping approach stores values such as {2,1,3,2} and no information is lost as the original space can be constructed precisely. The overhead incurred is however, the extra computation cost which is highly compensated by reducing I/Os.

The gap-keeping approach is not very useful. When gaps are very smaller only it might be useful to some extent. The compressed SI-Index has a triplet to represent each point. Therefore it is not straight forward solution. A space filling curve is built for efficient processing. It is also known as Z-curve and the values considered are known as Z-values. Pseudo ids are used instead of real ids for representing object details. This is done for better sorting procedures. Sorting values in Z-curve also sorts values automatically. The compression scheme has less complexity and saves space this increasing the speed of query processing when applied to nearest neighbor search on spatial databases. The SI-index works in tandem with two algorithms which are based on distance browsing and merging. Both synthetic and real time data is used for experiments. The dimensionality is fixed at 2 while the axis ranges 0 to 16383. Census is the real dataset used while synthetic data is constructed for the purpose of experiments. More details can be found on SI-index in [22]. The space cost is less when SI-index is used. As algorithms make use of the index, the query processing is efficient and consumes much less space.

IV. PROTOTYPE APPLICATION AND RESULTS

We built a prototype application using JAVA/J2EE platform. It is a web based application tested with environment such as a PC with 4 GB RAM, core 2 dual processor running Windows 7 operating system. The application was tested using web server Tomcat 7. The application has provision for building spatial database and also spatial queries. Figure 2 shows a snapshot of spatial query with multiple keywords.

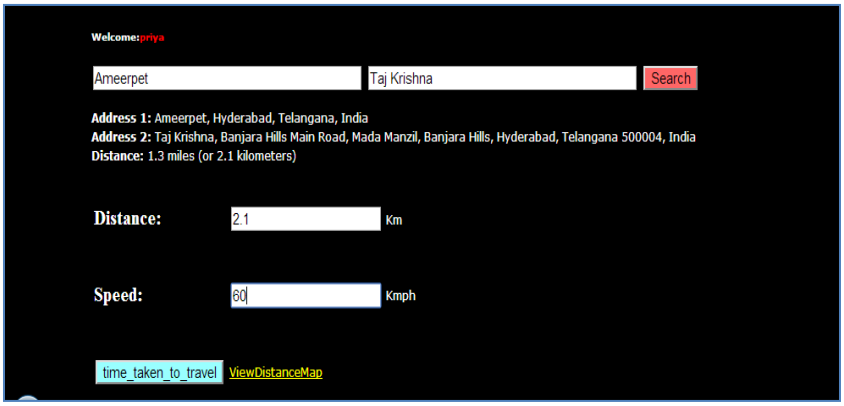

Figure 2 – Spatial nearest neighbor query with multiple keywords

As can be seen in Figure 2, it is evident that the query has both spatial and non-spatial predicates. Multiple keyword searches are supported. Distance and speed are the non-spatial predicated provided with KM, KMPH as measures. Ameerpet and Tajkrishna are the two keywords given at a time. The results are presented in Figure 3.

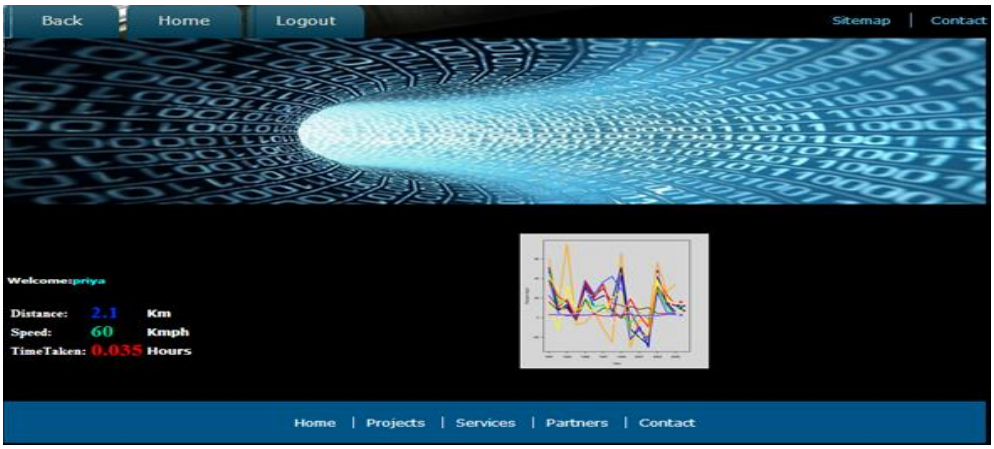

Figure 3 – Query results with a visual distance map

As can be seen in Figure 3, the results are presented textually and also with intuitive visual distance map. This makes sense as the spatial query is given with spatial and non-spatial predicates and the search process mechanism provided in the previous section was carried out.

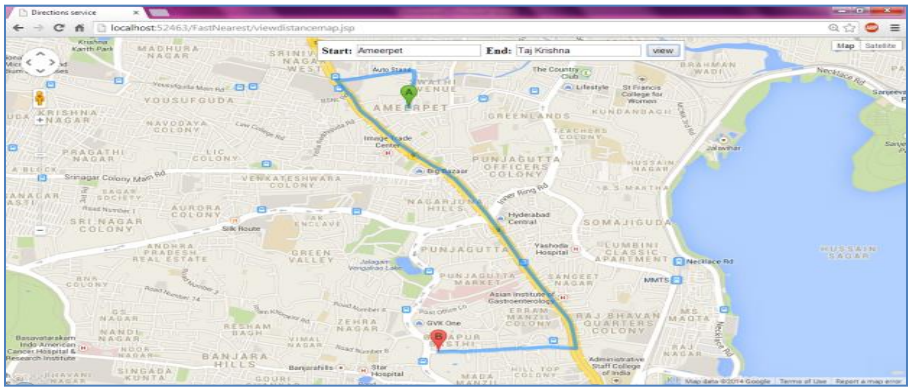

Figure 4 - The search results through Google maps

As can be seen in Figure 4, it is evident that the results are presented using web mapping service known as "Google Maps" powered by Google.

V. CONCLUSION AND FUTURE WORK

In this paper our focus is more on spatial data mining with multiple keyword search. The existing spatial query processing makes use of spatial query with predicates that are related to geographical objects. However, the real world applications do need the spatial and non-spatial predicates in order to obtain more useful information. IR2-tree proved to be a good solution for many years. However, it has certain drawbacks. Tao and Sheng [6] proposed spatial inverted index which improved the performance of spatial queries pertaining to nearest neighbor. However, it supports search with a single keyword. In this paper we built a prototype that supports spatial approximate string search with multiple keywords. The empirical results are encouraging. In future we improve the prototype to support features of precision agriculture.

REFERENCES

[1] Y. Zhou, X. Xie, C. Wang, Y. Gong, and W.-Y. Ma. Hybrid index structures for location-based web search. In Proc. of Conference on Information and Knowledge Management (CIKM) , pages 155–162, 2005

[2] R. Hariharan, B. Hore, C. Li, and S. Mehrotra. Processin g spatial- keyword (SK) queries in geographic information retrieval (GIR) systems. In Proc. of Scientific and Statistical Database Management(SSDBM) , 2007.

[3] I. D. Felipe, V. Hristidis, and N. Rishe. Keyword search on spatial databases. In Proc. of International Conference on Data Engineering (ICDE) , pages 656–665, 2008.

[4] C. Faloutsos and S. Christodoulakis. Signature files: A n access method for documents and its analytical performance evalua tion. ACM Transactions on Information Systems (TOIS) ,
[5] N. Beckmann, H. Kriegel, R. Schneider, and B. Seeger. The R*-tree:An efficient and robust access method for points and rectangle es. In Proc. of ACM Management of Data (SIGMOD), pages 322–331, 1990.

[6] Yufei Tao and Cheng Sheng, "Fast Nearest Neighbor Search with Keywords", IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, 2013, p1-13.

[7] 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 16–29, 2012.

[8] G. Cong, C. S. Jensen, and D. Wu. Efficient retrieval of th e top-kmost relevant spatial web objects. PVLDB , 2(1):337–348, 2009.

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

[10] D. Zhang, Y. M. Chee, A. Mondal, A. K. H. Tung, and M. Kitsu re- gawa. Keyword search in spatial databases: Towards searching by document. In Proc. of International Conference on Data Engineering (ICDE) , pages 688–699, 2009.

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

[12] J. Lu, Y. Lu, and G. Cong. Reverse spatial and textual k ne arrest neighbor search. In Proc. of ACM Management of Data (SIGMOD), pages 349–360, 2011.

[13] I. D. Felipe, V. Hristidis, and N. Rishe. Keyword search on spatial databases. In Proc. of International Conference on Data Engineering (ICDE) , pages 656–665, 2008.

[14] G. R. Hjaltason and H. Samet. Distance browsing in spatial databases. ACM Transactions on Database Systems (TODS) ,24(2):265–318, 1999.

AUTHORS

Premkumar Reddy K. received the B.Tech degree in computer science and technology from Jawaharlal Nehru Technological University, Hyderabad, India in 2011. He is currently working towards his M.Tech degree in Sreyas Institute of Engineering and Technology, Hyderabad, India. His research interests include data mining and cloud computing.

Email: kattapremkumarreddy@gmail.com

Vennu Santosh Kumar received the Masters degree in Computer Science and Engineering in the year 2010. He is Microsoft Certified System Engineer & CISCO Certified Network Administrator, he worked as a System Engineer in WIPRO Technologies (INDIA). In 2011 he joined as an Associate Professor at Sreyas Institute of Engineering and Technology in Computer Science Department. He has been involved in several tutorials, workshops, technical paper presentations .His research interests are focused on Computer Networks, Network Security & Mobile Computing.

Sirisha K L S received the Masters degree in Computer Science and Engineering in the year 2010. In 2010 she joined as an Assistant Professor at Keshav Memorial Institute of Technology in Computer Science and Engineering Department. Her research interests are focused on Network Security, Information Retrieval & Machine learning.

