iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 3 **Issue:** Issue I **Month of publication:** May 2015

DOI:

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)

Fast Accurate Mining on Spatial Database Using Keywords

D.Amutha Priya[1], Dr. T.Manigandan[2]
[1]PG Scholar, [2]Principal and Professor, Department of Computer Science and Engineering
P.A College of Engineering and Technology, Pollachi, Tamil Nadu, India

Abstract-Data mining is the knowledge discovery in databases. It is the analysis of data for relationships that have not previously been discovered. Spatial data mining is the application of data mining methods to spatial data. Weighted Set Cover is a straight forward method for searching the object based upon their spatial and textual features. Weighted Set Cover starts with the user queries which decompose it into partial queries. These partial queries are get executed and result objects are get merged together. Weighted Set Cover significantly reduces the false hits. Weighted Set Cover is built along with the spatial inverted index to avoid the multiple scan. Ranking spatial object can be done based upon its features. Each object has its own features. Influence score method gives score to the objects based upon their distance from the query location. Resultant objects are get sorted by object score and listed to the user.
Keywords: Weighted Set Cover, Ranking, Influence score

I. INTRODUCTION

Spatial data mining is the process of discovering interesting and previously unknown, but potentially useful patterns from large spatial datasets. Extracting interesting and useful patterns from spatial datasets is more difficult than extracting the corresponding patterns from traditional numeric and categorical data due to the complexity of spatial data types, spatial relationships and spatial autocorrelation. Conventional spatial queries, such as range search and nearest neighbor retrieval, involve only conditions on objects geometric properties. Many modern applications call for novel forms of queries that aim to find objects satisfying both a spatial predicate and a predicate on their associated texts. For example if the users are interested to find the nearest restaurants that offers Chinese and South Indian food all the same time. There are two ways to support this queries that combine spatial and text features. For the above query, we could first fetch all the restaurants whose menu contains the set of keywords {Chinese Food, South Indian Food} and then from retrieved restaurants, find the nearest one. Similarly the same query result can also find out reversely by targeting first spatial condition and then textual condition. Nearest neighbor search is to find the nearest object contains the set of keywords. Information Retrieval R-Tree efficiently retrieves the object containing set of keywords and it has main drawback of false positive. The Weighted Set Cover with spatial inverted index is used to remove the false positive. Ranking spatial object can be done based upon its features. Each object has its own features. Influence score method gives score to the objects based upon their distance from the query location. Resultant objects are get sorted by object score and listed to the user.

II. RELATED WORK

In [1] Norbert Beckmann et al. (1990) focused on R*-Tree, popular access methods for rectangle in spatial database. R*-Tree incorporates a combined optimization of area, margin and overlap of each enclosing rectangle in the directory. It efficiently supports point and spatial data at the same time. R*-Tree is completely dynamic, insertion and deletions can be inter mixed with queries and know periodic global reorganization is required. R*-Tree structures must allow overlapping directory rectangles.

In [5] Bernard Chazelle et al. (2006) introduced the Bloomier filter, a data structure compactly encoding a function with static support in order to support approximate evaluation queries. In Bloomier filter, the classical bloom filters hashing schemes main attribute space efficiency is achieved at the expense of a tiny false positive rate. The bloom filters can handle only set membership queries. Bloomier filter can deal with arbitrary function.

In [11] Chen li and Biji Hore (2007) focused on location based information retrieval. Geographical Information System (GIS) database is invaluable for many applications such as disaster response, national infrastructure protection and crime analysis. A framework is proposed for Geographical Information Retrieval (GIR) system and focus on indexing strategies can process Spatial Keywords (SK) queries effectively. Two type of indexing mechanism is used. First method is separate index for spatial and text attribute. Second method is hybrid indices techniques combine the spatial and inverted file indices.

In [14] Xin Cao and Gao Cong (2011) focused on retrieving a group spatial web objects such as the group's Keywords cores the query's Keywords and the objects are nearest to the query location and have the lowest inter object distances. Cao and Cong aimed to find a group of objects cover the keywords and the some of their spatial distances to the query is minimized. Approximation algorithm is used. It is a straight forward method of adapting the greedy algorithm is decompose the given user query q dynamically into a sequence of partial queries, each containing a different set of keywords. These partial query's results are merged together to find relevant objects.

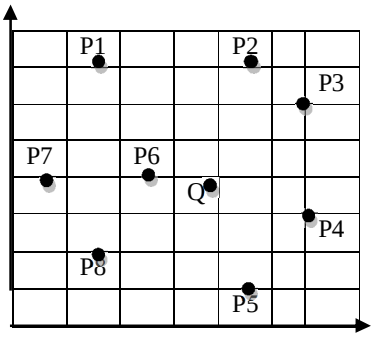
In [9] Ian De Felipe et al. (2008) focused on finding objects closest to a specified location contains set of keywords. A method to answer top k spatial queries is effectively presented the method has tight of data structure and algorithms used in spatial database search and Information Retrieval. I.D.Felipe builds a method consist of an Information Retrieval R-Tree. It is the structure based on R-tree integration with information retrieval technique.

III. PROBLEM DEFINITION

Let P be the set of multidimensional points. Each point $p \in P$ is associated with a set of words denoted as W_p. For example, P stands for a restaurant, W_p can be its menu. User entered query contains both the spatial and textual features. Nearest neighbor query specifies a point q and a set word W_q of keywords. The relevant point or object to the given query is given by P_q.

$P_q=\{p \in P \mid W_q \subseteq W_p\}$

$P_q$is the set of objects in P, document contains all the keywords in W_q.



P	W_p
P_1	{a,b}
P_2	{b,d}
P_3	{d}
P_4	{a,e}
P_5	{c,e}
P_6	{c,d,e}
P_7	{b,e}
P_8	{c,d}

Figure 1.Location of place in spatial database and description of the place.

IV. WSC WITH SPATIAL INVERTED INDEX

A geographic query is composed of query keywords and a location. A geographic search engine retrieves documents that are most textually and spatially relevant to the query keywords and the location and ranks the retrieved documents according to their joint textual and spatial relevance to the query. Spatial Inverted Index (SI-Index) together with a weighted set algorithm facilitates four major tasks in document searches, namely spatial filtering, Textual filtering, Relevance computation and Document Ranking.

Nearest neighbor search is an optimization problem for finding closest points of the objects current location. Weighted Set Cover is a greedy algorithm technique. It is the straightforward method decompose dynamically the given user query into sequence of partial queries. These queries are get evaluated and the results are found out by using spatial inverted index. The resultant objects are merged together and further scanning is done to retrieve correct object.

Spatial inverted index is one in which the spatial data are indexed with the help of their description. Inverted Index is an index data structure storing a mapping from content, such as words or numbers to its locations in a database file. The purpose of inverted index is the fast full text search. The answer to the spatial query is a list of objects ranked according to a combination of their distance to the query keywords. Objects are ranked by distance and keywords are applied as to eliminate objects. Search result can be viewed as per the distance from user's location. In this user search location has been depicted in map. User can also get the information like the food festival, offers in particular restaurant or shop nearby his current location.

A. Spatial Inverted Index With Wsc Algorithm

Input : User entered keyword.
Output : List of object C satisfies user entered keyword denotes by C.
Step1 : Initially C = ∅.
Step2 : Decompose the given user query set cover Q= {Q_1,Q_2, Q_n}.
Step3 : Pick set Qi in the cover.
Step4 : If Qi is the set of elements aren't covered yet.
Step5 : If Qi ≠ ∅
C=C ∪ Si
Si is the set of objects satisfying the Qi keyword.
Step6 : Repeat from step3 until all the set is covered.
Step7 : Return C.

User entered keywords is decomposed into partial queries. The partial queries are get executed. The partial queries answers are merged together. All the resultant objects are added to C. The retrieved objects should contain the user entered keywords.

V. RANKING OBJECTS

Ranking objects is very important task in various applications. For a given query in some location, needs to get a set of nearest objects that quality a particular condition. Ranking spatial objects can be done based on their features. Every object has its own features. Based on the features, the objects could be provided with some scores. In this paper we need to rank the data with its distance and score.

A. Ranking Algorithm

Spatial data are denoted by D, Features of the objects are denoted by F.

D={D1,D2,D3….Dn}
F={F1,F2,F3….Fn}

Step1 : Initialize D,F.
Step2 : If (D1 contains feature F1) then add the data D1 to R.
Step3 : For all features in F apply step 2.
Check all the data in D.
Step4 : Insert result into R.
Step5 : Sort the result with the objects score.
Step6 : The objects in the result set R are listed to the user.

B. Object Score

Influence score is the best method that gives the score to the objects. This method assigns higher weights to the data which are closest to the query locations. The final score of the objects is computed by multiplying its quality with weights. In this paper we are going to search nearby hotels and tourist place that should also contains the user entered keywords.

In this figure q represent the query locations and P_1, P_2, P_3, $P_4$and P_5 is the point in the spatial database. Influence score method assigns highest score to the object that is very closest to the query location. P_2 and P_4 are given with high score. The lower score are assigning to the P_1, P_3 and P5.

User may give review about each hotels and tourist place. The data are ranked based upon both the user reviews and distance to the query locations.

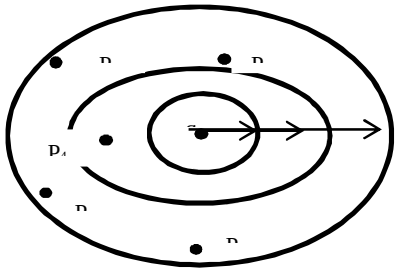


Figure 2. Influence Score Method

VI. RESULT

Weighted Set Cover is applied on a database which had more than 500 spatial data entries with their latitude and longitude points. Materialized datasets, materialization refers to collecting the data that are having specific features. The data was taken from Google map. Some additional information for the data like hotels and tourist are added to the database. The full data is shown in Figure 3.

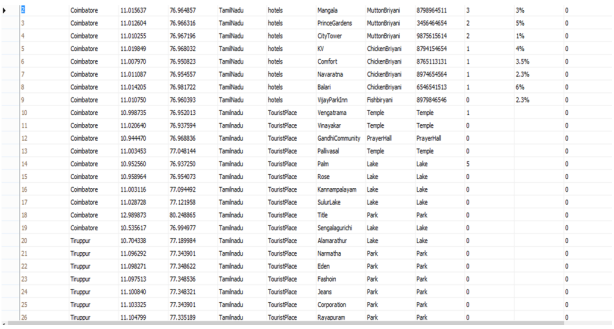


Figure 3. Data from the Google map.

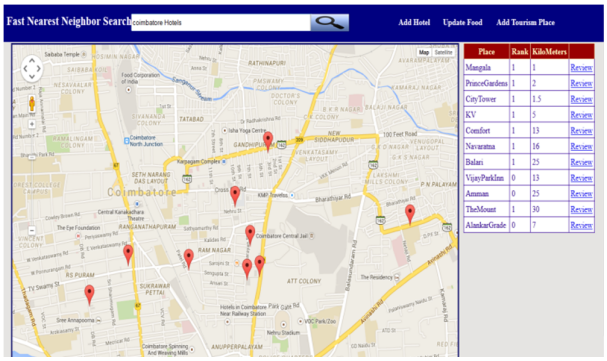


Figure 4. Hotels with rank and distance

The following displays the best tourist place in the specified area.

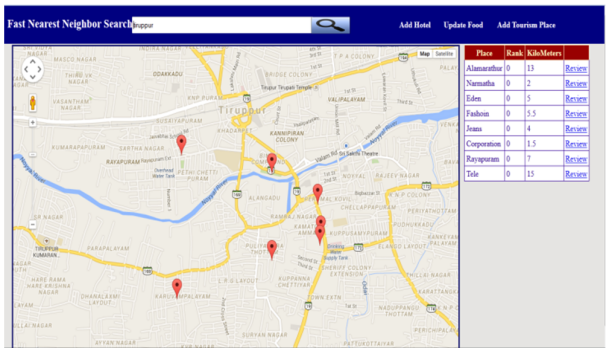


Figure 5. Result showing best tourist place

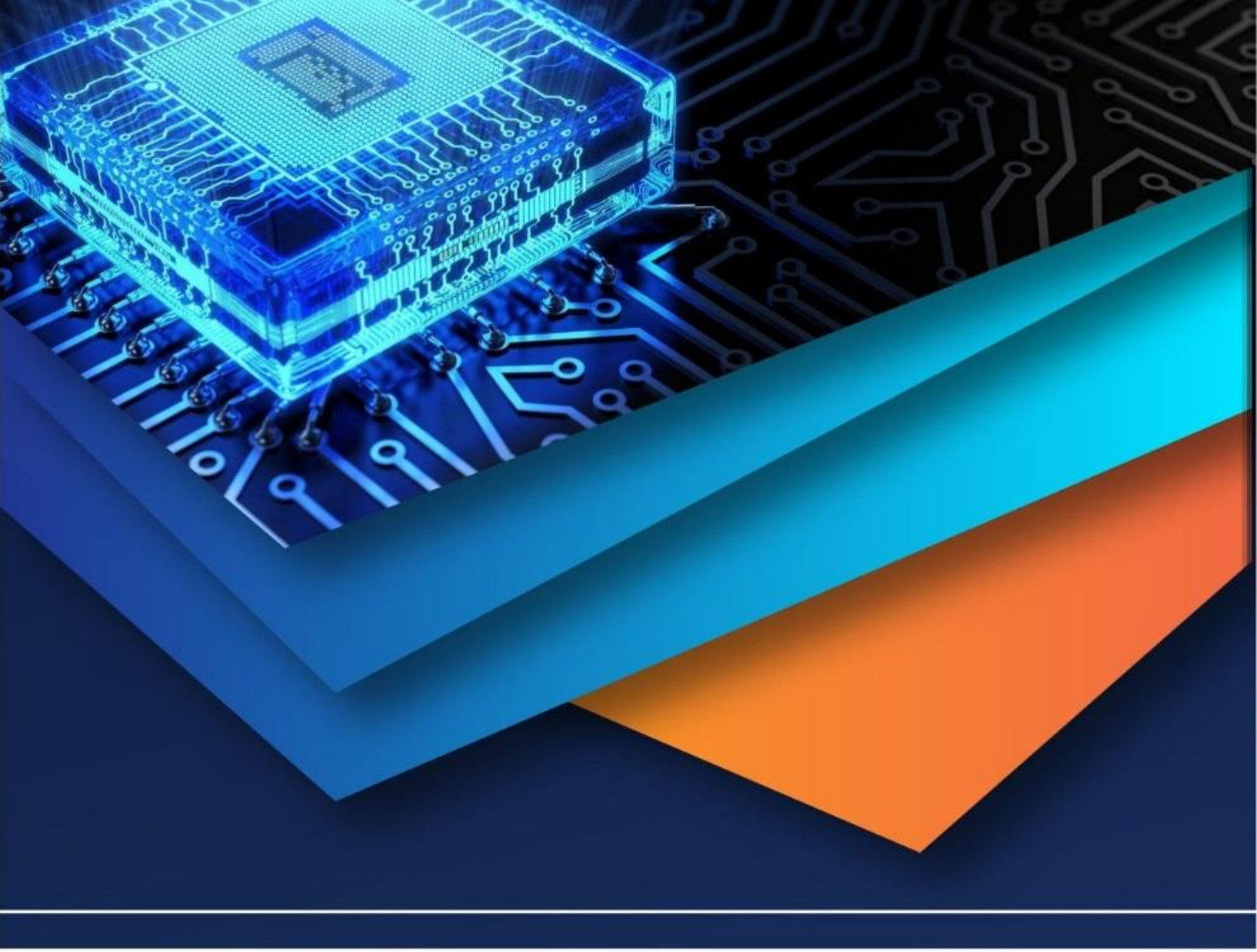
VII. CONCLUSION

By giving score to all the data in the database, the resultant data become more accurate. It provides most useful and accurate data that makes useful to make the right decision. In this paper WSC are used for both the content and location searching. WSC resultant objects are get ranked according to the score of the objects. WSC algorithm reduced the false hits that are

appeared in the conventional location searching algorithm.

REFERENCES

[1] 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 322–331,1990.
[2] Kamel and C. Faloutsos, "Hilbert R-Tree: An Improved R-Tree Using Fractals", Proc. Very Large Data Bases (VLDB), pp. 500-509, 1994.
[3] V. Hristidis and Y. Papakonstantinou. "Discover: Keyword search in relational databases", In Proc. of Very Large Data Bases (VLDB),pages 670–681, 2002.
[4] S. Agrawal, S. Chaudhuri, and G. Das, "Dbxplorer: A System for Keyword Based Search over Relational Databases", Proc. Int'l Conf. Data Eng. (ICDE), pp. 5-16, 2002.
[5] 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 30–39, 2004.
[6] N. Bruno, L. Gravano, and A. Marian, "Evaluating Top-k Queries over Web-AccessibleDatabases", Proc. IEEE Int'l Conf. Data Eng. (ICDE), 2002.
[7] 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.
[8] 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 277–288, 2006.
[9] 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
[10] 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.
[11] R. Hariharan, B. Hore, C. Li, and S. Mehrotra. "Processing spatial-keyword(SK) queries in geographic information retrieval (GIR)systems", In Proc. of Scientific and Statistical Database Management(SSDBM), 2007.
[12] D. Zhang, Y.M. Chee, A. Mondal, A.K.H. Tung, and M. Kitsuregawa, "Keyword Search in Spatial Databases: Towards Searching by Document", Proc. Int'l Conf. Data Eng. (ICDE), pp. 688-699, 2009.
[13] E. Chu, A. Baid, X. Chai, A. Doan, and J. Naughton, "Combining Keyword Search and Forms for Ad Hoc Querying of Databases", Proc. ACM SIGMOD Int'l Conf. Management of Data, 2009.
[14] X. Cao, G. Cong, and C.S. Jensen, "Retrieving Top-k Prestige- Based Relevant Spatial Web Objects", Proc. VLDB Endowment, vol. 3, no. 1, pp. 373-384, 2010.
[15] Man Lung Yiu; Hua Lu; Nikos Mamoulis; Vaitis, M., "RankingSpatial Data by Quality Preferences," Knowledge and DataEngineering, IEEE Transactions on , vol.23, no.3, pp.433,446,March 2011
[16] G. R. Hjaltason and H. Samet, "Distance Browsing in Spatial Databases," TODS, vol. 24(2), pp.265–318, 1999.
[17] N. Mamoulis, K. H. Cheng, M. L. Yiu, and D. W. Cheung.Efficient Aggregation of Ranked Inputs. In ICDE, 2006.



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY