A Review: Data Mining Technique Used for Searching the Keywords

Rasika Ashokrao Dugane
Computer Science and Engineering
HVPM's COET
Amravati, Maharashtra, India
dugane.rasika@gmail.com

Prof. A. B. Raut
Computer Science and Engineering
HVPM's COET
Amravati, Maharashtra, India

Abstract **-** Convential Spatial queries contains range search, nearest neighbor retrival involve only conditions on object geometric properties. Today, many modern applications call for innovative kind of queries that aim to find objects satisfying both a spatial predicate, and a predicate on their associated texts. For example, instead of considering all the restaurants, a nearest neighbor query would instead ask for the restaurant that is the closest among those whose menus contain "Dosa, Idli, Wadapav" all at the same time. Currently the best solution to such queries is based on the IR2-tree, which, such type of queries can be efficiently handled by IR2tree.

In the proposed work, we are developing a system searching is done on the basis of methods like nearest neighbor search with keywords is done by IR2 tree and spatial inverted index.We could first fetch all the restaurants whose menus contain the set of keywords {Dosa, Idli, Wadapav}, and then from the retrieved restaurants, find the nearest one The IR2-tree combines the R tree with signature files. Inverted indexes (I-index) have proved to be an effective access method for keyword based document retrieval.

*****

I. INTRODUCTION

A spatial info manages four-dimensional objects (such as points, rectangles, etc.), and provides quick access to those objects supported completely different choice criteria. The importance of spatial databases is mirrored by the convenience of entities in an exceedingly geometric manner. As an example, locations of restaurants, hotels ,hospitals so on are typically depicted as points in a map, whereas larger extents like parks, lakes, and landscapes typically as a mixture of rectangles. Many functionalities of a spatial info are helpful in different ways in specific contexts. For example, in an exceedingly earth science data system, vary search is employed to seek out all restaurants in an exceedingly sure space, whereas nearest neighbor retrieval will discover the eating place nearest to a given address.

We tend to haveseen some trendy applications that decision for the ability to choose objects supported each of their geometric coordinates and their associated texts. As an example, it wouldbe fairly helpful if a research engine is often accustomed find the nearest eating place that provides "Dosa, Idli, Wadapav" all at an equivalent time. As an example for the above query, we tend to might 1st fetch all the restaurants whose menus contain the set of keywords, so from the retrieved restaurants, find the nearest one. Similarly, one might additionally mate reversely by targeting 1st the spatial conditions – browse all the restaurants in ascending order of their distances to the query purpose till encountering one whose menu has all the keywords.

It's till recently that attention was amused to four-dimensional knowledge [12], [13], [21].The best technique up to now for nearest neighbor search with keywords is thanks to Felipe et al. [12]. They nicely integrate 2 well-known concepts: R-tree [2], a popular spatial index, and signature file [11], a good method for keyword-based document retrieval. By doing thus they develop a structure referred to as the IR2-tree [12], that has the strengths of each R-trees and signature files. Like R-trees, the IR2-tree preserves objects' spatial proximity, that is that the key to determination spatial queries with efficiency. On the opposite hand, like signature files, the IR2-tree is in a position to filter a considerable portion of the objects that don't contain all the question keywords, therefore considerably reducing the number of objects to be examined.

II. LITERATURE REVIEW & RELATED WORK

Many applications want finding objects that area unit nearest to a given location that contains a cluster of keywords. associate increasing selection of applications want the economical execution of nearest neighbor (NN) queries affected by the properties of the spatial objects. Owing to the recognition of keyword search, notably on the web, many of those applications modify the user to supply a list of keywords that the spatial objects need to contain, in their description or various attribute. As an example, land websites modify users to travel longing for properties with specific keywords in their description and rank them in line with their distance from a given location. We tend to tend to call such queries spatial

keyword queries [12]. A spatial keyword question consists of a question area and a gaggle of keywords. The answer may be a listing of objects stratified in line with a mixture of their distance to the question area and conjointly the affiliation of their text description to the question keywords.

A straightforward withall widespread variant that's used is that the distance initial spatial keyword question, where objects area unit stratified by distance and keywords area unit applied as a conjunctive filter to eliminate objects that don't contain them. Sadly there's no economical support for top-k abstraction keyword queries, where a prefix of the results list is required. Instead, current systems use ad -hoc combos of nearest neighbor (NN) and keyword search techniques to tackle the matter. For associate example, associate points R-Tree is used to get out the nearest neighbors associate points for each neighbor associate inverted index is utilized to see if the question keywords unit of measurement contained.

The economical methodology to answer top-k spatial keyword queries is based on the mixture of data structures and algorithms utilized in spatial in – formation search and knowledge Retrieval (IR). Significantly, the strategy consists of building associate data Retrieval R- Tree (IR2-Tree) that would be a structure supported the R- Tree. At question time associate progressive algorithmic program is utilized that uses the IR2-Tree to with efficiency turn out the high results of the question. The IR2-Tree is a R-Tree where a signature is supplementary to each node v of the IR2- Tree to denote the matter content of all spatial objects at intervals the sub tree motionless at „v". The top-k spatial keyword search formula that is affected by the work of Hjaltason and Samet [14] exploits this knowledge to search out the very best question results by accessing a bottom portion of the IR2-Tree. Spatial queries with keywords haven't been extensively explored. Within the past years, the community has sparked enthusiasm in learning keyword search in relative databases.

It's till recently that focus was entertained to dimensional knowledge. Existing works principally specialize in finding top-k Nearest Neighbors, wherever every node must match the complete querying keywords .It doesn't contemplate the density of information objects within the spatial area. Conjointly these ways area unit low economical for progressive question. Spatial information manages dimensional objects (such as points, rectangles, etc.), and provides quick access to those objects supported completely different choice criteria. The importance of spatial databases is mirrored by the convenience of modeling entities of reality during a geometric manner. As an example, locations of restaurants, hotels, hospitals and then on area unit typically depicted as points during a map, whereas larger extents like parks, lakes, and landscapes typically as a mixture of rectangles. Several functionalities of a spatial information area unit helpful in varied ways that in specific contexts.

As an example, during an earth science data system, vary search will be deployed to search out all restaurants during a sure space, whereas nearest neighbor retrieval will discover the building nearest to a given address. Section 3.1 reviews the knowledge retrieval R-tree (IR2-tree) [12] that is that the state of the art for responsive the nearest neighbor queries outlined in Section two. The IR2-tree [12] combines the R-tree with signature files. Next, we are going to review what's a signature file before explaining the main points of IR2-trees.Our discussion assumes the data of R-trees and the best-first algorithmic program [14] for NN search, each of which are well-known techniques in spatial databases.

III. ANALYSIS OF PROBLEM

Inverted indexes (I-index) have proved to be a good access technique for keyword primarily based document retrieval. Within the abstraction context, nothing prevents U.S. from treating the text description Wp of some extent p as a document, and then, building associate I-index. Each word within the vocabulary has an inverted list, enumerating the ids of the points that have the word in their documents. Note that the list of every word maintains a sorted n order of purpose ids that provides goodish convenience in question process by permitting associate efficient merge step. For instance, assume that we would like to find the points that reproof c and d. this can be essentially to figure the intersection of the 2 words' inverted lists. As each list square measure sorted within the same order, we tend to can do thus by merging them, whose I/O and computer hardware times are both linear to the full length of the lists. Recall that, in NN process with IR2-tree, some extent retrieved from the index should be verified (i.e., having its text description loaded and checked). Verification is also necessary with I-index, except for precisely the opposite reason. For IR2-tree, verification is as a result of we tend to do not have the elaborated texts of some extent, whereas for I-index; it is as a result of we tend to don't have the coordinates. Specifically, given associate NN question letter of the alphabet with keyword set Wq, the query algorithm of I-index initial retrieves (by merging) the set Pq of all points that have all the keywords of Wq , and then, performs |Pq | random I/Os to induce the coordinates of every purpose in Pq so as to judge its distance to letter of the alphabet. According to the experiments of [12], once Wq has only one word, the performance of I-index is very bad, that is anticipated as a result of everything within the inverted list

of that word should be verified. Curiously, as the size of Wq will increase, the performance gap between I index and IR2-tree keeps narrowing such I-index even starts to surmount IR2-tree at |Wq | = four. This can be not as stunning because it could seem. As |Wq| grows giant, not many objects ought to be verified as a result of the quantity of objects carrying all the question keywords drops quickly. On the opposite hand, at now a plus of I index starts to pay off. That is, scanning associate inverted list is comparatively low-cost as a result of it involves solely consecutive I/Os1, as critical the random nature of accessing the nodes of associate IR2-tree.

IV. CONCLUSION

We have seen many applications vocation for a search engine that's ready to expeditiously support novel forms of abstraction queries that area unit integrated with keyword search. The prevailing solutions to such queries either incur prohibitive area consumption or area unit unable to give real time answers. During this paper, we've got remedied the state of affairs by developing associate degree access methodology called the abstraction inverted index (SI-index). Not solely that then SI-index is fairly area economical, however conjointly it's the ability to perform keyword-augmented nearest neighbor search in time that's at the order of dozens of milliseconds. Moreover, because the SI-index is predicated on the conventional technology of inverted index, it's readily incorporable in an exceedingly industrial computer programme that applies massive similarity, implying its immediate industrial merits

V. 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] N. Beckmann, H. Krieger, 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.

[3] 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.

[4] X. Cao, L. Chen, G. Cong, C. S. Jensen, Q. Qu, A. Skovsgaard, D. Wu, And M. L. Yiu. Spatial Keyword Querying. Inner, Pages16–29, 2012

[5] X. Cao, G. Cong, And C. S. Jensen. Retrieving Top-K Prestige-Based Relevant Spatial Web Objects. Pvldb, 3(1):373–384, 2010.

[6] X. Cao, G. Cong, C. S. Jensen, And B. C. Ooi. Collective Spatial Keyword Querying. In Proc. Of Acm Management Of Data (Sigmod), Pages 373–384, 2011.

[7] B. Chazelle, J. Kilian, R. Rubinfeld, And A. Tal. Thegloomier 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.

[8] Y.-Y. Chen, T. Suel, And A. Markowitz. Efficient Query Processing In Geographic Web Search Engines. In Proc. Of Acm Management Of Data (Sigmod), Pages 277–288, 2006.

[9] E. Chu, A. Baid, X. Chai, A. Doan, And J. Naughton. Combining Keyword Search And Forms For Ad Hoc Querying Of Databases. In Proc. Of Acm Management Of Data (Sigmod), 2009.

[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] 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.

[12] 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.

[13] R. 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(Ssdbm), 2007.

[14] G. R. Hjaltason And H. Samet. Distance Browsing In Spatial Databases. Acm Transactions On Database Systems (Tods), 24(2):265–318, 1999.

[15] V. Hristidis And Y. Papakonstantinou. Discover: Keyword Search In Relational Databases. In Proc. Of Very Large Data Bases (Vldb), Pages 670–681, 2002.

[16] I. Kamel And C. Faloutsos. Hilbert R-Tree: An Improved R-Tree Using Fractals. In Proc. Of Very Large Data Bases (Vldb), Pages 500–509, 1994.

[17] J. Lu, Y. Lu, And G. Cong. Reverse Spatial And Textual K Nearest Neighbor Search. In Proc. Of Acm Management Of Data (Sigmod), Pages 349–360, 2011.

[18] S. Stiassny. Mathematical Analysis Of Various Superimposed Coding Methods. Am. Doc., 11(2):155–169, 1960.

[19] J. S. Vitter. Algorithms And Data Structures For External Memory. Foundation And Trends In Theoretical Computer Science, 2(4):305–474, 2006.

[20] D. Zhang, Y. M. Chee, A. Mondal, A. K. H. Tung, And M. Kitsuregawa. Keyword Search In Spatial Databases: Towards Searching By Document. In Proc. Of International Conference On Data Engineering (ICDE), Pages 688–699, 2009.

[21] 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