Efficient Indexing of Spatial Query

Vikas Patil, Madhumati Unde, Manjusha Jagtap

Abstract- In past few years the Geographical Information Retrieval is very active field for research. Due to this research a new type of search engine came into existence called as Geographical Search Engine. Geographical search engine help to retrieve document which more textually and spatially relevant to our query. Indexing structure for spatial relevant is the main goal of this field and also to store and retrieve document having spatial scope of the given query. In this context we give an efficient tree structure called IR-tree, which allows searches to adopt different scope on textual and spatial relevance of document.

Index Terms: Geographical Search Engine, Spatial Relevance, IR-Tree.

I. INTRODUCTION

In today's era internet has become a convenient tool for finding solution for any of the user's query. The information on Internet is stored in the form of documents. Whenever we required to access these documents it leads to scanning of large number of documents. The scanning results in memory overhead problem and increases search cost. To overcome this problem an efficient search engine is necessary. The characteristics of efficient search engine are that it should retrieve the most relevant document in minimum latency. To organize the documents according to the both textual and spatial relevance's an efficient indexing technique is required. To index the document efficiently Information Retrieval tree (IR tree) indexing can be used. The document set is huge so there may exist problems like storage overhead and access overhead. These problems are overcome with the help of IR tree. IR tree performs following functions:1)Filtering of spatially irrelevant documents; 2) Filtering of textually irrelevant documents;and3)computation and ranking relevance's.

The user query can fall under following categories: 1) Queries containing geographic terms such as city, state or country.2) Queries that do not contain such geographic terms.3) Non geographic queries that contain a geographic term.4) Non-geographic queries. Example1.Consider candidate is seeking admission in an institute and the address is given as XYZ institution near hotel Taj Diamond, Pune. So to get most relevant information search engine must consider complete keyword XYZ institution Taj Diamond hotel Pune into consideration. Let there be 6 documents in the web server. D = {D1, D2, D3... D6}. And D4 and D6 are not within the spatial scope. Here each document contains some textual words also. The frequencies of the keywords are shown in the fig 1. Now the task is to find most relevant documents.

Manuscript received October, 2013.

Vikas Patil, Dept. Of Computer, .B.E. (Computer), M.E.(CSE) , Zeal Institute's Dnyanganga College of Engineering and Research, Narhe ,Pune 41,INDIA ,University of Pune, INDIA.

Madhumati Unde, Dept. Of Computer , Student, Zeal Institute's Dnyanganga College of Engineering and Research, Narhe, Pune 41,INDIA,University of Pune, INDIA.

Manjusha Jagtap, Dept. Of Computer, Student, Zeal Institute's Dnyanganga College of Engineering and Research, Narhe, Pune 41,INDIA, University of Pune, INDIA.

A document is said to be relevant if at least one keyword is matched along with that it should have spatial relevance too. A document is said to be more relevant if number of textual keywords matched are more and location of document is within the spatial scope. In all the documents,D4 and D6 are discarded as they are not within the spatial scope and from fig 1.Document D3 is more relevant to the user query as in this document frequency of each keyword is more matched. Then in this approach D3 is first retrieved and other documents are retrieved on the basis of their relevance's. As numbers of documents are more, efficient index structure is needed.

Fig 1. Showing occurrence of words in the document

	Document Words	
	XYZ institute	Taj Diamond
D1	1	1
D2	2	1
D3	4	3
D4	2	1
D5	1	2
D6	0	1

II. RELATED WORK

In this we are going to review the existing work done in Geographical Search engine, textual index, and spatial index.

A.GEOGRAPHICAL SEARCH ENGINE

From past few years, because of increase in the application demand and fast growth of technology in the geographical information system, the geographical search engine has been receiving a lot of attention from both research and industry [12],[13].

Exiting Geographical search engines use two type of approaches i.e. Method I and Method II. Where Method I use separate indexes for spatial and textual information of query and Method II uses combined index [1],[2],[3] . Method I is an extend of convention textual search engine with spatial filtering capabilities of Quad-tree, R-tree, Grid Index logically [16][17].

Based on two indexes, a search generally follows a three step process :

Step 1: retrieving textually relevant documents with respect to query keywords via a conventional textual index.

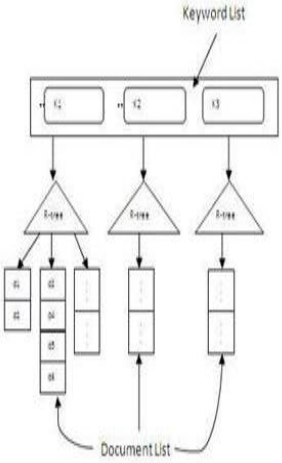
Step 2: filtering out the documents obtained from Step 1 that are not covered by the query spatial Scope.

Step 3: ranking the documents from Step 2 based on the joint textual and spatial relevances in order to return the ranked results to the user.

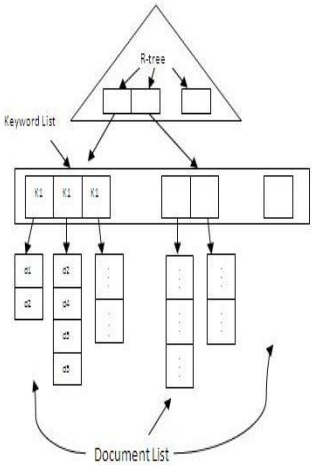
Whereas Method II uses one index for spatial and textual content of document, in context with it there are method proposed in [4], which are namely, 1) an inverted file on top of R-tree referred as $Hybrid_I$, and 2) an R-tree on the top of inverted files, referred as $Hybrid_R$. But both are not able to integrate the textual Filtering and spatial Filtering of query.

KR^*-tree is another hybrid indexing structure which supports both simultaneously textual and spatial filtering. This approach is extension of $Hybrid_R$ by augmenting with set of words in the internal nodes.

Finally there is other proposed model like IR^2-tree, the proposed structure for indexing IR-tree are also based on the R-tree ,they are not similar in their structure, extensibility and functionality.



(a) $Hybrid_I$



(b) $Hybrid_R$

Fig 2: Two Hybrid indexing Scheme

B. TEXTUAL RELEVANCE

We assume each document d in a given document set D is composed of a set of words Wd, and is associated with a location Ld. Given a query q that specifies a set of query keywords Wq and a query spatial scope Sq, the textual relevance and spatial relevance of a document d to q are formalized in Definitions of it. A document d said to be textually relevant to a query q if d contains some (or all) of queried keywords, i.e. $Wd \cap Wq \neq \emptyset$ .To quantify the relevance of d to q, a weighting function denoted by $\phi_q(d)$ is adopted. Thus, for a given q, $\phi_q(d_1) > \phi_q(d_2)$ means document d_1 is more textually relevant to q than d_2. In this sense we have to calculate TE and IDF of the documents. The terms can be defined as, a term frequency $tf_{w,d}$ measures which indicates the importance of the word within the document. On the other hand, the inverse document frequency $idf_{w,D}$ measures the specificity (importance) of a word w in the document set D. In this context to facilitate the mathematical calculation of TF/IDF of document, inverted files, i.e. a collection of inverted lists. To eliminate the overload on the system we first search document according to spatial relevance and then with respect to the textual relevance.

III. SPATIAL RELEVANCE

In the classification of information and data, geographic location has an important role. Most of the user queries are directly or indirectly geo-referenced. In information retrieval, the evaluation of relevance's is an important task. Spatial relevance focuses on some geographic regions having well-defined name (eg. "Show me an *Indian restaurant")* it will show those documents that are relevant to the concept of "Indian Restaurant".

There are many situations in which the geographic based results are as important as conceptual based results. For example, a person finding place to eat through his mobile device is interested in all those documents that are respect to his location and also those are conceptually relevant.

The spatial relevance of a document d, denoted $\varphi(d)$ as depends on the types of the spatial relationships defined between a document location L_d and a spatial scope S. Commonly adopted relationships as discussed in [4] include:

Enclosed: $\varphi(d)$ is set to 1 if the corresponding location is fully enclosed by query scope i.e

.

$$\varphi(d) = \begin{cases} 1, if\ L_d \subseteq S \\ 0, otherwise \end{cases} \quad (1)$$

Overlapping: is set to the fraction of the document location that is covered by the spatial scope, i.e.

$$\varphi(d) = \frac{Area(L_d \cup S)}{Area(L_d)} \quad (2)$$

Proximity: $\varphi(d)$ is representation by the inverse distance between the center of L_d and that of S i.e.

$$\varphi(d) = \begin{cases} \frac{1}{dist(L_d S)}, if\ L_d \subseteq S \\ 0, otherwise \end{cases} \quad (3)$$

Without loss of generality, we focus on the proximity in the following discussion. Other types of spatial relevance's can be supported by substituting proximity with a desired spatial relevance calculation.

Considering the spatial indexing scheme used, we classify the indices into three categories, namely R-tree based indices, grid based indices, and space filling curve based indices. R-tree based This category of indices use the R-tree [5] or a variation (e.g., the R*-tree). Most geo-textual indices belong to this category and use the inverted file for text indexing. In early work [6], the R-tree based indices loosely combine the R-tree and inverted files to organize the spatial and text data separately. In contrast, recent indices tightly combine the R-tree with a text index.

Grid based this category of indices combine a grid index with a text index (e.g., the inverted file). The grid indices divide space into a predefined number of equal-sized square or rectangular cells. The grid index and the text index can be organized [7] either separately or combined tightly [8]. Space filling curve based these indices combine inverted files with a space filing curve, and they include a Hilbert curve based index [9] and a Z-curve based index [10]. These indices are based on the property that points close to each other in the native space are also close to each other on the space filling curve.

IV.IR-TREE

In this part, we present IR-tree, an efficient index that carter the following required functions for geographical document search and ranking 1) spatial filtering: all the spatially irrelevant document have to be filtered out as early as possible to shrink the search space;2) textual filtering: all the textually irrelevant documents have to be discarded as early as possible to cut down the search cost; and 3) relevance computation and ranking

In a way to give an efficient geographic document search, we use an IR-tree structure. In IR-tree, clustering of set of documents is done into disjoint subsets of documents and abstraction of them is done into various granularities. By doing this, it enables the elimination of those (textually or spatially) irrelevant subsets. The efficiency of IR-text depends on its elimination power, which in turn, is highly related to the effectiveness of the document clustering and the search algorithm.

In IR-tree, the initial task is spatial clustering followed by textual filtering. There are many documents that are textually related but only few of those are within query's spatial scope so spatial filtering is done first in order to reduce the search space. And textual filtering is done to minimize the search cost. Finally, depending upon the joint relevance (spatial and textual) and the ranking that our system will do, then the top-k document searched will be returned. Thus IR-tree is designed in such a way that storage and access overhead our considered.

IR-Tree Structure to calculate the relevance of documents to user's query we are using TF-IDF [14][15] values. TF-IDF weighs a term in a document based on term frequency (tf) and inverse document frequency (idf) [11]. The *term frequency* measures the importance of word within *document*. Term frequency indicates the number of time the word has occurred in that particular document. Whereas *the inverse document frequency* calculates the importance of word within *document set.*

In IR tree, the set of document are clustered into disjoint subsets of documents which are then abstracted in various granularities. This leads to pruning of irrelevant subsets (textually or spatially).IR tree clusters the documents which are spatially related so that the documents which are not related to user's query location can be discarded. The textual words are represented with help of inverted files. Each leaf entry of an IR-tree contains an inverted file and each non-leaf node contains a document summary so that the tf and idf values of document words can be found at nodes without scanning individual documents. The document summary consists of minimal bounding box (MBB), cardinality of documents that come under the particular non-leaf node and TF-IDF pair values. The MBB covers all the locations of the document under that non-leaf node i.e. it is a small rectangular region covering all the locations in the document set under that particular non-leaf node.

V. CONCLUSION

To improve the working of geographical search engine it is important that the spatial content of the query has to considered and processed efficiently so that the result displayed must be related to the spatial part of the query also. Thus in order to make geographical search engine efficiently there is indexing and ranking of the spatial content has to done effectively.

In this paper, we are proposing an efficient indexing method for the spatial query entered by the user in any search engine. We also proposed an efficient structure namely IR-Tree for this. In addition, IR-tree makes it possible to adopt different weights for textual and spatial relevance of the document at the execution time and thus provide for a wide variety of application.

We are also trying to make a geographical search engine with IR-Tree structure indexing. We are also planning to enhance the indexing of spatial content of query using various access pattern.

ACKNOWLEDGMENT

This is a great pleasure & immense satisfaction to express our deepest sense of gratitude & thanks to everyone who has directly or indirectly helped us in completing our project work successfully. We express our gratitude towards Project guide Prof. Vikas Patil and Prof. S.M.Sangve, Head of Computer Science Department, Dnyanganga college of Engineering and Research, Narhe, Pune who guided & encouraged us in completing the project work in scheduled time. We would like to thank our Principal, for allowing us to pursue our project in this institute. No words are sufficient to express our gratitude to our parents for their unwavering encouragement. We also thank all friends for being a constant source of our support.

REFERENCES

1. R. Hariharan, B. Hore, C. Li, and S. Mehrotra, "Processing Spatial-Keyword (SK) Queries in Geographic Information Retrieval (GIR)Systems," Proc. 19th Int'l Conf. Scientific and Statistical Database Management (SSDBM '07), pp. 16-25, 2007.
2. C.B. Jones, A.I. Abdelmoty, D. Finch, G. Fu, and S. Vaid, "The SPIRIT Spatial Search Engine: Architecture, Ontologies and Spatial Indexing," Proc. Third Int'l Conf. Geographic Information Science (GIS '04), pp. 125-139, 2004.

3. R. Lee, H. Shiina, H. Takakura, Y.J. Kwon, and Y. Kambayashi,"Optimization of Geographic Area to a Web Page for Two-Dimensional Range Query Processing," Proc. Fourth Int'l Conf. Web Information Systems Eng. Workshops (WISEW '03), pp. 9-17, 2003.
4. Y. Zhou, X. Xie, C. Wang, Y. Gong, and W.-Y. Ma, "Hybrid Index Structures for Location-Based Web Search," Proc. 14th ACM Int'l Conf. Information and Knowledge Management (CIKM '05), pp. 155-162, 2005.
5. A. Guttman. R-trees: A dynamic index structure for spatial searching. In SIGMOD, pages 47–57, 1984.
6. Y. Zhou, X. Xie, C. Wang, Y. Gong, and W.-Y. Ma. Hybrid index structures for location-based web search. In CIKM, pages 155–162,2005.
7. S. Vaid, C. B. Jones, H. Joho, and M. Sanderson. Spatio-textual indexing for geographical search on the web. In SSTD, pages218–235, 2005.
8. A. Khodaei, C. Shahabi, and C. Li. Hybrid indexing and seamless ranking of spatial and textual features of web documents. In DEXA, pages 450–466, 2010.
9. Y.-Y. Chen, T. Suel, and A. Markowetz. Efficient query processing in geographic web search engines. In SIGMOD, pages 277–288, 2006.
10. M. Christoforaki, J. He, C. Dimopoulos, A. Markowetz, and T. Suel. Text vs. space: efficient geo-search query processing. In CIKM, pages 423–432, 2011.
11. K.S. Jones, "A Statistical Interpretation of Term Specificity and Its Application in Retrieval," J. Documentation, vol. 28, no. 1, pp. 11-21, 1972.
12. A. Markowetz, Y.-Y. Chen, T. Suel, X. Long, and B. Seeger,"Design and Implementation of a Geographic Search Engine,"Proc. Eighth Int'l Workshop Web and Databases (WebDB), pp. 19-24,2005.
13. K.S. McCurley, "Geospatial Mapping and Navigation of the Web,"Proc. Int'l Conf. World Wide Web (WWW '01), pp. 221-229, 2001.
14. D. Hiemstra, "A Probabilistic Justification for Using TF x IDFTerm Weighting in Information Retrieval," Int'l J. Digital Libraries,vol. 3, no. 2, pp. 131-139, 2000.
15. G. Salton and C. Buckley, "Term-Weighting Approaches inAutomatic Text Retrieval," Information Processing & Management,vol. 24, no. 5, pp. 513-523, 1988.
16. Y.-Y. Chen, T. Suel, and A. Markowetz, "Efficient QueryProcessing in Geographic Web Search Engines," Proc. ACM SIGMOD '06, pp. 277-288, 2006.
17. R. Lee, H. Shiina, H. Takakura, Y.J. Kwon, and Y. Kambayashi, "Optimization of Geographic Area to a Web Page for Two- Dimensional Range Query Processing," Proc. Fourth Int'l Conf. Web Information Systems Eng. Workshops (WISEW '03), pp. 9-17, 2003.

AUTHORS PROFILE

Vikas Patil, B.E. (Computer), M.E (CSE) Professor at Zeal Institute's Dnyanganga College of Engineering and Research, Narhe, Pune 41,INDIA ,University of Pune, INDIA.

Madhumati Unde, perusing B.E(Computer) at Zeal Institute's Dnyanganga College of Engineering and Research, Narhe, Pune 41,INDIA ,University of Pune, INDIA.

Manjusha Jagtap, perusing B.E(Computer) at Zeal Institute's Dnyanganga College of Engineering and Research, Narhe, Pune 41,INDIA ,University of Pune, INDIA.