DOI: http://dx.doi.org/10.26483/ijarcs.v8i7.4507

ISSN No. 0976-5697

Volume 8, No. 7, July – August 2017

International Journal of Advanced Research in Computer Science

REVIEW ARTICLE

Available Online at www.ijarcs.info

A REVIEW ON HYBRID GEO-TEXTUAL INDEXING TECHNIQUES

Sulbha Powar
Research Scholar, P. G. Department of Computer Science
SNDT Women's University
Mumbai, India

Dr. Ganesh Magar
Associate Professor, P. G. Department of Computer Science
SNDT Women's University
Mumbai, India

Abstract: The integration of a Geographic Information System (GIS) and internet technology has revolutionized the use of geospatial data and its applications in planning and implementation of strategies for a wide range of activities. Growth of location-based services has given new direction for development as it has increased location, textual and temporal information. Various techniques are developed that enable the indexing of data that contains both text descriptions and geo-locations to support the efficient processing of spatial keyword queries that take a geo-location and a set of keywords as arguments and return relevant contents that matches the arguments. The nature of spatial keyword queries has evolved over the time, index structures also have evolved depending on the nature of the query. A single index structure does not suit the needs of all types of queries. This paper presents a comprehensive study of hybrid geo-textual indices.

Keywords: Geo-Textual, Hybrid, Indexing, Query Processing, Spatial, Textual

I. INTRODUCTION

Location based data is available on huge scale with the wide availability of satellite, radio frequency identification (RFID), global positioning system (GPS), sensor technologies, smart phones and other mobile and stationary devices. The efficient management and analysis of such data is of great interest in a wide range of application domains. Few examples amongst wide range of applications of location based services are weather forecasting, disaster management systems, telecom and network services, urban planning, transportation planning, agriculture, asset management, tourism, water management and wildlife management.

Geographic Information System (GIS) is a large domain that provides a variety of capabilities designed to capture, store, manipulate, analyse, manage, and present all types of geographical data, and utilizes geospatial analysis in a variety of contexts, operations and applications. Many applications involve large collections of geo-spatial objects with their latitude, longitude and the attribute data. Queries that arise over geo-spatial data are of three main types: spatial range queries, nearest neighbor queries and spatial join queries [1]. These kinds of queries are very common in most applications of spatial data. Efficient algorithms for answering location based queries, which will consider both locations as well as attribute data are needed.

A. Geo-textual indices

A spatial keyword query takes a user location and user supplied keywords as arguments and returns objects that are spatially and textually relevant to these arguments. To answer these queries efficiently, geo-spatial data needs to be indexed. There are several techniques available to index data either based on location or based on text [1]. Most popular spatial indexing techniques include R-tree based indices, grid based indices and space filling curve based indices. Commonly used text indexing techniques are inverted files based indices and signature file based indices. While answering geo-textual queries spatial indexing and text indexing techniques can be applied one after the other in any order. But this method is not efficient as it does not support filtering out early enough the objects which do not satisfy either of location or text criteria.

The geo-textual indices combine spatial and text indexing [1]. The indices are categorized according to how they combine the two, namely text-first loose combination, spatial-first loose combination, text-first tight combination and spatial-first tight combination. A text-first loose combination index usually employs the inverted file as the top-level index and then arranges the postings in each inverted list in a spatial structure, which can be an R-tree, a grid or a space filling curve. In contrast, the top level of a spatial-first index is a spatial structure, and its leaf nodes contain inverted files or bitmaps for the text information of objects contained in the nodes. On the other hand, the tight combination index combines a spatial and a text index tightly such that both types of information can be used to prune the search space simultaneously during query processing. Two types of tight combinations have been used. One integrates a text summary into every node of a spatial index, and other integrates the spatial information into each inverted list

II. GEO-TEXTUAL INDICES - METHODOLOGY REVIEW

For creating geo-textual hybrid index, one needs to pick up a spatial index and a text index.

A. Hybrid indices based on inverted file and R tree

Various techniques are developed for hybrid indexing which use inverted file and R-tree. Different combining schemes of hybrid indexing structure which integrate inverted files and R*-tree was proposed [2]. Inverted file and R*-tree double index, first inverted file then R*-and first R*-tree then inverted file scheme are the three schemes experimented. These have almost same storage cost but second is better than first and third in query time. Indexes based on R*-trees are proven to be more efficient than indexes based on grid structures. KR*-tree were implemented for processing spatial-keyword queries with the AND semantics by capturing the joint distribution of

keywords appearing in space [3]. In this implementation, the pruning power of both space and text is exploited simultaneously, thus merging the two steps into one and hence greatly enhancing the performance. Information retrieval R-tree (IR2-Tree) combines an R-tree with superimposed text signatures [4]. To efficiently answer top-k spatial keyword queries, the tight integration of data structures and algorithms is used with each node containing both spatial and keyword information.

The IR-tree is studied and used extensively in variations of spatial keyword queries [5]. The framework proposed leverages the inverted file for text retrieval and the R-tree for spatial proximity querying [6]. Several indexing approaches are explored within the framework for computing the top-k query. The IR-tree and its variants, document similarity IR tree (DIR-tree), the cluster enhanced IR-tree (CIR-tree) and the cluster enhanced DIR-tree (CDIR-tree) are included in the framework. In an index node, tighter text relevancy scores can be estimated for a group of similar documents than for diverse documents that belong to different categories. Clustering improves the query performance. Hybrid spatial-keyword index (SKI) efficiently processes top-k spatial Boolean (k-SB) queries [7]. It combines an R-tree with an inverted index by the inclusion of spatial references in posting lists. Spatial inverted index (S2I) was proposed to improve the performance of top-k spatial keyword queries [8]. Index maps each distinct term in the vocabulary into a distinct aR-tree or block that stores all objects with the given term. The objects are stored differently according to the document frequency of the term and can be retrieved efficiently in decreasing order of keyword relevance and spatial proximity. The spatial inverted list (SI-index) is essentially a compressed version of an I-index with embedded coordinates [9].

User may not always enter the correct keyword and then approximate solution needs to be searched. In location based approximate-keyword (LBAK) queries approximate string search is used to identify for each query keyword those strings that are similar [10]. To answer such queries, a tree-based spatial index structure is augmented with approximate-string indexes such as a gram-based inverted index or a trie-based index and resulting structure is called as LBAK-tree index. In another implementation on spatial approximate string (SAS) query, the MHR-tree is proposed, which embeds min-wise signatures into an R-tree [11].

A mobile user needs to be continuously aware of the k spatial web objects that best match a query with respect to location and text relevancy. The moving top-k spatial keyword (MkSK) query, considers a continuously moving query location [12]. Solutions for moving queries employ safe zones that guarantee the validity of reported results as long as the user remains within a zone.

Varieties of variations are seen in the IR tree implementation depending upon the need of the query. The problem of retrieving a group of spatial web objects such that the group's keywords cover the query's keywords and objects are nearest to the query location and have the lowest inter-object distances are studied [13], [14]. Joint processing of multiple top-k spatial keyword queries, processing top-k spatial keyword queries on road network, distributed solution to answering spatial keyword queries on road networks, aggregate index search (AIS), summarizing both social and spatial information, location-aware top-k prestige-based text retrieval (LkPT) query, the generic location-aware rank query (GLRQ), top-k point-of-interest group retrieval called GroupFinder, why-not queries with a keyword count R-tree (KcR-tree) based algorithm and bR*-tree based m-closest keywords (mCK) query are some examples of these variations [15] - [23].

B. Hybrid indices based on kd-tree/space-filling curve/grid/quad tree

Though much research is done on R-tree index, hybrid index created using other spatial indexing style are also proposed. Hybrid indices are implemented using a spatial indexing based on coarse-space partitioning (CSP) and a space-filling curve (SFC) [24]. Index structure that combines K-d tree and inverted file for spatial range keyword query was also proposed [25]. Grid spatial index were studied and proved that query times for tightly coupled indexing structures were faster than loosely coupled indexing structures [26]. Spatial-keyword inverted file (SKIF) and length-constrained maximum-sum region queries were developed using grid spatial index [27], [28]. Hybrid index, called inverted file quad-tree (IQ-tree) was proposed to answer Boolean range continuous queries (BRCQ) over a stream of incoming geo-textual objects in real time [29]. A scalable integrated inverted index, named I3 was proposed, which adopts the quadtree structure to hierarchically partition the data space into cells [30]. Top k spatial keyword search (TOPK-SK), and batch top k spatial keyword search (BTOPK-SK) were studied based on the inverted index and the linear quad tree, called inverted linear quadtree (IL-Quadtree) [31]. Grid-based signatures hybrid structure was proposed to handle regions-of-interest (ROIs) queries [32].

III. DISCUSSION

Both textual and spatial queries have been extensively studied within their respective communities, as well as combining and forming hybrid index structures. If spatial and text indices are combined tightly, it prunes out the objects early enough which do not satisfy the query criteria. In previous methods the keywords are maintained separately. Hence queries are answered by the intersection of object satisfying query criteria from the textual index file and spatial index structure.

Although different hybrid indexes are proposed, they are very different in terms of structures, functionalities, and extensibility to searches with various relevant requirements. The performance of query evaluation depends on the indexing structure used for pruning the spatial and textual data space. It also depends on splitting technique used for the node splitting. Indexing structure also needs to take care of distance function like Euclidean, road networks used for finding the nearest object. It is seen that query format have changed over the period like moving query, group query, approximate query, collective query, predicate query, joint query, why not query and many other.

Much research is done on hybrid indexing techniques using R-Tree index and inverted files index. The comparison of these techniques is shown in Table I. It is also observed that most of the research is done mainly on the top-k queries and hence evaluation for other types of queries is also essential. Various parameters are used to evaluate query performance. It is shown experimentally that parameters like keyword size, signature length, k, average response time,

storage overhead, index size, construction time, update time, recall, precision, relative error affect the query performance.

Table I : Comparison of Hybrid Indexing Techniques based on Inverted file and R-Tree

Sr. No.	*Indexing structure*	*Spatial Index*	*Text Index*	*Advantages*	*Disadvantages*
1.	IR Tree	R*-tree	Inverted file	Structure of first inverted file then R*-tree is the most efficient in query time	Extra disk costs, High overhead in merging process, Generates many candidate object ids
2.	Keyword-R*-tree (KR*-tree)	R*-tree	Inverted file	Keyword list at each node reduces the disk IOs incurred during spatial filtering of objects	Suffers from unnecessary overhead when there are many candidates. Not efficient due to separation of document search and document ranking
3.	Keyword-R*-tree (KR*-tree)	R*-tree	Inverted file	Keyword list at each node reduces the disk IOs incurred during spatial filtering of objects	Suffers from unnecessary overhead when there are many candidates. Not efficient due to separation of document search and document ranking
4.	Multi-level IR^2-Tree (MIR^2-Tree)	R-Tree	Signature files	Combines an R-Tree with superimposed text signatures with optimal signature length for each level	Significantly increases the complexity of the tree maintenance operations
5.	IR Tree	R-tree	Inverted file	Ability to perform document search, document relevance computation and document ranking in an integrated fashion	Many pseudo-documents must be accessed when the query keywords are frequent, incurring high I/O cost
6.	DIR Tree CIR Tree	R-tree	Inverted file	Ranking based on location proximity and text relevancy Document similarity to build a more advanced R-tree namely DIR-tree clustering the nodes of DIR-tree to further improve the query processing performance	The additional cost required for clustering the nodes (pre-processing), and the extra storage space demanded by CDIR-tree Keeping a CDIR-tree updated is more complex
7.	hybrid Spatial-Keyword Index (SKI)	R-tree	Inverted file	Inclusion of spatial references in posting lists Effective pruning of the search space	Restricted to Boolean queries
8.	Spatial Inverted Index (S2I)	aR-tree	Inverted file	An aggregated R-tree is built for spatial pruning scalable as only the related inverted lists will be accessed	Difficult to do spatial aggregation across different R-trees
9.	spatial inverted list (SI-index)	R-tree	Inverted file	Fairly space economical, Efficient query processing	Must scan an inverted list from its beginning even though the point of interest lies deep down the list
10.	synopses tree	R-tree	Inverted file	Indexes synopses of objects Enables efficient pruning and estimation	Maintenance overheads

IV. CONCLUSION

This paper presents comprehensive study of different hybrid geo-textual indexing techniques. The indexing methods previously proposed used the pruning power of space and text, either separately or one followed by the other. Consequently, spatial keyword queries were answered in a two-step filtering process, space followed by text or vice-versa. Later as the tight hybrid indexing structure were developed the query performance was improved. It is further found that that query performance depends on the nature of the query and supporting index structure used for answering the query.

V. REFERENCES

[1] L. Chen, G. Cong, C. S. Jensen and D. Wu, "Spatial Keyword Query Processing: An Experimental Evaluation," in VLDB, 2013.

[2] Y. Zhou, X. Xie, C. Wang, Y. Gong and W.-Y. Ma, "Hybrid Index Structures for Location-based Web Search," in CIKM, Bremen, Germany , 2005.

[3] R. Hariharan, B. Hore, C. Li and S. Mehrotra, "Processing Spatial-Keyword (SK) Queries in Geographic Information Retrieval (GIR) Systems," in SSBDM, Banff, Alta., Canada, 2007.

[4] I. D. Felipe, V. Hristidis and N. Rishe, "Keyword Search on Spatial Databases," in ICDE, 2008.

[5] Z. Li, K. C. K. Lee, B. Zheng, W.-C. Lee and D. L. Lee, "IR-Tree: An Efficient Index for Geographic Document Search," IEEE Transactions on Knowledge and Data Engineering , vol. 23, no. 4, pp. 585 - 599, September 2011.

[6] G. Cong, C. S. Jensen and D. Wu, "Efficient Retrieval of the Top-k Most Relevant Spatial Web Objects," Proceedings of the VLDB Endowment, vol. 2, no. 1, p. 337–348, August 2009.

[7] A. Cary, O. Wolfson and N. Rishe, "Efficient and Scalable Method for Processing Top-k Spatial Boolean Queries," in SSDBM, 2010.

[8] J. B. Rocha-Junior, O. Gkorgkas, S. Jonassen and K. Nørv°ag, "Efficient Processing of Top-k Spatial Keyword Queries," in SSTD, Minneapolis, MN, USA, 2011.

[9] Y. Tao and C. Sheng, "Fast Nearest Neighbor Search with

Keywords," IEEE Transactions on Knowledge and Data Engineering, vol. 26, no. 4, pp. 878-888, April 2014.

[10] S. Alsubaiee, A.Behm and C.Li, "Supporting Location-Based Approximate-Keyword Queries," in ACM GIS' 10, San Jose, CA, USA, 2010.

[11] F. Li, B. Yao, M. Tang and M. Hadjieleftheriou, "Spatial Approximate String Search," IEEE Transactions on Knowledge and Data Engineering, vol. 25, no. 6, pp. 1394 - 1409, March 2010.

[12] D. Wu, Y. M. L. C. S. Jensen and G. Cong, "Efficient Continuously Moving Top-K Spatial Keyword Query Processing," in ICDE, Hannover, Germany, 2011.

[13] X. Cao, G. Cong, C. S. Jensen and B. C. Ooi, "Collective Spatial Keyword Querying," in SIGMOD, Athens, Greece, 2011.

[14] X. CAO, G. CONG, T. GUO, C. S. JENSEN and B. C. OOI, "Efficient Processing of Spatial Group Keyword Queries," ACM Transactions on Database Systems, vol. 40, no. 2, pp. 01-48, June 2015.

[15] D. Wu, M. L. Yiu, G. Cong and C. S. Jensen, "Joint Top-K Spatial Keyword Query Processing," IEEE Transactions on Knowledge and Data Engineering , vol. 24, no. 10, pp. 1889 - 1903, October 2012.

[16] Nørvåg, J. B. Rocha-Junior and Kjetil, "Top-k Spatial Keyword Queries on Road Networks," in EDBT , Berlin, Germany , 2012.

[17] S. Luox, Y. Luox, S. Zhoux, G. Congy, J. Guanz and Z. Yongx, "Distributed Spatial Keyword Querying on Road Networks," in EDBT, Athens, Greece:, 2014.

[18] K. Mouratidis, Y. T. Jing Li and N. Mamoulis, "Joint Search by Social and Spatial Proximity," IEEE Transactions on Knowledge and Data Engineering, vol. 27, no. 3, pp. 781-793, March 2014.

[19] X. Cao, G. Cong and C. S. Jensen, "Retrieving Top-k Prestige-Based Relevant Spatial Web Objects," Proceedings of the VLDB Endowment, vol. 1, no. 3, pp. 373-384 , September 2010.

[20] X. Liu, L. Chen and C. Wan, "LINQ: A Framework for Location-aware Indexing and Query Processing," IEEE Transactions on Knowledge and Data Engineering, vol. 27, no. 5, pp. 1288-1300, 2014.

[21] K. S. Bøgh, A. Skovsgaard and C. S. Jensen, "GroupFinder: A New Approach to Top-K Point-of-Interest Group Retrieval," Proceedings of the VLDB Endowment, vol. 6, no. 12, pp. 1226-1229 , August 2013.

[22] L. Chen, J. Xu, X. Lin, C. S. Jensen and H. Hu, "Answering Why-Not Spatial Keyword Top-k Queries via Keyword Adaption," in ICDE, Seoul, South Korea, 2015.

[23] D. Zhang, Y. M. Chee, A. Mondal, A. K. H. Tung and M. Kitsuregawa, "Keyword search in spatial databases: Towards searching by document," in IEEE International Conference on Data Engineering, Shanghai, China, China, 2009.

[24] M. Christoforaki, J. He, C. Dimopoulos, A. Markowetz and T. Suel., "Text vs. Space: Efficient Geo-Search Query Processing," in CIKM, Glasgow, Scotland, 2011.

[25] A. S. Nandar and S. M. Mint, "Hybrid Geo-Textual Index Structure for Spatial Range Keyword Search," Computer Science & Engineering: An International Journal, vol. 4, no. 5/6, pp. 21-28, December 2014.

[26] S. Vaid, C. B. Jone, H. J. and M. Sanderson, "Spatio-textual indexing for geographical search on the web," in SSTD, Angra dos Reis, Brazil, 2005.

[27] A. Khodaei, C. Shahabi and C. Li., "Hybrid Indexing and Seamless Ranking of Spatial and Textual Features of Web Documents," in DEXA, Berlin, Heidelberg, 2010.

[28] X. Cao, G. Cong, C. S. Jensen and M. L. Yiu, "Retrieving Regions of Interest for User Exploration," Proceedings of the VLDB Endowment, vol. 7, no. 9, pp. 733-744 , May 2014.

[29] L. Chen, G. Cong and X. Cao, "An Efficient Query Indexing Mechanism for Filtering Geo-Textual Data," in SIGMOD '13, New York, USA, 2013.

[30] D. Zhang, K. Tan and A. K. Tung, "Scalable Top-K Spatial Keyword Search," in EDBT, Genoa, Italy, 2013.

[31] C. Zhang, Y. Zhang, W. Zhang and X. Lin, "Inverted Linear Quadtree: Efficient Top K Spatial Keyword Search," IEEE Transactions on Knowledge and Data Engineering, vol. 28, no. 7, pp. 1706-1721, July 2016.

[32] J. Fan, G. Li, L. Zhou, S. Chen and J. Hu, "Seal: spatio-textual similarity sea," Proceedings of the VLDB Endowment, vol. 5, no. 9, pp. 824-835, May 2012.