ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Visit: www.ijirset.com

Vol. 7, Issue 5, May 2018

Search nearest Keyword Set in Multi-Dimensional Datasets

Nayana N. Agwan, Prof. N. G. Pardeshi

Department of Computer Engineering SRES COE, Kopargaon, Maharashtra, India

ABSTRACT: In this paper consider objects like images, docu-ments are tagged with keywords. These objects characterized by a collection of features and represented as points in a multidimensional feature space. In this paper it look at nearest keyword set queries on textual content high multidimensional dataset. Existing techniques using tree-based indexes suggest viable solutions to NKS queries on multidimensional datasets but performance of these algorithms deteriorates with the growth of length or dimensionality in datasets. In this paper it propose Projection and Multi Scale Hashing (ProMiSH) to permit fast processing for NKS queries and also it uses random projection and hash-based index structures to achieves high scalability.

KEYWORDS:multi-dimensional data, indexing, hashing.

I. INTRODUCTION

Data mining is that the process of determine patterns in massive information sets involving strategies at the inter-section of artificial intelligence, machine learning, statistics, and information systems. It is a knowledge base subfield of technology. The overall goal of data mining method is to extract information from an inform- ation set and transform it into an evident structure for additional use. In this paper multidimensional dataset in which objects such as images are represented as points by extracting its color histogram and having descriptive text information associated with them. These data point has a set of keywords. The presence of keywords in feature space allows for the development of new tools to query and extend these multi-dimensional datasets.

In this paper nearest keyword set refer as NKS queries on text-rich multi-dimensional datasets. An nearest keyword set query is a set of user-provided keywords and the result of the query include data points each of which contains all the query keywords and gives the result as top-k tightest cluster in the multi-dimensional space. For obtaining top-k tightest cluster result in multidimensional space it uses ProMiSH . ProMiSH-E uses a set of hashtables and inverted indexes to perform a localized search. ProMiSH-E creates hashtables at multiple bin-widths, called index levels. The hashing technique is inspired by Locality Sensitive Hashing (LSH) which is a state-of-the-art method for nearest neighbor search in high-dimensional spaces. Unlike LSH-based methods that allow only approximate search with probabilistic guarantees, the index structure in ProMiSH-E supports accurate search. A single round of search in a hashtable yields subsets of points that contain query results, and ProMiSH-E explores each subset using a fast pruning-based algorithm. NKS queries are useful for many applications, such as photo-sharing in social networks, graph pattern search, geolocation search in GIS systems .

II. REVIEW OF LITERATURE

Vishwakarma Singh, Bo Zong, and Ambuj K. Singh[1] In this paper, nearest keyword set referred to as NKS queries on text-rich multi-dimensional datasets. NKS query is a set of user-provided keywords, and the result of the query include k sets of data points each of which contains all the query keywords and forms one of the top-k tightest

ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Visit: www.ijirset.com

Vol. 7, Issue 5, May 2018

cluster in the multi-dimensional space NKS queries are useful for many applications, such as photo-sharing in social networks, graph pattern search, geolocation search in GIS systems and so on.

D. Zhang, B. C. Ooi, and A. K. H. Tung[2] it explain about Web 2.0,it focus on the fundamental application of lo-cating geographical resources.In Web 2.0, tagging is a popular means to annotate various resources, including news, blogs, speeches, photos and videos. Users are encouraged to add extra textual terms as semantic description or summarization for the objects. With human intelligence involved, the tags are well phrased so that much cost can be saved from handling term ambiguities. An inverted index is built along with the R-tree. It maintains inverted lists for all the tags in the database. Each element in the list consists of the node label derived from the construction of the R-tree and the actual location. Note that the list of locations are ordered by the label so that the data points close to each other in geographical space are probably still close in the inverted lists. Such an index is scalable in terms of both the number of locations and tags. Advantage of this system is an efficient tagcentric query processing strategy but drawback is the number of tags associated with each object is typically small,making it difficult for an object to capture the complete semantics in the query objects.

X. Cao, G. Cong, C. S. Jensen, and B. C. Oo[3] it explain about the Collective spatial keyword querying con-sidered as standard spatial keyword queries. These all involve different conditions on the spatial and textual aspects of places. In spatial databases, the arguably most fundamental queries are range queries and k nearest neighbor queries. In text retrieval, queries may be Boolean, requiring results to contain the query
keywords, or ranking-based, returning the k places that rank the highest according to a text similarity function.

R. Hariharan, B. Hore, C. Li, and S. Mehrotra[4] it explain the location based information is stored in GIS database. Location-specific keyword queries on the web and in the GIS systems [4], [5],[7] were earlier answered us-ing a combination of R-Tree [8] and inverted index. Tree-based indexes, such as R-Tree [8] and M-Tree [10], have been extensively investigated for nearest neighbor search in high-dimensional spaces.These information entities of such databases have both spatial and textual descriptions. This method introduces a framework for GIR system and focus is on indexing strategies that can process spatial keyword query. It introduces two index structures to store spatial and textual information. 1)Separate index for spatial and text attributes 2) Hybrid index. But by using first structure that is separate index for spatial and text attributes, if filtering is done first, many objects may lie within a query is spatial extent, but very few of them are relevant to query keywords. This increases the disk access cost by generating a large number of candidate objects. The subsequent stage of keyword filtering becomes expensive. And by using second structure that is hybrid index there are high overhead in subsequent merging process.

A. Khodaei, C. Shahabi, and C. Li[5] this paper explain the how large amount of location-based information generated and used by many applications. The Internet is the most popular source of data with location-specific information, such as documents describing schools at certain regions, Wikipedia pages containing spatial information and images with annota-tions and information about the places they were taken. Users of such a web-based application often need to query the system by providing requirements on a location as well as keywords in order to find relevant documents there is a significant commercial and research interest in location based web search engines. Given a number of search keywords and one or more locations that a user is interested in, a location-based web search retrieves and ranks the most textually and spatially relevant web pages. In this type of search, both the spatial and textual information should be indexed. Currently, no efficient index structure exists that can handle both the spatial and textual aspects of data simultaneously and accurately.In this paper, it propose a new index structure called Spatial-Keyword Inverted File to handle location-based web searches in an integrated/ efficient manner. To seamlessly find and rank relevant documents, develop a new distance measure called spatial tf-idf. Advantage is to perform top k searches but give poor performance.

V. Singh and A. K. Singh[6] In 2012, V. Singh and A. K. Singh proposed SIMP for r-NN queries in a high dimensional space. SIMP offers both 100 percent accuracy and efficiency for any query range unlike state-of-the-art meth-ods. SIMP uses projection,spatial intersection, and triangle inequality to achieve a high rate of pruning, and thus

ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Visit: www.ijirset.com

Vol. 7, Issue 5, May 2018

gains high performance. It efficiently implemented the spatial inter-section approach by hashing.It also developed statistical cost models to measure SIMPs performance.SIMP captures data distribution through its viewpoints. Datar et al.[9] used ran-dom vectors constructed from p-stable distributions to project points, computed hash keys for the points by splitting the line of projected values into disjoint bins, and then concatenated hash keys obtained for a point from m random vectors to create a final hash key for the point.

III. PROPOSED WORKS

Admin first generate dataset of images and document which containing text file. Generating dataset for images admin crawl the images from online through flicker and store into database so every image store with tagged and view count and the url of images.Generating dataset for document admin first upload the text file and store in database with file name,file description and file use. User enters the query and apply the ProMiSH search algorithm. NKS query is a set of user-provided keywords, and the result of the query may include k sets of data points each of which contains all the query keywords and gives the nearest keyword set result in the multi-dimensional space.

ProMiSH algorithm include following steps:

1: To construct hashtable inverted index.

2: To obtain the data points from extracted feature.

3: To search point having keywords.

4: To check duplicate candidate in subset of points.

5: To perform search in subset.

6: To find candidates.

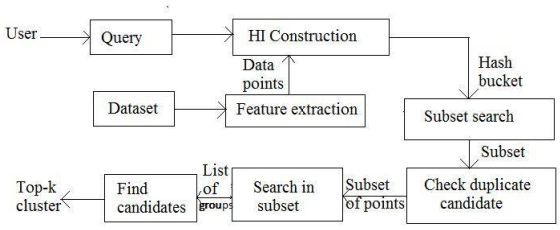

Fig. 1.System Architecture

Start with the index for exact ProMiSH . This index consists of two main parts. First Inverted Index Ikp in which treat keywords as keys, and each keyword points to a set of data points that are associated with the keyword. Second part is Hashtable-Inverted Index Pairs HI consists of multiple hash tables and inverted indexes .It having three paremeters index level,Number of unit random vectors and hashtable size.Project all the data points in D on a unit random vector and partition the projected values into overlapping bins of bin-width.

ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Visit: www.ijirset.com

Vol. 7, Issue 5, May 2018

To obtain the data points first images are transformed into grayscale and consider d be the desired dimensional-ity.Then convert each image into a d-dimensional point by extracting its color histogram, and associate each data point with a set of keywords.

To search point having keywords starts with the HI structure at index level s =0. ProMiSH-E finds the buckets in hashtable each of which contains all the query keywords by inverted index.

To check duplicate candidate in subset of points. So first subset contains points so it check the explored or not.Then uses hashtable to check duplicates for subset .Generate the hash values and concatenated to get a hash key for subset or element wise match of subset is performed.Otherwise subset is stored in hashtable using hash key.

The kth smallest diameter rk is retrieved from the priority queue PQ.For a given subset and a query Q, all the points are grouped using query keywords.A pairwise inner join of the groups is performed.An adjacency list AL stores the distance between points which satisfy the distance threshold rk.An adjacency list M stores the count of point pairs obtained for each pair of groups by the inner join.

An intermediate tuple curSet is checked against each point of list of group and determined using adjacency list whether the distance between the last point in curSet and a point in list of groups is at most rk . Then, the point is checked against each point in curSet for the distance predicate update diameter of curSet. If a point satisfies the distance predicate with each point of curSet, then newCurSet is formed .A candidate is found if curSet has a point from every group.

IV. CONCLUSION

In this paper, it gives solutions to the problem of top-k nearest keyword set search in multi-dimensional datasets.So for this purpose it developed ProMiSH which is based on ran-dom projection and hashing.ProMiSH-E is the exact ProMiSH which gives the optimal subset of points and approximate ProMiSH that searches near optimal result with better effi-ciency.So from above algorithm result show that ProMiSH is faster than state of the art tree method. In the future, we plan to explore other scoring schemes for ranking the result sets. In one scheme, techniques like tf-idf assign weights to the keywords of a point .

ACKNOWLEDGEMENT

I would like to thank my esteemed guide Prof.N.G.Pardeshi whose interest and guidance helped meto complete the work successfully.Also under the valuable help of Dr.D.B.Kshirsagar(HOD Comp.Dept.) and Prof. P.N.Kalavadekar ”PG Co-ordinator” who has provided facilities to explore the subject with more enthusiasm.

REFERENCES

[1] Vishwakarma Singh, Bo Zong, and Ambuj K. Singh, “Nearest Keyword Set Search in Multi-Dimensional Datasets”, VOL. 28, NO. 3, MARCH 2016.
[2] D. Zhang, B. C. Ooi, and A. K. H. Tung, “Locating mapped resources in web 2.0 ”in Proc. IEEE 26th Int. Conf. Data Eng., 2010, pp. 521532.
[3] X. Cao, G. Cong, C. S. Jensen, and B. C. Ooi, “Collective spatial keyword querying”,in Proc. ACM SIGMOD Int. Conf. Manage. Data, 2011, pp. 373384.
[4] R. Hariharan, B. Hore, C. Li, and S. Mehrotra, “Processing spatial keyword (SK)queries in geographic information retrieval (GIR) systems ”, in Proc. 19th Int. Conf. Sci. Statistical Database Manage., 2007, p. 16.
[5] A. Khodaei, C. Shahabi, and C. Li, “Hybrid indexing and seamless ranking of spatial and textual features of web documents ”, in Proc. 21st Int. Conf. Database Expert Syst. Appl., 2010, pp. 450466.
[6] V. Singh and A. K. Singh, “SIMP: Accurate and efficient near neighbor search in high dimensional spaces”, in Proc. 15th Int. Conf. Extending Database Technol., 2012, pp. 492503.
[7] Y. Zhou, X. Xie, C. Wang, Y. Gong, and W.-Y. Ma, “Hybrid index structures for location-based web search ”, in Proc. 14th ACM Int. Conf. Inf. Knowl. Manage., 2005, pp. 155162.
[8] A. Guttman, “R-trees: A dynamic index structure for spatial searching,”, in Proc. ACM SIGMOD Int. Conf. Manage. Data, 1984, pp. 4757.
[9] M. Datar, N. Immorlica, P. Indyk, and V. S. Mirrokni, “Locality sensitive hashing scheme based on p-stable distributions,”, in Proc. 20th Annu. Symp. Comput. Geometry, 2004, pp. 253262.
[10] P. Ciaccia, M. Patella, and P. Zezula, “M-tree: An efficient access method for similarity search in metric spaces,”, in Proc. 23rd Int. Conf. Very Large Databases, 1997, pp. 426435.