A Novel Hybrid Keyword Set Approach in Multi-dimensional Datasets

P.Vijaya Kumari [1], Dr. Mohammed Ali Hussain[2],Ravi Kumar Tenali[3]

[1]*M.Tech, Dept. of CSE, AndhraLoyolainstitute of Engineering and technology, Vijayawada.*

[2]*Professor, Dept. of CSE, Andhra Loyola institute of Engineering and technology, Vijayawada.*

[3] *Assistant Professor, Dept. of CSE, Andhra Loyola institute of Engineering and technology, Vijayawada.*

Abstract**:** *Keyword-based search in text-rich multidimensional datasets facilitates many fresh applications and tools. In this paper, we Ponder objects that are tagged with keywords and are embedded in a vector space. For these datasets, we perusal queries that ask for the tightest groups of points satisfying a given set of keywords. We propose a novel method called Promise (Projection and Multi Scale Hashing) that uses Irregular projection and hash-based index structures, and achieves high scalability and speedup. Our adaptability tests on datasets of sizes up to 10 million and measurements up to 100 for solicitations having up to 9 catchphrases demonstrate that ProMiSH scales directly with the Dataset estimate, the dataset measurement, the inquiry estimate, and the outcome measure. We show a Particular and an inexact adaptation of the calculation. Our experimental results on real and synthetic datasets show that ProMiSH has up To 60 times of speedup over state-of-the-art tree-based techniques. We are handling the spatial queries jointly and returns the only user specified number of optimal results; we implemented a cache based approach for efficient results.*

Index Terms**:***Querying, multi-dimensional data, indexing, hashing*

I. INTRODUCTION

In today's digital world the amount of data which is developed is increasing day by day. There is different multimedia in which data is saved. It’s very difficult to search the large dataset for a given query as well to archive more accuracy on user query. In the same time query will search on dataset for exact keyword match and it will not find the nearest keyword for accuracy. Ex: Flickr.

The amount of data which is developed is increasing day by day, thus it is very difficult to search large dataset for a given query as well to achieve more accuracy on user query. So we have implemented a method of efficient search in multidimensional dataset. This is associated with images as an input. Images are often characterized by a collection of relevant features, and are commonly represented as points in a multi-dimensional feature space. For example, images are represented using colour feature vectors, and usually have descriptive text information (e.g., tags or keywords) associated with them. We consider multi-dimensional datasets where each data point has a set of keywords. The presence of keywords in feature space allows for the development of new tools to query and explore these multidimensional datasets.

Our main contributions are summarized as follows.

(1) We propose a novel multi-scale index for exact and approximate NKS query processing.

(2) We develop efficient search algorithms that work with the multi-scale indexes for fast query processing.

(3) We conduct extensive experimental studies to demonstrate the performance of the proposed techniques.

1. Filename: It is based on image filename.

2. CBIR (Content based picture seek): Content-based picture recovery (CBIR), otherwise called inquiry by picture content (QBIC) and substance based visual data recovery (CBVIR) is the utilization of PC vision systems to the picture recovery issue, that is, the issue of looking for computerized pictures in expansive databases. Contentbased picture recovery is against

customary conceptbased approaches (see Concept-based picture ordering).

3. TBIR (Text based image search): Concept-based image indexing, also variably named as “description-based” or “text-based” image indexing/retrieval, refers to retrieval from text-based indexing of images that may employ keywords, subject headings, captions, or natural language text. It is opposed to Content-based image retrieval. Indexing is a technique used in CBIR**.**

Table -1: Comparison Table

	Filename	*CBIR*	*TBIR*	*NKS (Extended TBIR)*
No. of Result	*Highest*	*Low*	*High*	*Low*
Accuracy	*Low*	*High*	*Medium*	*High*
Performance	*Highest*	*Low*	*High*	*High*
User Satisfaction	*<50%*	*90-100 %*	*60-80%*	*90-100%*

2. LITERATURE SURVEY

We study nearest keyword set (referred to as NKS) queries on text-rich multi-dimensional datasets. An 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 forms one of the top-k tightest clusters in the multi-dimensional space. Illustrates an NKS query over a set of two-dimensional data points. Each point is tagged with a set of keywords. For a query the set of points contains all the query keywords and forms the tightest cluster compared with any other set of points covering all the query keywords. Therefore, the set is the top-1 result for the query Q.NKS queries are useful for many applications, such as photo-sharing in social networks, graph pattern search, geo-location search in GIS systems and so on.

We present an exact and an approximate version of the algorithm. Our experimental results on real and synthetic datasets show that the method has more speedup over stateof-the-art tree-based techniques.

Other related queries include aggregate nearest keyword search in spatial databases, top-k preferential query, top-k sites in a spatial data based on their influence on feature points, and optimal location queries. Our work is different from these techniques. First, existing works mainly focus on the type of queries where the coordinates of query points are known. Even though it is possible to make their cost functions same to the cost function in NKS queries, such tuning does not change their techniques. The proposed techniques use location information as an integral part to perform a best first search on the IR-Tree, and query coordinates play a fundamental role in almost every step of the algorithms to prune the search space. Moreover, these techniques do not provide concrete guidelines on how to enable efficient processing for the type of queries where query coordinates are missing. Second, in multi-dimensional spaces, it is difficult for users to provide meaningful coordinates, and our work deals with another type of queries where users can only provide keywords as input. Without query coordinates, it is difficult to adapt existing techniques to our problem.

Finding closest neighbors in expansive multi-dimensional information has dependably been one of the examination interests in information mining field. In this paper, we introduce our nonstop research on similitude seek issues. Previous work on exploring the meaning of K nearest neighbors from a new perspective in Pan KNN. It redefines the distances between data points and a given query point Q, efficiently and effectively selecting data points which are closest to Q. It can be applied in various data mining fields. A lot of genuine informational collections have immaterial or impediment data which enormously influences the viability and proficiency of discovering closest neighbors for a given question information point. In this paper, we show our way to deal with tackling the likeness seek issue within the sight of hindrances. We apply the idea of deterrent focuses and process the comparability look issues in an unexpected way. This approach can help to enhance the execution of existing information examination approaches. The closeness between two information guides utilized toward be founded on a likeness capacity, for example, Euclidean separation which totals the distinction between each measurement of the two information focuses in customary closest neighbor issues.

In those applications, the closest neighbor issues are illuminated in light of the separation between the information point and the inquiry point over a settled arrangement of measurements (components). Likewise early techniques experience the ill effects of the "scourge of dimensionality". In a high dimensional space the information are typically meager, and generally utilized separation metric, for example, Euclidean separation may not function admirably as dimensionality goes higher. Late research [8] demonstrates that in high measurements closest neighbor questions end up noticeably insecure: the distinction of the separations of most distant and closest indicates some inquiry point does not increment as quick as the base of the two, along these lines the separation between two information focuses in high dimensionality is less significant. Some methodologies are proposed focusing on fractional likenesses. However, they have limitations such as the requirement of the fixed subset of dimensions, or fixed number of dimensions as the input parameter(s) for the algorithms. Keyword-based search in text-rich multi-dimensional datasets facilitates many novel applications and tools. We consider objects that are tagged with keywords and are embedded in a vector space. For these datasets, we study queries that ask for the tightest groups of points satisfying a given set of keywords. We propose a method that uses random projection and hash-based index structures, and achieves high scalability and speedup. However, none of these algorithms considers detecting outliers simultaneously with clustering process. As a rule, anomalies are as essential as bunches, for example, charge card extortion location, disclosure of criminal exercises, revelation of PC interruption, and so on. Dissecting the information circulation with the thought of snags is basic for some informational collections.

As of late, different general strategies for investigation of development information and human exercises specifically were proposed. Distinctive methods for 3D geo-perception of space-time examples of individuals' travel involvement and portability is displayed in .Two sorts of calculations for mining intriguing examples from directions gained by GPSenabled gadgets are proposed. In the first type, the trajectories are converted into a sequence of stops or important parts (regions in which an object stayed more than a predefined time interval) before the algorithm for mining interesting patterns is applied. In the second type, the identification of important parts in a trajectory is part of the algorithm for mining patterns. Progressive clustering of trajectories of moving objects is presented. The authors combined clustering with visual interaction to let the analyst apply different distance functions based on the particular characteristics of trajectories under investigation. Visualization techniques (aggregations, ring maps) of daily repeating activities like travel, work, shopping are presented. An algorithm for finding interesting places and mining travel sequences from GPS trajectories is proposed. The algorithm detects frequent sequences on different scales, taking into account the interestingness of the visited place and the experience of a user. Research on movement data is usually done on trajectories acquired by GPS-enabled devices. However, large scale GPS datasets, which would allow us to perform qualitative analysis on the level of a city or country, are still not available. On the other hand, geo tagged photo collections could be obtained on the world scale, which makes them a valuable resource for the analysis of people’s activities. Concentration and movement of tourists at the scale of a city is analyzed using Flickr geo tagged photos. For this, the identified tourists in the city of Rome using user profiles and built heat maps to visualize regions of high tourist concentration. The heat maps were created by dividing a region into cells, counting then number of people who took photos in every cell and smoothing the visualization by interpolating between values of every cell. Nonetheless, no point by point examination of the technique, its favorable circumstances and detriments was given. Moreover, stream maps were utilized to imagine traveler development between went to places. These spots were associated by lines whose widths were relative to the quantity of vacationers. Mean-move, a non-parametric bunching calculation, was utilized as a part of to locate the most appealing spots on Earth on a neighborhood and city scale utilizing Flickr photographs. The spoke to cases of maps with developments of individuals. In any case, no point by point examination of the development was introduced.

Photograph sharing sites, for example, Flickr and Panoramio contain a large number of geo labeled pictures contributed by individuals from everywhere throughout the world. Qualities of these information posture new difficulties in the area of spatio-fleeting examination. In this paper, we characterize a few distinct assignments identified with investigation of appealing spots, purposes of premium and examination of behavioral examples of various client groups on geo labeled photograph information. We perform investigation and examination of transient occasions, rankings of touring spots in a city, and study versatility of individuals utilizing geotagged photographs. We adopt an efficient strategy to finish these errands by applying versatile computational systems, utilizing factual and information mining calculations, consolidated with intuitive geo-representation. We give exploratory visual examination condition, which permits the expert to recognize spatial and fleeting examples and concentrate extra information from expansive geo-labeled photograph accumulations. We exhibit our approach by applying the techniques to a few locales on the planet.

Gigantic measure of information have been produced in many teaches these days. The similitude seek issue has been considered in the most recent decade, and numerous calculations haves been proposed to comprehend the K closest neighbor look. Previously proposed Pan KNN which is a novel technique that explores the meaning of K nearest neighbors from a new perspective. It redefines the distances between data points and a given query point Q, and selects data points which are closest to Q efficiently and effectively. In this paper, first a brief introduction about previous work on Pan KNN and discuss the Fuzzy concept; then, we propose to use the Fuzzy concept to design OPan KNN algorithm that targets solving the nearest neighbors problems in the presence of obstacles.

3. EXISTING SYSTEM

In this paper, we propose ProMiSH (short for Projection and Multi-Scale Hashing) to enable fast processing for NKS queries. In particular, we develop an exact ProMiSH (referred to as ProMiSH-E) that always retrieves the optimal top-k results, and an approximate ProMiSH (referred to as ProMiSHA) that is more efficient in terms of time and space, and is able to obtain near-optimal results in practice. ProMiSHE uses a set of hash tables and inverted indexes to perform a localized search. Based on this index, we developed ProMiSHE that finds an optimal subset of points and ProMiSH-A which searches near-optimal results with better efficiency. ProMiSH is faster than state-of-the-art tree-based techniques, with multiple orders of magnitude performance improvement. Tree-based indexes, such as R-Tree and M-Tree have been extensively investigated for nearest neighbor search in high-dimensional spaces. These indexes fail to scale to dimensions greater than 10 because of the curse of dimensionality. Random projection with hashing has come to be the state-of-theart method for nearest neighbor search in high-dimensional datasets. Data ret al. used random 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. Our problem is different from nearest neighbor search. NKS queries provide no coordinate information, and aim to Find the top-k tightest clusters that cover the input keyword set. Meanwhile, nearest neighbor queries usually require coordinate information for queries, which makes it difficult to develop an efficient method to solve NKS queries by existing techniques for nearest neighbor search. In addition, multiway distance joins for a set of multidimensional datasets have been studied in tree based index is adopted, but suffers poor scalability with respect to the dimension of the dataset. Furthermore, it is not straightforward to adapt these algorithms since every query requires a multi-way

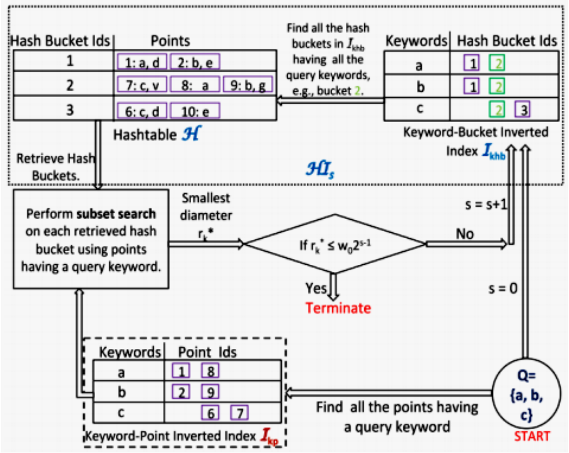
Hash Bucket Ids
Points
1
2
3
1: a, d
2: b, e
7: c, v
8: a
9: b, g
6: c, d
10: e
Hashtable $\mathcal{H}$
Find all the hash buckets in I_{khb} having all the query keywords, e.g., bucket 2.
Keywords
Hash Bucket Ids
a
b
c
1
2
1
2
2
3
Keyword-Bucket Inverted Index I_{khb}
Retrieve Hash Buckets.
$\mathcal{HI}_s$
Perform subset search on each retrieved hash bucket using points having a query keyword.
Smallest diameter r_k^*
If $r_k^* \leq w_0 2^{s-1}$
s = s+1
No
Yes
Terminate
s = 0
Keywords
Point Ids
a
b
c
1
8
2
9
6
7
Keyword-Point Inverted Index I_{kp}
Find all the points having a query keyword
Q= {a, b, c}
START

Fig 3: Index structure and flow of execution of ProMiSH.

distance join only on a subset of the points of each dataset. Scored based on distance between points and weights of keywords. Furthermore, the criteria of a result containing all the keywords can be relaxed to generate results having only a subset of the query keywords. We plan to explore the extension of ProMiSH to disk. ProMiSH-E sequentially reads only required buckets from Ikp to find points containing at least one query keyword. Therefore, Ikp can be stored on disk using a directory-file structure. We can create a directory for Ikp. Each bucket of Ikp will be stored in a separate file named after its key in the directory. Moreover, ProMiSH-E sequentially probes HI data structures starting at the smallest scale to generate the candidate point ids for the subset search, and it reads only required buckets from the hash table and the inverted index of a HI structure. Therefore, all the hash tables and the inverted indexes of HI can again be stored using a similar directory- file structure as Ikp, and all the points in the dataset can be indexed into a B+-Tree using their ids and stored on the disk. In this way, subset search can retrieve the points from the disk using B+-Tree for exploring the final set of results.

IV. CONCLUSION

We proposed solutions to the problem of top-k nearest keyword set search in multi-dimensional datasets. We proposed a novel index called ProMiSH based on random projections and hashing. Based on this index, we developed ProMiSHE that finds an optimal subset of points and ProMiSH-A that searches near-optimal results with better efficiency. Our empirical results show that ProMiSH is faster than state-of-theart tree-based techniques, with multiple orders of magnitude performance improvement. Moreover, our techniques scale well with both real and synthetic datasets. Ranking functions. In the future, we plan to explore other scoring schemes for ranking the result sets. In one scheme, we may assign weights to the keywords of a point by using techniques like tf-idf. Then, each group of points can be relaxed to generate results having only a subset of the query keywords.

REFERENCES

[1] W. Li and C. X. Chen, “Efficient data modeling and querying system for multi-dimensional spatial data,” in Proc. 16th ACM SIGSPATIAL Int. Conf. Adv. Geographic Inf. Syst., 2008, pp. 58:1 58:4.

[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. 521-532.

[3] V. Singh, S. Venkatesha, and A. K. Singh, “Geo-clustering of images with missing geotags,” in Proc. IEEE Int. Conf. Granular Comput., 2010, pp. 420-425.

[4] V. Singh, A. Bhattacharya, and A. K. Singh, “Querying spatial patterns,” in Proc. 13th Int. Conf. Extending Database Technol. Adv. Database Technol., 2010, pp. 418-429.

[5] J. Bourgain, “On lipschitz embedding of finite metric spaces in hilbertspace,”Israel J. Math., vol. 52, pp. 46-52, 1985.

[6] H. He and A. K. Singh, “GraphRank: Statistical modeling and mining of significant subgraphs in the feature space,”in Proc. 6th Int. Conf. Data Mining, 2006, pp. 885- 890.

[7] 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. 373-384.

[8] C. Long, R. C.-W. Wong, K. Wang, and A. W.-C. Fu, “Collective spatial keyword queries: A distance owner driven approach,” in Proc. ACM SIGMOD Int. Conf. Manage. Data, 2013, pp. 689-700.

[9] 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. IEEE 25th Int. Conf. Data Eng., 2009, pp. 688699.

[10] M. Datar, N. Immorlica, P. Indyk, and V. S. Mirrokni, Localitysensitive hashing scheme based on p-stable distributions, in Proc. 20th Annu.Symp.Comput. Geometry, 2004, pp. 253262.

[11] 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.

[12] R. Hariharan, B. Hore, C. Li, and S. Mehrotra, Processing spatialkeyword (SK) queries in geographic information retrieval (GIR) systems, in Proc. 19th Int. Conf. Sci. Statistical Database Manage., 2007, p. 16.

[13] S. Vaid, C. B. Jones, H. Joho, and M. Sanderson, Spatio-textual indexing for geographical search on the web, in Proc. 9th Int. Conf. Adv. Spatial Temporal Databases, 2005, pp. 218235.

[14] 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.

[15] A. Guttman, R-trees: A dynamic index structure for spatial searching, in Proc. ACM SIGMOD Int. Conf. Manage. Data, 1984 , pp. 4757.