IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

'SEARCHING NEAREST KEYWORD SET' IN MULTIDIMENTIONAL DATASETS USING PROJECTION AND MULTISCALE HASHING

[1]Miss.Patil Trupti Satish, [2]Prof.Mr.Mophare A.V.

[1]Student of M.E.CSE, [2] Professor in Computer Science & Enigineering
[1] Department Computer Science ,
N. B. Navale Sinhgad College of Engineering, Pune Road, Kegaon, Solapur, India

Abstract: This study has been undertaken to improve the searching methodologies for text, images and Multidimensional Datasets in modern search engines using Projection as well as Hashing For these datasets, study queries that ask for the tightest groups of points satisfying a given set of keywords. Here we propose a method called Projection & Multi-Scale Hashing that uses random projection, hash-based index structures which would achieve high scalability and speedup. We also propose a deterministic version of the Projection & Multi-scale hashing.

Index Terms **- Searching, Hashing, Datasets, Projection, NKS, Color, Vector, Tags, Keywords**

I. Introduction

The Objects we can consider text, images, documents, classified or organized by a collection of relevant features, and are commonly represented as points in a multi-dimensional feature space. For example, images are represented using **color** feature vectors, and generally have descriptive textual contents (e.g., tags or keywords) associated with them. In this article, 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 multi-dimensional datasets. In this article, we study and understand **nearest keyword set known 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 cluster in the multi-dimensional space. Fig. 1 illustrates an NKS query over a set of two-dimensional data points. Each point is tagged with a set of keywords. For a query Q ¼ fa; b; cg, the set of points {7; 8; 9} contains all the query keywords fa; b; cg and forms the tightest cluster compared with any other set of points covering all the query keywords. Therefore, the set f{7; 8; 9} is the top-1 result for the query Q.

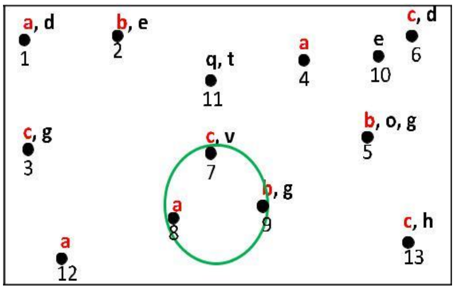

Fig:1 An example of an NKS query on a keyword tagged multi-dimensional dataset. top-1 result for query {a,b,c} is the set of points {7,8,9}

II. Literature Review

Location-specific keyword queries on the web and in the GIS systems were earlier answered using a combination of R-Tree and inverted index. Felipe et al. developed IR2-Tree to rank objects from spatial datasets based on a combination of their distances to the query locations and the relevance of their text descriptions to the query keywords. Cong et al. Integrated R-tree.

Existing work mainly focuses on the type of query where the coordinates of query points are known [7], [8]. 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.

Develop a novel index structure based on random projection with hashing. Unlike tree-like indexes adopted in existing works, our index is less sensitive to the increase of dimensions and scales well with multi-dimensional data.

III. Methodology

Multidimensional Data Model and Data Sets

A multidimensional database (MDB) is a type of database that is optimized for data warehouse and online analytical processing (OLAP) applications. Multidimensional databases are frequently created using input from existing relational databases. Whereas a relational database is typically accessed using a Structured Query Language (SQL) query, a multidimensional database allows a user to ask questions like "How many Aptivas have been sold in Nebraska so far this year?" and similar questions related to summarizing business operations and trends. An OLAP application that accesses data from a multidimensional database is known as a MOLAP (multidimensional OLAP) application.

A multidimensional database - or a multidimensional database management system (MDDBMS) - implies the ability to rapidly process the data in the database so that answers can be generated quickly. A number of vendors provide products that use multidimensional databases. Approaches to how data is stored and the user interface vary.

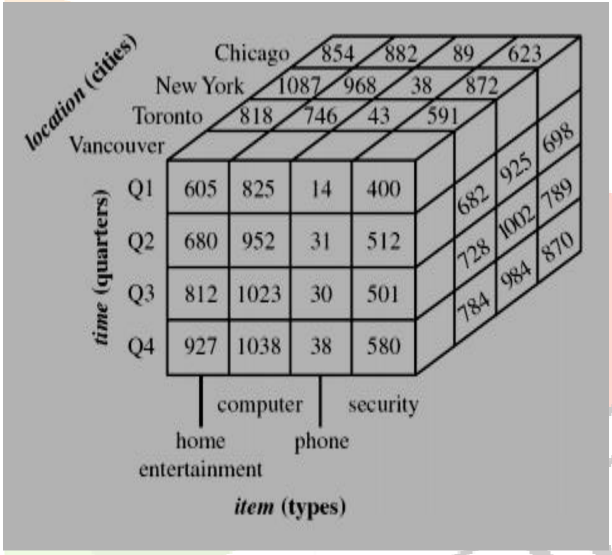

Fig 2: Multidimensional Data Model

Novel Multi Scale Hashing Approach Algorithm

Novel Multi Scale Hashing Approach Algorithm utilizes the syntactic data set is composed from various data resources. In which we gather location and item examine with distance calculation between the source and destinations by using a boot strapping algorithm, with an established keywords that are resultant from tags. The question co-ordinates play an essential role in each phase of the algorithm to trim investigation space. Our effort deals with if keyword as an input. In the proposed approach which is a new multi-scale catalogue for precise and near NKS enquiry dispensation and progress effective search procedures that effort with the multi-scale directories for fast query processing. Distance looking is also very informal with R-trees. In detail, the best-first process is accurately considered to yield data ideas in rising directive of their reserves. In order to run the submission resourcefully, the user must have subsequent features. User offers the keyword i,e items as an input.

IV. Implementation and Result

1. Propose a novel multi-scale index for exact and approximate NKS query processing.

2. Develop efficient search algorithms that work with the multi-scale indexes for fast query processing.

3. Conduct extensive experimental studies to demonstrate the performance of the proposed techniques.

1. Filename: It is based on image filename.

2. CBIR (Content based image search): Content-based image retrieval (CBIR), also known as query by image content (QBIC) and content-based visual information retrieval (CBVIR) is the application of computer vision techniques to the image retrieval problem, that is, the problem of searching for digital images in large databases. Content-based image retrieval is opposed to traditional concept-based approaches.

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

V. RESULTS AND DISCUSSION

Following table shows the comparison among Filename Based Search, Content based image search & Text based image search based on parameters like No. of Result, Accuracy, and Performance &User Satisfaction. After implementation performance of Nearest Keyword Search (NKS) system would be validated by using above parameter which is given by ‘X’.

Comparison Table

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

VI. CONCLUSION AND FUTURE WORKS

This topic proposes solution to the problem of top-k nearest keyword set search in multi-dimensional datasets and proposes a novel index called Projection & Multi-Scale Hashing based on random projections and hashing. Based on this index, develop Projection & Multi-Scale Hashing-Exact that finds an optimal subset of points and Projection & Multi-Scale Hashing-Approximate that searches near-optimal results. Projection & Multi-Scale Hashing utilizes less indexing time than tree-based techniques. We are in hope that this techniques would scale well with both real and syntheticdatasets.

VII. ACKNOWLEDGMENT

I profoundly grateful to Prof. A.V. Mophare for his expert guidance and continuous encouragement throughout to see that this project rights its target since its commencement to its completion. I would like to express my deepest appreciation towards Principal Dr. S.D. Navale, Prof. B.R.Solunke, HOD department of computer engineering and PG coordinator Prof. B.R.Solunke. I must express my sincere heartfelt gratitude to all staff members of computer engineering department who helped me directly or indirectly during this course of work. Finally, I would like to thank my family and friends, for their precious support.

REFERENCES

[1] W. Li and C. X. Chen, “Efficient data modeling and queryingsystem for multi-dimensional spatial data,” in Proc. 16th ACMSIGSPATIAL 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 resourcesin 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 ofimages with missing geotags,” in Proc. IEEE Int. Conf. GranularComput., 2010, pp. 420–425.

[4] V. Singh, A. Bhattacharya, and A. K. Singh, “Querying spatialpatterns,” 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 Hilbert space,” Israel J. Math., vol. 52, pp. 46–52, 1985.

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

[7] X. Cao, G. Cong, C. S. Jensen, and B. C. Ooi, “Collective spatialkeyword 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, “Collectivespatial keyword queries: A distance owner-driven approach,” inProc. 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 searchingby document,” in Proc. IEEE 25th Int. Conf. Data Eng., 2009,pp. 688–699.

[10] M. Datar, N. Immorlica, P. Indyk, and V. S. Mirrokni, “Localitysensitivehashing scheme based on p-stable distributions,” inProc. 20th Annu. Symp. Comput. Geometry, 2004, pp. 253–262.

[11] Y. Zhou, X. Xie, C. Wang, Y. Gong, and W.-Y. Ma, “Hybrid indexstructures for location-based web search,” in Proc. 14th ACM Int.Conf. Inf. Knowl. Manage., 2005, pp. 155–162.

[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-textualindexing for geographical search on the web,” in Proc. 9th Int.Conf. Adv. Spatial Temporal Databases, 2005, pp. 218–235.

[14] A. Khodaei, C. Shahabi, and C. Li, “Hybrid indexing and seamlessranking of spatial and textual features of web documents,” in Proc.21st Int. Conf. Database Expert Syst. Appl., 2010, pp. 450–466

[15] A. Guttman, “R-trees: A dynamic index structure for spatialsearching,” in Proc. ACM SIGMOD Int. Conf. Manage. Data, 1984,pp. 47–57.

[16] I. De Felipe, V. Hristidis, and N. Rishe, “Keyword search on spatialdatabases,” in Proc. IEEE 24th Int. Conf. Data Eng., 2008,pp. 656–665.

[17] G. Cong, C. S. Jensen, and D. Wu, “Efficient retrieval of the top-kmost relevant spatial web objects,” Proc. VLDB Endowment, vol. 2,pp. 337–348, 2009.

[18] B. Martins, M. J. Silva, and L. Andrade, “Indexing and ranking inGeo-IR systems,” in Proc. Workshop Geographic Inf., 2005, pp. 31–34.

[19] Z. Li, H. Xu, Y. Lu, and A. Qian, “Aggregate nearest keywordsearch in spatial databases,” in Proc. 12th Int. Asia-Pacific WebConf., 2010, pp. 15–21.

[20] M. L. Yiu, X. Dai, N. Mamoulis, and M. Vaitis, “Top-k spatial preferencequeries,” in Proc. IEEE 23rd Int. Conf. Data Eng., 2007,pp. 1076–1085.

[21] T. Xia, D. Zhang, E. Kanoulas, and Y. Du, “On computing top-tmost influential spatial sites,” in Proc. 31st I
nt. Conf. Very LargeDatabases, 2005, pp. 946–957.