**Imperial Journal of Interdisciplinary Research (IJIR)**
**Vol-3, Issue-8, 2017**
**ISSN: 2454-1362,** http://www.onlinejournal.in

# Survey on Multi-Dimensional Datasets

Anju K, Dhanasree K S & Vincy Rajan
Department of Computer Science, RIT Kottayam

***Abstract**: Keyword-based search in text-rich multi-dimensional datasets facilitates many novel applications and tools. In this paper, the objects that are tagged with keywords are embedded in a vector space. For these datasets, queries where asked for the tightest groups of points satisfying a given set of keywords. A novel method called ProMiSH (Projection and Multi Scale Hashing) which uses random projection and hash-based index structures, and achieves high scalability and speedup. Thus, the paper introduces an exact and an approximate version of the algorithms.*

## 1. Introduction

Objects (e.g., images, chemical compounds, documents, or experts in collaborative networks) are often characterized by a collection of relevant features, and are commonly represented as points in a multi-dimensional feature space. 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.

In this survey there are around five methods. 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.

Keyword search is the most popular information discovery method because the user does not need to know either a query language or the underlying structure of the data. The search engines available today provide keyword search on top of sets of documents. When a set of query keywords is provided by the user, the search engine returns all documents that are associated with these query keywords. Solution to such queries is based on the $IR^2$-tree, but $IR^2$- tree having some drawbacks. Efficiency of $IR^2$-tree badly is impacted because of some drawbacks in it. The solution for overcoming this problem should be searched. Spatial inverted index is the technique which will be the solution for this problem. Spatial database manages multidimensional data that is points, rectangles.

## 2. Multi-dimensional datasets approach

### 2.1 Querying

Given a set of d-dimensional data points D, we assume data points are uniformly distributed in the buckets of a hash table, and keywords of each data point are uniformly sampled from the dictionary.

Suppose D has N data points, each data point has t keywords, and the keywords are sampled from a dictionary of U unique keywords. Let Nv be the number of data points with keyword v. The expectation of Nv is computed as follows,

$$E[N_v] = \sum_{i=1}^{N}(1-(1-\frac{1}{U})^t) = N(1-(1-\frac{1}{U})^t).$$

### 2.2 Multi-Dimensional Data

A multi-way distance joins for a set of multidimensional datasets. 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 distance join only on a subset of the points of each dataset.

Also 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.

### 2.3 Indexing

Here, they 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.

This index consists of two main components.

- Inverted Index $I_{kp}$ : The first component is an inverted index referred to as $I_{kp}$. In $I_{kp}$,

keywords is keys, and each keyword points to a set of data points that are associated with the keyword. Let D be a set of data points and V be a dictionary that contains all the keywords appearing in D.

- Hash table-Inverted Index Pairs HI : The second component consists of multiple hash tables and inverted indexes referred to as HI. HI is controlled by three parameters: (1) (Index level) L, (2) (Number of random unit vectors) m, and (3) (hash table size) B. All the three parameters are non-negative integers.

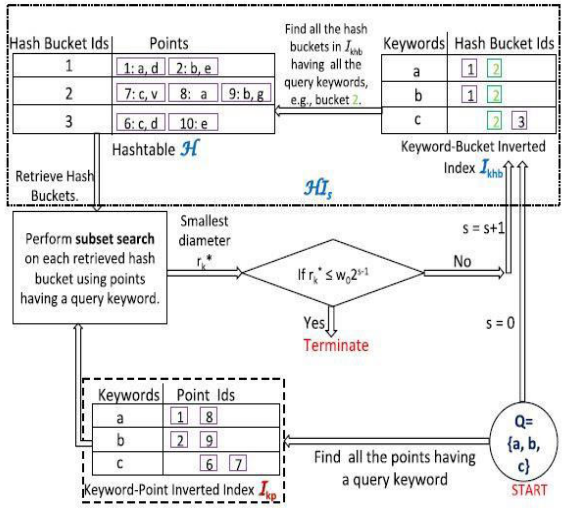


Figure 1. Index structure and flow of execution of ProMiSH.

### 2.4 Nearest Neighbour Search

Nearest neighbour search (NNS), also known as closest point search, similarity search. It is an optimization problem for finding closest (or most similar) points. Nearest neighbour search which returns the nearest neighbour of a query point in a set of points, is an important and widely studied problem in many fields, and it has wide range of applications. We can search closest point by giving keywords as input; it can be spatial or textual.

A spatial database use to manage multidimensional objects i.e. points, rectangles, etc. Some spatial databases handle more complex structures such as 3D objects, topological coverage's, linear networks. While typical databases are designed to manage various NUMERIC'S and character types of data, additional functionality needs to be added for databases to process spatial data type's efficiently and it provides fast access to those objects based on different selection criteria.

### 2.5 IR$^2$-tree

Keyword search is the most popular information discovery method because the user does not need to know either a query language or the underlying structure of the data. The search engines available today provide keyword search on top of sets of documents. When a set of query keywords is provided by the user, the search engine returns all documents that are associated with these query keywords. Solution to such queries is based on the IR$^2$-tree, but IR$^2$- tree having some drawbacks. Efficiency of IR$^2$-tree badly is impacted because of some drawbacks in it. The solution for overcoming this problem should be searched. Spatial inverted index is the technique which will be the solution for this problem. Spatial database manages multidimensional data that is points, rectangles.

## 3. Conclusion

In this paper, we proposed solutions to the problem of top-k nearest keyword set search in multi-dimensional datasets. A novel index called ProMiSH based on random projections and hashing . Based on this index an optimal subset of points and ProMiSH-A that searches near-optimal results with better efficiency. The empirical results show that ProMiSH is faster than state-of-the-art tree-based techniques, with multiple orders of magnitude performance improvement. Moreover, these techniques scale well with both real and synthetic datasets. An efficient incremental algorithm was presented that uses the IR$^2$-Tree to answer spatial keyword queries.

## 4. References

[1] Vishwakarma Singh, Bo Zong, and Ambuj K. Singh "Nearest Keyword Set Search in Multi-Dimensional Datasets", Ieee Transactions On Knowledge And Data Engineering, Vol. 28, No. 3, March 2016.

[2] W. Li and C. X. Chen, "Efficient data modelling 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.

[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. 373–384.

[4] 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. 253–262.

[5] I. De Felipe, V. Hristidis, and N. Rishe, "Keyword search on spatial databases," in Proc. IEEE 24th Int. Conf. Data Eng., 2008, pp. 656–665.

[6] 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.