Advances in Computer Science and Information Technology (ACSIT)
p-ISSN: 2393-9907; e-ISSN: 2393-9915; Volume 5, Issue 2; April-June, 2018, pp. 73-76
© Krishi Sanskriti Publications
http://www.krishisanskriti.org/Publication.html

Nearest Keyword Set Search Survey

Twinkle Pardeshi[1] and Prof. Pradnyakulkarni[2]

[1]Department of Computer Engineering Institute of Technology Pune, Maharashtra, India
[2]Department of Computer Engineering Maharashtra Maharashtra Institute of Technology Pune, Maharashtra, India
E-mail: [1]twinkle.pardeshi@gmail.com, [2]pradnya.kulkarni@mitpune.edu.in

Abstract—*In the field of research, objects can be thought of as images, chemical compounds, documents, or experts in collaborative networks. These objects can be categorized based on their features and can be represented as points for further analysis. Data mining deals with discovering unknown and useful patterns from a huge set of data. The major goal is extraction and understanding of data. Keyword based retrieval can improve the searching and indexing of documents. It helps in developing new tools. Research in the paper considers objects to be tagged with keywords and to be embedded into vector space. A keyword query is to be submitted by the user and based of given query the tightest group is returned. Tightest group consist of points that are closer to each other to provide relevant result set. Hashed based index structure is been introduced for efficient retrieval of data. A new type of query called as Nearest Keyword Set Queries is been introduced.*

1. INTRODUCTION

Keyword based search in multi-dimensional datasets facilitates many novel applications and tools. Objects are considered to be tagged with keywords and are embedded in space vector. Objects are known to be the collections of various features and are represented points in multi-dimensional feature space. Current research on queries goes well beyondpure spatial queries such as nearest neighbor queries, range queries, and spatial joins. Queries on spatial objects arerepresented by sets of keywords that are beginning to receive attention from the spatial database researchcommunity and the industry. Queries on spatial objects are associated with textual information that are represented by aset of keywords, have received significant attention. The main focus is on the Nearest Keyword Set (NKS) queries text rich multi-dimensional datasets. In NKS query, user provides set of keywords and query retrieves set of points that contains those keywords. These queries are used in the applications such as in photo sharing social network, where photos are tagged with people names and even location. NKS query is used to find similar photos based on tagged data.

- Keyword Mutex

For any given node set, which consist of keywords. There exists two keywords whose distance is greater than the given diameter than it is called as keyword mutex. There are two properties of the keyword mutex those are:

1. If any node set is keyword mutex then it can be pruned.
2. Keyword mutex is a monotone property.

- Distance Mutex

For any given node set, if there exists two nodes whose distance is greater than the given diameter then it is called as distance mutex. Same as the keyword mutex, distance mutex also have two properties:

1. If a node set *N* is distance mutex, then it can be pruned.
2. Distance mutex is a monotone property.

2. TYPES OF QUERIES

1. Spatial-Keyword (SK) Queries

GIS database contains location information which is useful in many application such disaster management, crime analysis, etc. Such information also plays a very important role in natural disaster. There are two main components in location based database those are location and textual information. Location information includes the shape, size of the entity and the textual information involves the query keywords describing the entity. The queries that requires such type of information are called as Spatial-Keyword (SK) Queries. GIS database are accessed using such type of query.

2. m-closest keywords (mCK) query

The mCK query tends to find the closest set of keywords to the provided query. The query finds the spatially closest tuples among all the tuples which match *m* user-provided keywords. Given a set of keywords from a document, *m*CK query can be useful in geotagging the document by comparing the keywords to other geotagged documents. To make mCK query efficient, for indexing BR-Tree is been used which provide faster access to the database.

3. Top-k Spatial Preference Queries

A spatial query ranks objects based on the features in their spatial neighborhood. For example, consider a agency office that holds a database with flats for lease. A customer may want to rank the flats with respect to the appropriateness of their location, defined after aggregating the qualities of other features within a distance range from them. Such type of retrieval is called as top-k spatial preference query. Such query returns the top k objects or entities in the database with the highest score. The specified score is defined by the features of the object.

4. Optimal-Location Query

Given a set of sites and weight of each objects with spatial location of that object, the Optimal-Location Query will return the location with the maximum influence. Influence here is described as the total weight of an object that are close to the given location to any given site. To get the optimal location, R* -Tree is used for optimal retrieval of optimum objects.

3. SEARCH ENGINES

Search engines are the software systems that enables a user to search information of any field. The results of search are displayed on the screen are mostly known as search engine results pages. There are various process that helps search engine to work efficiently. Those processes are web crawling, indexing, and searching. Two types of search engines are enlist below.

1. Keyword Search Engine

Features:

a) Traditional search engine.
b) Provides with the results related to input query.
c) Results depends upon keyword search and ranking system.
d) No stop words are used.

2. Semantic Web Search Engine

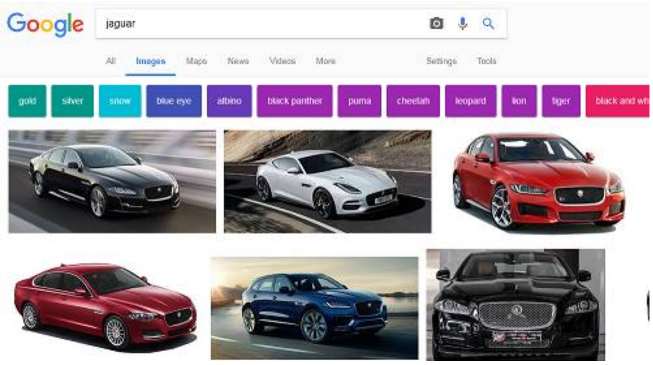
Features:

a) Semantic based approach
b) Accurate and relevant information
c) Results are not depended on keywords and ranking system.
d) Focuses on stop words.

4. SURVEY RESULTS OF DIFFERENT SEARCH ENGINES

All of the following search engines shows the results of one keyword search that is “Jaguar”. Results are as follows:

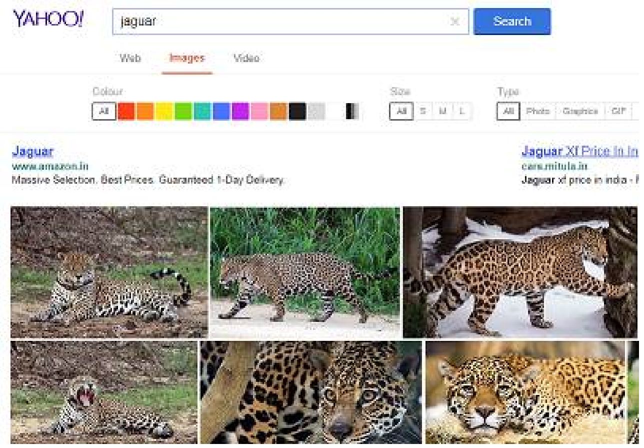
1. Google Search Engine



jaguar

Figure 1: Google Search Engine

2. Yahoo Search Engine



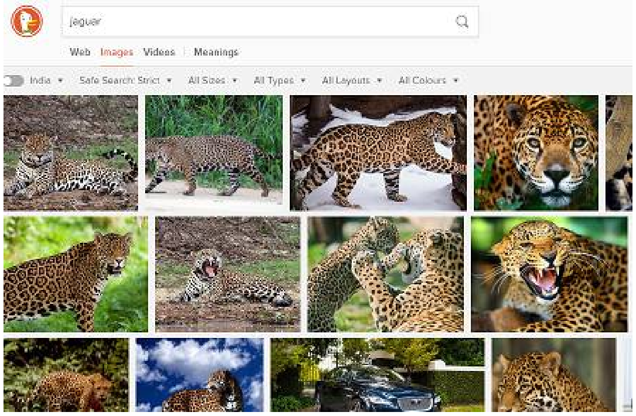
jaguar

Jaguar
www.amazon.in
Massive Selection. Best Prices. Guaranteed 1-Day Delivery.

Jaguar Xf Price In In
cars.mitula.in
Jaguar xf price in india - I

Figure 2: Yahoo Search Engine

3. DuckDuckGo Search Engine



jaguar

Figure 3. DuckDuckGo Search Engine

4. Bing Search Engine

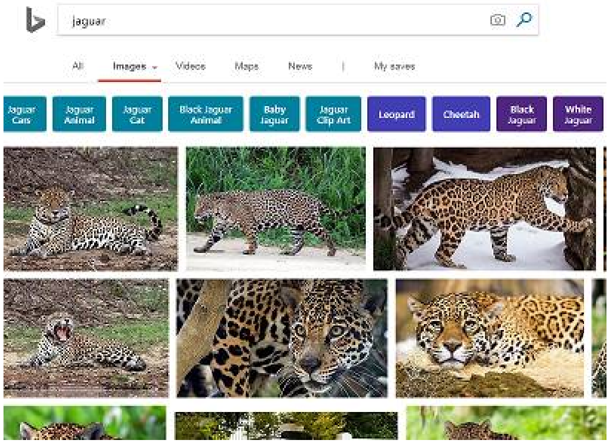


Figure 4: Bing Search Engine

5. MULTIWAY SPATIAL JOIN

A multiway spatial join is defined as set of relations. Each relation consist of attributes and set of spatial predicates. The problem is to find n- tuples. Most of the times the predicate are intersect predicate but at the same time other predicates can be involved such as near, nearest, meet any others. When there are two tuples so in many applications pair wise join is utilized. The queries can be in the form of as "Find all the villages which are crossed by river and also the industrial area". Solution to this can be obtain by computation of one pair join using R- Tree and appropriate spatial join algorithm. The structure of R-Tree is height balanced tree. It also introduces the minimum bounding rectangle. The most common queries are window queries in which the predicate is overlap. Processing of such query is done by R-Tree itself. The spatial join in such a manner that it selects from the two object sets that satisfy the given predicate. The multiway spatial join can be represented as graphs. /in the same manner, the graph can be represented as a constraint network corresponding to binary constrained problem.

6. RELATED RESEARCH

1. WordNet

Cognitive Science Laboratory of Princeton University has developed WordNet which is the large collection of English words. It helps to group the words into set of words and to analyze the relationship between those words. Building a good combination of dictionary and thesaurus that would be useful for n no. of research is the main aim. The database contains 150,000 words and total of 207,000 word-sense pairs. It has a major disadvantage to comparing concepts based on World Wide Web. Due to this limitation, it is rarely used now-a-days.

2. Google Distance

Google Distance is used to calculate the relation between two given concepts with the help of the co-relation provided by Google search engine. Google is a wide collections of web pages and are indexed by Google very efficiently to improve the search quality. A normalized Google distance (NGD) is defined to provide the semantic relations with the help of calculating the correlation from Google search results.

$$\mathrm{NGD(x,y)} = \frac{\max(log\ f(x)\ ,log\ g(y)) - \log f\ (x,y)}{\log N - \min(log\ f(x)\ ,log\ g(y))}$$

In the above equation, NGD(x,y) denotes the Normalized Google distance between concepts x and y. f(x),f(y), and f(x,y) denotes number of pages containing x, y, both x and y. N is the total number of web pages indexed by Google. There is one limitation of this method that by using this it becomes difficult to detect the concept relationship meronymy and concurrence in day to day life.

3. Flickr Distance

Flickr dataset is used in no. of application development and analysis. It is a pool of image database with large number of tags included with each image. It has its own distance measure other than Google's distance measure. It becomes very important to understand and detect the co-relation between two given words. And becomes more important when words are associated with an image.Visual language model (VLM) is one the method adopted for visual statistical analysis. Those images that will contain the given concept will probably share the same visual features that contributes to a model. To deal with distance measurement the square root of Jensen Shannon (JS) divergence between the VLMs is been calculated. This method is used because it deals with symmetric and satisfies triangle inequality property.

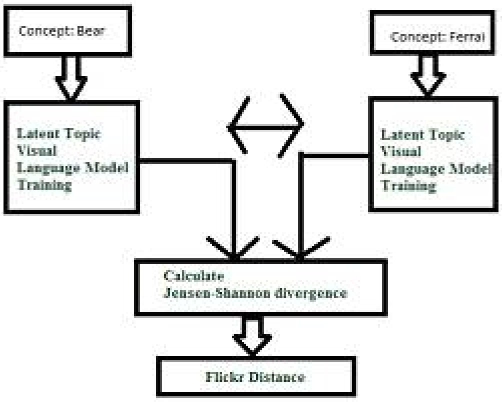

Concept: Bear
Concept: Ferrai
Latent Topic Visual Language Model Training
Latent Topic Visual Language Model Training
Calculate Jensen-Shannon divergence
Flickr Distance

Figure 5: Flickr Distance Example

7. DATASETS

Various datasets have been used to demonstrate the performance of various algorithm to search for a keyword. Various parameters are been tested using the following datasets. The parameters can be efficiency, effectiveness, index efficiency, performance, reliability and many more factors.

- Flickr

Flickr contains millions of images which are updated daily. Huge number of images are added daily. Images and their metadata information are to be crawled from flickr website using API service provided by flickr

- TIGER

TIGER (Topologically Integrated to the real dataset. It is been used by many authors in thefield of data mining domain.

- GOV data

The dataset is the collection of resources of webfrom USA government sites were the top domain is ".gov". The data iscrawled in the year 2002.

8. CONCLUSION

Keyword based search in multi-Dimensional dataset is involved in many applications in today's world. Therefore, solutions to the problem of Top-k nearest keyword set search in multi-Dimensional dataset are studied. Efficient search algorithms are studied that work with indexes for fast processing. A novel model called as Projection and multi-scale Hashing (ProMiSH) that uses random projection and hash based index structures and achieves high scalability and speedup is introduced.

REFERENCES

[1] Vishwakarma Singh, Bo Zong, and Ambuj K. Singh, "Nearest Keyword Set Search in Multi-Dimensional Datasets" ,in IEEE Transactions on knowledge and Data Engineering, vol. 28, no. 3, march 2016.

[2] 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.

[3] 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.

[4] 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.

[5] 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.

[6] 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.

[7] 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. 688-699.

[8] 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.

[9] 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. 450-466.

[10] A. Guttman, "R-trees: A dynamic index structure for spatial searching," in Proc. ACM SIGMOD Int. Conf. Manage. Data, 1984, pp. 47-57.

[11] 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.

[12] G. Cong, C. S. Jensen, and D. Wu, "Efficient retrieval of the top-k most relevant spatial web objects," Proc. VLDB Endowment, vol. 2, pp. 337-348, 2009.

[13] M. L. Yiu, X. Dai, N. Mamoulis, and M. Vaitis, "Top-k spatial preference queries," in Proc. IEEE 23rd Int. Conf. Data Eng., 2007, pp. 1076-1085.

[14] N. Beckmann, H.-P. Kriegel, R. Schneider, and B. Seeger, "The R*- tree: An efficient and robust access method for points and rectangles," in Proc. ACM SIGMOD Int. Conf. Manage. Data, 1990, pp. 322-331.

[15] R. Agrawal and R. Srikant, "Fast algorithms for mining association rules in large databases," in Proc. 20th Int. Conf. Very Large Databases, 1994, pp. 487-499.

[16] J. M. Kleinberg, "Two algorithms for nearest-neighbor search in high dimensions," in Proc. 29th Annu. ACM Symp. Theory Comput., 1997, pp. 599-608.

[17] A. Gionis, P. Indyk, and R. Motwani, "Similarity search in high dimensions via hashing," in Proc. 25th Int. Conf. Very Large Databases, 1999, pp. 518-529.

[18] D. Papadias, N. Mamoulis, and Y. Theodoridis, "Processing and optimization of multiway spatial joins using r-trees," in Proc. 18th ACM SIGMOD-SIGACT-SIGART Symp. Principles Database Syst., 1999, pp. 44-55.