EFFICIENT KEYWORD SET SEARCH WITH PMHR

Y. KATYAYANI[1], K. GOVARDHAN REDDY[2]

[1] PG Scholar, Dept of CSE, Gprec, Kurnool (District), Andhra Pradesh, INDIA.

[2] Associate Professor, Dept of CSE, Gprec, Kurnool (District), Andhra Pradesh, INDIA

ABSTRACT

Multi-dimensional datasets are datasets in which each data point consists of set of keywords in abstract space that concede to develop a new mechanisms for querying and explore these multi-dimensional dataset. In this proposal, we learn nearest keyword set search queries on text-rich, multi-dimensional datasets with ranking functions. A peculiar method called PMHR (Projection and Multi-scale Hashing with Ranking) that uses random projection, hash-based index structure, and ranking. Ranking is done based on repentance of keywords. Ranking is done by using tf-idf technique and provide efficient results. Keyword-based search is done on text-rich multi-dimensional datasets which facilitates many peculiar applications and mechanisms. We considered objects that are appended with keywords and are impacted in a vector space. From these datasets, we are going to study queries that are from the tightest groups of points by gratifying a given set of keywords.

Keywords: Querying, Multi-dimensional data, Indexing, Hashing, Ranking

1. Introduction

Let us consider multi-dimensional datasets from which each data point has a set of keywords. The existence of keywords in vector space allows us to develop new tools for querying and analyse these multi-dimensional datasets [1]. The NKS query consist a set of user-provided keywords, and the results of the query may includes keyword sets of data points where each point consists all the query keywords and finally the output will be the top-k tightest clusters in the multi-dimensional space [1]. Now we are adding ranking to this to get efficient results. In various GIS dig into fields, multi-dimensional spatial data are recurrently generated. Multi-dimensional spatial data is obtained when data attainment devices are located at different locations and these retrieved data is measured based on certain set of attributes which helps to provide spatial data. For example, scientists are interested in studying and analysing the weather conditions in a particular region. Thousands of data points, with the measurements of the different aspects, associated with their spatial signature and raise many challenging questions: What are the ways to analyze the data and to interpret the information? How to manipulate the data to support efficient data querying? [2].
We propose PMHR (Projection and Multi-scale Hashing with Ranking) to get efficient keyword queries. PMHR retrieves the top-k results. PMHR contains the following three desired characteristics for searching the keywords:

1) High quality of results
2) High efficiency and
3) Good scalability

PMHR uses hash tables and inverted indices to provide localized search for getting efficient outputs. PMHR creates hash tables at multiple bin-widths, called index levels. Searching in a hash table yields subsets of points which contains query results at single round, and PMHR explores each subset using a fast pruning-based algorithm. We propose ranking in this paper so that we can get the outputs efficiently. Ranking is done by using tf-idf technique.tf-idf abbreviated as Term Frequency and Inverse Document Frequency and is frequently used for getting the information and text mining. Ranking is a statistical measurement used to evaluate that how much the word is important for a document in a group of data. The importance increases directly proportional to the number of epoch a word appears in the total document but it is offset by the frequency of the word in the data. Variations of the tf-idf ranking technique are generally carried out by search engines as a central mechanism in scoring and ranking a document's importance which is given by a user queries.

The simple ranking technique is tf-idf for every query term; many more complicated ranking functions are alternative to this simple model. Tf-idf can be fruitfully used for filtering the stop-words in heterogeneous subject areas including text summarization and classification.

tf-idf technique:
The term tf-idf technique is a combination of two terms: The first term tf abbreviated as the normalized Term Frequency (TF) – How much times a word repeated in a document, divided by the total number of words present in that document; and the second term is the Inverse Document Frequency (IDF), expressed as the logarithm of the number of documents in the whole divided by the total number of documents from which the specific term appears.

2. Related work

Various queries are used to get the text rich outputs from the large amount of data .reasonably we have got through the NKS queries, Location related queries, spatial queries in the internet through the GIS systems. W. Li and C. X. Chen express spatial data - as geometric information. It identifies the geographical location such as boundaries of earth, natural or constructed features, oceans and more. In Multi-dimensional [3], spatial data is obtained by positioning a number of data retrieving devices at different regions to measure a certain set of attributes to study objects.

To represent the spatial locations we are using 3D format. Here we establish a different data model for handling multi-dimensional spatial data with three spatial dimensions. In this X. Cao uses clustering algorithms to group the dataset called "point clouds" each cloud is considered as a 3D spatial convex object and translated into a set of tetrahedrons, from these clustering algorithms the data is provided to the next two phases.
Geo-positioning: Geographic location is identified by using technology
Geo-tagging: Geo-tagging is the process of counting up geographical related metadata to various media such as geographic photograph or videos.

In this G. Cong uses the GPS or systems that exploit the wireless communication infrastructure; by using infrastructure accurate user location is highly available [4]. Similarly, increasing number of objects is available in the network that has an associated geographical location and textual description. We address the need for collective answers to spatial keyword queries, author assume a database of spatial web objects and then consider the problem for retrieving a group of spatial objects that collectively meet the user's needs by giving the location and a set of attributes as keywords. Some **keywords are:**

1) The group of objects textual description must co-inside with the query keywords,
2) The objects must be close to the query data point,
3) The objects in the group are near to each other.

This work addresses a unique spatial keyword query called the m-closest keywords (mCK) query by Chee, A. Mondal, considers a database with spatial objects, each tuple is related with some textual information and it is characterised in the form of keywords. The mCK query aims to group the spatially closest tuples that match according to the user-specified keywords [5]. Give a set of keywords which are present in the document, mCK query is very useful in geo-tagging the document with the comparison of the keywords with other geo-tagged documents which are present in a database. In this D. Zhang searching is performed on document, given a record that contains these three keywords, the user is interested to find a spatial location that the record is likely to be relevant. This can be done by issuing a mCK query on the three keywords. The measure of compactness for a set of m tuples and is defined as the maximal distance between any of the two tuples. Finding m closest keywords and searching in a subset of nodes algorithms are used to retrieve the document.

Efficient repossession of the top-k spatial objects put forward a new indexing framework for location conscious top-k text retrieval. The framework clout the inverted file for the retrieval of text and the R-tree for spatial contiguity querying. Many indexing methods are researched within the frame. The frame encloses algorithms for computing the top-k query by utilizing the proposed indexes, thus taking into consideration both text relevancy and location proximity to prune the search space. G. Cong works on hybrid in frame using IR tree to get the most relevant objects [6]

Detecting the objects in the geographical scope is very interesting and should be refer with good knowledge. Here D. Zhang focuses on the fundamental application for locating geographical resources and proposes an efficient tag-centric query processing strategy [7] [8]. The aim is to find a set of nearest co-located objects which together match the query tags. Firstly two specifications are required. First one aims to get the details of the location and processing the data to information. The second one is to hyperlink the data to the web page sites. B. C. Ooi wants to adopt tagging as a way to build a uniform data model for the mapped resources within the context of our marcopolo system. In Web2.0 tagging is a popular, because it provides various resources, including news, blogs, speeches, photos and videos. Users can also add extra textual terms such as semantic description or summarization for the objects, with human intelligence involved; the tags are well explained so that we can save much cost for handling term.

3. Problem statement

3.1 Existing system:

Existing system contains ProMiSH (Projection and Multi-Scale Hashing) is done on keyword queries that enables fast processing. We have an exact ProMiSH (referred to as ProMiSH-E) that continuously gives the minimal top-k results, and an approximate ProMiSH (referred to as ProMiSH-A) and is more efficient in terms of time and space, and is able to obtain results in practice.

3.2 Proposed system:

Amount of data generated everyday is expanding in dramatic manner. To get the data efficiently we use PMHR. It is abbreviated as projection and multi-scale hashing with ranking. Here we consider the inverted index and hashing techniques for getting the results. To improve the results efficiently we use ranking in the proposed system. Ranking gives the results in such a way that the top one will be first. Ranking provides the result based on keyword ratio. Ranking is done by tf-idf technique. Tf-idf defines term frequency and inverse document frequency.

3.2.1 Modules:

1. Multi-dimensional data
2. nearest Keyword
3. Indexing
4. Hashing.
5. Ranking

Multi-dimensional Data

Multi-dimensional datasets is data analysing procedure that covers data into two categories. They are data dimensions and measurements. Keyword-based search in text-rich multi-dimensional datasets provides many peculiar applications and tools. Multi-dimensional datasets has data points and each point has a set of keywords. The presence of keywords in feature space allows for the development of new tools for querying and analysing these multi-dimensional datasets. Algorithms may take huge amount of time to terminate a multi-dimensional dataset which consists of millions of points. Therefore, there is a necessary with an efficient algorithm that scales with dataset dimensions, and yields practical query efficiency on huge datasets. In multi-dimensional spaces, it is very difficult for users to provide meaningful coordinates, and our main challenge is to provide keywords as input.

Nearest Keyword

Let us consider multi-dimensional datasets where each and every data point has a set of keywords. The existence of keywords in vector area allows for the development of new mechanism for querying and analysing these multi-dimensional datasets. An NKS query is a query in which user will provide set of keywords, and the outcome of that query may includes k sets of data points from which it contains all the query keywords and forms one of the top-k thick clusters in the multi-dimensional space. Location-specific keyword queries on the internet and in the GIS systems were earlier answered using R-Tree and inverted index. In nearest keyword to rank objects from spatial datasets, it is done based on a combination of their distances to the query locations and the relevance of their text descriptions based on the query keywords is done by IR-tree.

Indexing

Indexing time is the metrics to evaluate the index size for PMHR. Indexing indicates the amount of time used to build PMHR variants. The memory usage and indexing time of PMHR is good. Memory handling increases slowly in PMHR when the number of scope in data points increases. PMHR is more efficient and it takes 80% less memory and 90% less time, and is able to obtain near-optimal results.

Hashing

The hashing technique is influenced by Locality Sensitive Hashing (LSH), which is a state-of-the-art method for nearest neighbor search in high-dimensional spaces. Hashing is done based on the records and methods used for hashing. Based on the index value obtained, we use to place the record at the particular bucket. Unlike LSH-based methods that allows only rough search with probabilistic guarantee, the index structure in PMHR supports accurate search. Random projection by hashing has become the state-of-the-art method for nearest neighbor search in the high-dimensional datasets.

Ranking

Ranking is done based on the keywords. Number of times the keyword appears in a document and number of times the keyword appeared throughout the whole dataset [9]. In this ranking is done by using Tf-Idf technique.

4. Implementation

Keyword set search is done by using PMHR, here we will get the results. To improve the results we add ranking technique to the PMHR so we can still get more efficient results based on keywords. Ranking is done by tf-idf technique.

Tf (Term Frequency) states that how much times a term repeated in a document [11] [10]. Generally every document has its own length, so it is probable that the visible of a term in long documents than shorter ones is more. Finally, the term frequency is generally divided by the total document (the total number of terms in the document)

tf(t) = (How many times a term t appears in a document) / (Total number of terms in the document)

In the case of the **term frequency** tf(*t*,*d*), the simple choice is to use the *raw count* of a term in a document, i.e. the number of times that term *t* occurs in document *d*. If we denote the raw count by $f_{t,d}$, then the simplest tf schemes are tf (*t*,*d*) = $f_{t,d}$. Different chances are

- Boolean "frequencies": tf(t,*d*) = 1 if *t* occurs in *d* and 0 otherwise;
- Term frequency regulated for document length : $f_{t,d}$ / (number of terms in d)
- Logarithmically scaled frequency: tf(*t*,*d*) = 1 + log $f_{t,d}$, or zero if $f_{t,d}$ is zero;
- Increased frequency, to avoid a bias towards longer documents, e.g. raw frequency divided by the raw frequency of the most occurring term in the document:

$$tf(t,d) = 0.5 + 0.5 * \ f(t,d)/\max\{ft^\wedge{}',d: t^\wedge{}' \in d\} \qquad (1)$$

IDF: Inverse Document Frequency, which states how important a term is. While solving TF, all the terms in a document considered as important. However it is known that certain terms, such as "is", "of", and "that", may appear a lot of times but have little importance. Thus we need to check down the frequent terms while scale up the rare ones, done by the following formula:

IDF (t) = log_e(Total number of documents / Number of documents with terms w in it).

$$idf(t,D) = \ log \frac{N}{|\{d \in D: t \in d\}|} \qquad (2)$$

See below for a simple example.

Example:

Let us consider a record which contains 100 words from that the word *apple* appears 3 times. The term frequency (i.e., tf) for *apple* is then (3 / 100) = 0.03. Now, we assume that the word apple appears one thousand times from 10 million records. Then, the inverse document frequency (i.e., idf) is calculated as log (10,000,000 / 1,000) = 4. Thus, the Tf-idf weight is the product of these quantities: 0.03 * 4 = 0.12.

5. Results

Many of the applications are using the ranking technique now we use ranking technique that is term frequency and inverse document frequency in projection and multi scale hashing, so that we can get the results efficiently. In the keyword rank graph the x-axis represents keywords and the y- axis represents the rank range. Here the bar graph states how frequently the keyword appears.

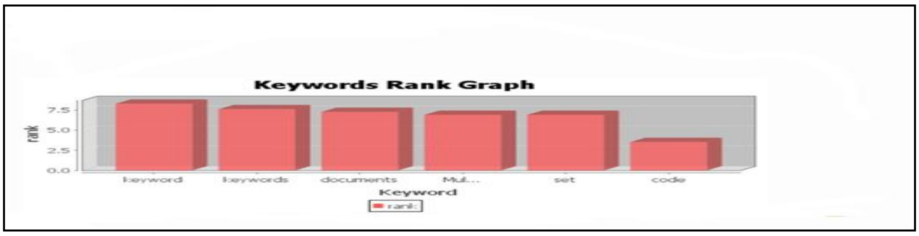

Figure 1: Keyword Rank Graph

Now we can get the document which is having the highest priority of keyword ranking and that document will be appeared first in the results, in such a way the remaining documents which contains all the keywords in the descending order will be displayed.

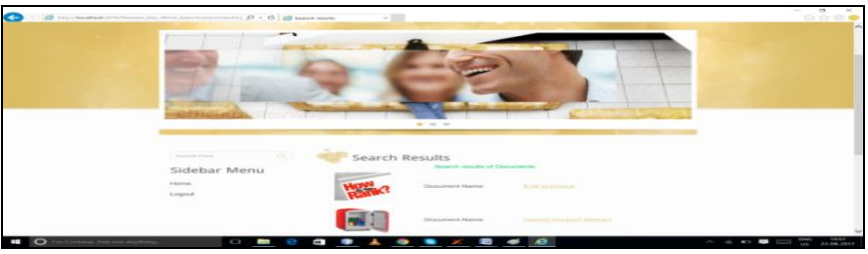

Figure 2: Documents retrieval

Here we can see the documents, and they came in the ranking order.

The sequence diagram for the entire system works

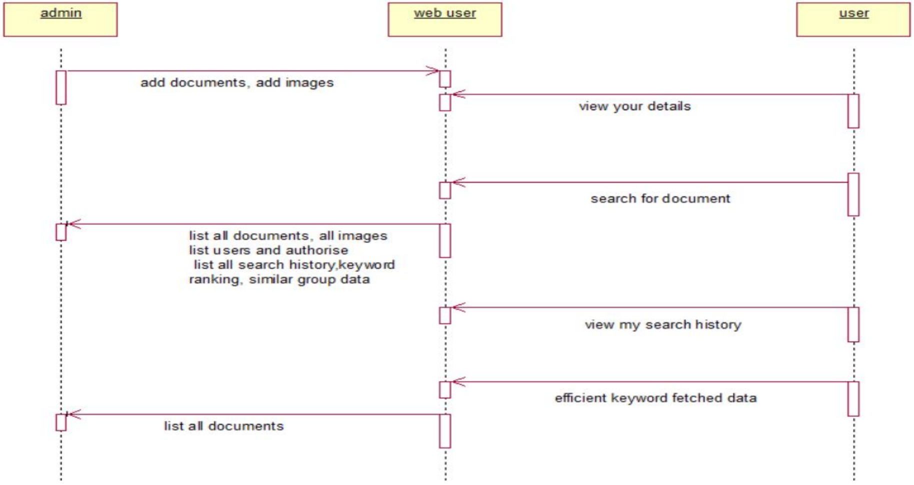
admin
web user
user
add documents, add images
view your details
search for document
list all documents, all images
list users and authorise
list all search history,keyword
ranking, similar group data
view my search history
efficient keyword fetched data
list all documents

Figure 3: Sequence diagram

6. Conclusion

In this paper we propose tf-idf ranking technique to the PMHR algorithm to get efficient results. In this paper efficient keyword set search is done to get efficient results. By using tf-idf technique the keywords that appear rapidly will be scored and are checked to get the results. Searching the documents with the help of keywords is done to get a resultant document. Document retrieval is done based on the keyword ranking. To this PMHR we can include disk extension in the future expansion. In this directory file is expanded for storing the keywords

References

[1]. Vishwakarma Singh, Bo Zong and Ambuj K. Singh, “Nearest Keyword Set Search in Multi-Dimensional Datasets by Group nearest query search in multi-dimensional spatial data”.

[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]. 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]. 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.

[5]. 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.

[6]. 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

[7]. J. Ding, L. Gravano, and N. Shivakumar. Computing geographical scopes ources. In VLDB ’00: Proceedings of the 26th International Conference on Very Large Data Bases, pages 545–556, San Francisco, CA, USA, 2000. Morgan Kaufmann Publishers Inc

[8]. Rajendra Kumar, Roul, Omanwar Rohit Devanand ,S. K. Sahay BITS, Pilani.,” Web Document Clustering and Ranking using Tf-Idf based Apriori Approach

[9]. Doreswamy and Hemanth K.S. A Novel Design Specification(DSD) based K-mean clustering performance Evaluation on Engineering material’s database, IJCA, Vol 55, No.15, Oct-2012.

[10].https://janav.wordpress.com/2013/10/27/tf-idf-and-cosine-similarity/

Author

Y.Katyayani received the bachelor's Degree in Computer Science and Engineering from Jawaharlal Nehru Technological University-Ananthapur and currently pursuing Master's in Computer Science and Engineering from G. Pulla Reddy Engineering College, Kurnool, A.P. Her research interests include Data Mining.

Sri K. Govardhan Reddy received bachelor's Degree, Master's Degree, Ph.D from Jawaharlal Nehru Technological University-Hyderabad in Computer Science & Engineering. He is currently an Associate Professor in Dept. of Computer Science & Engineering in G.Pulla Reddy Engineering College, Kurnool. He has 16 years of teaching experience. His research interests include Wireless Communication