INTERNATIONAL JOURNAL OF RESEARCH IN COMPUTER APPLICATIONS AND ROBOTICS
ISSN 2320-7345

MAINTAIN TOP-K RESULTS USING SIMILARITY CLUSTERING IN RELATIONAL DATABASE

Syamily K.R[1], Belfin R.V[2]

[1]PG student, Department of Computer Science and Engineering, Karunya University, Tamilnadu,
syamilykraju@gmail.com
[2] Assistant Professor, Department of Computer Science and Engineering, Karunya University,
belfin@karunya.edu

Abstract

Now a day's keyword extraction in Relational database containing large amount of dataset can be easily used by the user without knowing any query language. Scalable continual top-k keyword search using Lattice pipeline algorithm and Maintain algorithm are some techniques which are used. Lattice pipeline algorithm is used to find the greatest ten keywords from the whole database. But in real life content in database are updated frequently in order to solve this problem the system should handle the database updation to maintain the results. That is why the technique called maintains top-k result was introduced. Maintain the results avoid finding redundant results but in same time it increases the overload of memory to store the results after each updation. And also time cost is high because of keyword search from the maintained results. So the new enhanced method of maintaining the results is introduced. Similarity clustering is used to cluster the maintained results. Multi-viewpoint based clustering is one similarity clustering which use multiple point of similarity instead of single point. That increases the informative assessment of similarity. Effective clustering help the keyword searching process by reducing the time cost for searching the keyword in the maintained results. .

Keywords*:* Top-k results, Multi-viewpoint based clustering, Cosine similarity, similarity vector.

1. Introduction

Continual top-k keyword search processing in relational database is mainly used for data extraction in updating database. For example database for publication, which store details about the journal, author, title, year of publication etc. All details in a database of real time application are updated with time. So the continual search can efficiently report the top-k results of every keyword query while the database is being updated continually [5, 9]. In this system initially a search query is evaluated in a pipelined manner. After database updation the results can be maintain and new tuples can be inserted.

As an enhancement the system I proposed has an improved technique to maintain the top-k results [11]. The updated top k results limit on storing main memory and does not maintain properly in that scenario. So I propose clustering methodologies to maintain results. In this paper, I introduce a novel multiview point-based similarity measure [1, 3]. The major difference between a traditional similarity measure and the new proposed one is that the common clustering ids done using only a single viewpoint (origin), while the multi-viewpoint based clustering utilizes many different viewpoints (objects), which are in different clusters. Using multiple viewpoints, more informative assessment of similarity could be achieved. So instead of searching for all the maintained result directly go to the corresponding cluster and easily extract the data. The multi-viewpoint based similarity clustering [2, 4] is used in our system to cluster the maintained results and stored in an efficient way. And also it can reduce the keyword extraction time as an effect of clustering. Key contributions are as follows: (i) clustering and similarity vector calculation; (ii) cosine similarity calculation; (iii) dissimilarity object and multi-view point; (iv) multi-viewpoint based clustering and cluster movement.

The paper is organized as follows. In section 2, some basic concepts [4] are introduced. Section 3 presents the details of the proposed method. Section 4 gives performance comparison. Finally, in section 5 conclude the paper.

2. Preliminaries

2.1 Initial top-k results

Top-k keyword searching is a famous method used in data mining area. From the large database the relevant keyword can be extracted based on some ranking mechanism [5,8]. And the priority is determined using the term frequency and inverse document length and is called as TF-IDF weighting schema [7]. And for finding the particular number of high scored keyword an algorithm called Lattice Pipeline [6] is used. “Figure:1” shows the flow of working. It is worked in a pipelined manner. After issuing a query score value is calculated and gets the top-k results as output of LP algorithm.

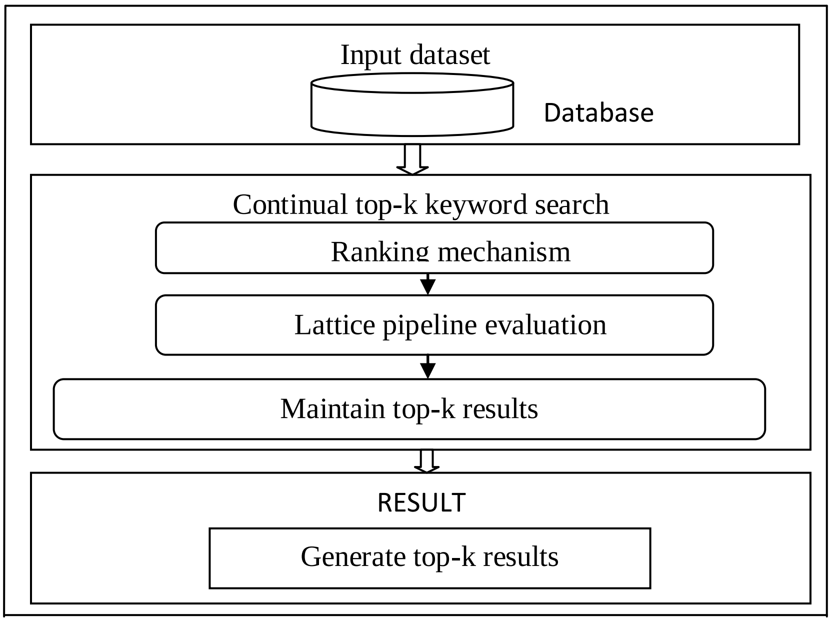

Figure 1: Flow chart for the continual keyword search

2.2 Maintain the top-k results

Database updation can be handled using the maintain algorithm [5]. If the current results are same as previous one this can be easily taken from the database. So it can avoid the re evaluation. In the same way if new results are appeared it can be stored in database. And also doesn’t remove the old one which is not the current top-k results because it may become the future results. So over all database updation can be solving by maintain the results separately.

3. Multi-viewpoint similarity based clustering

The new proposed system can be designed as “figure 2”. The problem with already exciting system is that memory over head in maintaining the results in database. This can be reduced and new clustering technique increase the efficiency of searching the resultant keyword from the whole dataset.

Maintaining the top-k results and used that result for future data extraction is the main feature of the new system. Maintain algorithm is used for handling the database updation to maintain the top-k results. Hence a clustering technique for improving the performance of the system is introduced. Through which easily asses the data from the clusters instead of unordered large number of documents.

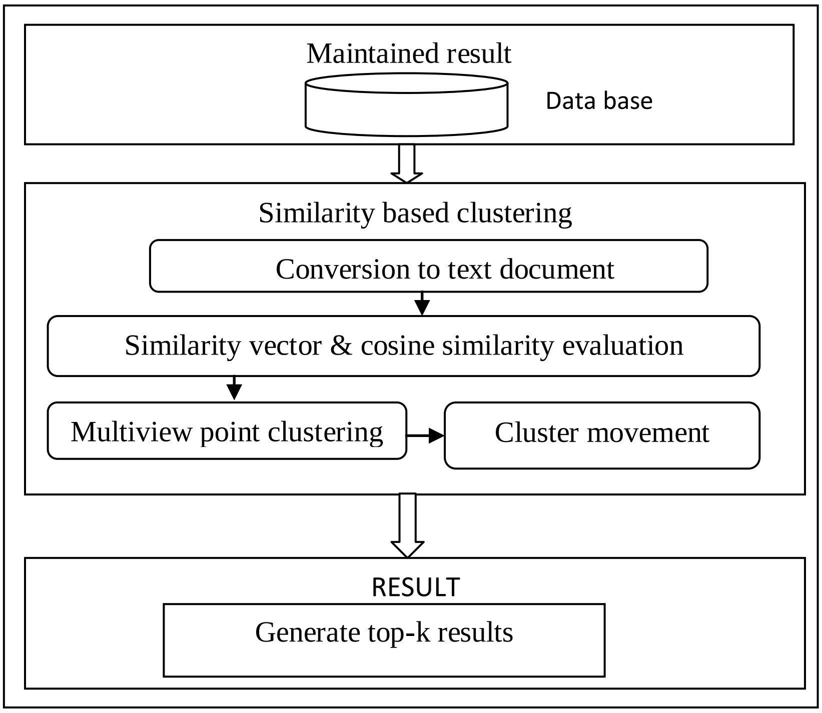
Maintained result
Data base
Similarity based clustering
Conversion to text document
Similarity vector & cosine similarity evaluation
Multiview point clustering
Cluster movement
RESULT
Generate top-k results

Figure 2: Flow chart of proposed system

3.1 Similarity vector and Cosine similarity

Cosine similarity for a term is calculated based on number of times that term appears in the document. Similarity is a useful measure of how similar two documents are likely. It help the system to cluster the terms which having maximum similarity. First consider small amount of text data which are distributed in different clusters. Then calculate the similarity vector [1] for each text document. Initially the loaded text data are distributed in different clusters and the similarity vector for each data in different clusters are calculated separately. The similarity vector is used to calculate the cosine similarity. Similarity vector means the common pattern or any other informational content similarity of the documents. Based on the important of specified data in a given context the similarity measure can change. So using similarity vector data's can be easily extracted and grouped together which having equal important.

After getting the similarity vector, cosine similarity [2] for each pair of document will be calculated. In this stage the three clusters are treated separately and in each cluster consider separate pair of document to find the cosine similarity. If d_i and d_j are two text documents then *cosine*(d_i,d_j) is consider as the cosine similarity between d_1 and d_2. And the (1) shows the formula for calculating cosine similarity.

$$Sim\ (d_i,d_j) = \cos\ (d_i,d_j) = d_i^t\ d_j \qquad (1)$$

From the cosine similarity it can identify the object which has maximum dissimilarity among the documents it belongs. Cosine similarity is calculated based on the term frequency of the document. If the cosine value is less then similarity is high so take the document with high cosine similarity value as maximum dissimilar document.

3.2 Dissimilarity object and multi-viewponit

The main task of the system is multi view point clustering [1,3]. To construct a new point of similarity use more than one reference point and take it as d_h. So get more accurate assessment of how close or distant a pair of points are, if look at them from many different viewpoint. From a third point d_h the directions and distances to d_i and d_j are indicated respectively by the difference vectors $(d_i\text{-}d_h)$ and $(d_j\text{-}d_h)$. Here instead of one reference point using more reference point for viewing the document. Multi view point clustering is used for the vector which has maximum dissimilarity. The Object with maximum dissimilarity is identified in the previous step. That object is viewed from different point of reference. Usually the reference points are taken from other two clusters. Then take one of the two clusters and group the object to that cluster.

3.3 Multi-viewpoint based clustering

Using multiple viewpoints, more informative assessment of similarity could be achieved. So the object with maximum dissimilarity is again clustered using multi view point based cluster. As a result the object is moved from one cluster to the other.

During future evaluation of the top-k keyword after updation the already exciting results are taken from the database which maintains the top-k results. So by applying the proposed method the efficiency of keeping the results in database is increased. And also increase the effectiveness of extracting the keyword from the large amount of data. Since the more informative assessment is used it is easier to find out the keyword from the clustered document. review process.

4. Performance Comparison

Top-k keyword search in relational database is more commonly used method. Because of frequent updation in database maintaining process is very difficult. So in already exciting system illustrate the maintain algorithm for maintain the top-k result. By maintain the results for future evaluation it can be reused further. But there is a problem in maintaining the large amount of data. This problem of high memory usage can be reduced by the proposed clustering method. So the multi view point based clustering reduces data accessing time. And because of the effective clustering the accuracy will also increased.

Result analysis can be shown on the basis of accuracy and time taken to get the results. The comparison done between the scalable continual top-k keyword searches and maintain top-k result processing using similarity clustering in relational database. Figure 3 shows the performance matrix accuracy and its changes in the already exciting system and the new proposed system. Accuracy of multi-viewpoint based similarity measure in top-k results is higher than that of the scalable continual top-k results maintenance.

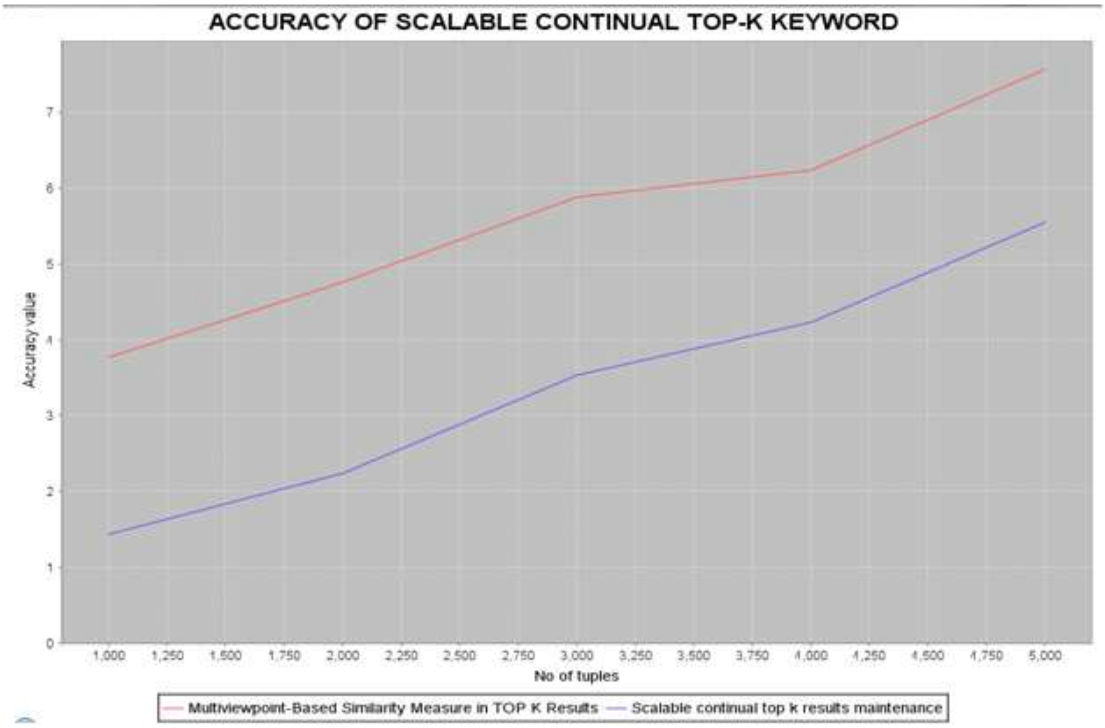

Figure 3: Comparison in terms of accuracy

Maintaining the top-k results in database cause large memory usage. The time for taking the results from already maintained result will change as an effect of the new technique. Instead search for the whole database the system searching needed only on the corresponding clusters. So the time is reduced as shown in the figure 4. The time usage increases with increase in number of tuples for both systems. But newly proposed or enhanced system use less amount of time for finding the results compared to the already exciting system.

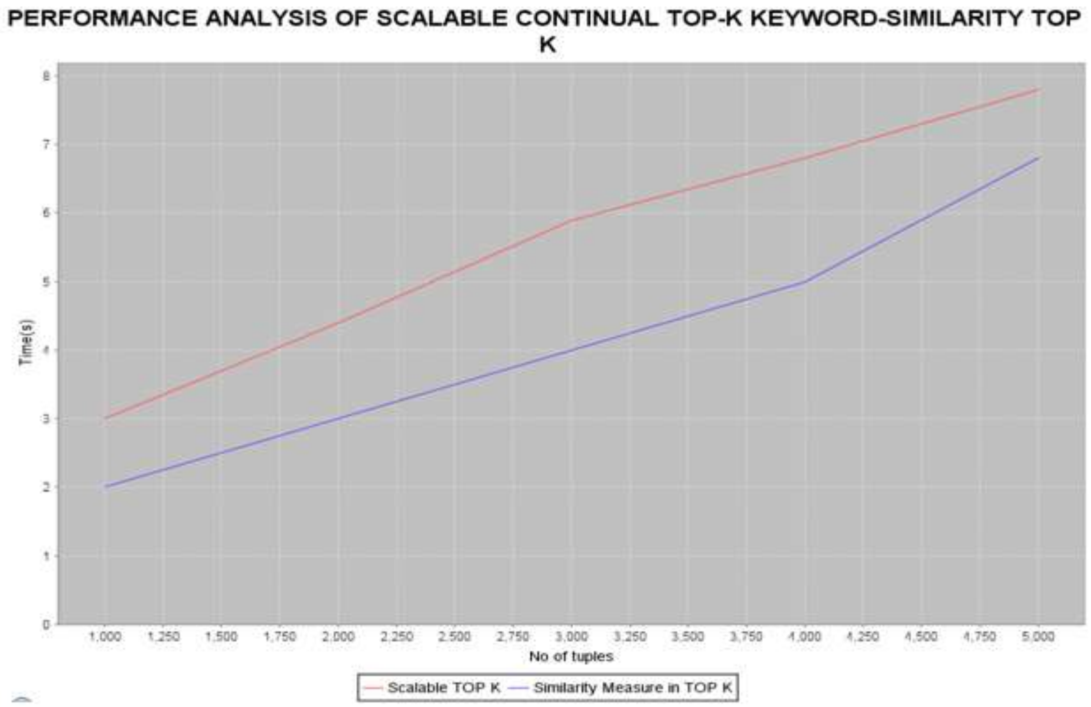

Figure 4: Comparison in terms of time

5. Conclusion

Continual top-k keyword search in relational databases can be used for answering continual keyword queries in databases that are updated frequently. In order to calculate the result before updation and maintaining the results by handling the database updation are done using the LP and maintain algorithm. For further improvement in the field of maintain the results new techniques are introduced. Multi view point based clustering is used to enhance the efficiency of the system.

REFERENCES

[1]. D. Nguyen, L. Chen, C. Chan, 2012, Clustering with multi-viewpoint based similarity measure, *IEEE*. 10.1109.

[2]. R. Saranya, P. Krishnakumari, 2013, Clustering with multi view point-based similarity measure using NMF, *IJSRM*. 2321-3418.

[3]. S. Chandrasekhar, K. Sasidhar, M. Vajralu, 2012, Study and analysis of multi-view point clustering with similarity measure, *IJETAE*. ISSN 2250-2459.

[4]. Yanwei Xu, Jihong Guan, Fengrong, 2012, Scalable continual top-k keyword search in relational database, *Elsevier*.

[5]. V. Hristidis, L. Gravano, Y. Papakonstantinou, 2003, Efficient IR-style Keyword Search Over Relational Databases, *VLDB*. 850861.

[6]. F. Liu, C. Yu,W.Meng, A. Chowdhury, 2006, Effective Keyword Search in Relational Databases, *ACMSIGMOD*. 563574,http://dx.doi.org/10.1145/1142473.1142536.

[7]. Luo, W. Wang, X. Lin, X. Zhou, J. Wang, K. Li, 2011, SPARK2: top-k keyword query in relational databases, *IEEE* Transactions on Knowledge and Data Engineering 23 (12) 1763–1780.

[8]. L. Qin, J.X. Yu, L. Chang, Y. Tao, 2009, Scalable Keyword Search on Large Data Streams, *ICDE*. 1199–1202.

[9]. Y. Luo, 2009, SPARK: Keyword Search System on Relational Databases. , Ph.D. thesis The University of New South Wales.

[10]. L. Qin, J.X. Yu, L. Chang, 2011, Scalable keyword search on large data streams, *VLDB* Journal 20 (1) 35–57.

[11]. S. Agrawal, S. Chaudhuri, G. Das, 2002, DBXplorer: a System for Keyword-based Search Over Relational Databases, ICDE. 5– 16.

A Brief Author Biography

Syamily K.R – She pursuing her M.Tech in Computer Science and Engineering from the Department of Computer Science in Karunya University, Tamilnadu, India. She received her Bachelor's degree from Cochin University of Science and Technology (CUSAT) in Computer Science and Engineering.

Belfin R.V – He is now a Assistant Professor at the School of Computer Science and Technology, Karunya University, Tamilnadu, India. He received his Master's degree from the Anna University in Computer Science and Engineering and Bachelor's degree from Karunya University in Computer Science. He was working with Wipro Technologies as an Oracle Technical Consultant from Jan 31 2011 to Jun 12 2013. And worked as project Trainee in Robert Bosch Engineering and Business Solutions India Ltd (RBEI).