SURVEY ON SCALABLE CONTINUAL TOP-k KEYWORD SEARCH IN RELATIONAL DATABASES

Syamily K R[1], G Naveen Sundar[2]

[1]*PG student, Computer Science and Engineering, Karunya University, Tamilnadu, India,* ***syamilykraju@gmail.com***
[2]*Assistant Professor, Computer Science and Engineering, Karunya University, Tamilnadu, India,*
naveensundar@karunya.edu

Abstract

Keyword search in relational database is a technique that has higher relevance in the present world. Extracting data from a large number of sets of database is very important .Because it reduces the usage of man power and time consumption. Data extraction from a large database using the relevant keyword based on the information needed is a very interactive and user friendly. Without knowing any database schemas or query languages like sql the user can get information. By using keyword in relational database data extraction will be simpler. The user doesn't want to know the query language for search. But the database content is always changing for real time application for example database which store the data of publication data. When new publications arrive it should be added to database so the database content changes according to time. Because the database is updated frequently the result should change. In order to handle the database updation takes the top-k result from the currently updated data for each search. Top-k keyword search means take greatest k results based on the relevance of document. Keyword search in relational database means to find structural information from tuples from the database. Two types of keyword search are schema-based method and graph based approach. Using top-k keyword search instead of executing all query results taking highest k queries. By handling database updation try to find the new results and remove expired one.

Index Terms: *Top-k, keyword search, relational database, information retrieval*

---***---

1. INTRODUCTION

Keyword search in relational database is a efficient information extraction method. This is a simple way of retrieving data from a large set of data with which the user need to give the keyword only without knowing any query language or database schemas.

In real life the database is updated frequently and so the result will be change for each updation. The method handles the database updation to maintain top-k result. Since the result change with time some scoring method is used to calculate the relevance finding the result. Based on the score the greatest k result will consider. But it will change for the next updation so future relevance of the result is taken. As a result of deletion already exciting result may expired and removed from the top results. And for insertion may cause addition of new results into the top-k results.

This paper deals with survey the different process and different types for each process of the keyword search in dynamic environment. Mainly contain three aspects keyword search in relational databases, top-k keyword search, and keyword search in relational data streams.

Keyword search in relational database means to find structural information from tuples from the database. Two types are schema-based method and graph-based approach. Using top-k keyword search instead of executing all query results taking highest k queries. By handling database updation try to find the new results and remove expired one.

2. KEYWORD SEARCH IN RELATIONAL DATA BASES.

Mainly two types are there schema-based and graph-based.

2.1 Schema-Based Keyword Search on Relational Databases.

In this method candidate networks are generated using database schema. Next CN are evaluated. For dynamic database schema based keyword search is used.

DISCOVER is used to create qualified joining networks of tuples. This mainly involve two steps

- Generate candidate networks of relations.
- Builds plans for efficient evaluation of the set of candidate networks.

DISCOVER system include architecture and a CN generation algorithm. Fig -1 shows the different steps included in it.

Candidate network generation the problem of generating redundant joining networks of tuple sets can be avoided by this system. Solutions are mainly analysis of the condition (check joining networks of tuples to be non-minimal) and candidate network generation algorithm. Here check the tuples that are joining to the candidate network are total and minimal. Minimal means condition doesn't contain any tuples with leaf as no keywords. Several theorems are introduced first theorem said about when a candidate network tuple set repeat in CN tuple.

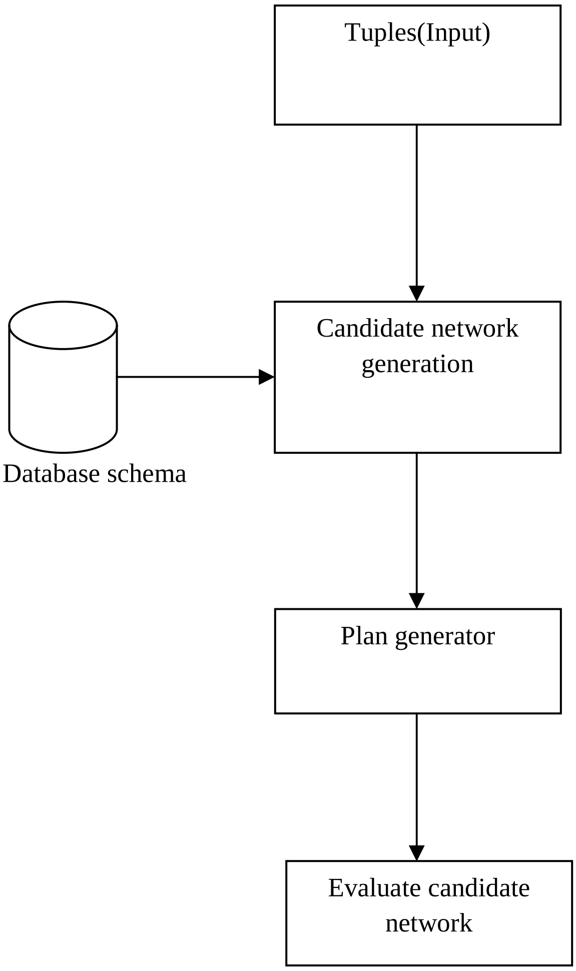

Fig -1: Steps for processing Candidate networks

Candidate network generation algorithm main feature of this algorithm is that it produce increasing size candidate network. That means the candidate network with smallest size and better solution are produced first.

In Evaluation of candidate networks the plan generator module receives input as candidate networks and evaluates them. During evaluation maximum optimization by storing subexpersion and reuse it. Space for the plan generator for tuple set is huge because this can be reduced by deleting the unused subexpersion and generates tuples for small relation. We take the intermediate result depend on the quantity which is inversely proportional to the size of the intermediate result and directly proportional to the frequency of occurrence. Greedy algorithm is used in DISCOVER.

In IR-Style keyword search it incorporate the relevance ranking of tuples trees into our query processing framework. This method increases the document-ranking quality and improves the quality of keyword query results over RDBMS. The ranking is done by combing scores of each attributes and tuples. IR-style also introduce some algorithms which work as per the query contain any AND or OR semantics. The hybrid algorithm decides the strategy for a given query at run time.

Along with IR-style m-keyword search the area over an open-ended relational data streams. In which the size of the connected tree is under user control. The problem is with the costly joins to be processed over time. In order achieve high efficiency reduce the no of intermediate results. For that a method used is new join processing approach. In this approach use selection/semi join instead of using joins directly.

2.2 Graph-Based Keyword Search

The enter database is represented as a graph with nodes as tuples and directed edges as foreign key references between the tuples.

The bidirectional expansion is an algorithm for generating result for keyword search on graphs. Both backward and forward search is used. Dijkstra's is used as backward search algorithm. The aim of this method is reduce the fraction of graph expansion. It create iterates only for not frequently occurring keywords and the forward path for frequent keywords.

Efficient and Adaptive keyword Search method integrate the database and IR techniques. The r-radius steiner graph problem is to avoid the complicity of large radius steiner graphs. R-radius steiner graphs are concise and non-Steiner nodes are excluded. And then an indexes is introduced in order improve the efficiency of extracting r-radius steiner graphs. The ranking function TF.IDF – based ranking is used. Using this function calculates the score value for each r- radius graph and combines the individual scores. This method takes textual relevancy based on term frequency.

Instead of trees which shows partial information about how the tuples are connected a community which contain all keyword within a given distance is used. So the memory usage is small. First find the all community then find top-k communities. The main advantage of this method is which allow the user to interactively reset the value of k during run time. And an index is used to project a small sub graph from the large whole graph.

3. TOP-K KEYWORD SEARCH

This is an efficient keyword search method which contains execution of the top-k queries by avoiding the production of all the query results. IR-style keyword search and SPARK are two methods. DISCOVER.

In IR-style keyword search it use Global Pipeline algorithm. It is an extension of Single Pipelined algorithm. It is used for efficiently answer a top-k keyword query over multiple candidate networks. Input of the GP algorithm is candidate networks and a set of non-free tuples. And the output is stream of joining trees of ranked tuples based on the score of the query. The idea of the algorithm is like process execution, in place of process consider candidate network. Concurrent execution of the CNs is done in round robin fashion and priority that combination testing is costly.

Another technique is Skyline-Sweeping algorithm proposed in SPARK. In advance to IR- style keyword search SPARK reduces the number of tested combinations. Then Global Pipeline algorithm include several unnecessary join checking that will be reduced and so database accessing is minimal. But this SPARK also produces testing of combinations which cannot produce results. The Lattice Pipeline and Maintain algorithm use the basic principles of both previous methods that are calculate the relevance score and processing tuples for pipelined fashion.

4. KEYWORD SEARCH IN RELATIONAL DATASTREAMS

Like relational database relational data streams also use the schema based frame work. Keyword search of relational data streams use two techniques S-KWS and KDynamics are two techniques. Operator mesh or L- Lattice is the concept used in it.

Operator mesh is used to reduce the CPU cost during the evaluation of joins in the CN evaluation step for database updation. Operator tree and Operator mesh are two concepts used in data streams. Operator trees are trees with source operators as leaf nodes that perform selection and joins as interior operators. During CN generation sources are added from left to right. The operator mesh created by integrating all operator tree so the CPU cost and memory overhead is reduced. The memory overhead is usually occurs for storing intermediate results.

Full mesh and Partial mesh are two query mesh is preprocessed in a single step. But in the case of partial mesh the preprocessing step is eliminated and the operator mesh is automatically shrinks and grows during run time.

In scalable continual keyword search on large database which use the L-lattice. In lattice the common sub trees are collapsed for sharing the computational cost of the CNs.

A novel approach for scalable m- keyword query contains two phases filter phase and join phase. In filter phase a candidate network is processed for filter the tuples that cannot be joined. And in the second phase include the joining of the tuples that can be joined in to the network. Next generate the L- lattice for the CNs.

Since the environment is dynamic it needs to respond to continual query. The KDynamic and S-KWS finding all query results while by taking top-k results the changing of the results can be avoids. In order to increase the efficiency

Table -1: Summary of the survey

Aspects	Methodology	Merits
Keyword search in RDB	Schema-based, CN evaluation [1]	Finding all trees
	Schema-based, Ranking [2]	Improve the quality
	Schema-based, Use selection and semi joins[7]	Reduce intermediate results
	Schema-based, Ranking and scoring method is used[9]	Calculate initial top-k results.
	Graph-based, Bidirectional expansion[3]	Reduce graph expansion.
	Graph-based, r-radius steiner graph[4]	Concise
Top-k keyword search	Global pipeline algorithm[2]	work over multiple CN
	Skyline-Sweeping algorithm[5]	Reduce db accessing
	LP and Maintain algorithm[9]	Handle the database updation
Keyword search in relational data streams	Operator mesh[6]	Reduce the CPU-cost and memory overhead
	KDynamics[7]	Share the computational cost.
	L-lattice	Reduce computational cost

scoring mechanism and pipelined evaluation will be added to the KDynamic method.

The table-1 shows the overall concept of this survey on the topic keyword search in dynamic environment.

CONCLUSIONS

In this paper the survey is conducted for comparing the performance of different techniques used for keyword search in dynamic environment. Keyword search in relational database can be performed in different ways. Based on the method used for keyword search on Relational database two types: schema-based and graph-based. In schema based approach sql query is converted to candidate network. DISCOVER and IR-style methods use the CN concept and ranking respectively. The graph-based approach use steiner graph and bidirectional expansion on it. Backward and forward expansion reduces the fraction of expansion and is concise. For scalable database schema based approach is efficient. For improving simplicity of keyword search take only top-k results instead of taking the whole search result. IR- style keyword search and SPARK use GP algorithm and Skyline-Sweeping algorithm respectively. But for updating database first calculate the initial top-k results and then maintain the top-k results for current after updation. Lattice pipeline and Maintain algorithm are the algorithms used. For further improvement scoring method and pipelined evaluation is used. And for keyword search in relational data streams S-KWS and KDynamics are two techniques which is for CPU cost reduction and decrease the database accessing. For improving the efficiency ranking mechanism is used.

ACKNOWLEDGEMENT

I feel it a pleasure to be indebted to my guide, **Mr. G Naveen Sundar, M.E,** Assistant Professor, Department of Computer Science and Engineering for his invaluable support, advice and encouragement and the references for his feedback.

REFERENCES

[1]. V. Hristidis, Y. Papakonstantinou, DISCOVER: Keyword Search in Relational Databases, VLDB, 2002. 670–681.

[2]. V. Hristidis, L. Gravano, Y. Papakonstantinou, Efficient IR-style Keyword Search Over Relational Databases, VLDB, 2003. 850–861.

[3]. V. Kacholia, S. Pandit, S. Chakrabarti, S. Sudarshan, R. Desai, H. Karambelkar, Bidirectional Expansion for Keyword Search on Graph Databases, VLDB, 2005. 505–516.

[4]. G. Li, B.C. Ooi, J. Feng, J. Wang, L. Zhou, EASE: an Effective 3-in-1 Keyword Search Method for Unstructured, Semi-structured and Structured Data, ACM SIGMOD, 2008. 903–914.

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

[6] A. Markowetz, Y. Yang, D. Papadias, Keyword Search on Relational Data Streams, ACM SIGMOD, 2007. 605–616.

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

[8] L. Qin, J.X. Yu, L. Chang, Y. Tao, Querying Communities in Relational Databases, ICDE, 2009. 724–735.

[9]. Yanwei Xu, Jihong Guan, Fengrong Li, Shuigeng Zhou Scalable continual top-k keyword search in relational databases, 2013 Elsevier, Data and Knowledge Engineering 86 (2013) 206-223.

BIOGRAPHIES

Syamily K R 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.

G. Naveen Sundar, received the B.E degree in Computer Science and Engineering from C.S.I Institute of Technology, Thovalai in 2002. He received the M.Tech degree from Karunya University; Coimbatore in 2006.He is currently working toward the PhD degree. He is working as an assistant Professor in Computer Science department of Karunya University. His main research interests are Association Rule Mining, Databases and Web Mining.