Keyword Search Algorithms: 
A Detailed Review

Khushbu Sharma, Sanjay Tiwari & Sayar Singh Shekhawat 
M-Tech ,Rtu Rajasthan , India 
Associate Professor (Arya College Of Engineering And Research Center)

Abstract: A keyword search scheme to relational database and unstructured data turns into a fascinating range of research framework. Keyword based search technique permits clients to recover specific records of intrigue. There exists many research attempts to tackle the issue of data recovery utilizing different keywords ranked search technique. This paper is a preparatory endeavors to overview the keyword searching techniques that utilizing the idea of keyword searching.

Keywords:— Structured and unstructured data, Keyword Search, Information Retrieval Query Workloads, Execution time, Graph-based Search.

1. Introduction

With the progress in the field of the information technology there are number of effective methods and efficient techniques of keyword search which are currently in use. Keyword search technique is largely used for searching unstructured data. With time it has resulted in development of number of techniques of rating and ranking of query results and to estimate the effectiveness of those techniques. In database community main focus is on enormous collection of the structured data which resulted in development of number of artificial techniques and methods for processing or executing the structured queries on the database. In today's era, the combination of database techniques and the information retrieval techniques is very vital. With the enormous growth of internet & increasing users of internet demanded requirement of keyword search techniques and to extend concept of keyword search over relational data. Keyword search techniques are very useful for analyzing both the structured as well as the unstructured data which contains the large amount of the textual information. In our research paper we will analyze various keyword search techniques and we will also try to analyze the areas on which we can work to improve performance of keyword search algorithms.

This paper describes Structured and Unstructured Data, Keyword Search, Relational Keyword Search, Keyword Search Technique Classification and Conclusion.

2. Structured and Unstructured Data:

Structured Data is one in which data is organized in terms of structures i.e. relations or tables and that structure will follow a strict database schema Like in SQL. These tables further organize the data in terms of rows and columns, where rows refer to tuples or records and columns refers to attributes and These all tables are bounded together with some cardinality or relationships e.g. one to many, many to many and so on as show in figure 1.

Unstructured data is totally Opposite to structured data. It contains data that do not organized in any predefined Schema. It can be in any form like Audio, Videos, JPEG Files, Pdfs , Text Files and so on and it is usually refers to information that doesn't reside in a traditional row-column database.

Example: As shown above in Figure 1: Structured data Contain data about Customer, Sales person and Monthly Record via Structured Data.

Figure 1: Relationship among Product, Customer, Sales person and Monthly Record via Structured Data.
3. **Keyword Search:**

Information retrieval is the process of gathering information by using keywords from the relevant document and that document can be unstructured or structured data. It hides its complexity from user by providing abstract view. As user don't have any knowledge about schema and any other query processing language, he can search through abstract interface by putting keywords. By using Keyword Search user can submit keyword to search engines (Internet Search) or structured data and in turn it returns a list of documents to user according to ranking. Ranking of documents are provided based on the keywords match and occurrence of keyword match in particular document. Ranking is provided in descending order of occurrence of keyword match and the document with maximum occurrence get higher priority.

4. **Keyword Search Techniques Classification:**

Keyword search techniques are classified into two main groups:

1. Schema Based Keyword Search
2. Graph based Keyword Search

4.1. **Schema Based Keyword Search:**

Schema based approaches support keyword search over relational database (like SQL) using execution of SQL commands [1]. These techniques are combination of vertices and edges including tuples and keys (primary and foreign key). Each tuple in database uses as vertex and edges define interdependency among tuples.

In the case of RDBMS, keyword search using the Schema Based Approach is performed via making use SQL. Schema Based approach working is divided into the two main steps:

i) Determine how to create and generate SQL queries in order to find out the structures among tuples.

ii) Determine how to evaluate the queries which are generated in step 1 efficiently.

**Discover:**

Discover is technique enabling their user to search into database via keywords with getting any query language Knowledge. According to searching keywords Discover First Create Candidate network graph of tuples and relations then graph output shortest sequence first. It performs all searching operations in two major steps as.

- Candidate Network Generator: It helps in generating all candidate networks of relations, which are known as join expressions one that generate the joining networks of tuples.

- Plan Generator: It builds plans for the efficient and the proper evaluation of the set of candidate networks, by making use of the opportunities to reuse common sub expressions of the candidate networks.[3]

In order to generate the optimal execution plan with respect to the actual cost, DISCOVER make use of the greedy algorithm. One main aspect of the DISCOVER is that, it performs keyword search without using the requirement of the user to know schema of the database.

For ranking the result, DISCOVER returns a monotonic score aggregation function. [5] The main drawback of this algorithm is that the cost of generating CNs set is very high. [4]

**Spark:**

The demand for RDBMs to support keyword search on text data is increasing as there is wide increase in the text data stored in the relational databases. The existing keyword search methods are not feasible for the text data search. The main aim of these techniques is to focus on effectiveness and efficiency of the keyword query search [7]. Spark concept devise a new ranking
formula by making use of the existing information retrieval techniques. The main use of the SPARK is that it is works on large scale real databases (Eg. Customer Relationship Management) by taking in consideration both the RDBMS Effectiveness and Efficiency.

It makes use of the Top-k Join algorithm which includes two efficient query processing algorithms for ranking function.

1. Dealing with Non-monotonic scoring function.
2. Skyline Sweeping Algorithm.

2. Graph Based Keyword Search:

A data graph DG is the clear representation of Relational Database. In this graph algorithm is extended to solve the keyword search queries. In this representation we have, DG (V, E), where DG is the directed graph, V is the set of the vertices represent data or record and E is the set of edges which also define relationship entity. In this graph, weights are assign to the edges in order to represent the proximity of the corresponding tuples, for example, we have two vertices u and v then the proximity of u and v is represented by a weight denoted by We{(u,v)}.

Types of Graph Based Search:

A. BANKs:

BANKS stands for Browsing and Keyword Searching. BANKS system represent relational model into data graph and according to matching keyword it activates graph nodes.

Structuring of BANKS algorithm:

In BANKS system, the database is represented using a directed graph and the record or tuple is represented as a node in the directed graph. Foreign Key or Primary keys are presented using the edge, which corresponds to the link between the corresponding tuples. The result processing of BANKS system will returns a sub-graph which is represented in form of connecting nodes, one which matches the query keyword. This sub-graph can also further refined in order to get the more accurate or more appropriate answer for the keyword we have searched.

The central node in this graph is one which connects all the keyword nodes and relationship between them. And this central node is known as root node and answer can be considered as rooted directed tree which contains directed path from root to each keyword node. The root node is information node and tree is connection tree.

Figure 3: The DBLP Bibliography Database (A Fragment of Database.)

BANKS system makes use of two types of datasets, DBLP and small thesis database. DBLP database is shown in Figure 3, which also shows how DBLP is converted into structured relational format. There were 124,612 nodes and 319232 edges on the graph.

Representation of the BANKS system consists of vertices, edges, edge weights, node weights. Each tuple T in database is represented by a node T in the graph. Suppose that there are two tuples T1 & T2 such that they have foreign key relationship, then in the graph this is represented by an edge from T1 to T2 and also a back edge from T2 to T1.

The weight for forward link along foreign key relationship reflects proximity relationship between the tuples and it is set to 1 by default. In the current implementation of BANKS the forward edge weight is set to s(R(u),R(v)) and the reverse edge weight is set to [s(R(v),R(u))*INv(u)] and the actual weight is the minimum of the two as follows:

\[ b(u,v)=\min(s(R(u),R(v)), s(R(v),R(u))*INv(u)) \]

BANKS is mainly used in order to publish organizational data, bibliographic data and electronic catalogs. BANKS helps us in extracting the relevant information without the knowledge of schema of the database [8]. A user can simply extract the required information by typing some keywords, then following hyperlinks and interacting with controls on the displayed results.

Keyword searching in BANKS is done by making use of the proximity based ranking on foreign key.
B. DataSpot

DataSpot is a database publishing tool and it lets the end user to explore the large database without making use of any query language. DataSpot make use of schema-less semi-structured graph which is known as hyperbase. According to the concept of the DATASPOT, the Search Server performs searches within the hyperbase and in return as a result it returns either HTML pages or object API[9]. The DataSpot used in electronic catalogs, yellow pages, classified ads, help desks and finance.

Model Description:

A graph structure which is used for the data representation in DataSpot is known as hyperbase. Hyperbase consists of nodes, edges, and node labels. Nodes are connected via directed edges. These directed edges are further classified as simple edge and identification edge. The simple edge is an edge in between the parent node and the child node. These set of children nodes are ordered while the parent nodes are not ordered. The simple edges cannot make a cycle.

Figure 4: The DataSpot Architecture

And the second type is identification edge, it is used to indicate reference and subject node relationship. The reference node uniquely identifies subject node and subject node can have maximum single reference node. Initially nodes in hyperbase represent data objects and edges represent association between them. DataSpot publisher translates the source data into hyperbase, which in turn efficiently queried by DataSpot server.

C. Proximity search

Proximity search works on general relationships among objects in order queries answers, these techniques helpful for the interactive query sessions. In text Processing Proximity Search looks for Documents or text files where two or more term occurrences according to match are within specified distances. Where distance is number of intermediate words. Internet Movie Site Database (IMDB) site makes use of the proximity search in order to answer its database queries. IMDB sites consist of 140,000 movies and information about over 500,000 film industry workers. The concept behind is that the database can be viewed as set of linked objects, where objects represent movies, actors, directors and so on. And the distance function based on links separating objects. [10]

Model Description:

The database is represented via graph in which data (objects) represented as vertices and relationship. Proximity works on the shortest distance between objects. For proximity searching database is viewed as collection of objects where objects are related by distance function. The figure 5 shows Proximity Search Model.

Figure 5 : Proximity Search Model

Conclusion:

In this paper we have analyzed various schema based and graph based keyword search techniques that provide an abstract view to user and in almost all approaches we have find the problem related to space and time complexity. In future we can reduce these deficiencies by introducing new models and concept to extend our research on implementing keyword search on the unstructured documents.

Reference:

1. A.Baid, I. Rae, J. Li, A. Doan, and J. Naughton “Toward Scalable Keyword Search over Relational Data,” Proceedings


6. Qi Zhou, Chong Wang, Miao Xiong, Haofen Wang and Yong Yu “SPARK: Adapting Keyword Query to Semantic Search” This work is funded by IBM China Research Lab.


