Available Online at www.ijcsmc.com

International Journal of Computer Science and Mobile Computing

A Monthly Journal of Computer Science and Information Technology

ISSN 2320–088X
IMPACT FACTOR: 6.017

IJCSMC, Vol. 7, Issue. 5, May 2018, pg.18 – 23

Searching in Database and Improving Output using Frequency Based Ranking

Ms. Unnathi Ladhe[1], Ms. Rohini Sable[2], Ms. Shital Tilekar[3], Prof. S. N. Ghotkar[4]

[1]Department of Computer Engineering and IT, College of Engineering, Pune-05
[2]Department of Computer Engineering and IT, College of Engineering, Pune-05
[3]Department of Computer Engineering and IT, College of Engineering, Pune-05
[4]Department of Computer Engineering and IT, College of Engineering, Pune-05
[1] unnathiladhe@gmail.com, [2] rohinisable1997@gmail.com, [3] shitaltilekar@gmail.com

Abstract -- Searching is one of the main topics being discussed these days, may it be in files, folders, big data or Database. The data is increasing as time passes; every second, minute data is added to data stores. Generally searching results in database gives multiple tuples as output of a query. To provide users with most appropriate output, frequency based ranking is used to improve the most likely solution of the query. Proposed methodology uses, query ranking mechanism based on frequency of keywords found in the query made by user. This approach helps to rank most relevant search results on the top of the result list.

Keywords-- SQL Like Operator, Database, Stopword, Keyword frequency, Search Engine

I. Introduction

Search engine can be used on database not only to retrieve output from world wide web but can also be used to search keywords corresponding to user specification on local database. Database can store enormous amount of data amounting to upto 10 GB or even more, for searching in such a big database it becomes difficult. Simply using 'AND' operator, 'LIKE' operator, truncation and wildcards cannot guarantee most relevant search output. Therefore using various operator and clubbing them with different natural language processing and information retrieval algorithms search output can be made relevant.

Stopword are words of no importance in the query which need to be omitted from the index of keywords. Stemming is the process of reducing inflected (or sometimes derived) words to their word stem, base or root form—generally a written word form. It can be used for indexing word to perform further process for searching. Stopword removal and stemming can be applied to the search query to break it down to keywords. Further the keywords can be used along SQL queries to find occurrence of them in the database. Sql query can also be used to keep a count of keywords present in particular tuple, marking it as the frequency for the tuple. The frequency of keyword can be stored in another table of database with primary key and its frequency as tuple. This table can be used finally to rank the output as per the max_count or the max frequency count for a particular primary key and then output the most relevant tuple as result.

What is natural language processing?

NLP is the processing that is performed on the language that we use in our day-to-day lives, i.e., the human language. Here, the language may be English, Pali or any other language. The processing of different languages is performed differently based on the grammatical and semantic rules of that particular language. It is used in variety of areas just like for speech recognition and artificial intelligence. Nowadays, Language Technology or Language Engineering are also the terms that are being used instead of NLP. Often, language and speech go hand in hand(e.g., Speech and Language Technology). NLP is essentially multidisciplinary: it is closely related to linguistics. It also has applications in research fields like cognitive science, psychology, philosophy and math (especially logic). Also, compiler techniques, theorem proving, formal language theory, machine learning and human-computer interaction use Natural Language Processing as a key factor. In this system, the question and answers on the QA portal will be in the natural language. In order to make the computer understand this natural language, we need natural language processing. When NLP processing is done on the questions and answers, the meaning can be extracted and that could be used in the recommendation part to recommend the appropriate software.

LIKE operator in SQL--

The SQL LIKE clause is used to compare a value to similar values using wildcard operators. There are two wildcards used in conjunction with the LIKE operator (the percent sign,%; and the underscore,_)The percent sign represents zero, one or multiple characters. The underscore represents a single number or character. These symbols can be used in combinations.[2]

II. Proposed Methodology

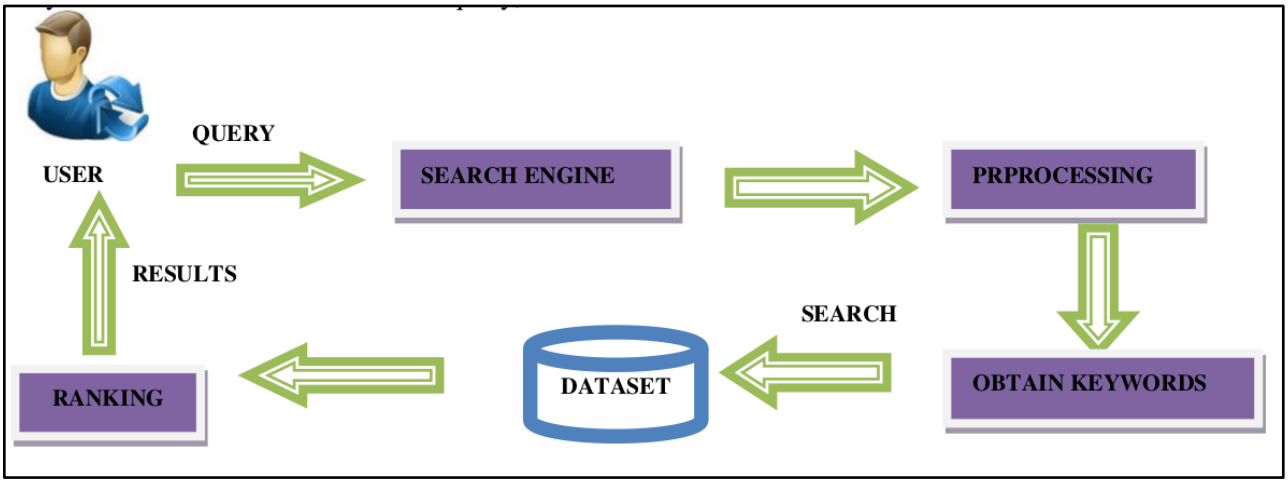

Fig: Proposed System Architecture

Working:

User inputs the query that needs to be searched in the search tab of search engine. The query input is then processed to obtain the keywords. Processing of query is as follows; query filtering, stemming, and stopword removal to get the final index of keywords. Once the keyword obtained searching is performed on the dataset(data present in database). Proposed methodology sets frequency of keywords on the searching table in another database table which contains primary key and related frequency(initially initialized to zero '0'). Further, the frequency from the frequency table is used to rank the most appropriate output at top of the list.

Query Filtering:

```
function test_input($data) {
        $data = trim($data);
        $data = stripslashes($data);
        $data = htmlspecialchars($data);
        return $data;
};
```
[3]

This function test_input() is used to filter the query input given by the user. trim() is used to remove whitespaces and predefined characters if any in the front and end of input.
stripslashes() is used to remove backslashes and clean up data if added by html code.
htmlspecialchars() encodes harmful html code inserted by user, ex sql injection attack, etc. It converts 5 characters to corresponding HTML entities wherever applicable.

Stemming:

There are various stemming algorithms available. The stemming algorithms can be classified into three types: Truncating, Statistical and Mixed.[4]

Stopword Removing:

```
<?php
        Function removeCommonWords($input){
                $commonWords = array(common words);//array containing common
words
                return preg_replace('/\b('.implode('|',$commonWords).')\b/','',$input);
        }
?>
```

Above function can be used to remove stopwords from a input string to this function.[5]

Searching every keyword and counting frequency:

```
//where $chars is array of keywords and $len is length of the array
for($x = 0; $x < $len; $x++){
        $query = "update <freq_table> set freq = freq + 1 where <primary_key> in (select
        <primary_key> from <table> where <search_column> like
        \"%".$chars[$x]."%\")";
        echo $query;
        $conn->query($query);
}
```

The above pseudo code depicts the use of LIKE operator to search keyword in required column of database. The query updates frequency table if the occurrence of keyword is found efficiently, though if the search column contains keyword multiple times frequency is marked as one, '1'. Here the freq of keywords is noted not frequency of particular keyword.

Further the output can be displayed using the frequency table; displaying the tuple with most frequency on the top.

System Design:(Data flow Diagram)--

The data flow diagram (DFD) is one of the most important modeling tools. It is used to model the system components. These components are the system process, the data used by the process, an external entity that interacts with the system and the information flows in the system.

The following gives diagram gives the data flow of the system using the data flow diagram modeling.

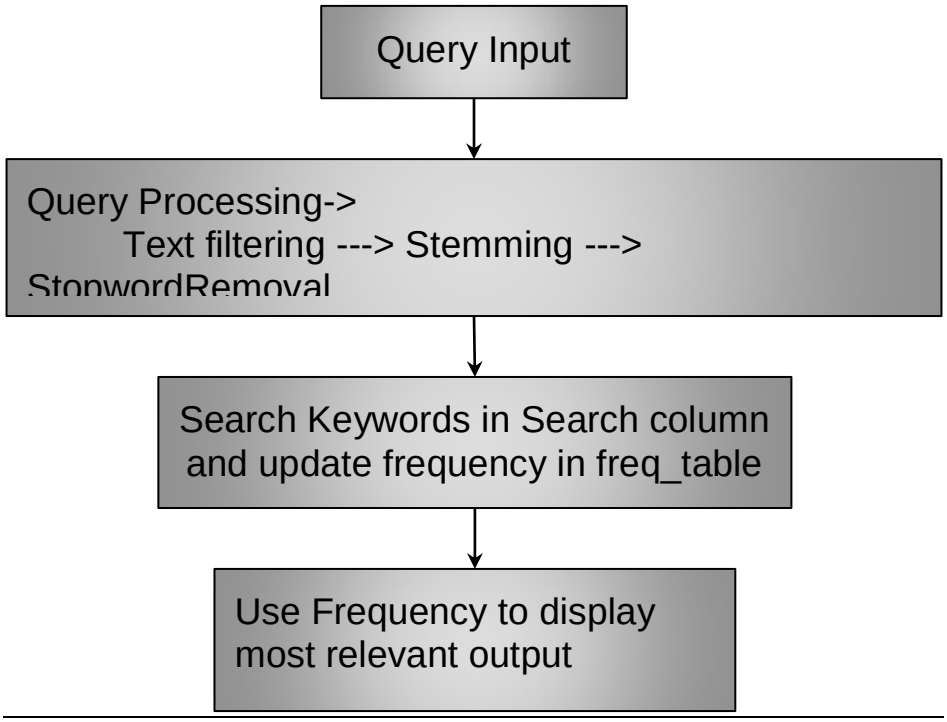

Fig: Data Flow Diagram of the Proposed System

III. Conclusions

Searching and search engine optimization have become a recent trend in today's research topic. Therefore the proposed method is to perform search on column of database, the method can be applied to on multiple column, and provide user with most appropriate and relevant output of the input query in search tab. Proposed methodology focuses on searching in database and displaying most suitable output on topmost of the output. Future work can be application of pthreads available in php api to perform search more speedily.

References

[1] User Intention Modeling in Web Applications Using Data Mining(2002), Zheng Chen, Microsoft Research Asia, 49 Zhichun Road, Beijing 100080, PR China.
[2] "tutorialspoint" URL https://www.tutorialspoint.com/sql/sql-like-clause.htm
[3] "techfry" URL https://www.techfry.com/php-tutorial/validate-form-data-with-php
[4] Ms. Anjali Ganesh Jivani "A Comparative Study of Stemming Algorithms" URL https://pdfs.semanticscholar.org/1c0c/0fa35d4ff8a2f925eb955e48d655494bd167.pdf
[5] KeithMorris, "github" URL https://gist.github.com/keithmorris/4155220

