International Journal for Research in Applied Science & Engineering Technology (IJRASET)

Improved Fuzzy Searching Technology

Mr. Shaikh Asharfali Ahmed[1], Prof. Ms. Megha Singh[2]
[1]M. Tech, [2] HOD and Asst. Professor, Dept. of Computer Sci. and Engg.
Central India Institute of Technology, Indore, MP, India

ABSTRACT-Instant searching technique finds answers to a query instantly when user types in keywords character-by-character. Fuzzy searching is advancement in instant searching which finds perfect match keywords of query keywords. User expects fasts results within few milliseconds and perfect match. This is the main computational challenge in this the high-speed requirement, i.e., each query needs to be answered within milliseconds to achieve an instant response and a high query throughput. In this paper, I propose fuzzy search by doing ranking to obtain results in efficient time and more accurate. Number of studies have been done like computing all answers which is slow and requires more space. Early termination technique may solve the above problem up to some extent. I proposed an approach that focuses on common phrases in the data and queries I also proposed searching on synonymous of keywords to find relevant answers by using edit distance function.

Keywords: - fuzzy search, proximity ranking, edit distance, dictionary, inverted index, trie index.

I. INTRODUCTION

Computers and computer networks are being used in each and every field. Information storing on large capacity is possible .Net- work provides resource sharing feature. Information is stored on various storage devices at different sites. As we store large amount of data it is necessary to have techniques to find out the required information from these storage devices/databases. Information searching must be fast and exact one. The user expects results as soon as he fires a query. Instant search is a system which finds answers to a query immediately while a user enters words character-by-character. Fuzzy improves user search experiences by finding resutls with keywords like to query keywords. A main thing in this that each query needs to be answered fast within less time to achieve an instant response

II. PROBLEM STATEMENT

Development of search system using advanced search features such as instant search, fuzzy search with edit distance idea and search based on synonymous words. Existing search schemes provide different types of ranking such as paging, ranking based on ranking based on term frequency and number of references of the documents and inverse document frequency. Ranking is very important as it determine the importance of the solutions as search queries usually contain similar keywords and user expects documents which have query keywords together.

III. RELATED WORKS

A. Fuzzy Search

The studies on fuzzy search can be classified into two categories, gram-based approaches and trie-based approaches. In the former approach, substrings of the data are used for fuzzy string matching . The second class of approaches indexes the keywords as a trie, and relies on a traversal on the trie to find similar keywords. This approach is especially suitable for instant and fuzzy search.

B. Indexing

Inverted Index is utmost usually used index; in this each keyword is mapped to a list of documents where the word is present. Indexing can play important role for faster retrieval of search results. Trie index is a form of hierarchy data structure, each node excluding a leaf node consists of many branches every branch signifies a specific character of the word. Forward Index is a kind of index in which a document is mapped to list of keywords existing in the document. As mentioned in Hyb indexing outclasses inverted indexing by a aspect of 15 – 20 in worst case. Hyb indexing is a difference of inverted lists in which data in compressed.

C. Ranking

Ranking is used to rank documents based on distance between query keywords. The entire set of documents which have all query keywords or words similar to query keywords are considered for ranking. If all words are in matching bin or neighboring bin then the document is ranked higher else the document is ranked lower.

IV. PROPOSED METHOD

A. Proposed System Architecture

Searching and ranking module is responsible for searching relevant result using keywords and indexes, this module ranks results based on proximity distance between query keywords. User uses the system to search significant results. File system stores data set as well as index files.

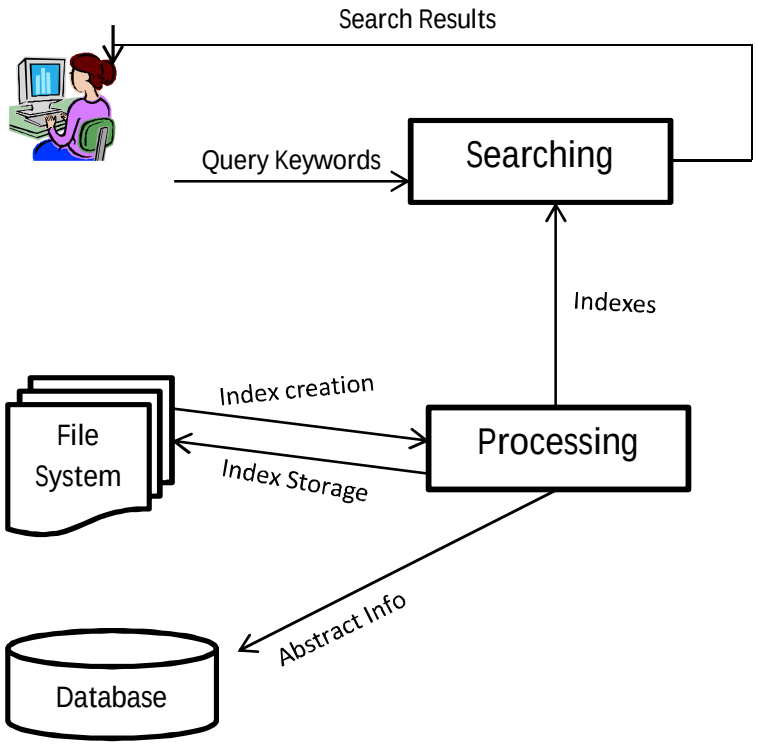

Fig.1. System Architecture

This system is implemented by using three novel algorithms such as -

1) FindSimilarWord: This algorithm is used to find list of similar words for a given query word with threshold distance "tr". During processing step all unique words in data set are represented using Trie data structure. "node" represent root node of a Trie. Child nodes of a Trie are traversed, while traversing if the edit distance of a node which represent the prefix of certain word in dictionary exceeds threshold distance then the remaining child nodes are skipped from checking for similar words. If the edit distance is within threshold distance then the word will be added to result list of similar words.

2) Proximity Ranking: This algorithm proximity Ranking, is used to rank documents based on distance between query keywords. The entire set of documents which have all query keywords or words similar to query keywords are considered for ranking.

3) Search Algorithm: This algorithm search, is used to search relevant documents. Initially preprocessing of query keywords is done, this involves removal of stop words from the keyword list. For each query keyword stemming is performed. To find list of similar words to each query keyword threshold distance is calculated based on length of query keyword. Similar words are found using Edit distance. Inverted lists are intersected to determine documents which contain all query keywords. Proximity ranking is applied to find documents with phrases. Documents without phrases i.e. Documents containing few query keywords but they do not form a phrase will be identified. The result will be displayed as union of documents with phrases and documents without phrases.

Volume 3 Issue IX, September 2015
ISSN: 2321-9653
International Journal for Research in Applied Science & Engineering Technology (IJRASET)

V. EXPERIMENTAL RESULT

This system is tested by using IMDB dataset which contains 1647 records. The efficiency is checked by considering precision and recall parameters.
In our system, Assume the following:
A database contains 1647 records on a particular topic
A search was conducted on comedy, action, short, funny and mini and 380, 203, 1, 380, 1 records were retrieved respectively.
Calculate the precision and recall scores for the search.

Solution

A = The number of relevant records retrieved,

B = The number of relevant records not retrieved, and

C = The number of irrelevant records retrieved.

A = 380, B = 380 and C = 0

Recall = (380 / (380 + 0)) * 100%

= 380/380 * 100%

= 100%

Precision = (380 / (380+0)) * 100%

= 380/380 * 100%

= 100

Sr. No	Keyword	Number Of Relevant Records Retrieved(A)	Number Of Relevant Records In The Database Not Retrieved (B)	Number Of Irrelevant Records Retrieved. (C)	Recall	Precision	Accuracy
1	Comedy	380	0	0	100	100	100
2	Action	203	0	0	100	100	100
3	Short	1	0	0	100	100	100
4	Funny	380	0	0	100	100	100
5	Mini	1	0	0	100	100	100

Volume 3 Issue IX, September 2015
ISSN: 2321-9653
International Journal for Research in Applied Science & Engineering Technology (IJRASET)

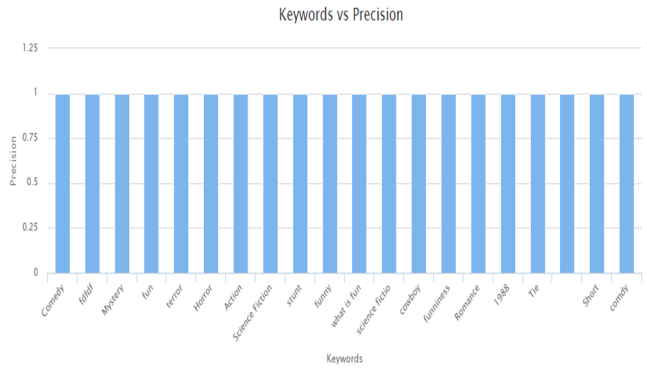

Fig: 2 Keyword vs Precision Graph

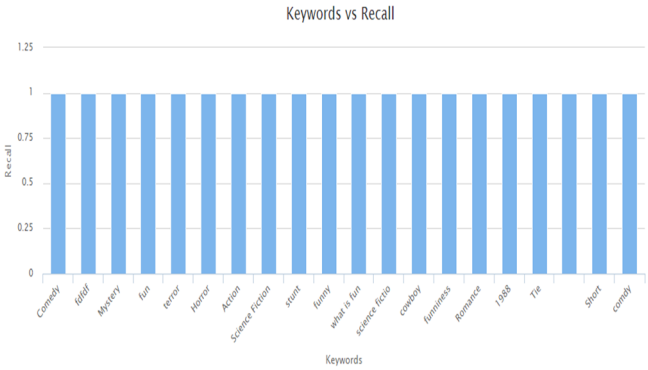

Fig: 3 Keyword vs Recall Graph

VI. CONCLUSION

This system is developed to provide advanced search features like fuzzy search and proximity ranking and synonymous wise search. Fuzzy search helps user in retrieving significant results even if there are few typing errors in the query keywords. Proximity ranking ranks records based on distance between query keywords. Edit distance is used in fuzzy search. Proximity ranking makes use of altered inverted list by storing word positions in the form of bin Id. There are other advanced search features like auto completion, page ranking etc. which will be considered in future for implementation.

REFERENCES

[1] Inci Cetindil, J. Esmaelnezhad, T. Kim and Chen Li “Efficient Instant-Fuzzy Search with Proximity Ranking,,” IEEE 30 International conference on data engineering year 2014.
[2] M. Persin, J. Zobel, and R. Sacks-Davis, “Filtered document retrieval with frequency-sorted indexes,” JASIS, vol. 47, no. 10, pp. 749–764, 1996
[3] A. Singhal. “Modern information retrieval: A brief overview.” Bulletin of the IEEE Computer Society Technical Committee on Data Engineering.