Documents Retrieval Using the Combination of Two Keywords

Rohitash Chandra Bhensle, Saikiran Chepuri, Menta Snjeeva Avinash
M. Tech. Scholar
(Software Technology)
VIT University Vellore, Tmilnadu, India

Abstract* - In the search engine, the NLP (Natural Language Processing) and statistically-based systems are used for making the query. The statistical system is recognizing the terms for searching and also it provides the stems and singular and plural forms of words. The statically based system may also provide the weights of every term. In the Natural Language processing system the parts of speech, identifies objects, verbs, subjects, agents and synonyms and alternating forms for appropriate nouns are tags. Then it is able for creating an unambiguous representation of submitted query and the term weights are computed. For the particular query request the list of the documents are retrieve on the search engine from the database. Using the keywords the search engine obtained the results for submitted query. The Stemming algorithms and Stop-lists/Stop-words are used for reducing the consuming of size of the disk. 'the', 'is', 'an' are the example of stop-words and 'reading', 'playing', 'watches' are the examples of stemming algorithms. In the Information Retrieval system the vector space model and the Boolean model are using for the documents ranking. The search engine optimization is started with submitting the keywords on the search engine that should be very clear and understanding for the query processing and also known that which keywords are more relevant and will performs well for better results. So, in this paper, for retrieving the documents from the database the new technique, combination of the 'two keywords' are proposed and rearranges the list of documents in the order of weight.

***Index Terms* - Search Engine, Information Retrieval (IR), Ranking, Natural Language Processing**

I. INTRODUCTION

The idea of searching the Information Retrieval using computers is popularized from the article "As We May Think" by "Vannevar Bush" [13]. In the information retrieval system the complete search process is depends on the query and the particular query is to go look for truly what information we want to retrieve in your search engine. It is the main base of the questions that produced to come close to get best retrieve the information in the search engine i.e. the information retrieval system in the search engine is responsible for retrieval, maintenance and storing of information results. For information retrieval the query is based on full-text or keywords.

A query is made up of the keywords which describe the search topic and the relevant terms of keywords are focused the retrieval process using operators. From a question like: 'Who is the best actor?' So here the keywords are (best and actor). We can add the additional keywords like Hollywood, bollywood or world.

The best approach of information retrieval is forever to use the *least* words possible while more than one keyword is generally required. The popular search engines like Google return significant results to natural language queries. Semantic search engines like Hakia are now accomplished of make intellect out of sentences although there are a lot of search engines that have not fixed up to this stage of complexity, which is why it is still good to give attention to *significance of* keywords. Now you have the building blocks to construct a successful query.

Formally, when the users are submitted the query on the search engine for information retrieval, the number of the documents results are listed down. The Retrieval of Information in the search engine has been categorized into several ways depends upon Text, Image, music and Video. The aim of the Information Retrieval is to manage the accurate and relevant searches which retrieve from the based on user query submission [10]. It is inquired of retrieving the structured information instead of the unstructured information which is happens for the users to submit query in their own sentences i.e. Natural Language processing is concerned. Traditional Information Retrieval systems are extracting more relevant documents in unambiguous manner on the other hand web information contains a lot of information in an ambiguity and uncertain structures. The ambiguous queries are having heavy of data i.e. circumstances that can have more relevant explanation. Where the modern search engine can exhibits the numerous blocks of data, some textual and multimedia types of data that are listed out in the ranking method.

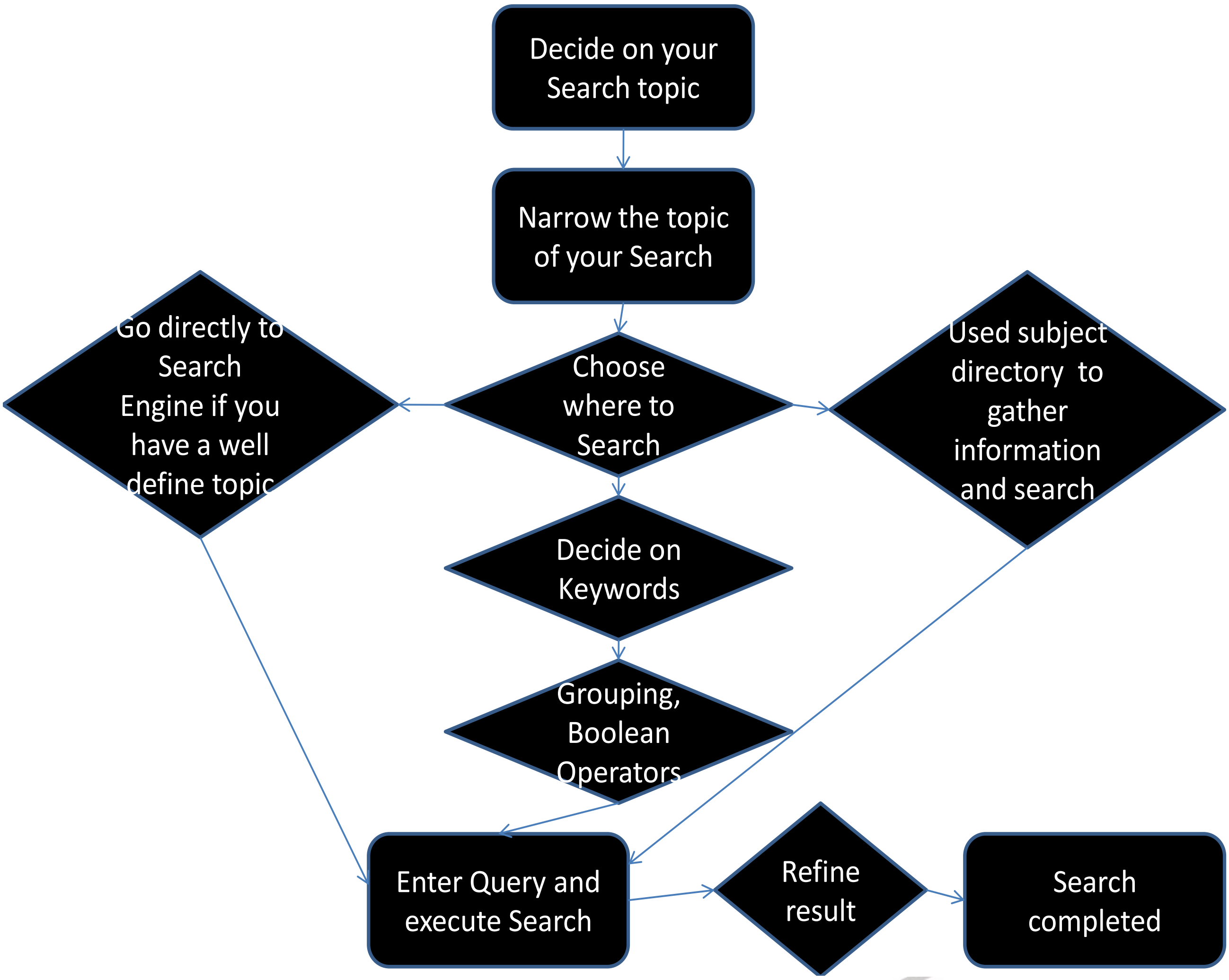

Decide on your Search topic
Narrow the topic of your Search
Go directly to Search Engine if you have a well define topic
Choose where to Search
Used subject directory to gather information and search
Decide on Keywords
Grouping, Boolean Operators
Enter Query and execute Search
Refine result
Search completed

Fig 1 flowchart diagram of query processing.

However, in this paper our study and research of area is rearranging the listed Text documents based on the combination of two Keywords. Text Document Retrieval is the collection of textual information which is the fundamental issue of Search engines [6]. It gets the information for submitted query through, extracting the keywords, in the query and listed documents i.e. Relevance Feedback [9]. One of the Information Retrieval concepts is that the query can be internally processed and judges that which documents are relevant or non-relevant for retrieving the information [3].

In recent last few years, a lot of IR results came along with significance judgment grow to be availability, supervised learning-based method have been devoted to automatically learning an efficient ranking function from training data.[19] See [14], [15], [16], [17], [18].

There will be a huge number of documents collection in the internet where as the problem of listing the documents as per structured is being complicated. Google uses Page Rank algorithm to rank their websites in the search engine. In the Larry's Page Ranking algorithm, its determine the ranking of web pages by counting the number of quality links on the search engine to a page determining how the web document page is important [20].

Page Rank spread out the scheme of back- link by "not counting links from all pages equally but also normalized by the number of links on a page." (Brin and Page, 1998).[20]

In the field of Search engine processing there are many categories for retrieve the information like Indexing, spider or crawler, Searching and ranking. For seeking the website, a program called crawler or spider are also scan the information of the documents in the order to form the replica of visited pages for information retrieval, in the search engine index to grant the fast and effortless search process. Through the Indexing method term frequent word or keyword can extracts from the listed documents and construct in a appropriate alphabetical order of the terms. [12].

These days the Search-Engines work as the collection of information requirements, strangeness and user friendly. When the users submitted the query, they expect to find "the more relevant information" i.e. find exactly what he would like, without being plagued with a number of riotous and unfeasible responses. In this paper, a new method is proposed for retrieving the more relevant information and Rearranging or re-ranking of the listed documents in search engine based on the "combination of two keywords." This is improving the excellence of the query process on the Web Search Engine.

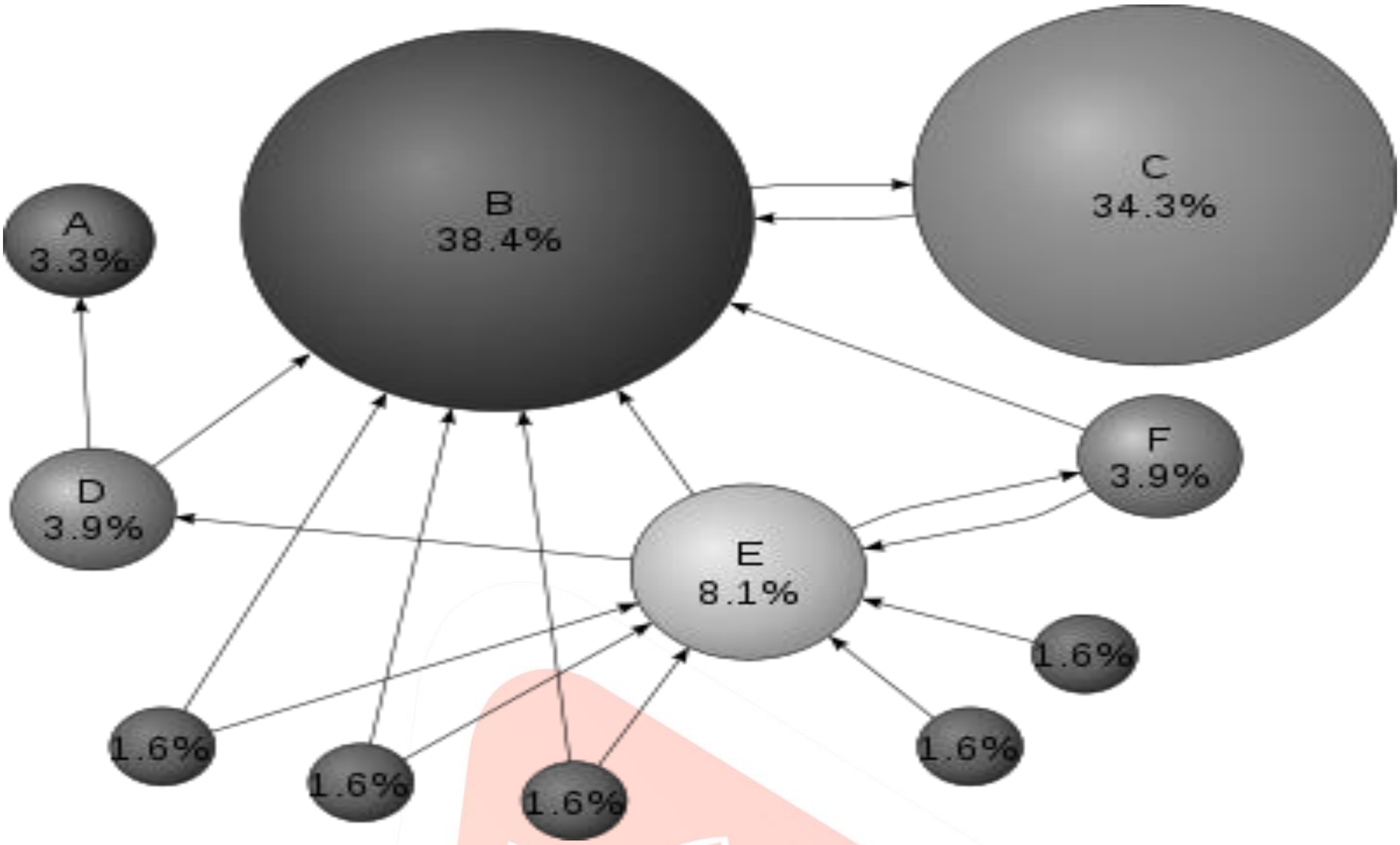


Fig 2 Example of PageRank from Wikipedia

II. LITERATURE SURVEY

In this paper, a analysis of prior work for retrieving the document which is processed through the use of keywords. Based on the submission of the query in the Search Engine the IR system response the some sort of list of documents. Since a query make with a logical combination of keywords, that can be found set of documents which are retrieved on search engine through different IR techniques such as Vector-space model, Boolean model etc. [10]. The performance evaluation of Information Retrieval can be measured through Precision and Recall process. Before going for all this, normalization and query refinements are done [2].

In this paper, for extracting the more relevant documents the combination of two keywords are taken based on the weight. For calculating the weight of the keyword the term Term-Frequency (TF) and Inverse Document Frequency (IDF) are used. The TF is gives the term which are more occurrence in the documents and the term IDF gives the percentage of importance of keyword in relevant documents i.e. the number of term "t" that occurred in a document "d" which is term freq (t, d). The formulae which is recommended by Cornell Smart for recognizing the term frequency as:

$$TF(t,d) = f(t,d)$$

Where f (t, d) is the no. of occurrence the terms 't' in the documents 'd'.

The Inverse document frequency which gives the scaling factor or importance of a keyword which is:

$$IDF(t) = log\frac{N}{|\{d \in D: t \in d\}|}$$

Where N = Total no. of Documents in the corpus.
$|\{d \in D: t \in d\}|$ = no. of Documents where the term t appears.
The TF-IDF measures:

$$TF - IDF(t,d,D) = TF(t,d) * IDF(t,D)$$

This gives the overall keyword weight [1] which is helpful for extracting the keyword.

An alternate work identified with Ontology connections that been utilized within Information Retrieval which enhances the exactness and review. Question extension is the significant vital which alludes to finding the terms in perspective of cosmology [4].

An alternate writing has distinguished about utilizing of Semantic door for discovering the semantic data from heterogeneous situations. The unrestricted documents from different areas can be scheduled for user which the system uses the accessible semantic information and gather other information from web pages. [5].

III. PROPOSED WORK

For extracting the information on the Search engine the Keyword significance is the most important concern that being followed by IR, therefore in this paper, obtained the documents that to be listed out using the combination of the "two keywords" which gives more relevant documents on the search engine rather than a "single keyword" search.

Based on the calculating of the weight of the frequent term from the listed documents, the most weight term taken as the keyword. For determining the occurrence of keyword's analyze the occurrence of the term in the one sentence of the paragraph, then one paragraph and then second paragraph and so on and whole document.

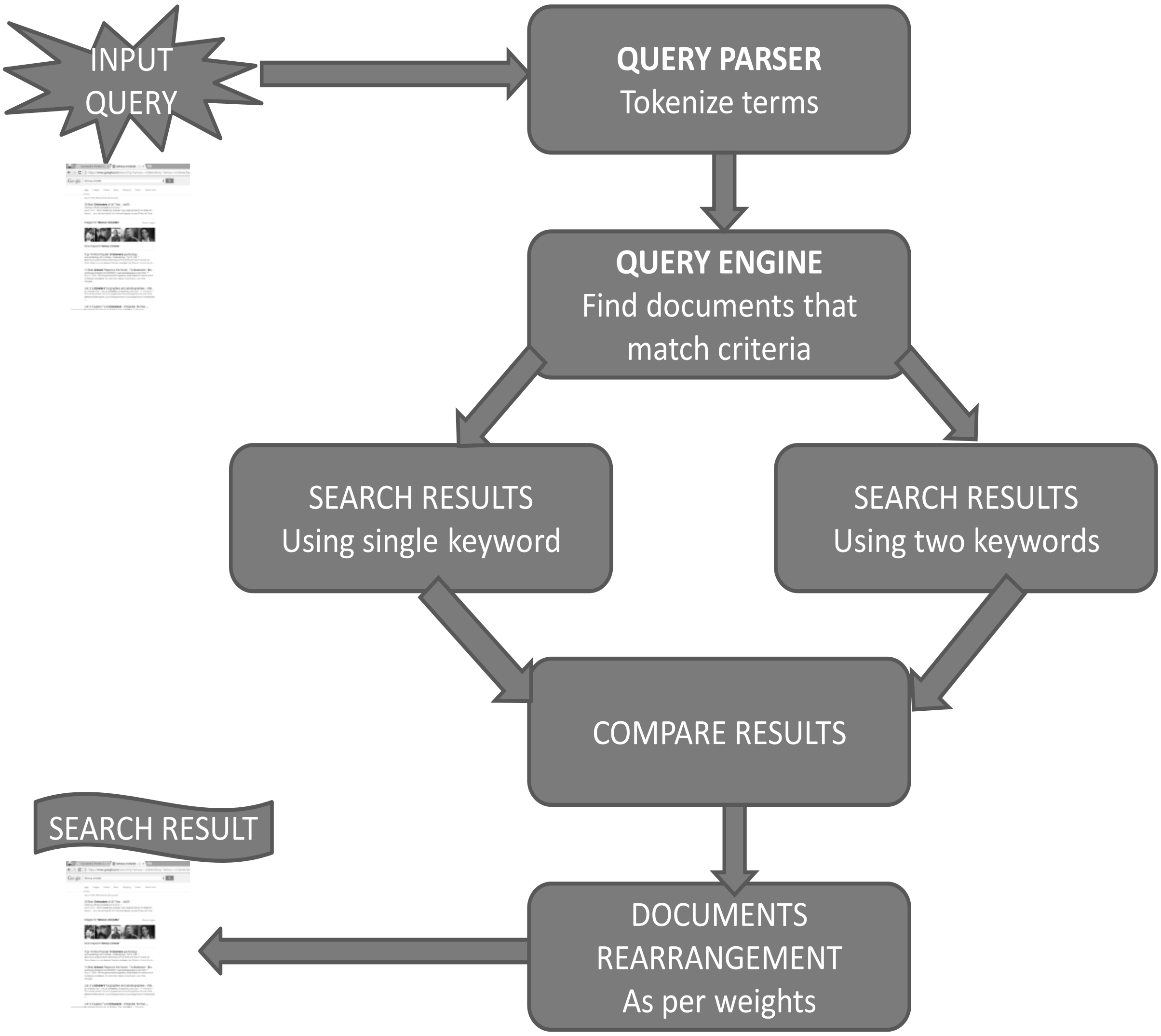
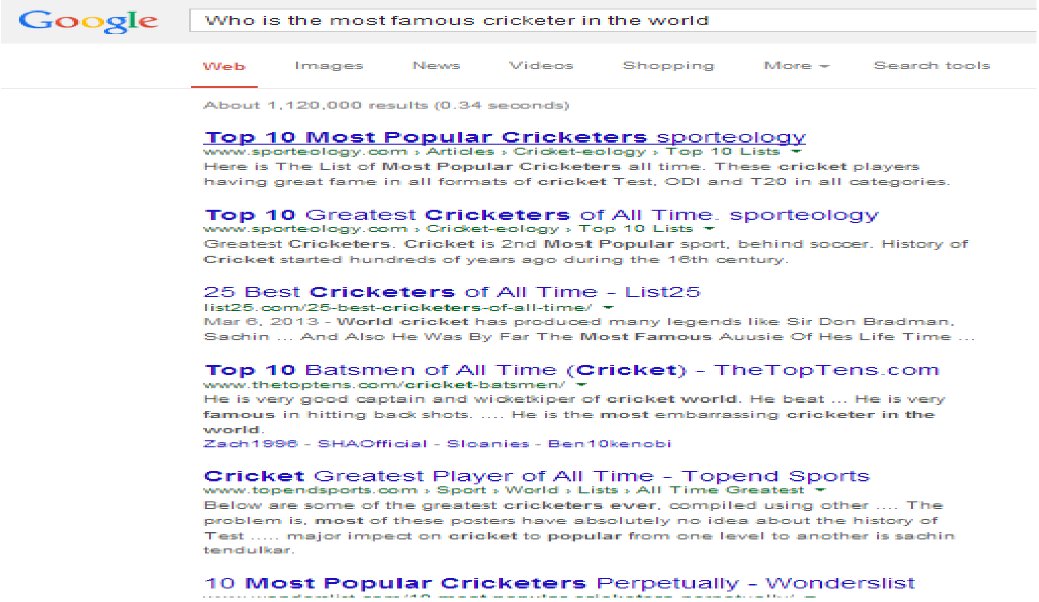


Fig 3 Diagram of information retrieving using combination of two keywords on the search engine.

For example, when users submitted the query on the search engine like "Who is the most famous cricketer in the world?"



Who is the most famous cricketer in the world

About 1,120,000 results (0.34 seconds)

Top 10 Most Popular Cricketers sporteology
www.sporteology.com › Articles › Cricket-eology › Top 10 Lists
Here is The List of Most Popular Cricketers all time. These cricket players having great fame in all formats of cricket Test, ODI and T20 in all categories.

Top 10 Greatest Cricketers of All Time. sporteology
www.sporteology.com › Cricket-eology › Top 10 Lists
Greatest Cricketers. Cricket is 2nd Most Popular sport, behind soccer. History of Cricket started hundreds of years ago during the 16th century.

25 Best Cricketers of All Time - List25
list25.com/25-best-cricketers-of-all-time/
Mar 6, 2013 - World cricket has produced many legends like Sir Don Bradman, Sachin ... And Also He Was By Far The Most Famous Auusie Of Hes Life Time ...

Top 10 Batsmen of All Time (Cricket) - TheTopTens.com
www.thetoptens.com/cricket-batsmen/
He is very good captain and wicketkiper of cricket world. He beat ... He is very famous in hitting back shots. .... He is the most embarrassing cricketer in the world.
Zach1996 - SHAOfficial - Sloanies - Ben10kenobi

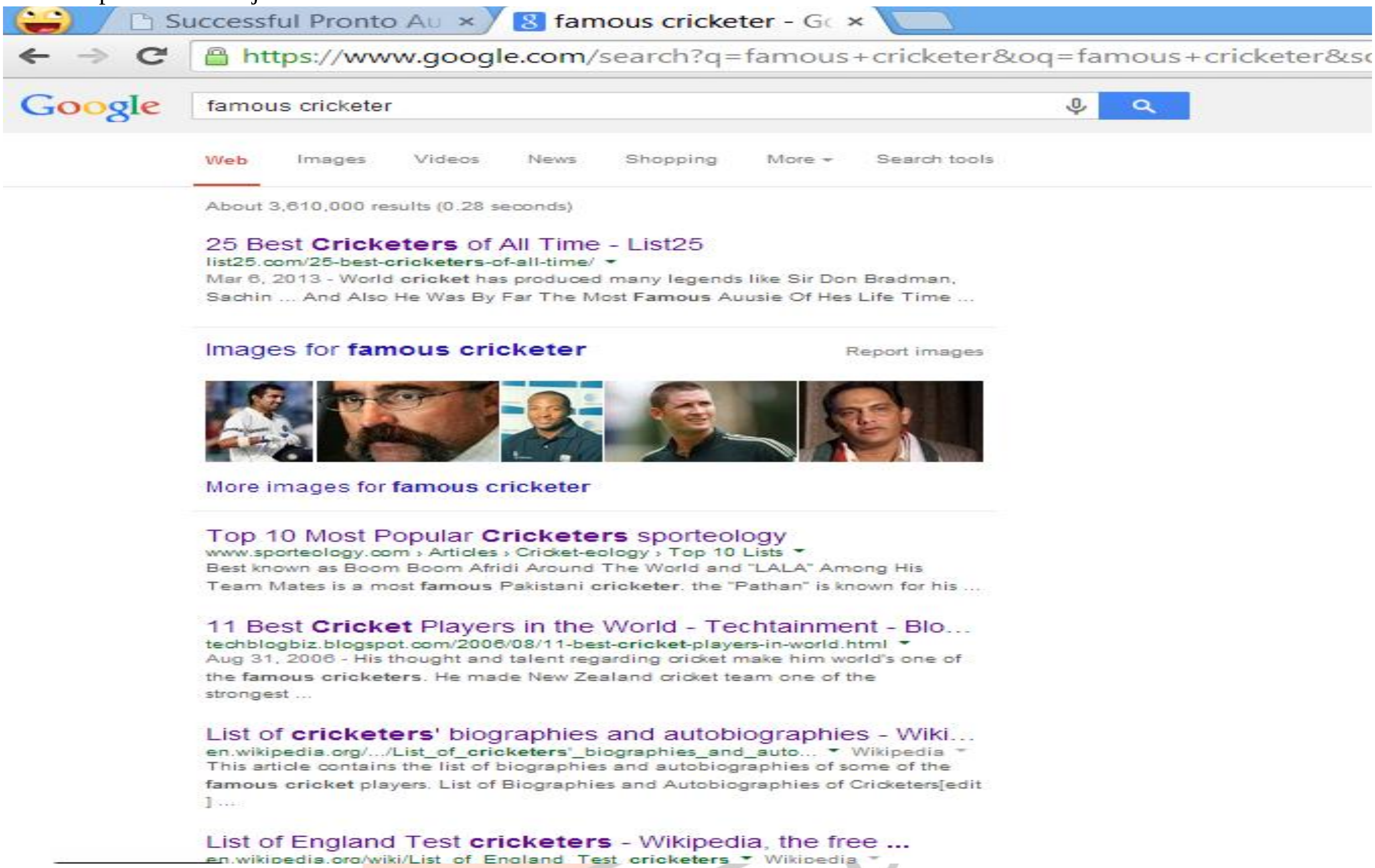
Cricket Greatest Player of All Time - Topend Sports
www.topendsports.com › Sport › World › Lists › All Time Greatest
Below are some of the greatest cricketers ever, compiled using other .... The problem is, most of these posters have absolutely no idea about the history of Test ..... major impect on cricket to popular from one level to another is sachin tendulkar.

10 Most Popular Cricketers Perpetually - Wonderslist

Fig 4 Listed documents using single keyword on the search engine.

Here the keyword “who” is question mark that can be avoided using stop-list and matches that can be termed as match as per stemming algorithm approach. After all the query elegance the keywords considered are: “famous”, “cricketer”, “world”. Then for searching the more relevant documents the combination of keywords are taken as a pairs such as: (famous, cricketer), (world, cricketer). The term “cricketer”, “world” and “famous” are to be searched for the most relevant documents. The figure number-(iv) shows list of documents on the Google page using single keyword.

We have considered the combination of two keyword as “noun” with “adjective” where the noun keywords are specifies the more importance for suggesting relevant documents. “famous cricketer” is taken as a example where crickerter specifies the noun and famous specified the adjective.



About 3,610,000 results (0.28 seconds)

25 Best **Cricketers** of All Time - List25
list25.com/25-best-**cricketers**-of-all-time/
Mar 6, 2013 - World **cricket** has produced many legends like Sir Don Bradman, Sachin ... And Also He Was By Far The Most **Famous** Auusie Of Hes Life Time ...

Images for **famous cricketer**

More images for **famous cricketer**

Top 10 Most Popular **Cricketers** sporteology
www.sporteology.com › Articles › Cricket-eology › Top 10 Lists
Best known as Boom Boom Afridi Around The World and "LALA" Among His Team Mates is a most **famous** Pakistani **cricketer**. the "Pathan" is known for his ...

11 Best **Cricket** Players in the World - Techtainment - Blo...
techblogbiz.blogspot.com/2006/08/11-best-**cricket**-players-in-world.html
Aug 31, 2006 - His thought and talent regarding cricket make him world's one of the **famous cricketers**. He made New Zealand cricket team one of the strongest ...

List of **cricketers**' biographies and autobiographies - Wiki...
en.wikipedia.org/.../List_of_**cricketers**'_biographies_and_auto... Wikipedia
This article contains the list of biographies and autobiographies of some of the **famous cricket** players. List of Biographies and Autobiographies of Cricketers[edit] ...

List of England Test **cricketers** - Wikipedia, the free ...
en.wikipedia.org/wiki/List_of_England_Test_**cricketers** Wikipedia

Fig 5 Listed documents using combination of two key words on the search engine.

Pseudo code

```
Algo_Two _Keyword Search
{
For (each document in local database)
{
String i,j; // i  &  j assumed  as the  keywords
w as weight
If (strcat ( i, j) !=NULL)
{
Retrieve “two keyword” combination search by going through:
        A, while  (Occurrence  of  keywords in one sentence of the paragraph.)
                W=w+4;
        B, while ( Occurrence of   keywords in a paragraph.)
                w=w+3.
        C, while ( Occurence of  keywords in whole document.)
                W=w+9.
}
Else
{
Retrieve “single keyword” search in a whole document.
}
}
```

Further, the more combination of keywords is not suitable to refine the search, like: consider a keyword like: “world famous cricketer” won’t give much sufficient effective results. It is because the term frequency of keyword in a document can be of

"NULL" and other thing that the whole query can be of shorthand (where users give small queries at search engine like: IPL most player, here the whole query can be regarded as only single keyword).

IV. CONCLUSION AND FUTURE WORK

In In this paper, for retrieving the more relevant documents of submitted query, the combination of two keywords technique are given more weight compare to single keyword weight. So the propose technique, the combination of the two keywords are list out the more relevant documents for submitted query. In the future, the framework of the combination of two keywords, that will help to raise the more and more relevant accurateness of results for submitted query on the search engine. The concepts of this technique can be used in the "Conversational searching process" for improving the search engine efficiency.

REFERENCE

[1] Jiawei Han, Micheline Kamber, "Data Mining: Concepts and Techniques", 2nd edition, Morgan Kaufmann Publishers

[2] Gerald J.Kowalski, Mark T.Maybury, "Information Stora ge and Retrieval Systems: Theory and Implementations", 2nd edition, Kluwer Academic Publishers.

[3] Stefan Buttcher, Charles L.A. Clarke, Gordon v.Cormack, "Information Retrieval: Implementing and Evaluating Search Engines", MIT Press, Cambridge, Mass, 2010.

[4] T.Andreasen, J.Nilsson, H.Thomsen, "Ontology-based quering", In Proceedings of Fourth International Conference on Flexible Query-Answering System, pp:15-26, Warsaw, Poland, Agosto 2000

[5] Fisnik Dalipi, Illia Ninka,"Semantic Information Retrieval from Heterogeneous environments", International Journal of Scientific & Engineering Research, ISSN: 2229-5518,volume 4,Issue 8, August-2013.

[6] Mark Sanderson, W.Bruce Croft," The History of Information Retrieval System" ,Proceedings of IEEE, vol.100,May 13th,2012.

[7] Kolikipogu Ramakrishna, B.Padmaja Rani, D.Subramanyam," Information Retrieval in Telugu Language Using Synset Relationships", 2013©IEEE, 978-1-4673-2818-0/13

[8] Yiyao Lu,Hai He,Hongkun Zhao,Weiyi Meng," Annotating Search Results from Web Databases", IEEE Transactions on Knowledge and Data Engineering, vol.25,March 2013

[9] Jin Young Kim, W.Bruce Croft,"A Field Relevance Model for Structured Document Retrieval", unpublished.

[10] Sandor Dominich," The Modern Algebra of Information Retrieval", Springer,pp:74-93.

[11] Gobinda G. Chowdhury," Natural Language Processing", unpublished

[12] DonnaHarman," RankingAlgorithms", http://orion.lcg.ufrj.br/Dr.Dobbs/books/book5/chap14.htm.

[13] Vannevar Bush Auther of the Article "As We May Think" published in Atlantic in july 1995.