Improvising Search Engine Efficiency by Prioritizing Query-String

Sugandha Dani
Lecturer, Department of MCA,
Priyadarshini College of Engineering,
Nagpur, India

Dr. C. S. Warnekar
Former Pricipal,
Cummins College of Engineering,
Karve nagar, Pune, India

ABSTRACT

With the recent burgeoning of websites and networks, retrieving the desired information from large warehouses has become a challenging task. This calls for efficient techniques to populate better search results, more relevant to the user's query. Normally, a system user invokes the search engines like Google, Excite or AltaVista to enter the query. However it is observed that often the given query string is vaguely interpreted by the search engine, there by failing to retrieve much appropriate or expected data. Also some times the semantics of query is overlooked by the search engine producing divergent results. To overcome the problem this paper suggests certain priorities via formalizing the underlying emphasis in the query string to improvise the search engine.

Keywords: Search Engine, Query string, Emphasis, font style, priority

1. INTRODUCTION

The rapid evolution of the cyber world has greatly facilitated the availability of data on web. Search Engines [1] retrieve the required information via internet by suitably answering a submitted Query. However it is observed that often the given query string is vaguely interpreted by the search engine, and so the SERPs (Search Engine Result Pages) are hardly populated by the desired information. Hence in order to increase the chance of getting the desired page in search list few new techniques have been recently introduced [3][4][5][6][7] Besides, as the semantics of query is generally overlooked by the Search Engine, typing a longer query-string instead of few key words may also produce irrelevant search-results, The prioritizing of query-string suggested in this paper tries to circumvent the problem.

The next section presents an overview of working of search engine, including various existing techniques used to populate a page in a search list. Section 3 discusses different techniques currently used for optimization of search engine. In section 4 we propose a way to prioritize the keywords in the entire query string thereby improvising the search engine efficiency. Avenues for future work are indicated towards the end.

2. STRUCTURE OF SEARCH ENGINE

Internet search engines are special sites on the Web that are designed to help people find information stored on other sites. There are differences in the ways various search engines work, but they all perform three basic tasks:

1. They search the internet by searching for specific words appeared in the query.
2. They keep an index of such words.
3. They allow the users to look for the information based on a single word or combination.

Search Engine is usually composed of crawler, indexer, searcher and inquiry, shown as Figure 1. For the great amount of data in web page, the particular rule of index has to be established to improve the search efficiency. The index is one of the core technologies of search engine, which directly influences the quality of result. So far, the most popular and effective index method is inverted file, i.e. the file with word splitter firstly forms the index data, and then these data will be inverted.

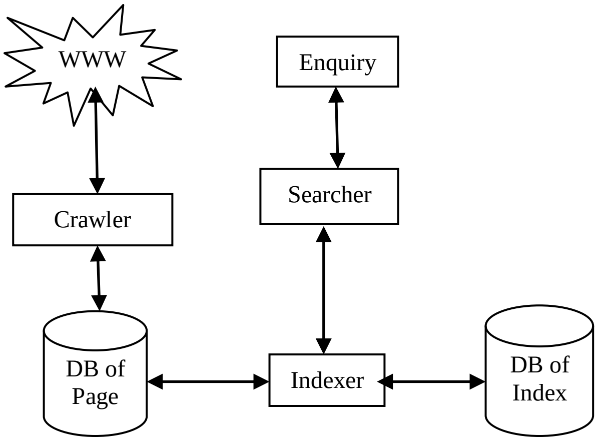

FIGURE 1: Structure of Search Engine

It has been observed that all search engines would not produce the same result for identical query. Even the order or sequence of the words presented to the same search-engine, affects the search-result. The question is why this happens? The possible reasons are:

- Some search engines index more web pages than others.
- Some search engines also index web pages more often than others.
- Different algorithms are used to compute relevance of the page to a particular query.
- Generally, populating large number of links happens to be the aim of Search Engine Optimization, instead of populating the list of pages which are more relevant.
- Typing entire query-string instead of few key words also produces irrelevant search-results.

- There is also a user-side-factor which influences the search result viz. the sequence or arrangement of words in a query. For example, a query typed in Google as “Wikipedia article on Pingalacharya’s Chhanda Shastra ” retrieves results which are not so relevant to the user. Adding quotation marks to the query in this example, results in “No match found”.

3. SEARCH ENGINE OPTIMIZATION (SEO)

Aiming at its retrieval capability of web pages, Search Engine Optimization or SEO technique demands the basic elements of a website should be constructed to fit the search engine retrieval principle. Accordingly, ten typical sorting parameters used in few conventional SEO techniques include:

Title Tags , Keyword Density, Site Structure , Internal Links,

URL structure , Domain, Indexing, etc.

Researchers in this field also make use of certain factors in order to rank the corresponding Search Engine on some point scale.[2] However, it is observed that the Search Engine Optimization itself has become an obstacle in this case as it produces false or unwanted result. The site owners can cheat SEO by creating link-farms containing hundreds and thousands of websites which increases its rating.

4. PRIORITIZING KEYWORDS

Various tactics like tactics [8] for keywords, for links, for domain name and hosts are used by site owners. Also various search engine optimization techniques are used to list more number of pages in the result thus the Search Engine Optimization itself has become an obstacle in this case as it produces false or unwanted result.

To avoid the above situation, we recommend use of emphasis on certain words in the query, the technique used by people while speaking natural language. For example, in a spoken language statement like “I never said that, she stole my purse” the emphasis on different words would alter the semantics as follows:

I never said that, she stole my purse. Emphasizing on I would mean it may be said by someone else but it was not said by me.

I never said that, she stole my purse. Emphasizing on never would mean I did not ever said.

I never said that, she stole my purse. Emphasizing on said would mean though I know the fact I have never explicitly said it. Similarly every other word emphasized would change the semantics of the statement accordingly.

Hence we suggest that priority-based Formal Emphasis may be introduced in a query-string by adding font style like bold, italics, underline etc. This would help to retrieve more relevant data, if not hitting the bull’s eye. Also sometimes the styles may be combined with compounded preference values in order to make the search even more pointed.

Suggested preferences are 1. Bold, 2. Italics, 3. Underline.

The query in the above example can then be augmented as “Wikipedia article on pingalacharya’s Chhanda Shastra” so that the bold type face ie Chhanda Shastra would be interpreted as word of first preference, pingalacharya’s as word of second preference and wikipedia as word of third preference. This is then passed on the web crawler in the same order, that is first word as “Chhand Shastra”, then ‘pingalacharya’s”, then “Wikipedia” and then “article on”, improvising the search result.

This procedure would first search the strongly emphasised word in query-string though it appears last and search specifically for that word in titles, headings, subheadings and meta data which will generate index list with more weightage on the specific word. The weightage will be calculated as per the preference level and accordingly the index will be generated thus improvising the search result.

When such a prioritized query is to be used we need to make following modifications to the search engine design.

1. Changing the simple text box to rich text box format so that typefaces can be applied.
2. Adding code for retrieving tokens from the query.
    a. The query will be sent using html tags like <b> </b> for bold, <i></i> for italics etc.
    b. These keywords will be arranged in a sequence of the preferences.
3. Adding code to generate layered index with preference list.
4. Searching for particular keyword considering the preferences given.

5. MODIFIED SEARCH PROCEDURE

The modified search-procedure is explained below:

5. The token list generated would contain the sequence for searching as Chhand Shastra first, then pingalacharya’s, then Wikipedia article.
6. This then can be passed on the web crawler in the same sequence that is first word as “Chhand Shastra” then “Pingalacharya” and lastly “Wikipedia article”
7. The search done with this preference list will populate the webpages which contain something about “Chhand Shastra” then it tries to serach “Pinglacharya” and so on.

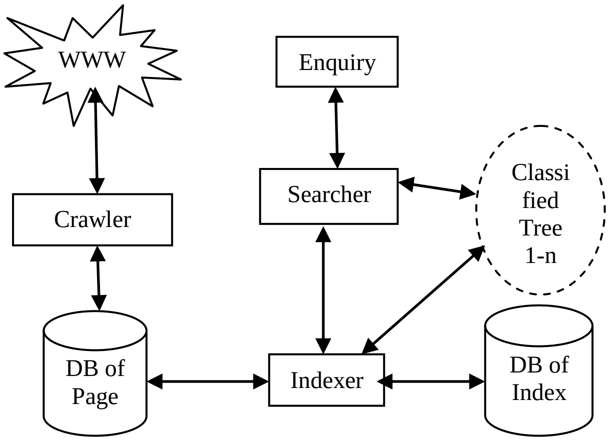

Figure 2: Structure of search engine with classified tree

Above figure shows search result after using special styles and the result is optimized. Between the crawler and indexer, one parallel layer is introduced, that is, classified tree, shown as elliptical area within the dotted line in Figure 2. In this layer, multiple classified trees form the classified forest are searched.

6. CONCLUSION AND AVENUES FOR FUTURE WORK

The paper suggests use of special styles in the query string while searching the information on the net. Adding such emphasis in the query-string generates a preference list of the words used in the query. This guides the search-engine towards more promising results. Obviously, the single-word querries are degenerate cases. The research can be futher enhanced by connecting the list of words to semantic net or a cluster of synonyms. This proposed enhancement is however, beyond the scope of present paper.

7. REFERENCES

[1] The Anatomy of a Large-Scale Hypertextual Web Search Engine By Sergey Brin and Lawrence Pagein

[2] Second International Workshop on Education Technology and Computer Science (ETCS), 2010, pages 673 - 675

[3] Shari Thurow. Search Engine Visibility[M] .(2nd Edition). 2005, Page 77-79.

[4] Ju Jiehui, "Chinese Search Engines' PageRank Algorithm and Implementation". Computer Engineering and Design, 2007,28 (7) Page 1632 - 1635.

[5] Shari Thurow. Search Engine Visibility[M] .(2nd Edition). 2005, 4 :77~79

[6] Richard John Jenkins. Search Engine Optimization[M] .lynda.com, Inc, 2006, :55~56 .

[7] Peter Kent. Search Engine Optimization For Dummies .2003, 2 :67~68 .

[8] International Conference on Computer Application and System Modeling (ICCASM), 2010, Pages V13-538 - V13-541