International Journal of Innovative Research in Engineering & Management(IJIREM)
ISSN: 2350-0557, Volume-1, Issue-2, September-2014

Review on Efficient Approach for Web Search Engine Using Page Level Keyword

Akanksha N. Upate
Computer Science Department,
S.G.B.A.U,
Amravati, India,

Surabhi G. Rathi
Computer Science Department,
S.G.B.A.U,
Amravati, India,

Ashvini D. Zade
Computer Science Department,
S.G.B.A.U,
Amravati, India,

Snehal S. Nimje
Computer Science Department,
S.G.B.A.U,
Amravati, India

ABSTRACT
Web has become integral part of our lives and search engines plays an important role in making user search the contents online using textual queries. A search engine provides services too. With amount of information available on web, it is important to verify whether search engines satisfy all the requirements of users by their search results. So it becomes necessary to evaluate search engines based on user's point of view. Evaluation of search engines is a method of determining how search engines meet the user's needed information. All Web search engine is a computer program that allows user to search and retrieve web documents with queries for their information needs. Page level keywords are the keywords occur in individual pages of websites. page level keyword measures paper relevancy of search engines results. So this can be a basis to provide more relevant Search results to the users. This paper proposes a model for dynamic construction of a resultant page from various results fetched by the search engine.

Keywords
Search Engine; Evaluation; Keyword count; Page Level keyword; Precision.

1. INTRODUCTION
WWW is a huge collection of information that includes text, audio, video etc. With huge increase in availability of information through WWW, it has become difficult to access desired information on Internet; therefore many users use Information retrieval tools like Search Engines to search desired information on the Internet. With a search engine, user types in the keywords for the information they wish to access and the search engine returns a list of clickable URL's that best match with the entered Keywords [2].

Page level keyword is an important factor to measure the relevancy of the search engine results [3]. During this research various queries are taken and a database has been created for them. Later the queries are run by the user(s) to calculate the page level keywords and the results are calculated. The average count and total count have to be calculated. A keyword can be any word on web page but stop words cannot be considered useful keywords such as a, an, the, of, from. For example, the word, "basically" in previous sentence is not a very useful keyword. Useful keywords would be search, search engine, search engine ranking etc [3].Page level Keyword are the keywords found on the individual pages of a web site such as in title, header tag and content. Page level keywords are those which are keywords for our internal pages [6].The usage of page level keywords includes keywords present in the title tag, keywords present in first word of title tag, keywords present in anchor text of an internal pointing link, keywords present in page H1 tag and ALT tags, keywords present in page's URL string [4]. Many evaluation strategies have been reported for search engine evaluation. Previous research has focused upon Human ranking based evaluation, automatic evaluation strategies, and Survey based evaluation strategies.

2. RELATED WORK
Many publications evaluate Web search engines (e.g. Notess, 2000). Perhaps the best known of these are Search Engine Watch

(http://www.searchenginewatch.com) and Search Engine Show down (http://www.searchengineshowdown.com). However, many of these publications did not employ formal evaluation procedures with rigorous methodology.

2.1. Human Relevance Judgments [8]

Not all search engine studies used human relevance judgments as the basis of evaluation, probably due to the difficulty and expense of such efforts. Courtois and Berry (1999) studied the first 100 items retrieved by five search engines. Instead, they used a computer program to automatically check the location, proximity, etc. of the search terms in the retrieved documents and used this information to compare the search engines in the study. When human relevance judgment was used, there was a variation in who makes the judgment [8].

2.2 Human Ranking Based Evaluation [8]

Vaughan et al.(2004) compared three search engines (Google, AltaVista, Teoma).The author measured quality of result ranking, ability to retrieve top ranked web pages and stability using queries and links from each search engine and graduate students to rate the links. The search engine, Google is best in quality, ability to retrieve top ranked pages and stability. Vaughan et al has a higher point of view for the evaluation strategy and argue discussed over the new evaluation methods [8].

2.3 Automatic Evaluation Strategies [4]

Jinbiao et al.(2009) suggested a simple, accurate, effective, automatic and safe system which is used to automatically evaluate search engines. It has four modules-sampler, crawler, refinery, evaluator. C# is used as a development tool in this work. Actions are implicit feedback interactions such as scrolling, saving, printing, bookmarking, adding to favorites and copying. At the end, Author gives score to search engine on the basis of coverage evaluation and sequence evaluation [4].

2.4 Survey Based Evaluation Strategies [4]

In this, features of five search engines are compared which are available to the user while searching the information. For this, they had taken a survey in which 263 participants participated and examined their interests in search engines. From this survey, they find out which search engine provides best utility and services to the user and most likely used by the people and they find out that users give highest rank to Google [4].

3. WEB SEARCH ENGINE [1]

When we do web search we are actually search in the web made up of over 60 trillion individual pages and its constantly growing. This is done with the software programming called as Spiders. Spiders start with fetching of few web pages and forward the link on those pages and fetch the pages they link to and so on. Algorithms get to work looking for clues to better understand what we mean. Based on these clues search engine pulls relevant documents from the index using amongst various factors. Finally it combines all these factors together to produce each page over the score and sent back to the search about half a second after we submit our search [1].

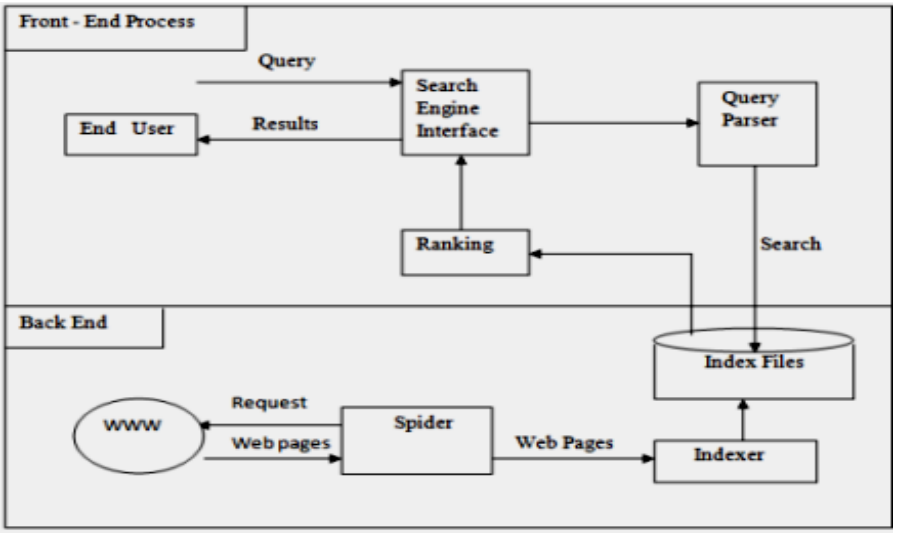

Fig 3.1 Search Engine Architecture [1]

3.1 Working of Search Engine: [1]

3.1.1 Query Processing [1]

Step 1: Tokenizing

As soon as a user inputs a query, the search engine — whether a keyword-based system or a full natural language processing (NLP) system — must tokenize the query stream, i.e., break it down into understandable segments. Usually a token is defined as an alpha-numeric string that occurs between white space and / or punctuation.

Step 2: Parsing.

Since users may employ special operators in their query, including Boolean, adjacency, or proximity operators, the system needs to parse the query first into query terms and operators. These operators may occur in the form of reserved punctuation (e.g., quotation marks) or reserved terms in specialized format (e.g., AND, OR). At this point, a search engine
may take the list of query terms and search them against inverted file.

Steps 3 and 4: Stop list and stemming.

Some search engines will go further and stop-list and stem the query. The stop list might also contain words like “a”, “the”, “of” etc in the input, are referred to as stop words. These are not useful keywords. For very short queries, search engine may drop these two steps.

Step 5: Creating the query.

How each particular search engine creates a query representation depends on how the system does its matching. If a statistically based matcher is used, then the query must match the statistical representations of the documents in the system. At this point, a search engine may take the query representation and perform the search against the inverted file.

Step 6: Query expansion

Since users of search engine includes only a single statement of their information needs in a query, there is possibility that the information, they need may be expressed using synonyms, rather than the exact query terms. Therefore, systems may expand the query into all possible synonymous terms and perhaps even broader and narrower terms.

Step 7: Query term weighing.

The final step in query processing involves computing weights for the terms in the query. This step indicates that either how much to weight each term or simply which term in the query matters most and must appear in each retrieved document to ensure relevance.

3.2 Search and Matching Function [1]

Searching the inverted file for documents meeting the query requirements, referred to simply as "matching," is typically a standard binary search, no matter whether

the search ends after the first two, five, or all seven steps of query processing. While the computational processing required for simple, non-Boolean query, matching is far simpler than when the model is an NLP-based query within a weighted, Boolean model, it also follows that the simpler the document representation, the query representation, and the matching algorithm.

4. MECHANISM [3]

This work attempts to evaluate and compare the performance of major search engines in an objective, fair and efficient manner with subject to general queries that may be posed by students in their everyday lives. In our proposed framework we evaluated three search engines on the basis of page level keywords using forty educational queries.

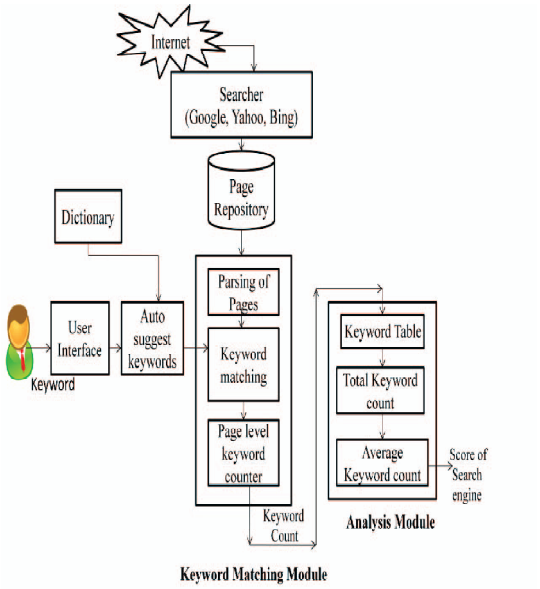

Fig.4.1 Framework for search engine evaluation based on page level keywords [3]

4.1 Selection of Search Engines [3]

The search engines chosen for evaluation are Google (www.Google.com), Yahoo (www.Yahoo.com) and Bing (www.Bing.com). Google is selected because it is the largest publicly available search engine and has the highest usage statistics of 80.65%. Bing and Yahoo are on second and third position respectively [3]

4.2 Selection of Query Set [3]

The proposed criterion for evaluation based on page level keywords identifies which search engine is providing better relevant results for educational queries.

Fig.4.1 shows the proposed framework for search engine evaluation. The framework consists of different modules each having its own function and significance.

The top ten pages retrieved by each search engine are saved in a page repository. These pages are parsed so as to match the input keyword with the words in the page. The page level keyword Counter is incremented for every hit. Page level keyword counter outputs keyword count. The total keyword count and Average keyword count is calculated. Average keyword count concludes the score of search engine.

4.2.1 Searcher: Searcher fetches top ten results for each query, executed on search engines Google, Yahoo and Bing and saves them in a page repository.

4.2.2 Page Repository: Page Repository is the database of1200 search results retrieved from Google, Yahoo and Bing for all the forty keywords.

4.2.3 Dictionary: Dictionary is a database of keywords taken for analysis. Dictionary is constructed by taking a sample of educational keywords from the set of most popular searched keywords on internet.

4.2.4 User Interface: User interface is the graphical user interface on which user performs search .User enters the input keyword in the text box of user interface.

4.2.5 Auto Suggest: Auto suggests keywords displays suggested keywords from dictionary in dropdown menu as user starts by typing a keyword into the text box. As letters are entered, relevant suggestions are provided. Auto suggest method can make the entering of keyword more efficient, time saving and reduce chances of spelling error.

4.2.6 Keyword Matching Module: Keyword Matching Module includes Parsing of pages, Keyword Matching and Page level Keyword Counter.

4.2.6.1 Parsing of Pages: It is a function that picks pages from page repository and parses the pages in order to determine the number of times the input keyword occurs in a particular web page.

4.2.6.2 Keyword Matching*:* Keyword matching is matching the input keyword with words in a particular web page.

4.2.6.3 Page level Keyword Counter*:* This counter is incremented for every hit, every time a particular web page contains the input keyword.

4.2.7 Analysis Module: Analysis Module includes Keyword Table, Total keyword Count, and Average Keyword count.

4.2.7.1 Keyword Table: Keyword Table displays the keyword count for each of the ten pages. The general structure of the table
is shown in Table I.

Table I Keyword Table [3]

Page 1	KWC1
Page 2	KWC2
Page 3	KWC3
Page 4	KWC4
Page 5	KWC5
Page 6	KWC6
Page 7	KWC7
Page 8	KWC8
Page 9	KWC9
Page 10	KWC10

Where KWC is keyword count

4.2.7.2 Total keyword Count (TKWC): From the keyword count table, we get the total keyword count. The summation of keyword counts for each page gives total keyword count. It means that total keyword count gives the number of times the keyword appears in all the ten pages for each search engine Google, Yahoo and Bing. The total keyword count is computed as:

TKWC = KWC1 + KWC2 + ... + KWC10 (1)

4.2.7.3 Average Keyword Count (AKWC): Average Keyword count is the arithmetic average of total keyword count of forty keywords. Average count concludes the relevance score of each search engine.

AKWC = (TKWC1 + TKWC2 + ...TKW(n)) / n (2)

4.2.7.4 Algorithm [3]

The algorithm describing the sequence of computation steps involved in present work is given below:-

Search evalu()
1. read a key from the user input.
2. match it from all the existing keyword list in the dictionary.
3. pick up the page from the page repository.
4. calculate the frequency of keyword in the page.
5. show it in keyword table.
6. repeat the step from 3 to 5
7. repeat the above steps from 3 to 6 for Google, Yahoo, and Bing repository.
8. calculate the keyword count in top ten pages and put it in Total keyword count table and find the average of total keyword count for all the queries.
9. final score depends on average of total keyword count obtained for all the queries taken into consideration to show which search engine has the most keywords based search results in their top ten pages.
10. show final score for each search engine.

5. CONCLUSION

The search engine evaluation based on page level keywords, clarifies the effectiveness of the search results. Hence page level keywords can be good criteria for search engine evaluation. In addition to the implementation of the basic methods of evaluation that is page level keywords based evaluation, an auto suggest field is also included to participate with a realistic and human friendly environment for the end user for efficient evaluation. So it is concluding that page level keyword measures relevancy of search engines results.

REFERENCES

[1] Ms. Nilima V. Pardakhe, Prof. R. Keole" Enhancement of Web Search Engine Results Using Keyword Frequency Based Ranking " IJCSMC May 2014

[2] Ashlesha Gupta, Ashutosh Dixit, A. K. Sharma ," Relevant Document Crawling with Usage Pattern and Domain Profile Based Page Ranking", IEEE 2013.

[3] S. Goel, “Search Engine Evaluation Based on Page Level Keywords”, 3rd IEEE International Advance Computing Conference (IACC), 2013.

[4] Shikha Goel Sunita Yadav " An Overview of Search Engine Evaluation Strategies"(IJAIS) ,February 2012

[5] http://www.monash.com/spidap4.html.

[6]http://forums.seochat.com/keywords-30/page-level-keywords-anddomain-level-keywords-437929.html.

[7]http://www.boxcarmarketing.com/blog/item/top-5-ranking-factors in- search/.

[8]http://www.seomoz.org/article/search-ranking-factors.