INFORMATION RETRIEVAL EVALUATION OF SEARCH ENGINES FOR BUSINESS AND RESEARCH

[1]Ogbu, V. I. and [2]Nwabueze, C. A.
[1]Department of Electrical/Electronic Engineering, Maritime Academy of Nigeria, Oron,
[2]Department of Electrical/Electronic Engineering, Chukwuemeka Odumegwu Ojukwu University, Uli, Anambra State.
Email: ogbuikechukwu2255@gmail.com, ca.nwabueze@coou.edu.ng.

ABSTRACT
Information retrieval is a vital aspect of online search. The rate of data retrieval is highly dependent on the search engine especially in the academia where online research is paramount. All search engines use web crawler which is sometimes called SPIDER. This is an internet boast that systematically browser the internet. It is also used for updating of websites. They work by copying all the pages visited for later processing by a search engine which indexes the downloaded pages for efficient search. This paper bring to the fore, various search engines, their mean capacity and proposes a model for fast data retrieval base on the position of ranking with discounted cumulative gain (DCG).

Keyword: Search Engine, Network, Information Retrieval, Crawler.

1.0 INTRODUCTION
In web search engines (SEs), a spider program fetches as many documents as possible. Another program (an indexer), then reads these documents and creates an index based on the words contained in each document. Each search engine uses a proprietary algorithm to create its indices such that, ideally, only meaningful results are returned for each query search engines. To view a Web page on the World Wide Web (www), the procedure begins either by typing the Universal Resource Locator (URL)of the page into a Web browser or by following a hyperlink to the page or resource. The Web browser then initiates a series of communication messages behind the scenes in order to fetch and display it. First, the server-name portion of the URL is resolved into an Internet Protocol (IP) address using the global distributed Internet database known as the domain name system or Domain Name Search (DNS). This IP browser then requests the resource by sending a Hypertext Transport Protocol (HTTP) request to the Web server at the particular address as depicted in figure 1 for Google query architecture [1].

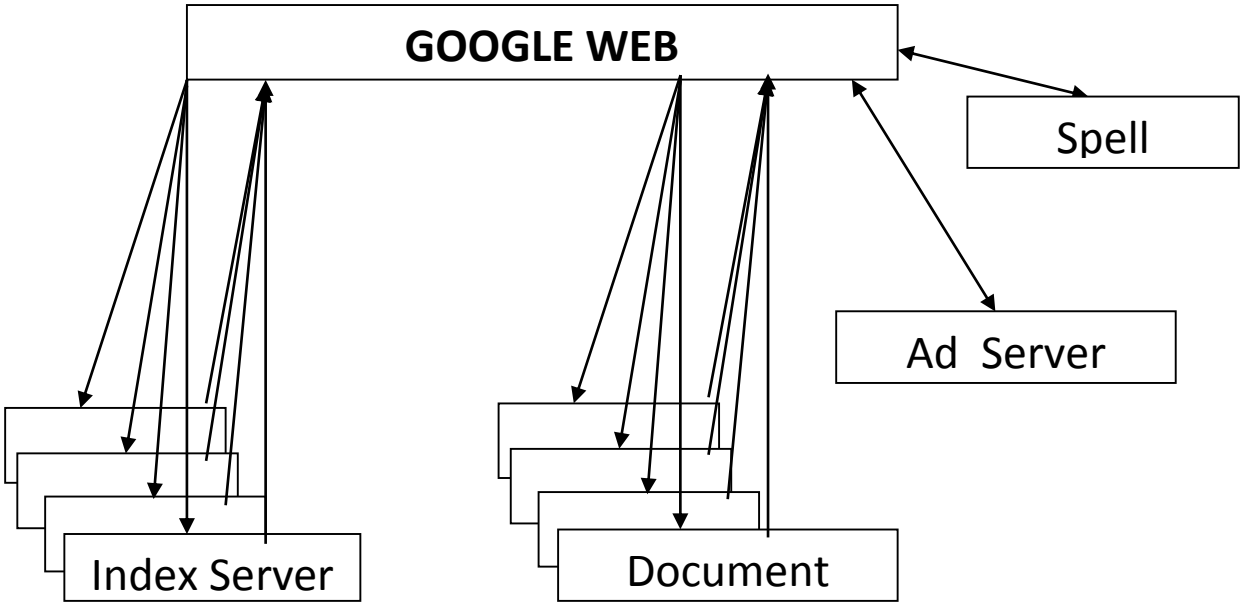

Figure 1: Google Query Searching Architecture [1].

This architecture arises from two basic insight facts: it provides reliability in software rather than in server-class hardware, secondly design is the best for aggregate request throughout, not peak server response as enumerated in this paper.

2.0 COMPONENT OF SEARCH ENGINE DEVELOPMENT

2.1 Component of a Search Engine Development Tool

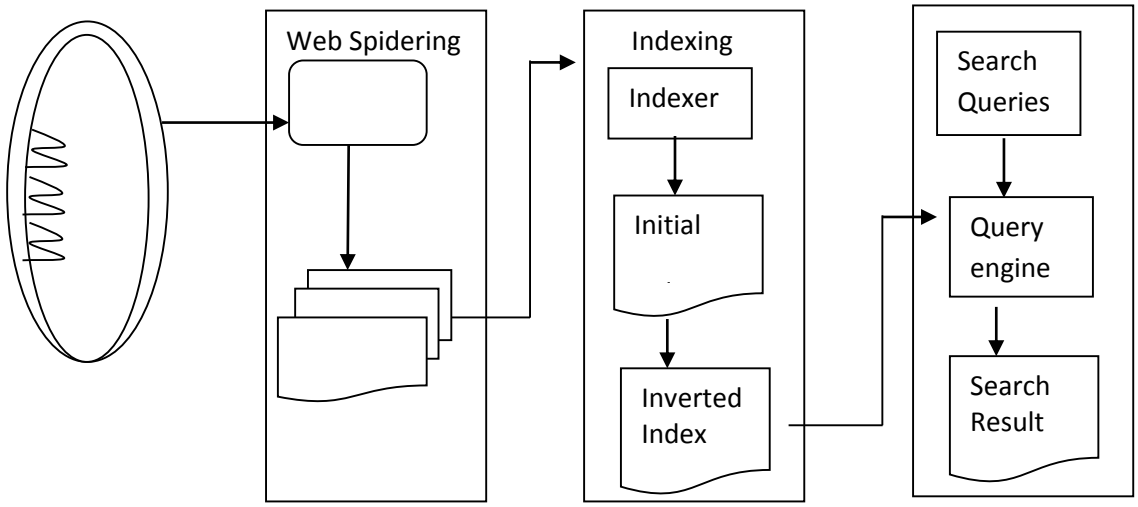

Figure 2: Architecture of Common Vertical Search Engine [2].

A common search engine development tool (as in figure 2) consist of a number of web spider indexer and a query engine, which helps the user in collecting, indexing and querying web pages respectively.

2.2 Information Retrieval Process

The major goal of an Information Retrieval (IR) system can be describe as the representation, storage, organization and access to information items. A description of the different resources, components and tasks involved in an information retrieval system is shown in figure 3.

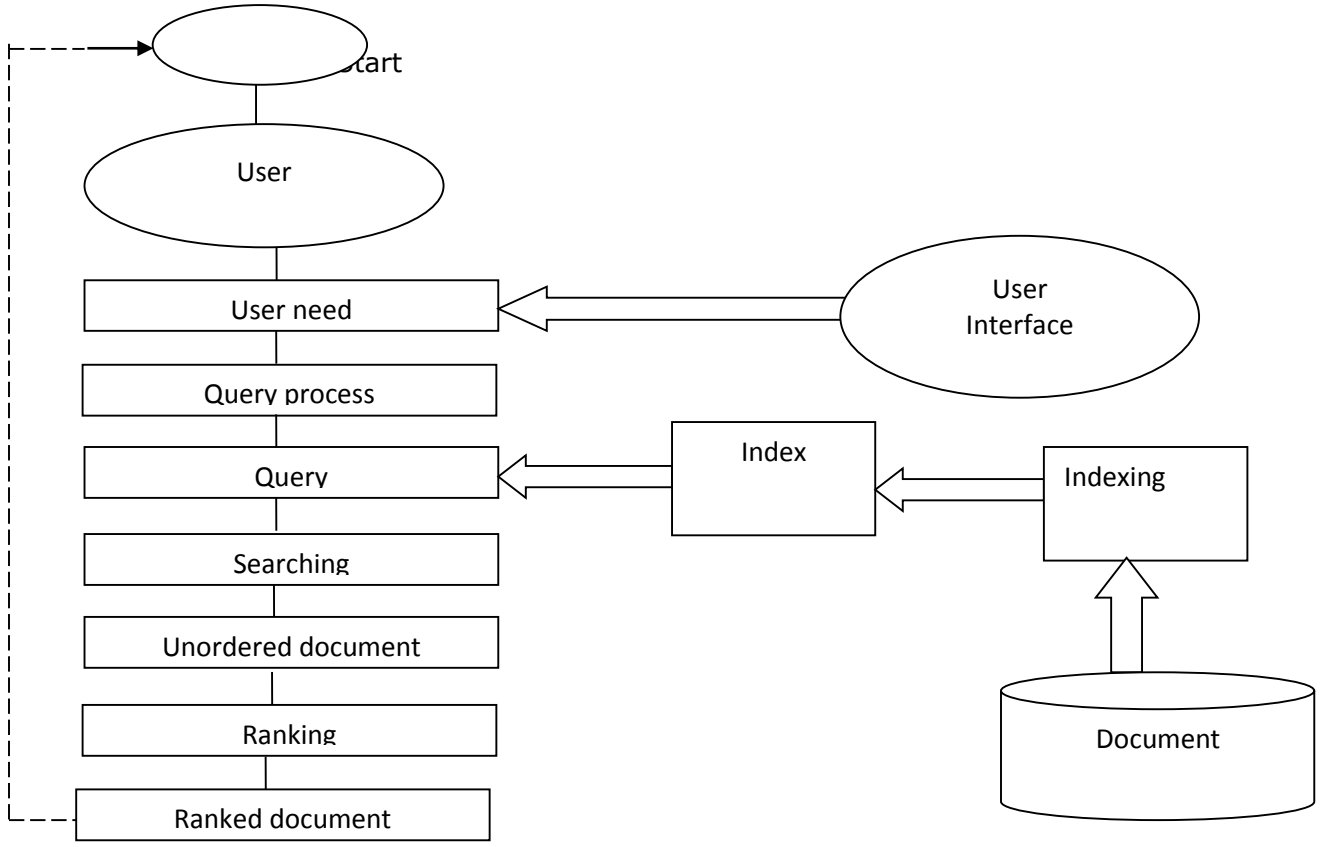

Figure 3: Information Retrieval Process

3.0 PROPOSED MODEL

For effectively retrieving relevant document by 112 strategies, the documents are typically transformed into a suitable representation. Each retrieval strategy incorporates a specific model for its document representation purpose [3, 4]. The models are classified into two dimensions:

3.1 Mathematical Basis

a) Set-theoretic model: this addresses document as set of words or phrases. The common model here are:

i. Standard Boolean Model
ii. Fuzzy Retrieved
iii. Extended Boolean Model

b) Algebraic Mode: this addresses queries as vectors, matrices or tuples. Vector is represented as a sealer value:

i. Extended Boolean model
ii. Topic based vector spaces model
iii. Vector spaces model.

c) Probabilistic Model: This addresses the process of document retrieval as probabilistic references. The model comprises of:

i. Binary Independence Model
ii. Probabilistic relevance Model
iii. Uncertain Interference
iv. Language Model
v. Divergence-from-randomness Model
vi. Latent Dirichlet Allocation.

3.2 Properties of the Model

a) Model without term: this is usually represented in vector space model by the orthogonality assumption of term vector on in probability Model by an independency assumption for term variables.
b) Model with transcendent term: this allows a representation of interdependencies between terms but they do not allege how the interdependency between two terms is defined.

The necessity of informative retrieval by the crawler is based on its PRECISION. It is analogous to positive predictive values.

Mean average precision(MAP) for a set of query is the mean of average precision scores for each query [5]:

$$\text{MAP}= \frac{\sum_{q=1}^{Q} \text{AVep(Q)}}{\text{Q}} \qquad (1)$$

where Q is the number of queries.

Base on the position of ranking, the discounted cumulative gain (DCG), uses a graded relevance scale of document from the result of the useful documents or gain. The DCG accumulated at a particular end position P is defined as:

$$\text{DCG} = \text{rel}_i + \sum_{1=2}^{P} \frac{\text{reli}}{\log \text{i}} \qquad (2)$$

4.0 DATA COLLECTION AND ANALYSIS

The evaluation of an information retrieval system is the process of assessing how well the system meets the needs of its users. Virtually all modern evaluation metrics e.g Mean, average, precision, discounted cumulative gain, are designed for ranking retrieval. Using keywords such as Hyper Pre-processor, Object Oriented Program, Engineering, Operating System, Linux, Traffic, Sequential Query Language, Software Paper, Thesis, Frequency, Micro-control, and

Mean can confirm and analyze the quantity, quality and the time of information retrieval. Table 1 shows the order of importance of the addresses [4].

Table 1: Addresses in Order of Importance of Search Engines

	QUERY SEARCH	Google	MSN	YAHOO	ALTA Vista	CLOUDSHARE SaaS
1	PHP	Query Qty 9240000	Query Qty 1583408	Query Qty 110000	Query Qty 224813	Query Qty 34
2	OOP	7270000	1739483	435115	7516787	11
3	JAVA	726000	207605	67500	125157	33
4	Engineering	11600001	6255462	10000000	3674348	29
5	OS	8820020	171494	105423	177320	34
6	LINUX	5370000	1305195	795112	1992876	31
7	DATABASE	1180120	243040	124133	266029	33
8	TRAFFIC	753000	113116	100000	161083	27
9	SQL	23500000	8123	1315000	48117366	38
10	SOFTWARE	5910000	364205	652000	525593	34
11	PAPERS	277000	74879	50000	58903	30
12	RESEARCH	22500000	3390131	1100000	6606411	7
13	THESIS	8230000	34783984	85000000	71309208	3
14	FREQUENCY	35900000	13588351	73000000	46	15
15	MICROCONTORL	3400000	776817	330000	625816	31
16	MEAN	23745076	4307019	10945619	9560338	28

Base on the order of importance, search engines are assessed by;

i. Quality of document retrieved by the engine
ii. Time taken for each keyword search to be completed by the engine.
iii. Per thousand precision of documents retrieved (by observing and recording the number of result that give us relevant information about the keyword used).

This reveals that Google search engine has the highest document retrieved efficiency in terms of total number of document retrieved for all the queries (key words)

Google (Mean = 23, 745,076)

Yahoo (Mean = 10,945,619)

Alta Vista (Mean=10,945,619)

MSN (mean =4,307,019)

CLOUSHARE (Mean=28)

This experiment was made possible via the ICT Research Laboratory of Chukwuemeka Odumegwu University, Anambra State. The data were analyzed using Microsoft Excel to examine the quality of the documents retrieved, time taken to retrieve the document.

Table 2: Experiment Showing Time Taken To Retrieve Information

Test case	Cloudse (sec)	ALTA Visita (secs)	Yahoo (secs)	MSN (Secs)	Google (Secs)
1	0.054	0.5	0.15	0.19	0.16
2	0.064	0.59	0.1	0.23	0.19
3	0.054	0.5	0.4	0.19	0.16
4	0.061	0.56	0.75	0.22	0.18
5	0.057	0.5	0.2	0.2	0.17
6	0.051	0.53	0.6	0.18	0.15
7	0.061	0.47	0.15	0.22	0.18
8	0.054	0.56	0.65	0.19	0.16
9	0.054	0.5	0.15	0.19	0.16
10	0.051	0.5	0.7	0.18	0.15
11	0.064	0.47	0.8	0.23	0.19
12	0.047	0.59	0.6	0.17	0.14
13	0.047	0.44	0.15	0.17	0.14
14	0.057	0.44	0.3	0.2	0.17
15	0.054	0.53	0.4	0.19	0.16
16	0.055	0.51	0.41	0.2	0.16

CONCLUSION

This work was intended to eliminate the problems encountered by researchers in using appropriate search engines for efficient information retrieved. It was observed that Google has large data composition with average retrieval time of 0.16 secs, which is good but adverts lows the engine. It is therefore good for businesses. Yahoo is second best with average retrieval time of 0.3 secs and is very good for advertisements. ALTA Vista is good for advertisement and information retrieved, while MSN is the least attractive. It is therefore advisable to use intelligent search engines with queuing-RSVP algorithm in SAN architecture (Intelligent cluster algorithm search engine), since they do not have advertisements and are tailored for high precision of information retrieval.

REFERENCES

1. Luiz, G. O. et al (2013), *Web Search For A Planet Google Cluster*, IEEE Computer Society, 2003 IEEE Conference, pp. 22-28.
2. Brin, B. N. et al (1998), *The Anatomy of a Large-Scale Hyper Textual Web Search Engine*, Proceeding of the Seventh International Conference on World Wide Web, Brisbane, Australia, pp. 234 – 241.
3. Meaning, C. D., Raghaveh, P. B. and Suhutze, H. (2009), *Evaluation in Information*, Introduction to Information Retrieval, Chapter 8, pp. 39 – 72.
4. Jaeyong, B. C. et al (2011), *A Cloud Portal With A Cloud Service Search Engine*, Proceedings of International Conference on Information and Intelligent Computing (IPCSIT), Vol. 18, pp. 87 – 93.

5. Miriam, J. Q. et al (2009), *Semantically Enhanced Information: An Ontology-Based Approach*, M.Sc Dissertation, University of Negred, Madrid.