Introduction to Information Retrieval Systems

Ms. Rashmi Janbandhu
Lecturer, Information Technology
Rajiv Gandhi College of Engineering & Research
Nagpur, India
rashmi.janbandhu@gmail.com

Mr.Viplove Karhade
Scholar, Computer Science and Engineering
North Eastern University, Boston
viplovekarhade@gmail.com

***Abstract*—**Database is a collection of Information. This information can be of various domains. To retrieve the information needed at the instinct requires a proper mechanism, this mechanism is known as Information Retrieval System. The paper considers the World Wide Web as its Database and discusses the Information Retrieval System in the same context.

Keywords*-Information, retrieval system, search engine, crawler,logic,relevancy.*

*****

I. INTRODUCTION

The Information Retrieval System basically deals with information and access to the information as and when needed. Information Retrieval System not only focuses on retrieving data quickly but also focuses on storing it in proper structure so as to support its quick retrieval.

Information Retrieval System in context with World Wide Web urges to introduce Search Engines. Search Engines are an interface to database. These engines ask user to enter the words basically known as keywords related to his search. These keywords are treated as query to database. There are many search engines currently working on net:

www.google.com
www.search.msn.com
www.search.yahoo.com
www.hotbot.com

All the search engines use different mechanism to retrieve data. Some of the search engines combine two or more search engines to retrieve accurate data according to user query.

II. SEARCH ENGINES

Early times the World Wide Web was very small in its size, the first search engine was just a simple program of retrieving information from database using query words. This program was nothing but the first search engine known as World Web Wanderer.

As the decades elapse the World Wide Web i.e the database went on increasing in its size, so was the complexity getting increased for the search engines. The commercial importance of web was getting into spotlight which encroached many developers to develop the search engines with new and accurate mechanisms.

Around in the year 2001, google came out as a well-known search engine. Many search engines were developed parallel to google but could not stand by with it. Each search engine keeps its mechanism secrete and copyrighted as so is of google and all the others.

The other search engines jus in low completion with google are Yahoo and MSN.

A. *Features*

The Search engines can limit their search on various parameter. The search can be limited on format of information. The various formats of information are:

- .pdf
- .doc
- .jpeg etc

The another feature that search engines takes care of is accuracy. It tries to limit its search in such a way that the results are genuinely accurate according to the users query. For the above purpose the search engines may use various methods or mechanism such as neural networks, semantic databases, lexical databases etc. The use of above mentioned techniques help to find relatedness amongst the query word which helps in increasing the accuracy.

B. *Architecture*

The diagram shows the architecture of simple search engine:

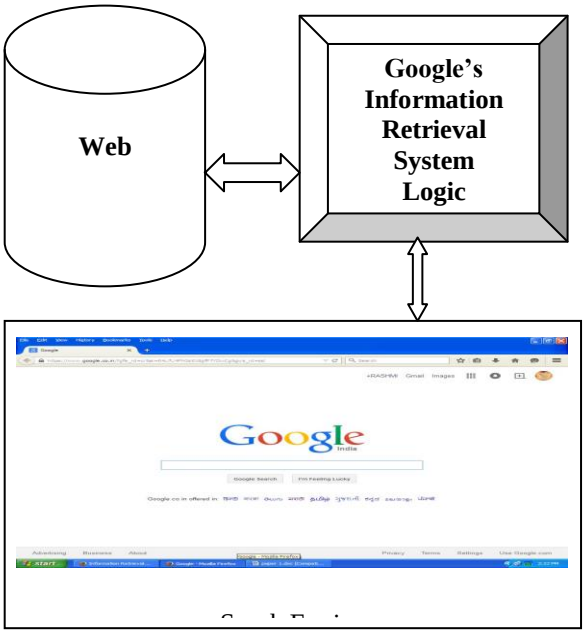

Figure 1. Architecture:Search Engine

The architecture shows that the Information System retrieval logic works in between Web i.e database and Search Engine. It takes the input from the search engine interface which is entered by the user. This input is nothing but the keywords for search. Logic processes the keyword on Web i.e Database and retrieves the data relevant to keywords as output back on the interface of search engine.

III. WEB CRAWLING

A. *Introduction*

Crawlers are the logic which works for months or years together to segregate data into focused domains. The crawlers are programs which work on World Wide Web to indentify each documents domain and place the document in that particular domain for easy retrieval. As and when the keywords are processed they are interrogated for their domain. Once the logic identifies the domain the search is narrowed or is directed only to that particular domain for further searches.

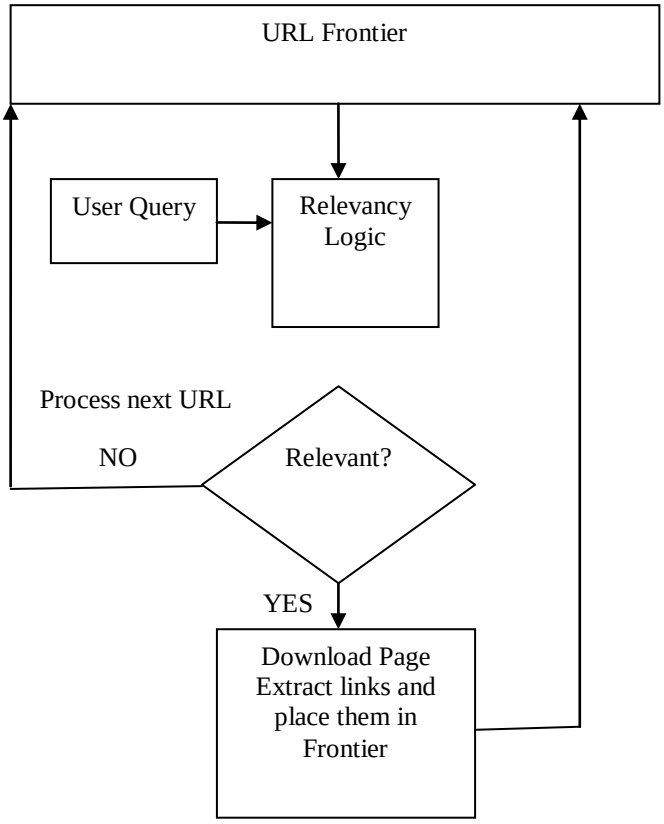

Figure 2. Architecture:Web crawler

The crawler has a queue known as URL frontier which is initialized using Seed URLs. The process starts with the first URL. The relevancy logic takes into account the user query and computes the relevancy cost. This cost is interrogated against a threshold value. If the relevancy cost is above threshold then the URL is considered to be relevant for the search and the links on that URL's page are placed in the queue so as to get processed to find more relevant documents. If the relevancy cost id below threshold then the next URL in the queue is processed. The threshold value is mostly computed on the basis of rigorous training and testing of the crawler. The process completes when all the URL in the queue are processed. The time taken by this process can be months or years.

B. *Robots*

Each URL has a file associated with it known as robot.txt. This file basically contains the information related to access to the web document. It contains the information which is read by web crawlers that the document can be parsed or not. Is the robot.txt disallows the crawler then the crawler will not process the document for search.

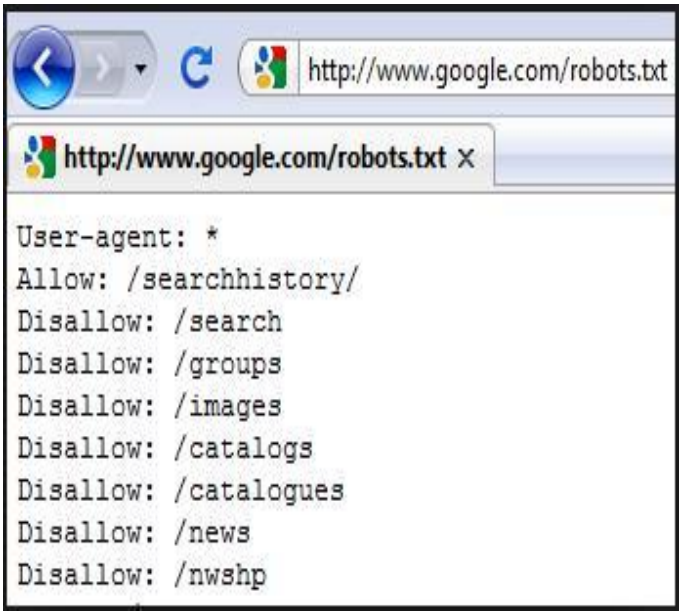

Figure 3. Robot.txt

The "User-agent: *" means this robot.txt is applicable to all crawlers. "Disallow: / " specifies which sites in the web document are not allowed to be parsed. Figure 3 shows a robot.txt of www.google.com. The robots of any site can be accessed by simply appending "/robot.txt" to the website URL.

C. *Multithreaded crawling*

The delay in segregation degrades the results value, so in order to compensate with this problem the crawlers were coded with multiple threads. Each thread worked as a individual crawler, but would be in co-ordance with other threaded crawler. Different threads could request different host for different web document at the same time reducing delay.

D. *Focused crawling*

In many situations the user are very focused to a particular topic or domain. So for such people the focused crawling came into existence. The search remains focused to the domain of the user. These kind of crawlers are organization oriented, where the organization is focused on one particular topic.

E. *Up-to-date crawling*

The Web changes each second so the crawlers must revisit the page for changes and value the changes accordingly which would help the user ultimately.

IV. INFORMATION RETRIEVAL SYSTEM

The complete Information Retrieval System consists of user interface, management module managing storage and retrieval, logic for cost calculation and ranking module.

The interface module is used to communicate with the user and outside world. The basic aim is to take input and show output to the user. The input taken is query which is used as keywords to direct the search.

The management module looks after the proper schema of database in which it is to be stored for easy and quick retrieval. The main aim of managing the data is quick access to information whenever needed.

Finally, the ranking module arranges the results in descending priority according to cost of relevancy calculated with the help of logic.

The ranking module provides its result to interface again which is shown to the user as output.

A. *Processing Document*

Processing is done before the collection of the document. The processing before storage are:

- Stop-Word Removal
- Stemming

The stop words are those words which are frequently appearing in the document but are of least importance. These words hamper the accuracy of the search and hence are liable to be removed. The standard list of such words is available with many reputed university research sites.

The stemming of the word helps in avoiding ambiguity for example if we are performing stemming of *goes*, *going* it would be stemmed as *go*, even *went* will be stemmed as *go*.

B. *Various model*

This section deals with various models of Information Retrieval System

- Boolean Model:
  This model deals with Boolean logic. The query is considered as in the following example: Red, yellow flowers found in Mahabaleswar or Pachmari- will have to be written as [[red&yellow]&[Mahabaleshwar|Panchmari]&flower. It is very hard to restrict the search in this model.

- Vector Model:
  The model is creates vector using the keywords of the query and document after the processing i.e stop word removal and stemming. The method employs similarity function over query vector and document vector.

- Semantic Model:
  This model is an expensive model. The model makes use of semantic databases in its logic to compute the relevancy cost.

- Probabilistic Model
  The model has ideal output ready which is perfectly the answer to user query. The query should be a description of the already made output. The effort taken is to match the query to the ideal output which is ready.

C. *Precision and Recall*

- *Precision* (P) measures the ability to retrieve which are mostly relevant.
- *Recall*(R)measures the ability of the search to find all of the relevant items in the corpus documents.

$$P = \frac{Number\ of\ relevant\ documents\ \ retrieved}{Total\ number\ of\ documents\ \ retrieved}$$

$$R = \frac{Number\ of\ relevant\ documents\ retrieved}{Total\ number\ of\ relevant\ documents}$$

D. *Retrieved Vs relevant Documents*

The figure below shows the relation between retrieved document and relevant document:

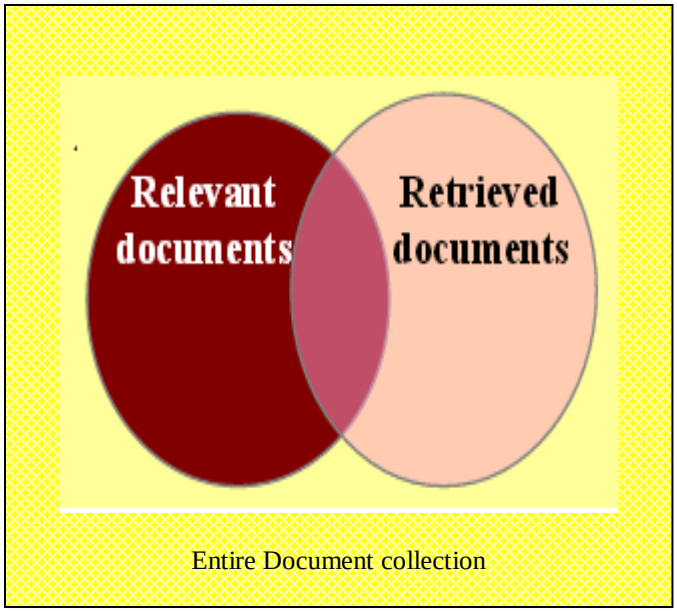

Figure 4. Relevant vs Retrieved Document

V. CONCLUSION

The paper provides a brief introduction on what Information Retrieval System consist of. It also addresses some modules of Web. It gives a brief idea of the database and processing of documents. The paper addresses about programs such as search engine and crawler which are important programs in context of Information Retrieval System on web.

ACKNOWLEDGMENT

The paper contains all the knowledgeable content best known to the authors.

REFERENCES

[1] Boca Raton, "Understanding Information Retrieval System", 2012.

[2] Introduction To Information Storage And Retrieval Systems by W. B. Frakes

[3] http://nordbotten.com/ADM/ADM_book/Ch1_InfoRetrieval.htm .

[4] Miller, G. A. "WordNet: A Lexical Database for English," Communications of the ACM (Vol. 38, No. 11), 1995, pp. 39-41.

[5] Liu, H. & Singh, P. (2004) ConceptNet: A Practical Commonsense Reasoning Toolkit. BT Technology Journal, To Appear. Volume 22, forthcoming issue. Kluwer Academic. Publishers.

[6] Fabian M. Suchanek, Gjergji Kasneci, and Gerhard Weikum. 2007. Yago: a core of semantic knowledge. In Proceedings of the 16th international conference on World Wide Web (WWW '07).

[7] Hull, D. A. and Grefenstette, G. "A Detailed Analysis of English Stemming Algorithms", Technical report, Xerox Research and Technology, 1996.

[8] Ho, T. K. "Stop Word Location and Identification for Adaptive Text Recognition, Int'l," J. of Document Analysis and Recognition (:3), 2000, pp. 16—26.