A Modern Information Retrieval at a Glance

[1]Narendra. M. Jathe, [2]Pratiksha S. Kalmegh, [3]Hemant Mahalle
[1]Assistant Professor, [2] PG Student, [3]Associate Professor
Department of Computer Science, Arts Commerce and Science College Kiran Nagar Amravati (India)

***Abstract* - Most IR systems compute a numeric score on how well each object in the database matches the query, and rank the objects according to this value. Our proposed research work based upon the concept of web data mining. This is the core domain of proposed work. The Information retrieval (IR) is the activity of obtaining information resources relevant to an information need from a collection of information resources. Searches can be based on non-full-text. Information Aggregation is a service that gathers relevant information from multiple sources to provide convenience and add value by analyzing the aggregated information for specific objective using Internet Technologies. We call the providers of these services aggregators. In proposed work try to implement a new retrieval methods. In which work on the problem for comparative analysis of two or multiple websites on same desktop screen. That facility cannot provide by the existing search engines. We have try to create framework for open number of websites at a time for easy access of contents readings from that multiple websites.**

***Index Terms* – Information Retrieval, Information Aggregation, Ranking.**

I. Introduction

Information retrieval is the hot topic for researchers. In this field day to day updating the things. Web retrieval also an interested area for data scientist or Researchers. The proposed research work based on the problem for comparative analysis of two or multiple websites. To accessing the multiple web pages at a same time or in same machine this is the critical task for user. If web user need to compare the Multiple web pages which relevant to each other than user need to find it separately and perform analysis on that site. If user use multiple computers for that purpose then that thing will be costly and time consuming and require multiple users for performing such types of task. So to solve this problem we were trying to create framework for opening the multiple WebPages at a same screen. To perform such types of research study the multiple web embedded\ related techniques. This is the basic work to implement the proposed algorithm. In following some points we will need to explain some core terms and previous work related to our research.

Information Retrieval

Information retrieval (IR) is the activity of obtaining information resources relevant to an information need from a collection of information resources. Searches can be based on full-text or other content-based indexing. Automated information retrieval systems are used to reduce what has been called "information overload". Many universities and public libraries use IR systems to provide access to books, journals and other documents. Web search engines are the most visible IR applications.

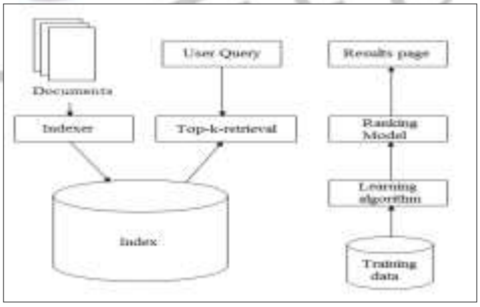


Figure 1 Information Retrieval

An information retrieval process begins when a user enters a query into the system. Queries are formal statements of information needs, for example search strings in web search engines. In information retrieval a query does not uniquely identify a single object in the collection. Instead, several objects may match the query, perhaps with different degrees of relevancy. IR refers to the method of extracting the information resources in a pre-defined automated manner, from an available lot of information resources. The search operations can be formulated on the basis of the metadata/full text or other indexing techniques. The main aim of IRS is to obtain relevant information by comparing the query with the associated and available documents [2].

Data Aggregation

Data aggregation is any process in which information is gathered and expressed in a summary form, for purposes such as statistical analysis. A common aggregation purpose is to get more information about particular groups based on specific variables such as age, profession, or income. The information about such groups can then be used for Web site personalization to choose content and advertising likely to appeal to an individual belonging to one or more groups for which data has been collected. For example, a site that sells music CDs might advertise certain CDs based on the age of the user and the data aggregate for their age group. Online analytic processing (OLAP) is a simple type of data aggregation in which the marketer uses an online reporting mechanism to process the information [3].

Web Base Information Retrieval

The web is a huge and fast growing collection of publicly available information over the internet. Searching the web for relevant information is becoming difficult as the new publican private networks, hardware devices and information formats being available at internet and the volume of data is growing every day. Deep web, privacy issues during crawling, identification and indexing of dynamic web pages are major concerns that the web search engines are considering as immediate challenges. Another dimension of the design challenges for a search engine is the interaction with novice users. Novice users are not trained for formal search engine query interface, unlike experts; they are unable to formulate exact information needs in the formal query language of search engine [3].

A reason of the uncertainty in a user query is the fact that human thinks and communicates in vague language. Natural languages queries regarding text, image and other information needs are infinitely capable to represent human thought process, uncertainty, incompleteness and imprecision, increasing the complexity of IR problem manifolds. The needs to handle the uncertainty in representation have resulted in approximated queries using weighted terms and fuzzy linguistic terms. Different types of document representation/indexing and explicit thesauruses are used to deal with problem of approximate search and fuzzy matching algorithms applied in the history of IR [3].

Web Crawling

When most people talk about Internet search engines, they really mean World Wide Web search engines. Before the Web became the most visible part of the Internet, there were already search engines in place to help people find information on the Net. Programs with names like "gopher" and "Archie" kept indexes of files stored on servers connected to the Internet, and dramatically reduced the amount of time required to find programs and documents. In the late 1980, getting serious value from the Internet meant knowing how to use gopher, Archie, Veronica and the rest. Today, most Internet users limit their searches to the Web, so we'll limit this article to search engines that focus on the contents of Web pages [6].

II. Literature Survey

Sr. No.	Author	Year	Proposals/Findings
1.	Khari, M., Jain, A., Vij, S., & Kumar, M.	2016	The author have present about the analysis of different information retrieval models [1].
2.	, M. W., & Ansari, M. A.	2012	The researcher is find out the survey in information retrieval [3].
3.	Suguna,S., Sundaravadivelu, V., & Gomathi, B.	2016	The research about the During e-learning the users can easily share, reuse, and organize the knowledge. Using the search engine the e-learners search the web pages by set of keywords [2].
4.	Ahmad, W., & Ali, R.	2016	The survey is about different types of social networking services and user's information shared on these services [8].
5.	Chandni Saini, Vinay Arora	2016	The survey is about strategies of information retrieval in web crawling has been presented that are classifying into four categories viz: focused, distributed, incremental and hidden web crawlers [9].
6.	Liang Yunjuan , Ma Lijuan	2011	The paper gave research and analysis on the inverted index technology and adopted the improved TFIDF weighting formula in order to improve the accuracy of retrieval [10].
7.	Chun Yi Liu, Chuan Yi Tang, D. Frank Hsu	2012	In this paper, they investigate these issues using five TREC datasets, TREC 2-6 (1993-97) [].
8.	Astha Tamrakar, Santosh K. Vishwakarma	2015	This work is about carried out to analyze and evaluate the retrieval effectiveness of various probabilistic models with use of new data set i.e; FIRE 2011 [12].

9.	Sanjib Kumar Sahu, D. P. Mahapatra, R. C. Balabantaray	2016	To gather the significant information from such a vast available resource Information Retrieval (IR); a method of retrieving such information resources which are relevant to an information need is applied [13].
10.	Pratibha Sharma, Brahmdutt Bohra, Surendra Yadav	2016	In this research, they will talk about the characterization of web information and its order and taking about the records (logs) maintained by the server [14].
11.	Martin Mehlitz, Christian Bauckhage	2007	In this paper, they analyze these flaws and introduce a new evaluation measure to overcome them. Based on a simple yet rigorous mathematical analysis of the effect of certain parameters in cluster based retrieval, we show that certain conclusions drawn in the recent literature must be taken with a grain of salt [15].
12.	D Evangelin Geetha G Krishna Naidu T V Suresh Kumar K Rajani Kanth	2008	In this paper, they describe procedure to calculate the response time early in the life cycle and simulate the performance metrics for an information retrieval system in different architectures namely, centralized and distributed [16].
13.	Massimo Melucci	2016	In this paper, they address the problem of query sample selection bias in machine learning terms and study experimentally how retrieval system rankings are affected by it [17].
14.	Mark Sanderson and W. Bruce Croft	2012	This paper describes a brief history of the research and development of information retrieval systems [18].
15.	Dr Kehinde K. Agbele, Eniafe F. Ayetiran	2016	The goal of this article is to develop algorithms that optimize the ranking of documents retrieved from IRS according to user search context [19].
16.	Sanjeev Patel, Kriti Khanna, Vishnu Sharma	2016	In this paper, they have evaluated different approaches for documents ranking [20].
17.	Zhang Shuxiang, Sri Devi Ravana	2016	The purpose of this study is to suggest a technique to estimate the reliability of the retrieval system effectiveness rank in a list of ranked systems based on its performance in previous experiments [21].

II. *Objectives*

i. Open multiple (Three in given case) websites.
ii. Open relevant Websites on the basis of Keyword.
iii. Maintain the rank of site (Accession) according to Google Rank.
iv. Create Simple framework to evaluate above mention objective.

III. *Proposed Methodology*

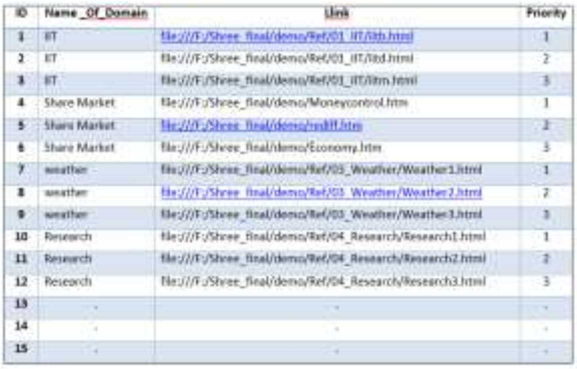
Step 1: Take a query from a user in the form of keyword.

An information retrieval process begins when a user enters a query into the system. A Query is a request for information from database. It is a precise request for information retrieval with database and information systems. Information Retrieval and web search domain revolves around search and retrieve Methodologies. Keyword search has been most popular and easy to used technique. Using this keyword take query from user. This query is in the form of keyword.

Step 2: Tally the respective keywords with our standard database.

In this phase tally the user query with standard database. This dataset contains the number of keywords which is the relevant words in our daily life. IIT, shear Market, weather, research.. etc. such a types of keyword will be frequently used for comparison in both sites for example qualification to professors in IIT Bombay and IIT Delhi, share market rate in multiple site (Difference occurs because of updating of site).

Following figure shows the keywords and their site link in tabular form. Which is created by using manual crawling and find the ranking on google trends. According to ranking created the database.



ID	Name _Of_Domain	Link	Priority
1	IIT	file:///F:/Shree_final/demo/Ref/01_IIT/iitb.html	1
2	IIT	file:///F:/Shree_final/demo/Ref/01_IIT/iitd.html	2
3	IIT	file:///F:/Shree_final/demo/Ref/01_IIT/iitm.html	3
4	Share Market	file:///F:/Shree_final/demo/Moneycontrol.htm	1
5	Share Market	file:///F:/Shree_final/demo/rediff.htm	2
6	Share Market	file:///F:/Shree_final/demo/Economy.htm	3
7	weather	file:///F:/Shree_final/demo/Ref/03_Weather/Weather1.html	1
8	weather	file:///F:/Shree_final/demo/Ref/03_Weather/Weather2.html	2
9	weather	file:///F:/Shree_final/demo/Ref/03_Weather/Weather3.html	3
10	Research	file:///F:/Shree_final/demo/Ref/04_Research/Research1.html	1
11	Research	file:///F:/Shree_final/demo/Ref/04_Research/Research2.html	2
12	Research	file:///F:/Shree_final/demo/Ref/04_Research/Research3.html	3
13	-	-	-
14	-	-	-
15	-	-	-

Figure 2:- Tally the respective keywords with our standard database

Step 3: Find the ranking on the basis of Google Ranking.

We find the ranking of websites using Google Trends. Google ranking is based on the websites accessed in one minutes second day etc.. parameters. When user search on the Google and website is search many more time than this website is in Google ranking these website rank is one and this all websites is related to one keyword. Then this websites is in ranking.

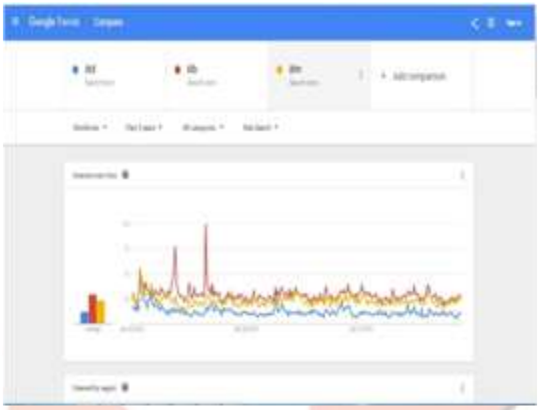
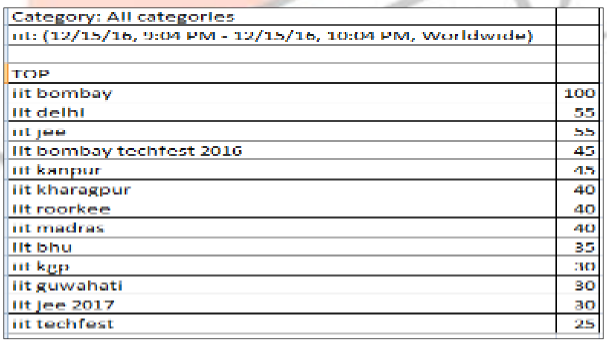


Figure 4:- Google Trends



Category: All categories	
iit: (12/15/16, 9:04 PM - 12/15/16, 10:04 PM, Worldwide)	
TOP	
iit bombay	100
iit delhi	55
iit jee	55
iit bombay techfest 2016	45
iit kanpur	45
iit kharagpur	40
iit roorkee	40
iit madras	40
iit bhu	35
iit kgp	30
iit guwahati	30
iit jee 2017	30
iit techfest	25

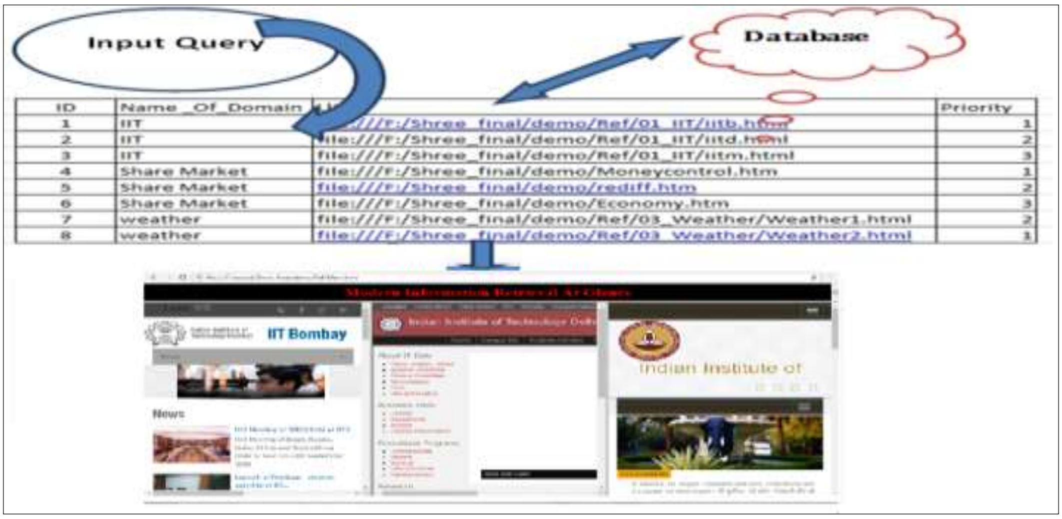
Figure 5:- Google Ranking

Step 4: Check Index.

Arranging the data in the form of sequence is called indexing. Our created database is useful for index. The database data is in the form of keyword. And this keyword is in index. In information retrieval to accessed data from web you need to use keyword. The index is the collection of keyword.

Step 5: Display three Web Pages at a time. For Comparison/Comparative Studies.

Comparative analysis of two or multiple websites that facility cannot provide by the conventional search engine. We have try to create framework for open number of websites at a time for easy access of contents readings from that multiple websites. To create this framework. In this framework three frames are open. Then search the similar type of website regarding one keyword. Then this three websites URL is linked with each other. Then this linked website is link with our keyword. Using this linking when we click on the keyword the three websites are open at a time. The three websites are open for comparative study. User can easily compare each and every contents of this three websites.



ID	Name _Of_Domain	U	Priority
1	IIT	file:///F:/Shree_final/demo/Ref/01_IIT/iitb.html	1
2	IIT	file:///F:/Shree_final/demo/Ref/01_IIT/iitd.html	2
3	IIT	file:///F:/Shree_final/demo/Ref/01_IIT/iitm.html	3
4	Share Market	file:///F:/Shree_final/demo/Moneycontrol.htm	1
5	Share Market	file:///F:/Shree_final/demo/rediff.htm	2
6	Share Market	file:///F:/Shree_final/demo/Economy.htm	3
7	weather	file:///F:/Shree_final/demo/Ref/03_Weather/Weather1.html	2
8	weather	file:///F:/Shree_final/demo/Ref/03_Weather/Weather2.html	1

Figure 5:- Proposed Methodology

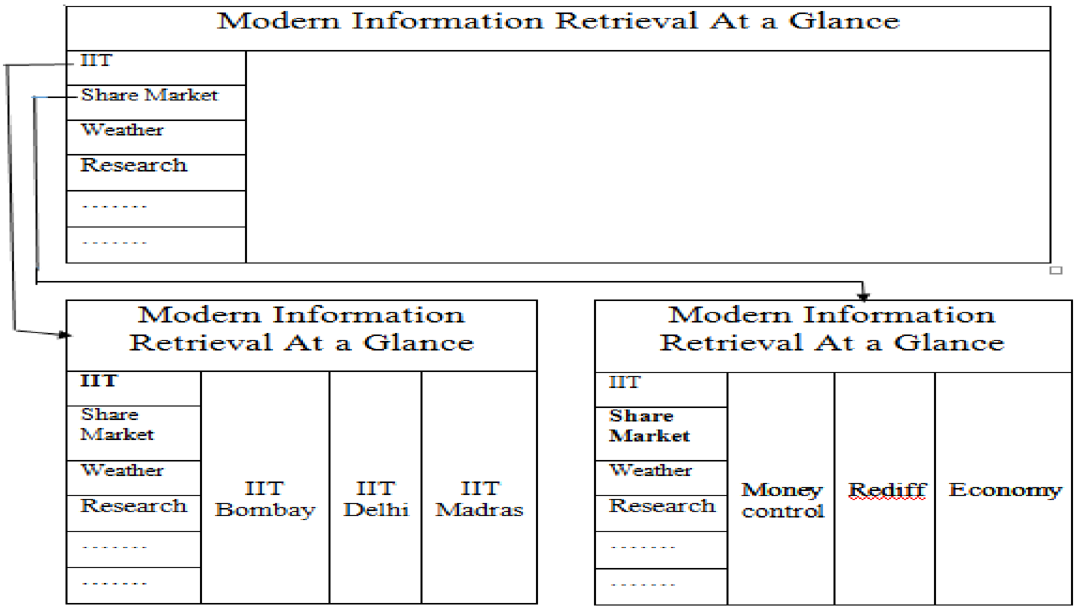


Figure 7:- Comparative Analysis

IV. *Result and Conclusion*

This model is useful for the industrial professional as well as students for performing there comparative study on different websites. This model also helpful to reduce the human resources for access multiple site at multiple machines, reduce resource coast and introduce new information retrieval techniques. This work is the initial work, To implement fully automated model of this research, need create web crawler for finding the related websites and manipulate calculate or find out its rank on Google.

V. *References*

[1] Khari, M., Jain, A., Vij, S., & Kumar, M. (2016, March). Analysis of various information retrieval models. In Computing for Sustainable Global Development (INDIACom), 2016 3rd International Conference on (pp. 2176-2181). IEEE.

[2]Suguna, S., Sundaravadivelu, V., &Gomathi, B. (2016, March). A novel semantic approach in E-learning information retrieval system. In Engineering and Technology (ICETECH), 2016 IEEE International Conference on (pp. 884-889). IEEE.

[3]Ahmad, M. W., & Ansari, M. A. (2012, June). A survey: Soft computing in intelligent information retrieval systems. In Computational Science and Its Applications (ICCSA), 2012 12th International Conference on (pp. 26-34). IEEE.

[4] https://en.wikipedia.org/wiki/Data_mining

[5] https://www.cs.uic.edu/~liub/WebContentMining.html

[6] http://computer.howstuffworks.com/internet/basics/search-engine1.html

[7] www.lisbdnet.com/online-information-retrieval-syste

[8] Ahmad, W., & Ali, R. (2016, March). Information retrieval from social networks: a survey. In Recent Advances in Information Technology (RAIT), 2016 3rd International Conference on (pp. 631-635). IEEE.

[9] Chandni Saini, Vinay Arora," Information Retrieval in Web Crawling: A Survey", 2016 Intl. Conference on Advances in Computing, Communications and Informatics (ICACCI), Sept. 21-24, 2016, Jaipur, India, 978-1-5090-2029-4/16/$31.00 @2016 IEEE, (pp. 2635- 2635)

[10] Yunjuan, L., Lijun, Z., Lijuan, M., & Qinglin, M. (2011, March). Research and application of information retrieval techniques in intelligent question answering system. In Computer Research and Development (ICCRD), 2011 3rd International Conference on (Vol. 2, pp. 188-190). IEEE.

[11] Liu, C. Y., Tang, C. Y., & Hsu, D. F. (2012, December). A comparative study on the combination of multiple retrieval systems. In Pervasive Systems, Algorithms and Networks (ISPAN), 2012 12th International Symposium on (pp. 169-181). IEEE.

[12] Tamrakar, A., & Vishwakarma, S. K. (2015, December). Analysis of Probabilistic Model for Document Retrieval in Information Retrieval. In Computational Intelligence and Communication Networks (CICN), 2015 International Conference on (pp. 760-765). IEEE.

[13] Sahu, S. K., Mahapatra, D. P., & Balabantaray, R. C. (2016, April). Analytical study on intelligent information retrieval system using semantic network. In Computing, Communication and Automation (ICCCA), 2016 International Conference on (pp. 704-710). IEEE.

[14] Sharma, P., Bohra, B., & Yadav, S. (2016, September). Comparative analysis of web-mining approaches for efficient mining of server log formats. In Reliability, Infocom Technologies and Optimization (Trends and Future Directions)(ICRITO), 2016 5th International Conference on (pp. 180-185). IEEE.

[15] Mehlitz, M., Bauckhage, C., Kunegis, J., & Albayrak, S. (2007, October). A new evaluation measure for information retrieval systems. In Systems, Man and Cybernetics, 2007. ISIC. IEEE International Conference on (pp. 1200-1204). IEEE.

[16] Geetha, D. E., Naidu, G. K., Kumar, T. S., & Kanth, K. R. (2008, May). Simulative Performance Evaluation of Information Retrieval Systems. In Modeling & Simulation, 2008. AICMS 08. Second Asia International Conference on (pp. 297-302). IEEE.

[17] Melucci, M. (2016, October). Impact of Query Sample Selection Bias on Information Retrieval System Ranking. In Data Science and Advanced Analytics (DSAA), 2016 IEEE International Conference on (pp. 341-350). IEEE.

[18] Sanderson, M., & Croft, W. B. (2012). The history of information retrieval research. Proceedings of the IEEE, 100(Special Centennial Issue), 1444-1451.

[19] Agbele, K. K., Ayetiran, E. F., Aruleba, K. D., & Ekong, D. O. (2016, October). Algorithm for Information Retrieval Optimization. In Information Technology, Electronics and Mobile Communication Conference (IEMCON), 2016 IEEE 7th Annual (pp. 1-8). IEEE.

[20] Patel, S., Khanna, K., & Sharma, V. (2016, April). Documents ranking using new learning approach. In Computing, Communication and Automation (ICCCA), 2016 International Conference on (pp. 65-70). IEEE.

[21] Ravana, S. D., & Shuxiang, Z. (2016, May). Estimating the Reliability of the Retrieval Systems Rankings. In Software Networking (ICSN), 2016 International Conference on (pp. 1-5). IEEE.