

Effective Web Search using HAPIT Algorithm

^[1]S. Gunasekaran, ^[2]V.B. Ajith Raghavan, ^[3]Avinash Nimmagadda, ^[4]T. Harsh Singhanian, ^[5]R. Anandhan
^{[1][2][3][4][5]}Department of Computer Science Engineering, VSB Engineering College, Karur, Tamil Nadu. INDIA.

Abstract – Nowadays, in IT industries several aspects of information retrieval with focus on ranking and user-friendly nature of web search engine are used. This proposed HAPIT work integrates Hubs & Authorities, Page Ranking, HITs, which has several concepts and components of an information retrieval system than existing system. This work also discusses an important theoretical model of IR and how IR systems are evaluated. This work compared with existing case studies of the available search engines and its problems in common. In future, the work is extracted with upcoming trends in search based on image search, location search.

INTRODUCTION

The World Wide Web is a popular and usual way of propagating information across the world. The web is huge and dynamic with lot of contents repeated and unique. With these contents across the web, it is difficult for the user to search the required document because; the web provides a hyperlink structure with which the user is deviated from the original content. The major goal of a web search engine is to find the relevant data that the user searches on the web but, finding the relevant content that matches the user's interest and behavior is the difficult task where most of the search engine failed. When a user searches a query on the web, he/she is provided with a list of documents that are relevant and irrelevant to the search. Most other search engines use keyword-based searches on the web to retrieve the document to the user. This search list has many irrelevant documents which are not needed by the user thus retrieving this is wastage of time.

There are many algorithms available for the retrieval of the data but the results are not satisfying. This paper provides a real-time search engine for the complex World Wide Web by providing key features like relevancy, ranking. This search engine makes use of most of the available algorithms as a combination of those algorithms the proposed web search engine provides a better satisfying result to the user.

Web Search Engine

The World Wide Web consists of billions of documents and huge amount of information in the form of web pages and web sites. But, sometimes the user is deviated from the original content. There are many types of search engines available. They differ from each other by their usage, structure, retrieval method, and retrieval data.

Components of Web Search

Text Operation: Operations that are performed on the user query to obtain the required keyword to be searched in the document over the web.

Query Processing: Removing the unwanted details from the query such as conjunctions, articles, prepositions and so on.

SearchBase: It is the basement of all kind of activities in the search engine.

Caté: It is the process identifying the relevant document for the given query.

Documents: These are the documents available in the web and are retrieved to the user based on the users request.

Indexing: It is the process of maintaining an index of the contents of the document and are done by the indexer.

Searching: The process of searching the required data over the web by using the keyword obtaining from the user.

Components of Search Engine

Text Acquisition: The major task of text acquisition is to make the documents available that are searched.

Text Transformation: It is the process of transforming the user query into a form understandable by the database.

The Index Creation: Crawler will take all document from the web and give it to the indexer.

Query Processing: The query obtained from the user is processed to extract the needed information and keyword from the query.

User Interaction: User will type the query as the need based on few basic interaction types to obtain the most relevant document

Ranking: Ranking is the process of providing ranks to the documents based on the frequency of usage and relevance.

RELATED WORKS:

1. Title : Enhancement of Web Search Engine Results Using Keyword Frequency Based Ranking.

Authors : Ms. Nilima V. Pardakhe¹, Prof. R. Keole²
Department of Comp. Sci. & Engg, S.G.B.A.U., Amravati, India² Department of Comp. Sci. & Engg, S.G.B.A.U., Amravati, India

Proposed System : The proposed system in this paper informs about the retrieval of user data based on the frequency of the keywords based on the ranking algorithms.

2. Title : Content Based Ranking for Search Engines.

Authors : P.Sudhakar, G.Poonkuzhali, R.Kishore Kumar, Member IAENG

Proposed System : The proposed system in this paper informs about the ranking based on the content searching is processing.

3. Title : A World Wide Web based image search engine using text and image content features.

Authors : Bo Luo, Xiaogang Wang, and Xiaoou Tang
Department of Information Engineering, The Chinese University of Hong Kong

Proposed System : The proposed system in this paper informs about the text and the image is processed in it.

4. Title : The Research of Search Engine Based on Semantic Web

Authors : Yi Jin, Zhuying Lin, Hongwei Lin School of Mathematics and Computer Science, Guizhou Normal University, Guiyang, Guizhou, 550001, China

Proposed System : The proposed system in this paper informs about the semantic which gives about the user exactly required.

5. Title : A Real-Time Search Engine for the Web of Things

Authors: Benedikt Ostermaier_, Kay R'omery_, Friedemann Mattern_, Michael Fahrmaiz and Wolfgang Kellererz

Proposed System : The proposed system in this paper informs about the web of things which represent the easy of accessing the documents in the web

6. Title : Content Based Ranking for Search Engines

Author : P.Sudhakar, G.Poonkuzhali, R.Kishore Kumar, Member IAENG

Proposed System : The proposed system in this paper informs about the ranking based on the search strategy.

7. Title : The Research of Search Engine Based on Semantic Web

Authors : Yi Jin, Zhuying Lin, Hongwei Lin
School of Mathematics and Computer Science, Guizhou Normal University, Guiyang, Guizhou, 550001, China

Proposed System : The proposed system about in this paper informs about the semantic web which and how effectively it works and the process of searching method it represents

8. Title : Enhancement of Web Search Engine Results Using Keyword Frequency Based Ranking

Authors : Ms. Nilima V. Pardakhe, Prof. R. Keole
Department of Comp. Sci. & Engg, S.G.B.A.U., Amravati, India

Proposed System : The proposed system provides high relevancy to the user using the concept of Caté. Caté is a kind of table which used to identify the document which is most relevant to the given user query. This is done to provide high end user satisfaction. The use of Caté table helps in providing the document of the required field, type and data. The other concepts like indexing for creating the index of the document, crawler for crawling the web to obtain the tags and Hubs and Authorities for obtaining the links to authorities from the hubs for a web link.

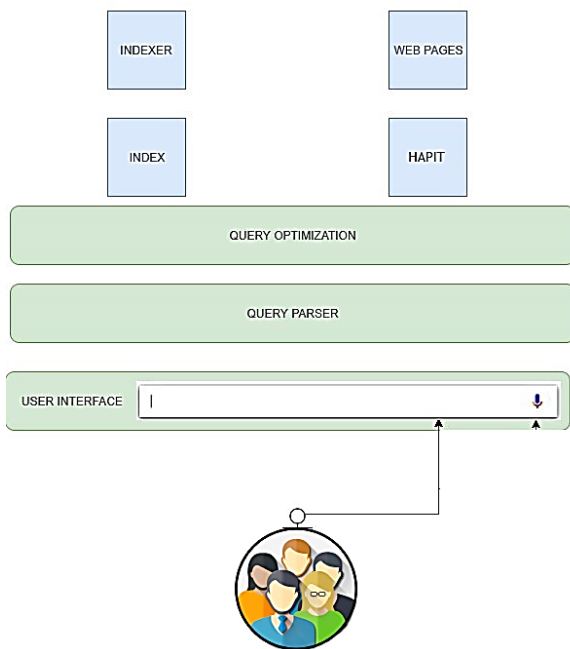
ARCHITECTURAL DESIGN

Architecture of the developed work uses the advantage of many available algorithms such as HITS, Page ranking, Hubs and Authorities. User query is processed by the Query parser to obtain the main keywords of the search query. The obtained keywords are then sent for pre-processing. In the context of web search the pre-processing plays an important role. The obtained query from the user tend to be very complex for the computer machine to understand because it contains many unwanted key terms like prepositions, articles and other items. This pre-processing is done by the Query optimizer

to improve the quality of the search. If the search is improved then the search engine accuracy is also improved. After obtaining the required keywords the proposed algorithm works. It will provide

Then this search engine uses a crawler to crawl the web pages available and an Index which contains the data processed by the crawler. This index is managed by the Indexer. Finally, the obtained results are provided to the user searching for the data.

BLOCK DIAGRAM



PROPOSED METHODOLOGY:

Where user need to search some document from there we compared the two type of users are there un-education and education the un-education will use the typing method where the educated will use the both the voice search and the typing method this process will be processed in the user side.

On the backend, crawler will crawl all the web pages with the seed given. All the web pages are stored in the

directory. Indexer will then take the page and do the process over it and collect the important term and key words. Then, the index is saved. Algorithm will take the user query as input then produce the output, as exactly.

ALGORITHM

- Assign the seed
- Read the page and give to index and page link to visits
- Replicate to do parallel process
- Procced third step
- Read the page
- Produce stop list
- Get and calculate the prepending of word=link number of the page
- Get the query
- Apply dictionary
- Try to combinations
- Produce the list
- Give to user

ALGORITHM EXPLANATION

Crawler is the first step in every search engine but crawler needs a seed to initiate. After which Crawler read all the web page and hand over to the indexer. If the page is visited again crawler will update the page if necessary. The indexer will index the index from the documents. The indexing process is done in the indexer phase. Then in the query optimization the user query will be optimized to exactly what the user wants. The SearchBase is the concept that is used to increase the relevancy for the use document that is retrieved. The user query and the document relevance should be matched, that will give the increase the relevancy. The SearchBase will use the Caté which is the best way to search the document. Caté will tell the component to search only the necessary part, to improve the relevancy to provide the fast access. The component will give the retrieved document to the user. Of course, UI (User Interface), user will interact through the user interface. The query will be given through the user interface and the relevant document will also displayed in the user interface.

EXPERIMENTAL ANALYSIS

Our project is deal about the web search engine process is represent about the effective search in the web search engine which make the use to be comfort and exact search is used to represent in the web and the motivate is to

represent the user satisfaction. Where it is represent based on the indexing.

CONCLUSION AND FUTURE WORK

Thus, the developed algorithm gives a better result than the available search-engines. However better results are needed to be developed further which could provide the user more efficient and user-friendly results. The current search of data is based on the images, texts and audio-video files future work will be done to provide search on all kind of data. This is because the data in the current era is of many forms.

REFERENCE

- [1] E. Miluzzo, N. D. Lane, K. Fodor, R. Peterson, H. Lu, M. Musolesi, S. B. Eisenman, X. Zheng, and A. T. Campbell, "Sensing Meets Mobile Social Networks: The Design, Implementation and Evaluation of the CenceMe Application," in *SenSys 2008*, pp. 337–350.
- [2] "Botanicalls." [Online]. Available: www.botanicalls.com
- [3] "Bicing." Available: www.bicing.com
- [4] D. Guinard, V. Trifa, and E. Wilde, "Architecting a Mashable Open World Wide Web of Things," *ETH Zurich, Tech. Rep. CS-663*, 2010. Available: www.vs.inf.ethz.ch/publ/papers/WoT.pdf,
- [5] Han, J. and Kamber, M. "Data Mining: Concepts and Techniques", Morgan Kaufmann Publishers, 2001.
- [6] Hongqi li, Zhuang Wu, Xiaogang Ji, Research on the techniques for Effectively Searching and Retrieving Information from Internet, International Symposium on Electronic Commerce and Security, IEEE 2008.
- [7] Jaroslav Pokorny, Jozef Smizansky, "Page Content Rank: An approach to the Web Content Mining",
- [8] Junghoo Cho Narayanan Shivakumar Hector Garcia-Molina, Finding replicated web collections , MOD 2000, Dallas, TX USA
- [9] Moldovan D I, Mihalcea R. Using wordnet and lexical operators to improve internet searches[J]. *IEEE Internet Computing*, 2000, 4(1): 34-43.
- [10] Guha R, McCool R. The tap knowledge base. <http://tap.stanford.edu/>
- [11] Anyanwu K, Sheth A. P-Queries: enabling querying for semantic associations on the semantic web[C]//*Proceedings of the 12th international conference on World Wide Web. ACM*, 2003: 690-699.
- [12] Popescu A M, Etzioni O. Extracting product features and opinions from reviews[M]//*Natural language processing and text mining. Springer London*, 2007: 9-28.
- [13] The Research of Search Engine Based on Semantic Web Authors : Yi Jin, Zhuying Lin, Hongwei Lin School of Mathematics and Computer Science, Guizhou Normal University, Guiyang, Guizhou, 550001, China.