Hybrid Indexing Technique for Web Crawler

Name : Mahak ,
Phd Scholar (Department of CSE, Kurukshetra University ,Kurukshetra)

***Abstract*: Search engines use indexes for fast accessing data in order to answer user queries. Data can be web documents, pictures, spatial data, etc. Usually, inverted indexing is used to index web documents. In this paper, we introduce a new indexing method for web documents, implementation of the Hybrid Indexing for the web crawler. It also described the better graphical user interface for crawler which provides the support for controlling the number of outputs to be displayed, different mime types supported by crawler which allowed the search engine to restrict or allowed the desired mime type for files to be downloaded. It also provide some advanced setting options for crawler using which different type of undesired url's or web content can be restricted from downloading.**

Keywords: inverted index, text based ranking, hybrid indexing, web documents, URL

I. INTRODUCTION

A. Indexing and querying web pages

The Web search process has two main parts: off-line and on-line. The off-line part is executed periodically by the search engine, and composed of downloading a sub-set of the Web to build a collection of pages, which is then converted into a searchable index. The on-line part is run every time a user query is executed, and uses the index to select some candidate documents that are sorted according to an estimation on how relevant they are for the user's need. This process is depicted in Figure 1.1

Web pages come in many different formats such as plain text, HTML pages, PDF Web documents, and other proprietary formats. For indexing Web pages, the first stage is to extract a standard logical view from the documents. The logical view for documents which is mostly used in search engines is the "bag of words" model, in which each and every web document is seen only as an unordered set of words. In today's Web search engines, this view is enhanced with more information concerning word frequencies and text formatting attributes, and also meta-information about Web pages including embedded descriptions and explicit keywords in the HTML markup.

There are several text normalization operations that are executed for extracting keywords, the most used ones are: tokenization, stopword removal and stemming. In information retrieval stopwords are usually discarded also for efficiency reasons, as storing stopwords in an index takes considerable space because of their high frequency.

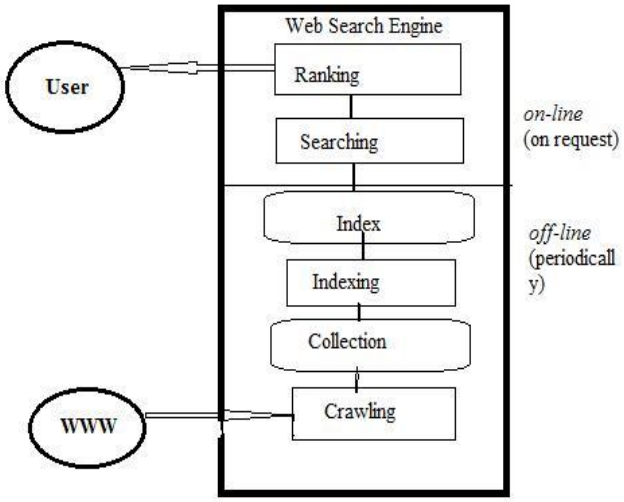


Figure 1.1: A Web search engine periodically downloads and indexes a sub-set of Web pages (off-line operation). This index is used for searching and ranking in response to user queries (on-line operation). The search engine is an interface between users and the World Wide Web.

PROPOSED METHODOLOGY

An inverted index is composed of two parts: a vocabulary and a list of occurrences.

The vocabulary is a sorted list of all the keywords and for each and every word in the vocabulary, a list of all the "places" where the keyword appears in the collection is kept. Figure 2.1 shows a sample of inverted index, considering all words including stopwords. When a user's query is considered, the lists are extracted from the inverted index and then merged. Queries are frequent because usually hashing in memory is used for the vocabulary.

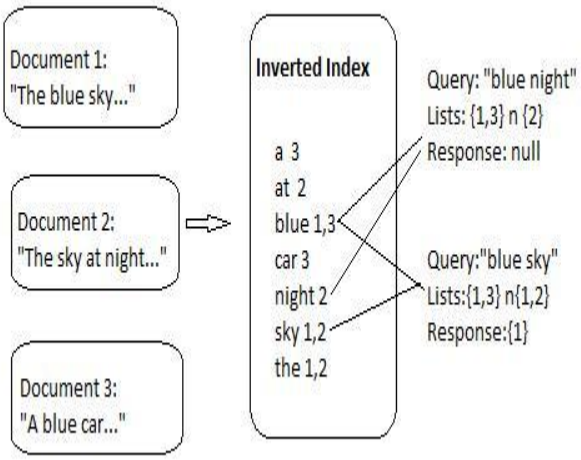


Figure 2.1: A sample inverted index with three documents. All tokens are considered for the purpose of this example, and the only text normalization operation is convert all tokens to lowercase. Searches involving multiple keywords are solved using set operations.

B. Distributing Query Load

Query response time in today's search engines requires to be very fast, and should be done in a parallel way involving number of systems. For parallelization, the inverted index is usually alloted among several physical computers. Basically two techniques are used to partition index.

Query processing involves a central "broker" that is assigned the task of distributing incoming queries and merging the results. As the results are usually shown in groups of 10 or 20 documents per page, the broker does not need to request or merge full lists, only the top most results from each partial list. Search engines exploit the fact that users seldom go past the first or second page of results. Search engines provide approximate result counts because they never perform a full merge of the partial result lists, so the total number of documents in the intersection can only be estimated.

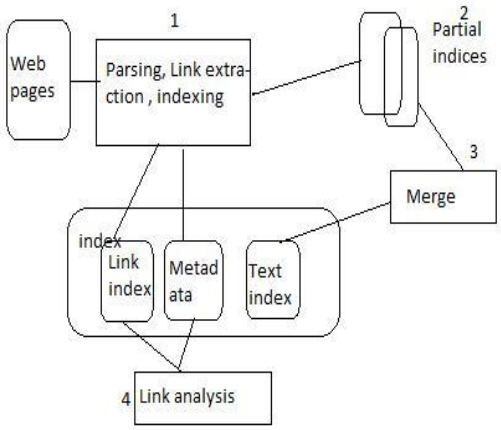


Figure 2.2: Indexing for Web search. (1) Pages are parsed and links and extracted. (2) Partial indices are written on disk when main memory is exhausted. (3) Indices are merged into a complete text index. (4) Off-line link analysis can be used to calculate static link-based scores.

When a global inverted file is used, the vocabulary is partioned into differentl parts containing roughly the same amount of occurrences. Each computer is alloted a part of the vocabulary and all of its occurrences. Whenever a query is received, the query is sent to the computers holding the query terms, and the results are merged afterwards. Hence, this task of load balancing is not basic .

Query processing involves a central "broker" that is assigned the task of distributing incoming queries and merging the results. The results are used to be shown in groups of 10 or 20 documents per page, the broker does not need to request or merge full lists, only from each partial list, the first teo or three or top most results are taken.

C. Text Based Ranking

A Page Rank is calculated from a mathematical algorithm based on the webgraph, created by all World Wide Web pages as nodes and hyperlinks as edges, taking into consideration authority hubs.

The vector space model is the standard technique for ranking documents which is basically based on a query. In this model, both a document and a query are seen as a pair of vectors in a space with as many dimensions as terms as the vocabulary. In this type of space model, the similarity of a query to a document is given by a formula that transforms each vector using certain weights and then calculates the cosine of the angle between the two weighted vectors:

$$sim(q,d) = \sum_t w_{t,q} \times w_{t,d} / \sqrt{\sum_t w^2_{t,q}} \times \sqrt{\sum_t w^2_{t,d}}$$

In pure text-based information healing systems, documents are shown to the users in decreasing order using this similarity measure.

TF stands for term frequency, and the idea is that if a term appears several times in a document it is better as for describing the contents of that document. The TF is normalized relative to document length, that is, the term *t* divided by the frequency of the most frequent term in document *d*:

$$t f t,d = f req_{t,d} / max_l \ f req_{l,d}$$

IDF stands for inverse document frequency and reflects how frequent a term is in the whole collection. The rationale is that a term that appears in a few documents gives more information that a term that appears in many documents. If there are N number of documents and *nt* if the number of documents containing the query term *t*, then $id f t = \log N/nt$

Using these measures, the weight of each term in given by:

$$W_{t,q} = (1/2 + 1/2\ tf_{t,q})idf_{t,q} \ , \qquad W_{t,d} = tf_{t,d}$$

The 1/2 factor is added to avoid a query term having 0 weight. Several substitute of weighting schemes have been designed, but the weighting scheme is one of the most used and gives good results in practice.

D. Proposed Architecture:

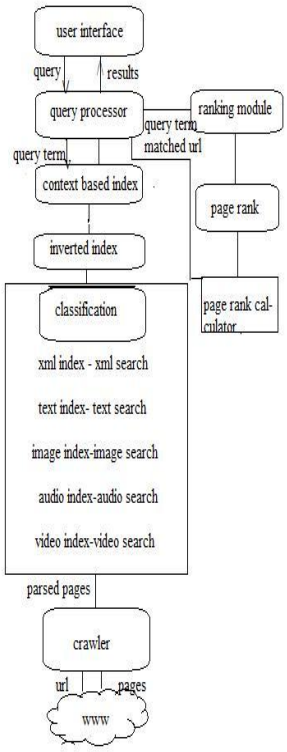
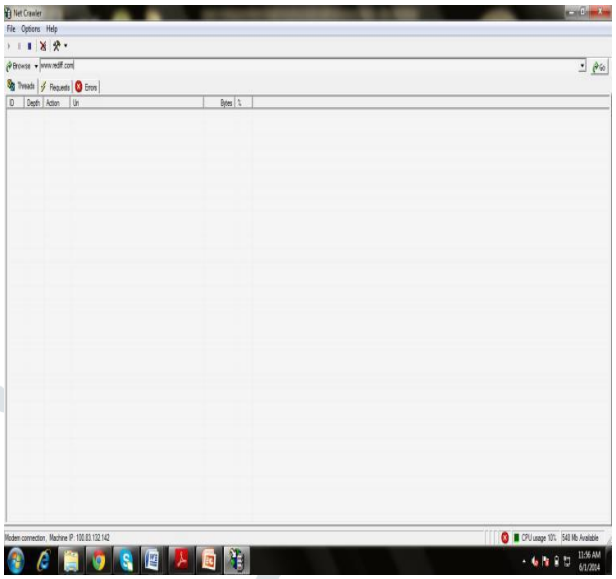


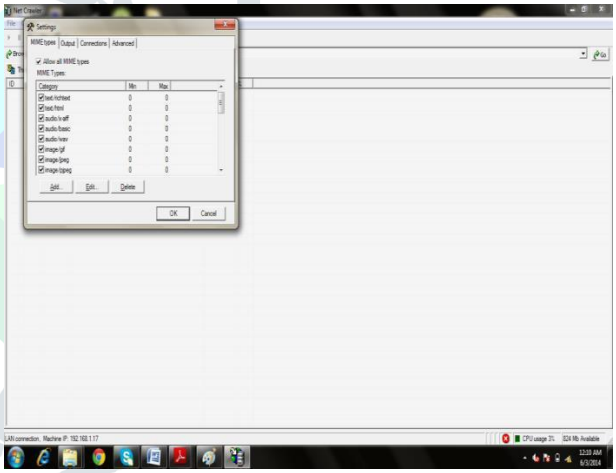
Fig 2.3: Hybrid Indexing Technique

III. ANALYSIS

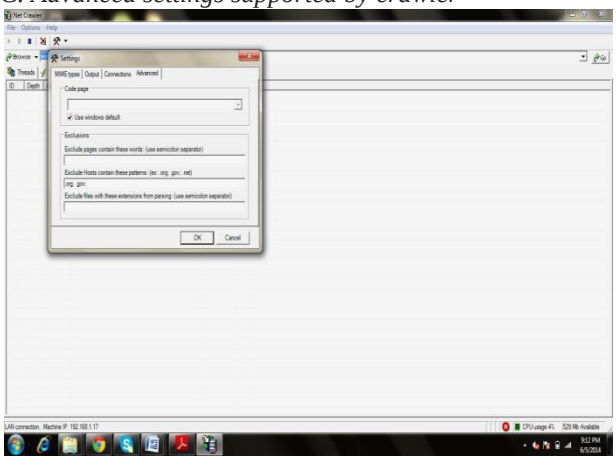
A. The Web Crawler



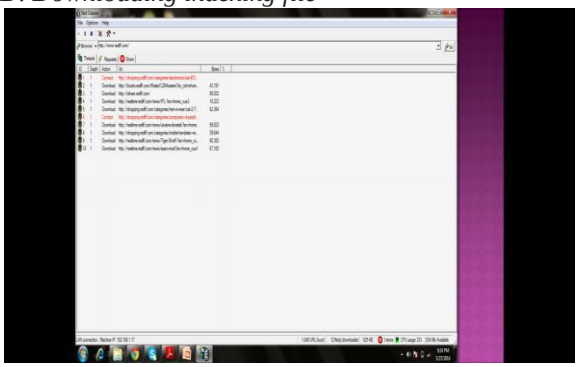
B. Allowed Mime types of documents to be downloaded



C. Advanced settings supported by crawler



D. Downloading indexing file



E. Classification within indexing file

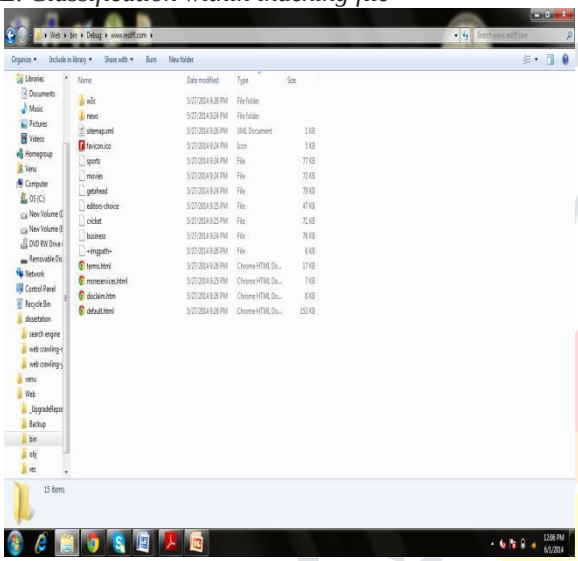


TABLE 1: COMPARISON OF VARIOUS INDEXING TECHNIQUES AND HYBRID INDEXING TECHIQUE

Proposed technique	Based on	Advantages
CBI technique	Technique arranges the documents in a database according to the contextual senses of the keywords present in them.	• It serves the different search strategies of both the casual and the professional users of the Internet
Clustering index	based on user preferences by recording the hit rate of the service to pick appropriate clusters for the index using the clustering method.	i. Based on user preferences by recording the hit rate of the service to pick appropriate clusters for the index using the clustering method. ii. It greatly improves the efficiency of semantic query.
Semantic index based on RDF triples	based on RDF(Resource Description Framework) triples, in order to catch and manage the semantics of a given textual document.	Managing efficiently and effectively very large amount of digital documents
Indexing Web pages using Web bots	• method uses concept that the content of the pages linked to or from the indexed page for indexing • demonstrate usage of a new method based on bots which analyze content of the pages linked to or from the page of interest.	The similarity of the word usage at the different link distance from the page of interest and demonstrate that a structure of words used by the linked pages enables more efficient indexing and search
Multidimensional indexing approach: MIWD	In MIWD, it just descend the tree from the root based on the most similarity path to find the first leaf item as the query answer.	retrieval time of MIWD method is better than traditional methods
Hybrid indexing technique	Based on text based ranking and also uses inverted indexing to provide index of different classified documents	The technique helps in managing effectively large amount of document based on MIME types

IV. CONCLUSION

The quality of result depends on the information stored in index. Performance of search engine can be enhanced by efficient indexing. There are various indexing techniques available which are based on different concepts to provide a better collection format for downloaded documents so that it become easy and fast for search engine to search for a query. The proposed Hybrid Indexing Technique takes the advantages of page ranking based indexing technique, inverted indexing techniques and context based indexing technique. The proposed technique arranges the documents downloaded based on the

mime types. The advantage of the proposed technique is it serves the different search strategies of both the casual and the professional users of the Internet. It provides more relevant documents in first top URLs according to the user interest.

V. FUTURE SCOPE

The Web Crawler of the search engines is expert in crawling various Web pages to gather huge source of information. The crawler go to a web page, reads through it, and then follows links to other pages within the site. The crawler will return to the site on a regular basis, such as every day or every week, to look for changes. This paper described the implementation of the Hybrid Indexing, which is based on the web crawler. It also described the better graphical user interface for crawler which provides the support for controlling the number of outputs to be displayed, different mime types supported by crawler which allowed the search engine to restrict or allowed the desired mime type for files to be downloaded. It also provide some advanced setting options for crawler using which different type of undesired url's or web content can be restricted from downloading. In future some more advanced options can be added in crawler so that it can provide more relevant and attractive results to be displayed to the user. Also performance analysis can be done for the proposed hybrid indexing techniques such as evaluation of computational cost and comparison with other existing techniques can be done.

REFERENCES

[1] Dr Rajender Nath, "Web Crawlers: Taxonomy, Issues & Challenges" Volume 3, Issue 4, April 2013, International Journal of Advanced Research in Computer Science and Software Engineering

[2] A. K. Sharma, "A Novel Context Based Indexing of Web Documents**",** 2012 International Conference on Communication Systems and Network Technologies

[3] MaoLi, "Efficient Clustering Index for Semantic Web Service Based on User Preference", 2012 International Conference on Computer Science and Information Processing (CSIP)

[4] Robin Sharma," Web Page Indexing through Page Ranking for Effective Semantic Search", Proceedings of7'h International Conference on Intelligent Systems and Control (ISCO 2013)

[5] Pooja Mudgil, "An Improved Indexing Mechanism to Index Web Documents", 2013 5th International Conference on Computational Intelligence and Communication Networks

[6] Ioannis Avraam, "A Comparison over Focused Web Crawling Strategies", 2011 Panhellenic Conference on Informatics.

[7] L. Huilin, K. Chunhua ; W. Guangxing, (2007)
"Efficiently Crawling Strategy for Focused Searching Engine", Advances in Web and Network Technologies and Information Management, Lecture Notes in Computer Science, Vol. 4537/2007, 25-36.

[8] Changshang Zhou,;Wei Ding; NaYang, "Double Indexing Mechanism of Search Engine based on Campus Net", Proceedings 2006 IEEE Asia-Pacific Conference on Services Computing (APSCC'06)

[9] Philipp Astm; Michael Kapfenberger; Stefan Hauswiesne (Nov 2008) , —Crawler Approaches And Technology.