View metadata, citation and similar papers at core.ac.uk

brought to you by CORE

provided by International Journal of Computer (IJC - Global Society of Scientific Research and...

International Journal of Computer (IJC)

ISSN 2307-4523 (Print & Online)

© Global Society of Scientific Research and Researchers

http://ijcjournal.org/

Context Based Indexing in Search Engines: A Review

Suraksha Singla[a*], Geetanjali Gandhi[b]

[a]*M.Tech Scholar of Computer Science & Engineering BSAITM, India*

[b]*Assistant Lecturer in Department of Computer Science & Engineering BSAITM, India*

Abstract

There are so many increasing amount of information in the today's World Wide Web. For these increasing amount of information we need efficient and effective index structure .Most indexing techniques directly matched terms from the documents and terms from query. Granting efficient and fast accesses to the index is a key issue for performance of web search engines. The main aim of search engine is to provide most relevant documents to the users in minimum possible time. Indexing is performed on the web pages after they have been gathered into a repository by the crawler. The existing architecture of search engine shoes that the index is built on the basis of the terms of the document. The context of the documents being collected by the crawler in the repository is being extracted by the indexer using the context repository, thesaurus and Ontology repository and then documents are indexed.

1. Introduction

The World Wide Web is the collection of large amount of information which is increasing day by day. It has become one of the most important resources. The internet contains hundreds of thousands of electronic collections that contain high quality information. The basic aim is to select the best collection of information for a particular information need. The indexing phase of search engine can be viewed as a Web Content Mining Process. The indexer extracts a large amount of number of all occurrences of each term within every document. This information is maintained in an index, which is usually represented using an Inverted File (IF). Modern Web search engine can cache, index and search several billion of web pages, which only includes a small part of all existing documents in the web. The context of the documents that is collected by the crawler in the repository is being extracted by the indexer using the context repository according to their respective context. The Crawler match the frequency of words from user's query.

---

* Corresponding author.

If the higher frequency words is matched then the document is relevant. But they do not analyze the context of the keyword. Searching of data relevant to a given query which is made from general language is called information retrieval system. The documents extracted during the indexing phase are compared with the query. The documents which resemble most are given to the users where they evaluate the relevance of document with respect to their need. Therefore, the major issue to be addressed in information selection is the development of a search mechanism that will help in getting maximum relevant documents. In the current scenario, the documents are retrieved if they contain keywords specified by the user. Indexes are automatically built without human intervention. The ideal indexing is to dynamically choose a set of features to represent documents given user's information needs.

2. Related Work

This section describes the previous works on this domain of indexing and usage of ontology in information retrieval systems. Firstly, related works on indexing are described. Many algorithms and techniques have already been proposed but they seem to be less efficient in accessing the index.

Reference [1] introduces a technique for indexing, the keywords extracted from the documents. It uses a height balanced binary search (AVL) tree.

Reference [2] proposes the context based indexing in search engine using ontology. The index is built on the basis of the context of the documents rather than the terms.

Reference [3] introduces a double indexing mechanism for search engines based on Campus Net. The Campus Net Search Engine (CNSE) is based on full-text search engine, but it is not general full-text search engine as it is a private net.

The CNSE consists of crawl machine, Chinese automatic segmentation, index and search machine. They proposed double indexing mechanism which consists document index as well as word index.

Reference [4] introduces a reordering algorithm for indexing, which partitions the set of documents into K ordered clusters on the basis of similarity measure. According to this algorithm, the biggest document is selected as centroid of the first cluster and n/K1 most similar documents are assigned to this cluster. This algorithm is not effective in clustering the most similar documents.

Reference [5] proposes the threshold based clustering algorithm in which the number of clusters is unknown.

Two documents are classified to the same cluster if the similarity between them. If the threshold is small, all the elements will get assigned to different clusters. If the threshold is large, the elements may get assigned to just one cluster.

Reference [6] proposed the context driven focused crawler(CDFC) that searches and downloads highly relevant web pages.The proposed design significantly reduces the storage space at the search engine side.

Parul Gupta and his colleagues [7] describes the pre –ranking of the similar documents after the formation of the index.

This paper proposes an ontology driven pre ranking of the documents with identical context and post ranking of the search results.

3. Architecture of Context Based Indexing

Search Engine processing involves the steps are carried out by three different modules. The Crawler collects web documents and stored in a repository. Every Web page has an information from HTML web pages.

1. **Webpage repository**- This is the collection of web documents that have been gathered by the crawler from the WWW.

2. **Indexer**- It maintains an index of the documents that being collected by the crawler which is the form of posting list that contain the term as well the document id of the documents.

3. **Document Preprocessing**- This step performs stemming as well as removal of stop words. A stop word is any word which has no semantic content. Others words are preposition and articles.

4. **Thesaurus**- It is a dictionary of words available on the World Wide Web from thesaurus.com which contains the words as well as their multiple meanings.

5. **Context Repository**- This is the database which contains the various contexts. Also the new contexts derived from thesaurus are stored in this repository.

6. **Ontology Repository**- This is the database of ontologies which contains the various relationships among objects in various domains. It contains various concepts with their relationship.

7. **Context of the documents-** This context represents the theme of the document that has been extracted using context repository and ontology repository.

8. **Index**- This is the final index that is constructed after extracting the context of the documents.

9. **Query Processing**- This module searches the result in the index and provides the relevant result to the user.

10. **Query Interface**- This is the module of the search engine that receives user queries.

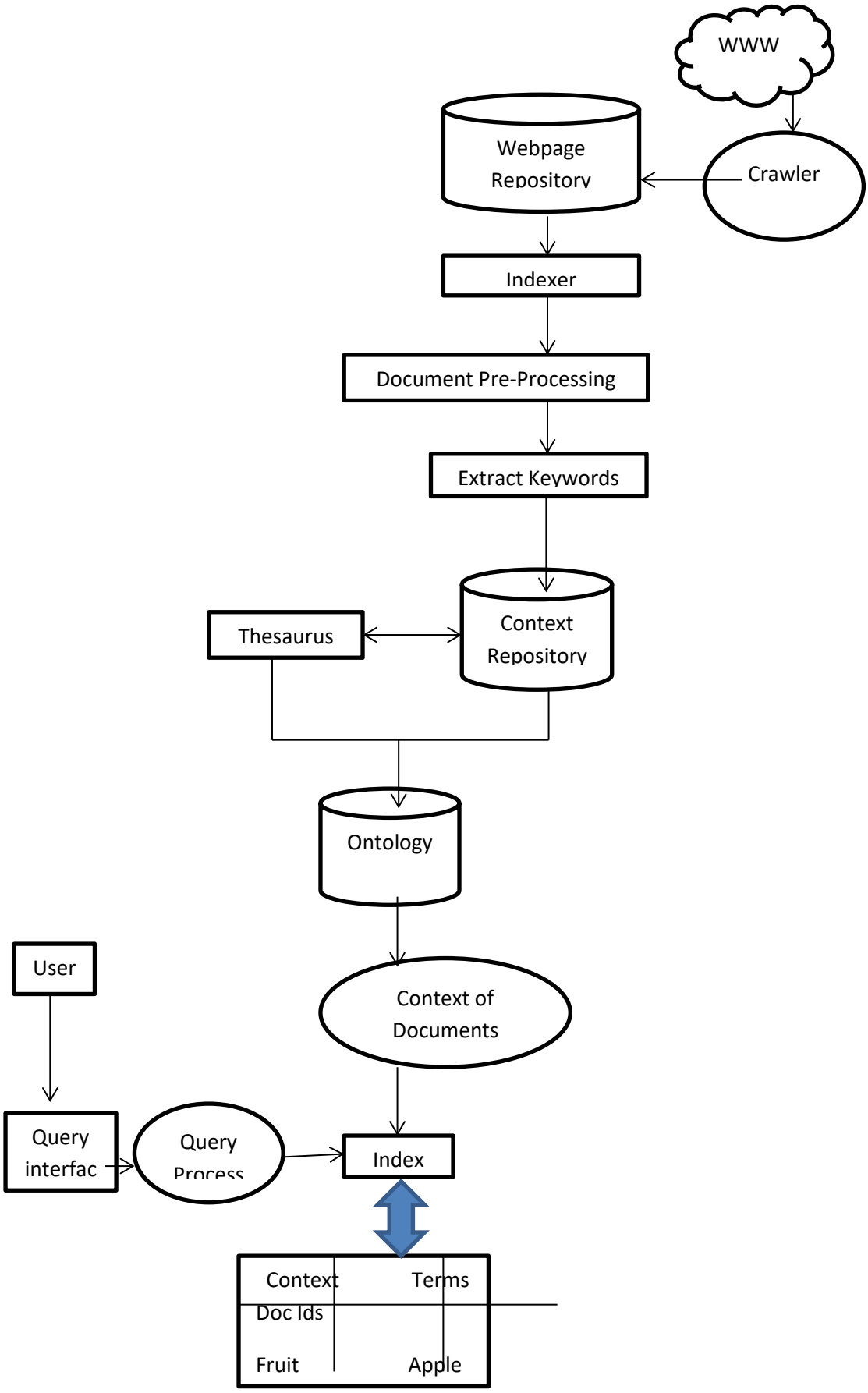

Figure 1: Architecture of context based indexing

Table 1.1: Comparison between Indexing techniques

Traditional Indexing	**Context Based Indexing**
It is arranged according to their appearance in the document	These are arranged alphabetically according to their context
No Professional Indexer	Many Professional Indexer
It might not contain the page no	It maintains the page number

4. Conclusion

This paper presents an indexing structure that can be constructed on the basis of the context of the document. The context of the documents can be extracted with the help of thesaurus and ontology repository that defines the concepts and relationships between the terms. So this paper uses ontology for context based index building. A rough estimate of support values for the existing and the proposed system clearly depicts the better performance of the existing system.

References

[1]. Nidhityagi, Rahul Rishi , R.P. Agarwal "Context based Web Indexing for Storage of Relevant Web Pages" International Journal of Computer Applications (0975 – 8887) Volume 40– No.3, February 2012 [2]. Parul Gupta and A.K.Sharma "Context based Indexing in Search Engines using Ontology", International Journal of Computer Applications, Volume 1 No. 14, pp 49-52, 2010.

[3]. Changshang Zhou, Wei Ding and Na Yang, "Double Indexing Mechanism of Search Engine based on Campus Net", Proceedings of the 2006 IEEE Asia-Pacific Conference on Services Computing (APSCC'06), 2006.

[4]. Fabrizio Silvestri, Raffaele Perego and Salvatore Orlando.Assigning Document Identifiers to Enhance Compressibilityof Web Search Engines Indexes. In the proceedings of SAC,2004.

[5]. O. Zamir, O. Etzioni, O. Madanim, and R.M. Karp, "Fast andIntuitive Clustering of Web Documents," Proc. Third Int'l Conf. Knowledge Discovery and Data Mining, pp. 287-290, Aug. 1997.

[6]. NareshChauhan and A. K. Sharma," Design of an Agent Based Context Driven Focused Crawler",BVICAM'S International Journal of Information Technology, pp 61-66, 2008.

[7]. Sajendra Kumar, Ram Kumar Rana ,Pawan Singh " Ontology based Semantic Indexing Approach for Information Retrieval System" International Journal of Computer Applications (0975 – 8887) Volume

49– No.12, July 2012.

[8]. B.Chandrasekaran and John R.Josephson, Ohio State University V.RichardBenjamins,Universityof Amsterdam "What are Ontologies,and Why do we need them?"IEEE INTELLIGENT SYSTEMS(1094-7167),Volume 14 No.1,pp20-26,1999.

[9]. S. Chakrabarti , M. van den Berg, and B. Dom. "Focused crawling: a new approach to topic-specific web resource discovery". In WWW-8, 1999