IOSR Journal of Computer Engineering (IOSR-JCE)
e-ISSN: 2278-0661,p-ISSN: 2278-8727, Volume 17, Issue 2, Ver. IV (Mar – Apr. 2015), PP 94-97
www.iosrjournals.org

Context Based Indexing in Search Engines Using Ontology: Review

Varsha Rathi[1], Neha Bansal[2]

[1]*M.Tech Scholar of computer science & Engineering BSAITM, India*
[2]*Senior Lecturer in Department of computer science & Engineering BSAITM, India*

Abstract: *Nowadays, the World Wide Web is the collection of large amount of information which is increasing day by day. For this increasing amount of information, there is a need for efficient and effective index structure. The main aim of search engines is to provide most relevant documents to the users in minimum possible time. This paper proposes the indexing structure in which index is built on the basis of context of the documents rather than on the terms basis using ontology. The context of the document that are being collected by the crawler is extracted using the context repository, thesaurus and ontology repository and then documents are indexed according to their respective context.*

Keywords: *Context, Context repository, Indexing, Ontology repository, Semantic web.*

I. Introduction

With the rapid growth of the Internet, the World Wide Web (WWW) has become one of the most important resources for obtaining information. Currently, there are huge amount of documents existing in the World Wide Web. Finding information from WWW according to user interest becomes very crucial task. It is the largest database in universe which is mostly understandable by human users and not by machines. It lacks the existence of a semantic structure. Modern web search engines can cache, index and search several billion of web pages, which only includes a small part of all existing documents in the web. When user submits a query, it produces large number of results that may or may not satisfy user's query. It provides irrelevant information to the users and the search quality could not meet a user's requirement.

The existing crawlers match the frequency of words with the keywords in the user's query. If higher frequency words match with the topic keyword, then the web document is relevant. But they generally do not analyze the context of the keyword in the web page before they download it.

The main aim of search engines is to provide most relevant documents to the users in minimum possible time. So indexing is performed on the web pages after they gathered into a repository by the crawler. The existing architecture of search engine shows that the index is built on the basis of the terms of the document. But the context based indexing allows the indexing structure in which index is built on the basis of the context rather than on the terms basis using ontology.

The context of the documents that is 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 according to their respective context.

1.1 Introduction to ontology

Ontology [6] is the study of the kinds of things that exists. Its mean "theory of existence". It is a representation vocabulary, often specialized to some domain typically some common sense knowledge domain. It forms the knowledge representation for that domain. There are different types of ontology component can be defined like concepts, instances etc. Concept is the main component of ontologies that can be defined in different manner:-

Textual definition: the concept "parrot" is defined by the sentence "as individual animal being" like Bird.

Logical definition using formula:-the bird is defined by the formula "Living entity U Nonliving entity".

Set of properties:-A concept "Bird" can have the property like "type", "color", "food". Finding concept can also be explained by the set of instances of a bird.

The concept of ontologies has contributed to the development of Semantic Web [5] where Semantic Web is an extension of the current World Wide Web in which information is given in a well-defined meaning that translates the given unstructured data into knowledgeable representation data. In other words, Semantic Web is an information that is machine understandable. It allows users to extract web pages according to the context rather than the matching of keywords in order to retrieve relevant web documents to the user's query.

II. Related Work

NidhiTyagi, R.P Agarwal [1] This paper proposes a technique for indexing [1] the keyword extracted from the web documents along with their contexts wherein it uses a height balanced binary search (AVL) tree, for indexing purpose to enhance the performance of the retrieval system.

P. Gupta and A. K. Sharma [2] worked on context based indexing in search engines using ontology. The index construction is done on the basis of the context using ontology. The context repository, thesaurus and ontology repository are used by the indexer to identify the context of the document.

C. Zhou, W. Ding and Na Yang [3], the paper introduces a double indexing mechanism for search engines based on campus Net. The CNSE consists of crawl machine, Chinese automatic segmentation, index and search machine. The proposed mechanism has document index as well as word index. The document index is based on, where the documents do the clustering, and ordered by the position in each document. During the retrieval, the search engine first gets the document id of the word in the word index, and then goes to the position of corresponding word in the document index. Because in the document index, the word in the same document is adjacent, the search engine directly compares the largest word matching assembly with the sentence that users submit. The mechanism proposed, seems to be time consuming as the index exists at two levels. The critical look at the available literature reveals that there is a requirement for a technique to organize the keyword and their contexts in a better fashion as storing in a linear fashion makes searching of a document a bit time consuming.

N. Chauhan and A. K. Sharma [4] proposed, the context driven focused crawler (CDFC) that searches and downloads only highly relevant web pages, thus, reducing the network traffic. A category tree has been used, which provides flexibility to the user for interacting with the system showing the broad categories of the topics on the web. The proposed design significantly reduces the storage space at the search engine side.

III. Architecture Of Context Based Indexing

Architecture of context based indexing is represented in Fig. 1. The web pages are gather by crawler and are stored in the huge repository. Each web page document is identified by its document id.

Various components of architecture are

Crawler: This is an Internet bot that systematically browses the World Wide Web, for the purpose of Web Indexing.

Crawled Webpage Repository: This is the collection of web documents that have been collected by the crawler from the WWW.

Indexer: It maintains an index of the documents that are being gathered by the crawler which is in the form of posting lists that contains the term as well the document identifiers of the documents which contain the given term.

Document Preprocessing: This step performs stemming as well as removal of stop words. A stop word is any word which has no semantic content. Common stop words are prepositions and articles, as well as high frequency words that do not help retrieval.

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.

Word Net: This is a lexical database for the English language. It groups English words into sets of synonyms called Syn Sets, each expressing distinct concept.

Context Repository: This is a database which contains the various contexts. Also the new contexts derived from thesaurus are stored in this repository. The context repository maintains a database of several types of context data.

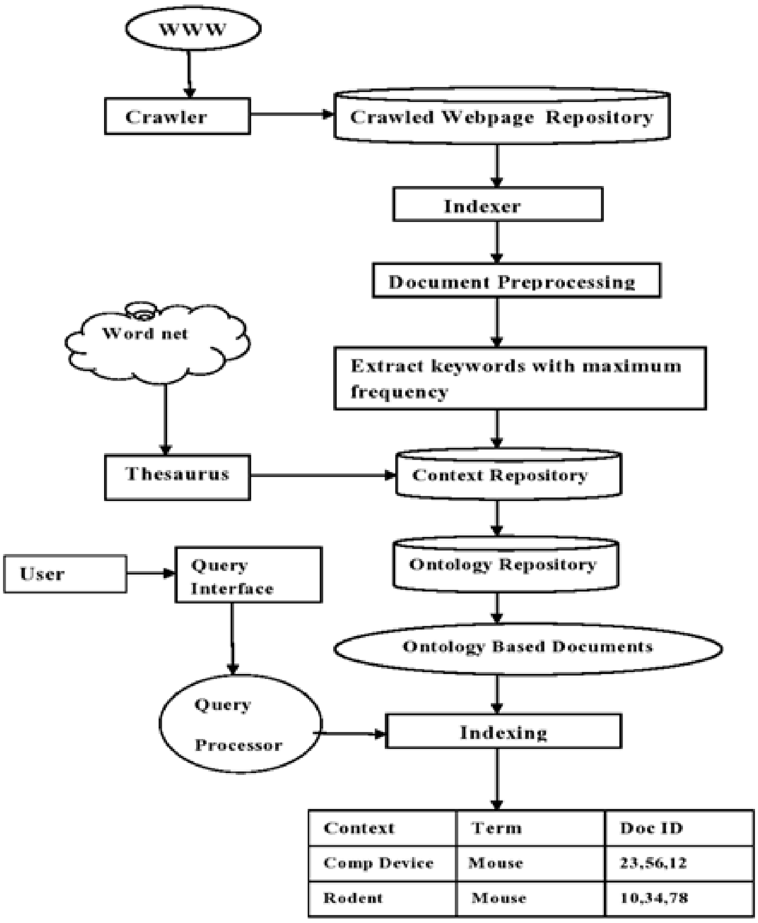
Ontology Repository: This is a database of ontology's which contains the various relationships among objects in various domains. Ontology repository contains various concepts with their relationships.

Ontology based context of the document: This represent the semantic or theme of the document that has been extracted using context repository, thesaurus and ontology repository.

Indexing: After extracting the context of the document on the basis of ontology this is e final index that is being constructed. Rather than being formed on the term basis, the index is constructed on the context basis with context as first field, term as second field and finally the document identifiers of the relevant documents.

Query Interface: This is the module of the search engine that receives user queries.

Query Processor: This module searches the result in the index and provides the relevant result to the user.



Context	Term	Doc ID
Comp Device	Mouse	23,56,12
Rodent	Mouse	10,34,78

Fig1: Architecture of context based indexing

IV. Algorithm For Constructing Index

The algorithm depicted in Fig.2 shows the various steps in the construction of the context based index and hence context based searching.

1. The web pages are collected by the web crawler from WWW and are stored in the webpage repository.
2. The indexer takes the web pages collected by the crawler and parses them into index.
3. In document preprocessing step, the crawled web documents are preprocessed and extract the keywords along with their frequency of occurrence.
4. Now, the context of the keywords with maximum frequency are being searched in the thesaurus (a dictionary of words available on WWW from thesaurus.com). This step helps in extracting the context of the document. As keyword may have multiple contexts, so multiple contexts are extracted.
5. Next step is to extract the specific context of the document from these multiple contexts and are stored in the context repository.
6. Now the multiple contexts and the terms of the document are compared with the ontology repository. The context of the document is extracted by matching the keywords of the document and the multiple contexts with the concepts and the relationship terms in the ontology repository.
7. Now the keywords along with the context are indexed using any indexing scheme such as B+ tree and AVL tree. The index consists of three columns, the one containing the context, the second one containing the terms related to the context and the third one contains the lists of documents that contain the term with that specific context.
8. When the user fires a query with the context explicitly specified, then the index is being searched first on the context basis rather than on the term basis.
9. After the context is matched, the keywords in user's query are matched with the terms related to that context in the index.
10. Now the document identifiers of the relevant documents are retrieved and the user is provided with best matching documents.
11. Thus the index provides a fast access to document contents and structure.

Fig 2: Algorithm for constructing index

V. Conclusion

This paper presents an indexing structure that can be constructed on the basis of the context of the document. The context of the document can be extracted with the help of thesaurus and ontology repository that defines the concepts and relationship between the terms. So this paper uses ontology for context based index building. This offers the retrieval from index on the basis of context rather than keywords. This will help in improving the web search quality by providing the most relevant documents to the user's query as a result.

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]. 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.

[5]. 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.

[6]. 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.

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