Extraction of Contextual Relevance of Web Documents

Nidhi Tyagi
Shobhit University,
Meerut, India.

Rahul Rishi
Maharishi Dayanand University,
Rohtak, India.

R.P. Agarwal
Shobhit University,
Meerut, India.

ABSTRACT
The crawled web pages should be organized in a fashion where they are more understandable to machine, for producing the results which are meaningful and relevant. The set of web pages can be categorized into different contextual sense if the crawler has the technique to understand their meaning and the domain identification. The contextual relevance of the web documents can be known, if the frequent occurring patterns of the keywords in the web page are identified. This can be achieved through data mining technique for generating frequent patterns, using FP- Growth. It will help in deducing the set of keywords of the documents and this knowledge is added in the knowledge store which will further facilitate in the building the ontology for the crawled web pages and organizing them and thus increasing the rank of the document.

Keywords
Context, ontology, frequent patterns, relevance, FP-growth.

1. INTRODUCTION
Context based retrieval system has received a great deal of attention in recent years. The practice of the context can minimize ambiguity in searching and provides foundation to effective information retrieval system. The document in the repository can be organized on the bases of the their respective contexts, to make the searching task simplified, since the documents stored in the repository without the context may lead to ineffective results leading to the wastage of time and resources. The information on the web is available in different formats [1] they are segregated and represented into semantically meaningful format. The set of web pages can be categorized into different contextual sense if the crawler has the technique to understand their meaning and the domain identification. This task can be achieved through a technique of data mining referred as FP- Growth.

1.1 F-P GROWTH
Mining plays an essential role in the process of knowledge discovery in databases, in which intelligent methods are applied in order to extract patterns. FP-Growth [2] approach is based on divide and conquers technique for producing the frequent item sets, without candidate generation. It firstly compresses the database showing frequent item set in to FP-tree. FP-tree is built using 2 passes over the dataset. Further, it divides the FP-tree into a set of conditional database and mines each database separately, thus extracts frequent item sets from FP-tree directly.
The technique can assist in extracting the frequent keywords from the crawled web pages which can further help in identifying the context of the documents. The method is suitable when ambiguity exist for keywords which are hyponyms, and exact context of the document is not known from the title of the web document.

1.2 CONTEXTUAL ONTOLOGY
The term ontology refers to the exact specification of a shared conceptualization. Domain ontology models a specific domain, which represents part of the real world. The domain ontology can be enhanced to as contextual ontology [3], for representing and understanding the real world entity in the more precise and clear-cut way. Such ontology provides a vocabulary for representing knowledge about a domain and for describing specific situations in the domain as it shares the context information in a pervasive computing domain and include machine-understandable definitions of basic concepts in the domain and relations among them.
The frequent keyword pattern identified from the FP-Growth and the relationship identified among the set of keywords helps in providing semantically meaningful representation to the web documents.

2. RELATED WORK
In this section, a review of previous work on context based retrieval system is discussed. The technique for the information retrieval in [4] creates the ecology of ontology which stores the ontological representation of the crawled documents. Further, the efficiency of ontology based indexing is enhanced as compared to the traditional key-word based retrieval system. The literature survey reveals there have been few significant attempts to merge information retrieval and ontological models, [5] proposed a text processing system to build ontological domain.
In the research paper [6], an architecture has been proposed for ontology based semantic web crawler that can exploit the semantic metadata to efficiently discover and extract information from the Semantic Web and the Semantic matching between content of downloaded web page and ontology is used to guide the crawler towards relevant information. The systems with context models, are discussed in [7], represents context in form of attribute-value tuples. [8] represents context in Gaia system as first-order predicates written in web

ontology language. Existing formal context models support formality and address a certain level of context reasoning. Though, none of them has addressed formal knowledge sharing, or has shown a statistical evaluation for the viability of context reasoning in pervasive computing environments, where we always have to face resource-constraint devices. Further research in, [9] present an ontology-based formal context model. It addresses issues including formal context representation, knowledge sharing and logic based context reasoning. Through performance analysis, it reveals quantitative evaluation for context reasoning in pervasive computing environments. The research paper [1] discussed the technique for providing semantic structure to the HTML documents to store it in the knowledge base as predicates, which simplify the task of providing more useful results to the user.

The critical look at the available literature reveals that the crawled web pages should be organized in a fashion where they are more understandable to machine, for producing the results which are meaningful and relevant.

Frequent patterns of the keywords, to find the contextual sense of the document can be achieved through Apriori algorithm, but FP growth improve the pitfalls of Apriori, that is candidate generation method to extract the frequent patterns. The drawback of Apriori is that, it needs to generate the huge number of candidates, so it required scanning the database and checking a large set of candidates repeatedly. The interesting method FP Growth can be used to reduce the drawback of the Apriori that mines the complete set of frequent item sets by pattern matching. The depth first searching method to find the matching patterns decreases the time complexity for searching.

3. PROPOSED WORK

The simplified architecture of context based crawler is represented in figure1. The context based crawler involves three tiers: the user tier, the resource tier and the context deduction interface tier. In this particular framework the resource tier consists of information collection from different resources and added ontology by the CBC. The context deduction interface tier is the semantic processing unit, which indexes the user query to the matches the relationship existing between them.

The overall architecture of the Context Based Crawler is represented in the figure 2. The architecture can be divided into two parts:

1. Firstly, the basic crawler which performs the task of retrieving the web documents according to the seed URLs are submitted initially. It includes the modules as: URL dispatcher, mapping manger, DNS resolver, crawl worker, URL-IP database and repository.

2. The second part performs the task of extracting contextual sense of the crawled web pages stored in the repository by the crawler. The major components of this part are: XML convertor, F-P keyword generator, ontology generator, context and synonym identifier, thesaurus and knowledge base.

Description of the various modules

i) URL Dispatcher

This component read the URL database and fills the URL-IP Queue, it may also be initiated by the user, who provides a seed URL in the beginning.

ii) DNS Resolver

The resources on the web are known by domain name server i.e. URL. The name of the domain name server must be translated into an IP address before crawler can communicate with server to downloading the resource. The DNS resolver uses this service offered by internet and returns back to calling mapping manager.

iii) Mapping Manager

Mapping manager gets a URL-IP set from the URL-IP Queue as input, creates multiple instance of mapper threads as URL Mappers.

iv) Segregator

This component identifies the XML/RDF/HTML documents crawled by the crawl worker and forwards the HTML documents to XML convertor to extract knowledge from the documents which will further help in generating the ontologies for the different domains. It downloads the .TOL and robots.txt file for resolved URL and segregates the internal and external links. It starts downloading pages for the URL in the queue. The RDF/XML pages are forwarded for the predicate generator .and the hyper text URLs are submitted for converting them into XML format.

v) XML converter

The preprocessed HTML document, are transformed to the XML format of the document is generated with the tool Light HTML to XML converter [8].

vi) F-P keyword generator

From the keywords extracted from the web documents stored in the repository , frequent keywords patterns are identified. The technique of F-P growth is used [2], which is quite common for the analysis of frequent item sets. The frequent patterns generated assist in finding the contextual sense of the web pages, as represented in figure 3.

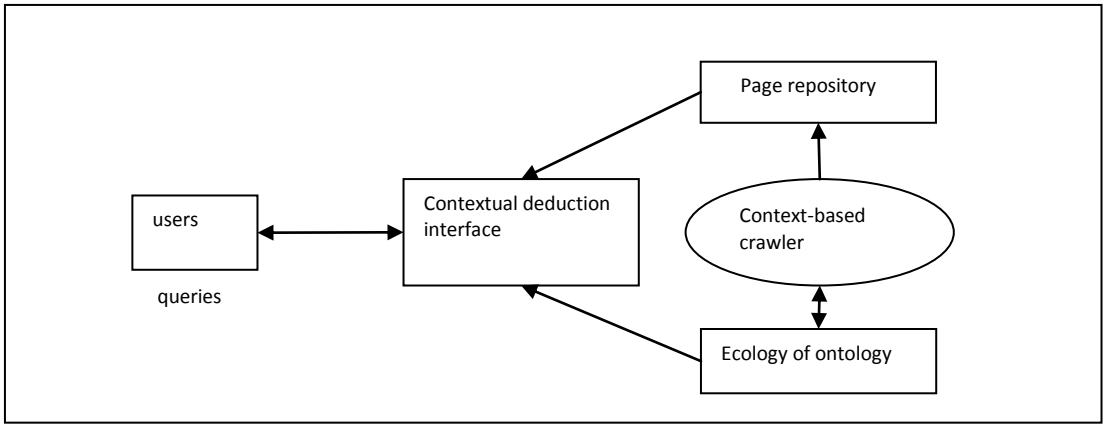

Figure1: Context Based Web Crawler Deployment Diagram

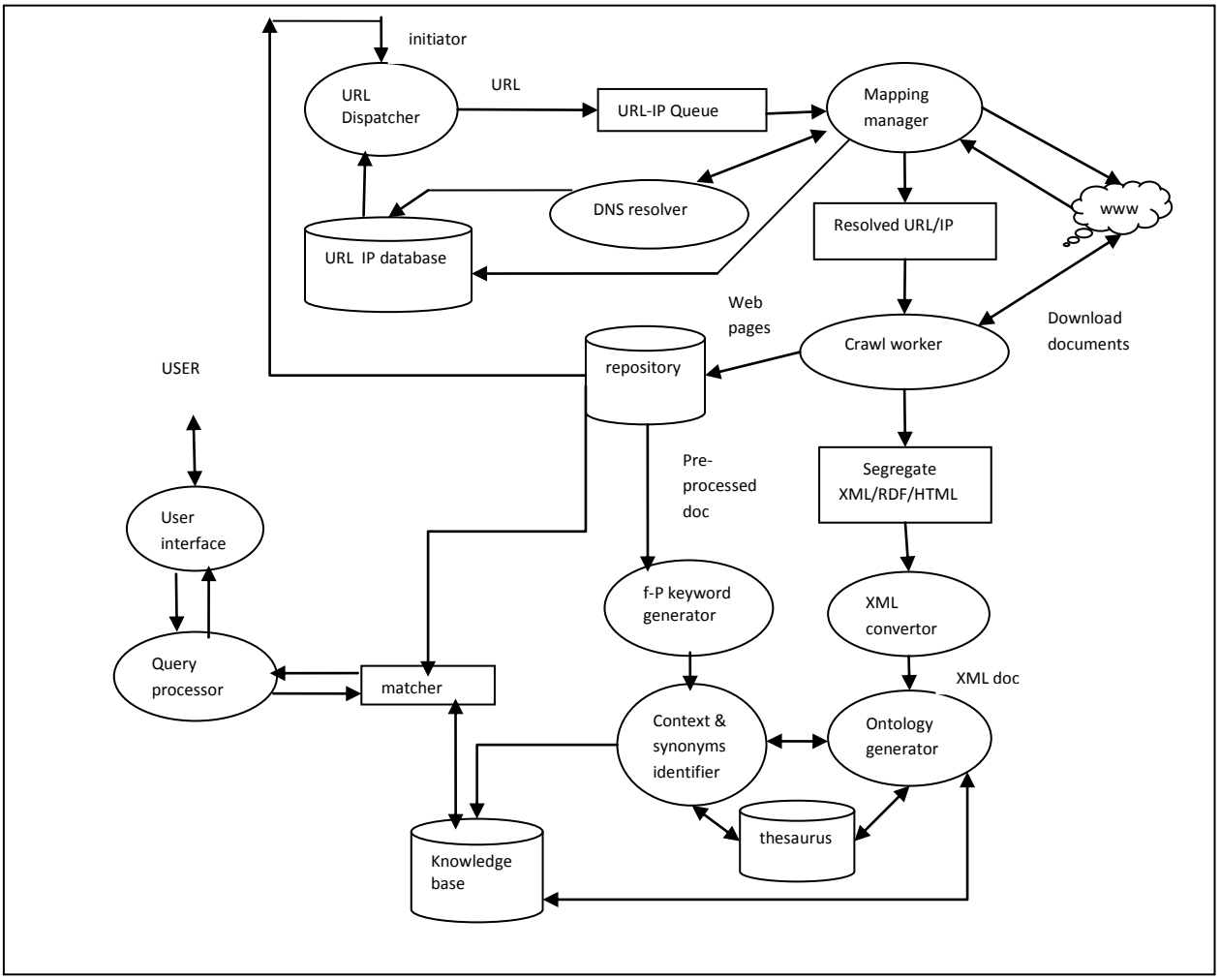

Figure 2: Architecture of the Context Based Crawler

Algorithm

Begin
Read a URL from the set of seed URLs
DNS resolver translate URLs into corresponding IP-address
Mapping manager gets a URL-IP set from the URL-IP Queue
Download the document
Save in repository
Begin
Segregate the XML/RDF/HTML documents
Convert all documents into XML
Generate the ontologies
Extract keywords from web documents
Begin
Identify support count for keywords
Discard keywords below threshold value
Construct FP tree

Generate frequent patterns using FP growth
End
Identify the context of document using thesaurus
Update contextual ontology
Store in knowledge base
End
End.

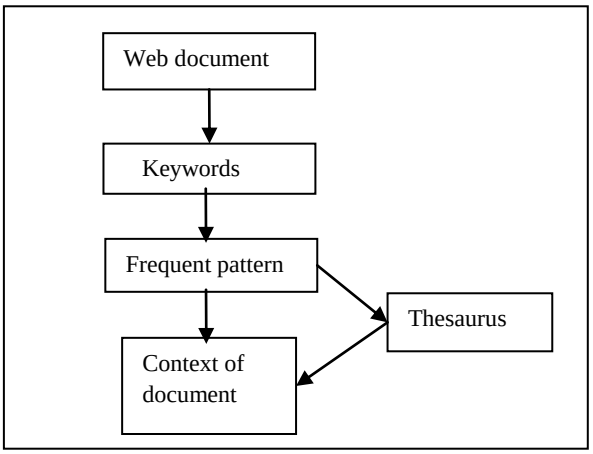

Figure 3: Representation of extracting context from web document.

4. EXAMPLE

Suppose nine documents are retrieved from the web. The keywords extracted from the documents are as shown below in table 1. Keywords are assigned the ID and corresponding support count is calculated for each as shown in table 2.

Document	Keywords
Document 1	Flow , Current , stream
Document 2	Current, clock
Document 3	Current, ions
Document 4	Flow, current, clock
Document 5	Flow, ion, grid
Document 6	Current, ion
Document 7	flow, ion, energy
Document 8	Flow, current, ions, stream
Document 9	Current, flow, ions

Table 1: Set of keywords extracted from documents.

Keyword	***Keyword ID***	***Support Count***
Flow	K_1	6
Current	K_2	7
Ion	K_3	6
Clock	K_4	2
Stream	K_5	2
Grid	K_6	1
Energy	K_7	1
Hours	K_8	1

Table 2: Set of keywords with the corresponding keyword-ID and support count.

If the minimum support count is 2, keywords with less value are discarded. F-P tree as shown in figure 4 is constructed using the contents of table 3 which maintain keyword-ID, support count and document-ID.

Keyword	***Support Count***	***Document s***
K2	7	Doc. 1,2,3,4,6,8,9
K1	6	Doc. 1,4,5,7,8,9
K3	6	Doc. 3,5, 6,7,8,9.
K4	2	Doc. 2,4
K5	2	Doc. 1,8

Table 3: Relevant keywords for FP tree (with support count and document-ID).

The FP-tree is mined by creating the conditional pattern bases and frequent pattern are generated as {K2, K1, K5} and {K2, K1, K3}. As the documents are to be organized on the bases of the hyponyms, the context of the frequent patterns derived from the above technique.
The set of keywords {K2, K1, K5} is related to the hyponyms ‘current’ having the contextual sense ‘water current’ and the keyword set {K2, K1, K3} is also related to hyponyms ‘current’ having the contextual sense ‘electric current’.

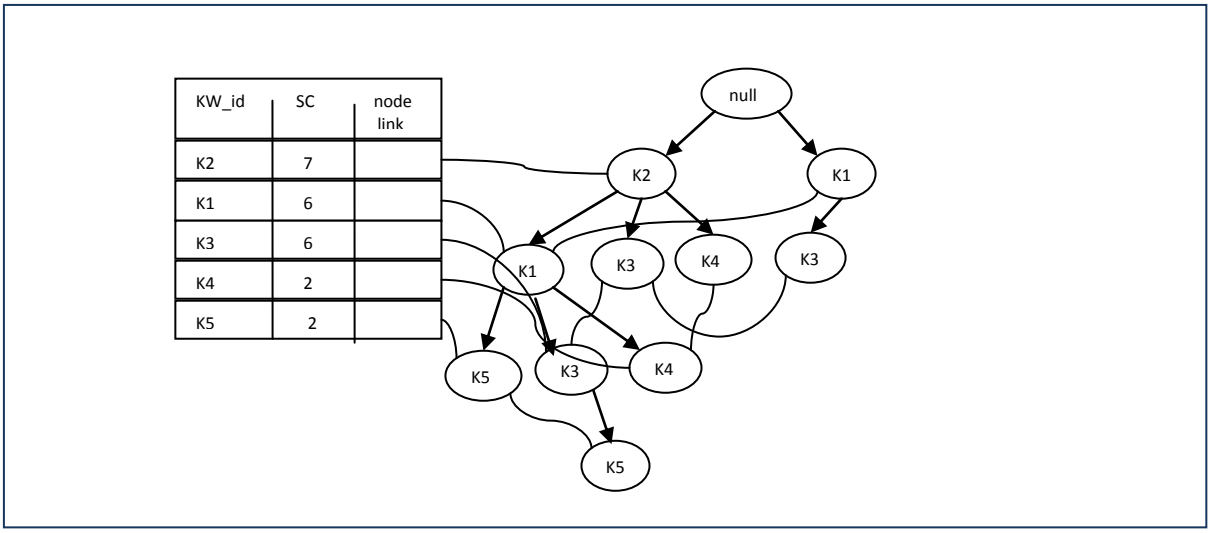

Figure 4: F-P tree.

5. CONCLUSION

The crawled web pages should be organized in a fashion where they are more understandable to machine, for producing the results which are meaningful and relevant. The set of web pages can be categorized into different contextual sense if the crawler has the technique to understand their meaning and the domain identification. This task is achieved through a technique of data mining referred as FP- Growth in the proposed work. The benefits of FP-growth can be summarized in two points: Completeness and compactness. Completeness, as it preserves complete information for frequent pattern mining for the keywords and it never break a long pattern. And compactness, because it reduces irrelevant info— infrequent items are gone, items in frequency descending order: the more frequently occurring, the more likely to be shared, is never larger than the original set of keywords (database).

This strategy is divide-and-conquer based and further leads to focused search as it represents the compressed form of the database(FP-tree structure) and does not require repeated scan of entire set of keywords.

REFERENCES

[1] Nidhi Tyagi , Rahul Rishi and R.P. Agarwal,"Semantic Structure Representation of HTML Document Suitable for Semantic Document Retrieval", International Journal of Computer Applications (0975 – 8887) Volume 46– No.13, May 2012.

[2] Jiawei Han and Micheline Kamber, "Data Mining Concepts and Techniques", Elsevier publication, Second edition, 2007.

[3] Ahmed Ab. Arara and Robert Laurini, "Formal Contextual Ontologies for Intelligent Information Systems", World Academy of Science, Engineering and Technology 11, 2007.

[4] Nidhi Tyagi, Rahul Rishi and R.P. Agarwal,"Contextual Ontology: A Storage Tool for Extracting Context from Web Pages", International Journal of Computer Applications (0975 – 8887) Volume 56– No.7, October 2012.

[5] L. Weihua, "Ontology supported intelligent information agent", proceeding IEEE Symp, on Intelligent Systems, pages383-387, IEEE, 2002.

[6] Ram Kumar Rana and Nidhi Tyagi," A Novel Architecture of Ontology-based Semantic Web Crawler", International Journal of Computer Application, Volume 44– No18, April 2012.

[7] A.K.Dey, et al. "A Conceptual Framework and a Toolkit for Supporting the Rapid Prototyping of Context- Aware Applications", Human- computer Interaction Journal, Vol. 16(2-4), pp. 97-166, 2001.

[8] Anand Ranganathan, et al. "A Middleware for Context- Aware Agents in Ubiquitous Computing Environments",USENIX International Middleware Conference, 2002.

[9] Xiao Hang Wang, Da Qing Zhang, Tao Gu1 and Hung Keng Pung "Ontology Based Context Modeling and Reasoning using OWL", Second IEEE Annual Conference on Pervasive Computing and Communications Workshops, 2004.