International Journal of Information and Communication Technology Research



http://www.esjournals.org

# Context Semantic Preprocessing for Indexing in Information Retrieval Systems

**Thinn Lai Soe, Kay Khaing Win**
University of Technology (Yatanarpon Cyber City), Pyin Oo Lwin, Myanmar

## ABSTRACT

Information retrieval (IR) is concerned with representing, searching, and manipulating large collections of electronic text and other human-language data. IR systems and services are now widespread, with millions of people depending on them daily to facilitate business, education, and entertainment. For these increasing amounts of information, we need efficient and effective index structure when we have to find needed information from the web. Therefore, indexing is one of the main parts of Information Retrieval system. IR system cannot well work without an accurate and efficient index. The role of semantic is the most important part of IR system because of the advance of intelligence system. Preprocessing step is also important part of indexing in IR system. This paper produces context semantic preprocessing using the concept of context ontology to provide indexing of IR system. This system can efficiently index the clusters with the aid of semantic and context preprocessing.

**Keywords:** *indexing, clustering, semantic and context, IR system*

## 1. INTRODUCTION

Every day, millions of people use the internet to answer questions of they wanted to know. Unfortunately, at present, there is no simple and successful means to consistently accomplish this goal. One common approach is to enter a few terms from a question into a Web search system and scan the resulting pages for the answer. Thus it is just a laborious process. Information retrieval (IR) is finding material (usually documents) of an unstructured nature (usually text) that satisfies an information need from within large collections (usually stored on computers) [1]. Regular users of Web search engines casually expect to receive accurate and near-instantaneous answers to questions and requests merely by entering a short query — a few words — into a text box and clicking on a search button. Underlying this simple and intuitive interface are clusters of computers, comprising thousands of machines, working cooperatively to generate a ranked list of those Web pages that are likely to satisfy the information need embodied in the query [2]. But IR system has so many problems. One of these is lack of semantic of a word. IR model governs how a document and a query are represented and how the relevance of a document to a user query is defined [14]. There are four main models: Boolean model, vector space model, language model and probabilistic model. But these models based on keyword or term matching, i.e., directly matches terms in the user query with those in the documents. If a user query uses different words from the words used in a document, the document will not be retrieved although it may be relevant because the document uses some synonyms of the words in the user query.

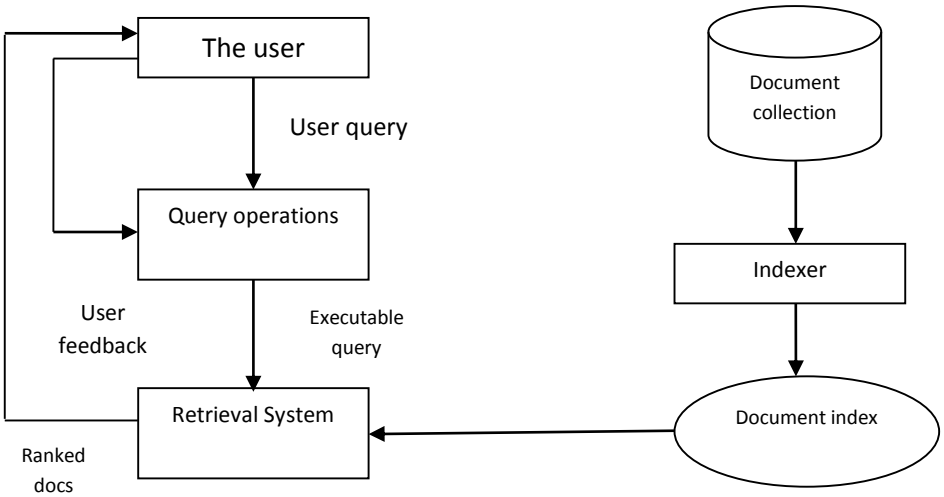


**Figure 1 Information System Architecture**

The following is the explanation of IR system components and brief description of the above figure. Text Operations forms index words (tokens) which includes stopword removal and stemming. Indexing constructs an inverted index of word to document pointers. Searching retrieves documents that contain a given query token from the inverted index. Ranking scores all retrieved documents according to relevance metric. Figure 1 is the architecture of the information retrieval system. In figure, Figure 1, we can easily see the main part of the IR system which is indexer. Some systems use the clustering method which is one of data mining method when the role of indexing needs. Preprocessing step also helps the indexer. When indexer doesn't use the preprocessing step, the indexer takes into account all of the words which may be no semantic content words or meaningful words included in the document. It may decline the quality of retrieval results. Moreover, such systems do more several works than the other system which does preprocessing. The most important advantage is that it may improve the percentage of recall.

This paper presents the preprocessing part of IR system that is based on the concept of context ontology. The preprocessing step of this system is quite difficult from the others. This paper is organized as follows. Related works are presented in Section 2. Section 3, describes the proposed system and advantages of the system are submitted in Section4 and then conclusion is in Section 5. In section 6, the limitation of this system is presented.

## 2. RELATED WORK

In this paper, a review of previous works on preprocessing step that supports in indexing of retrieval system is shown. There are many techniques that have been proposed already but such techniques produce inefficient and inaccurate results.

In paper [3], this system constructs the semantic suffix tree clustering algorithm (SSTC for short) and it is used in information retrieval system. Before constructing suffix tree, this system first does the preprocessing process. Preprocessing of this system consists of removal of stopwords, stemming, handling of digits, hyphens, punctuations, and cases of letters. Stopwords are frequently occurring and insignificant words in a language that help construct sentences but do not represent any content of the documents. In many languages, a word has various syntactical forms depending on the contexts that it is used. For example, in English, nouns have plural forms, verbs have gerund forms (by adding "ing"), and verbs used in the past tense are different from the present tense. These are considered as syntactic variations of the same root form. Such variations cause low recall for a retrieval system because a relevant document may contain a variation of a query word but not the exact word itself. Stemming refers to the process of reducing words to their stems or roots. A stem is the portion of a word that is left after removing its prefixes and suffixes. Preprocessing can reduce the number of nodes in the suffix tree but the phrase does not lose meaning. But this paper is quite different from the traditional preprocessing process. This paper uses the context ontology for preprocessing. Therefore, it doesn't need to do stopwords removal and stemming used in other system. The detail explanation is explained follows.

## 3. PROPOSED SYSTEM

This section presents the proposed system and detail explanation of the ontology that aids in preprocessing.

### 3.1 Proposed Preprocessing of the System

This section describes the proposed preprocessing step of the system. Preprocessing is the step which aids every in information retrieval system. Preprocessing steps for the text are removal of stopwords, stemming, removing digits, hyphen, punctuation marks and the case of letters. The following is the explanation of the system. This system helps the information retrieval system and it is fast in searching of the system after the system has used the preprocessing. First of all, there are HTML tags for web page which are served for input. These input HTML tags are passed through the HTML parser. Parsing or syntactic analysis is the process of analyzing a string of symbols, either in natural language or in computer languages, according to the rules of a formal grammar. A parser is a software component that takes input data (frequently text) and builds a data structure – often some kind of parse tree, abstract syntax tree or other hierarchical structure – giving a structural representation of the input, checking for correct syntax in the process. Parsing may imply simple terms extraction or it may involve the more complex process of tidying up the HTML content in order to analyze the HTML tag for context and meaning. HTML parser is used to parse the web page. HTML parser is used to parse the input HTML page, which produces a stream of tokens or terms to be

indexed. The following is the diagram for the preprocessing system which is used in IR system for celebrity domain.

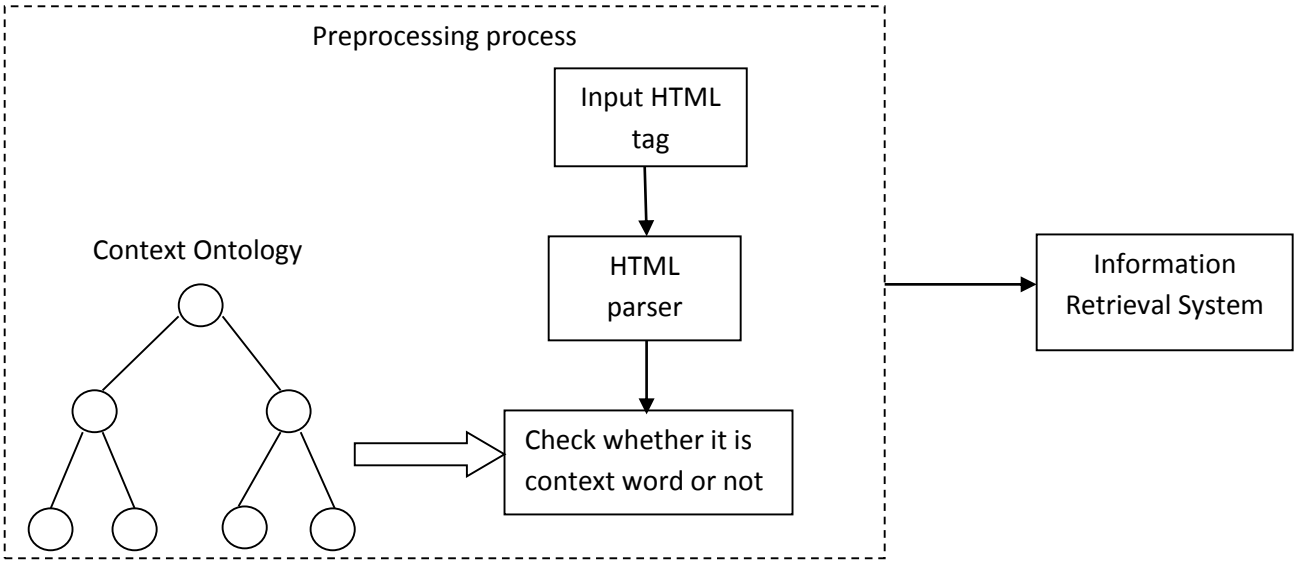


**Figure: Proposed System architecture**

After parsing, the system tests the parsed words and the context words which are in context ontology for the celebrity domain so as to know whether these words are equal or not. These steps are included in the preprocessing step of this system. E.g. For the perfume "Jennifer Lopez" which context names are "Glow after dark", "Blue Glow" is compared with the words in web page. E.g. For the singer "JLo" whose context names are "Dance again", "Papi", "Let's get loud", "On the floor" etc is also compared with the words in web page. If that words from the web page equal with context words in context ontology, these words don't do nothing and that web page are marked the name as that context name.

After naming, those web pages are stored in a specified place. After that, the next web page is taken for preprocessing described earlier. As presented above, the web page, which contained the context word 'On the floor', are stored in the file path named 'JLO/Song/ etc. In order to test whether they equal or not, the system use the edit distance for similarity. For the IR system, there are so many input queries. When the input query comes, the IR system tests the input words whether it equal with the words in the file that stored in the specified file path. If they are same, IR system lets these files to show the user.

### 3.2 Ontology Representation

This section describes the ontology representation of the system which aids in preprocessing of the system. The word "ontology" seems to generate a lot of controversy in discussions about AI. It has a long history in philosophy, in which it refers to the subject of existence. It is also often confused with epistemology, which is about knowledge and knowing. In the context of knowledge sharing, ontology means a *specification of a conceptualization*. That is, ontology is a description (like a formal specification of a program) of the concepts and relationships that can exist for an agent or a community of agents. This definition is consistent with the usage of ontology as set-of-concept-definitions, but more general [4]. Ontology can be viewed as the backbone to support various types of information management including IR, storage and sharing on web. It is used to reason about the entities within the domain, and may be used to describe the domain [6]. Context ontology include context of the term,

that term which create context, their relationships among context and the term and other terms which support to make the context. Ontologies often contain a model of a domain, its taxonomy the relationships between its entities. Context ontology defines a common vocabulary to share context information in a pervasive computing domain [5]. The classes alone will not provide enough information to search from the user desire questions. Once we have defined some of the classes, we must describe the internal structure of the concepts. In general, two types of properties are distinguished in OWL classes:

- Data type properties, relations between instances of classes;
- Object type properties, relations between instances of two classes;

This system constructs the celebrity ontology for the celebrity domain. In the foreign country, they produce the product like perfume, shower gel, body lotion, earphone and even sunglasses with the name of the famous celebrity. So, the name of one celebrity name can be the information of the real celebrity and the information of the product name that produces their name. The system constructs the ontology named "Celebrity" and its class name is "Person". Thus, the person class in this system has twenty-one data type properties. Some of which are earphone_url, earphoneName, filmName, filmName_url,footballClub, football_url etc. The data types for these data type properties are undefined. The following is the figure for the celebrity ontology and that ontology helps in the preprocessing of the system.

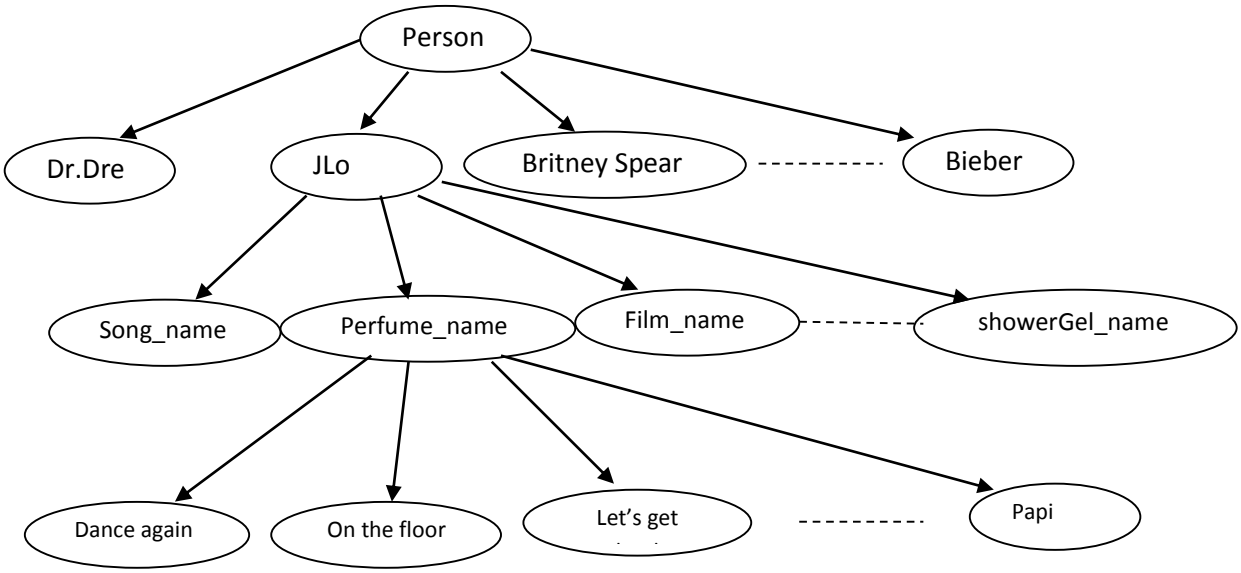


**Figure: Ontology Representation for the system**

## 4. ADVANTAGES OF THE SYSTEM

This section presents the advantages of the system. First of all, this system does not need to do any extra work such as removal of stopwords, stemming and other preprocessing tasks. Therefore, the time and complexity for doing such tasks are reduced by using that system. There is no need to store stopwords file and other algorithm like poter's stemmer. And the next is it gives more accurate search results than other IR system which does not use context ontology. Moreover, this system can give users that they really want either songs or perfume. And the third advantage of this system is it can solve IR problem. IR system has two problems which are synonymy (means that multiple words having the same meaning) [5] and polysemy (means a

word has multiple meanings). This system can solve one of information retrieval problem that is polysemy by the aid of constructing context ontology in celebrity domain form Hollywood.

## 5. CONCLUSION

This paper presents the preprocessing steps of any web document by using the concept of ontology which is constructed only for the celebrity domain. The information of the products using the names of celebrities and the information of them are used so as to construct context ontology. The advantages of construction of context ontology is that it can solve one of information retrieval problems that is polysemy (a word with different multiple meanings) and it cannot depend on any other system like thesaurus which can be taken online from thesaurus.com. Thus, this paper can be more faster time preprocess time others and can give more aid in indexing system of the information retrieval system.

## 6. LIMITATION OF THE SYSTEM

This system can only be used in searching the information about celebrity domain of information retrieval system. Furthermore, it needs to construct domain ontology so as to use this system. And the next one it needs to know is this system has to collect the information of all celebrities.

## REFERENCES

[1] Christopher D. Manning, Prabhakar Raghavan, Hinrich Schütze, Introduction to Information Retrieval.

[2] http://en.wikipedia.org/wiki/InformationRetrieval

[3] Janruang, J., Guha, S.: Semantic Suffix Tree Clustering. In: DEIT 2011, IEEE, Bali, Indonesia (2011).

[4] http://www-ksl.stanford.edu

[5] http://en.wikipedia.org/wiki/Synonym

[6] http://en.wikipedia.org/wiki/Ontology

[7] Soe, T.L, "Context based Indexing to Support Search Engline", 11th International Conference on Computer Application, Yangon, Myanmar.

[8] Soe, T.L, "Semantic Clustering with Context Ontology for Information Retrieval System", International Journal of Computer (IJC) (ISSN 2307-4523) Vol 11, 2013.

[9] Simone Santini and Alexandra Dumitrescu "Context based semantic data retrieval".

[10] Naresh Kumar Nagwani, Dr. Shrish Verma, "Software Bug Classification using Suffix Tree Clustering (STC) Algorithm"IJCST VO l. 2, Issue 1, March [11]

[11] Sajendra Kuar, Ram Kumar Rana, Pawan Singh, "A Semantic Query Transformation Approach Based on Ontology for Search Engine", International Journal on Computer Science and Engineering (IJCSE), May 2012.

[12] V'aclav Sn´aˇsel, Pavel Moravec, Jaroslav Pokorn, "WordNet Ontology Based Model for Web Retrieval".

[13] N Chen, Technical Report 2006-505 "A survey of Indexing and Retrieval of Multimodal Documents: Text and Images".