International Journal of Engineering Research and Applications (IJERA) ISSN: 2248-9622
International Conference on Industrial Automation and Computing (ICIAC-12th & 13th April 2014)

RESEARCH ARTICLE OPEN ACCESS

Ontology Based Approach for Domain Specific Semantic Information Retrieval System

Ms. Pratibha S. Sonakneware*, Prof. S. J. Karale**
*(Department of Computer Technology, Yeshwantrao Chavan College of Engineering, Nagpur (Maharashtra), India.
Email: psonakneware@gmail.com)
** (Department of Computer Technology, Yeshwantrao Chavan College of Engineering, Nagpur (Maharashtra), India.
Email: sjkarale@in.com)

ABSTRACT
In today's era, there is rapid growth of resources in the Internet. Therefore, improving the information retrieval technology has become important. The existing information retrieval technique is based on the keywords search. In keyword based system most of the times, irrelevant information is retrieved. It consumes more time than required. In order to overcome the limitation of keyword based approach, ontology based semantic search approach is implemented. In ontology based approach conceptual information retrieval is done which means that the system interprets the meaning of users' query and finds the relationship between different concepts and then retrieves the semantic information. This paper describes the information retrieval system in which user gives a natural language query which the system interprets as input query and extracts the semantic information by using ontology based approach. For conceptual search, user's query is expanded to SPARQL query using quepy tool and this SPARQL query is fired on knowledge base which is in the RDF data format thus retrieving the relevant answer. The IR system finds out semantic query terms for query expansion by using WordNet.
Keywords **-** Information Retrieval, Ontology, Quepy, RDF, Semantic Search, WordNet .

I. INTRODUCTION

There is a very huge number of documents in the World Wide Web, which are growing at a steady pace. Out of all of them finding a relevant one, which one needs, is a challenging task. To search the required document, traditional keyword-based information retrieval technique is used. In keyword based system, user is provided with too many results among which, most results are irrelevant. User faces difficulty in figuring out the exact one which he needs. Therefore, to overcome that limitation of keyword based system, technique of conceptual search is implemented [1], [2]. Conceptual search implies search by meaning instead of matching of keywords. In conceptual search technique, the system understands the meanings of the concepts, finds the relations between concepts that users specify in their queries and then retrieves the semantic answer [2]. As an example, consider the meaning of the word 'can' - taken either as 'container for storing water' or as 'be able to'. The keyword based system retrieves data for both the meanings. On the other hand, the ontological concept retrieves concept from our domain. This conceptual search technique is implemented by using the concept of ontology.

In the field of computer science and information technology, philosopher Gruber defines the concept of ontology as "an explicit specification of conceptualization". It means that ontology describes the relationship between different concepts, properties and their attributes. In semantic search system, the concept of ontology is used to search results by contextual meaning of input query instead of keyword matching [19].

Ontology provides a knowledge-sharing framework that supports the representation and sharing of domain knowledge [2]. An increasing number of ontologies are being developed, and their reuse and sharing offers several benefits. One important benefit is that we can significantly save time and effort by reusing existing ontologies instead of building new ones every time. Another advantage is that heterogeneous systems and resources can interoperate seamlessly by sharing a common knowledge [21].

In the proposed system, the meaningful concept is extracted from user's input query. Using this concept, query expansion is performed. Query expansion implies that the query is converted into more meaningful format. In the proposed system, input query is converted into a SPARQL query. SPARQL is an RDF database language. SPARQL query is then fired on to the RDF database and accesses the relevant information [15].

II. ONTOLOGY AND SEMANTIC VISION FOR INFORMATION RETRIEVAL

2.1 ONTOLOGY

Ontology is a framework for semantic search. It is the excellent semantic retrieval tool for

concept hierarchy and appropriately supports logical reasoning. In the field of Artificial Intelligence, ontology is defined as "the basic terms and relations comprising the vocabulary of a topic area, as well as the rules for combining terms and relations to define extensions to the vocabulary"[21].
Ontology has a vital role in accessing and interchanging the information, use and reuse of knowledge, sharing of information and common understanding of specific domain are communicated among people for developing their applications. Ontology can describe entities, their objects, and properties of these objects and shows the relationship in a specific domain in a way that computers can process and automate.

2.2 SEMANTIC VISION

The semantic search extends the current World Wide Web by improving the facilities of information retrieval by contextual meaning. In order to retrieve semantic information, ontology occupies a central role for semantic exchange of information. Logic of semantic search in information retrieval system has knowledge to present concepts and their interrelationship in semantic view . Semantic search system is a vocabulary of information retrieval system to show the concepts and their interrelationship and then show the corresponding source of that concept.
In ontology based semantic information retrieval system, the data is retrieved from different hierarchies, which means that from classes, subclasses, properties and instances of properties. After semantic concept extraction, IR system sorts the results according to relevancy and retrieves the semantic information [20].

III. PROPOSED SYSTEM

The proposed architecture of our semantic Information Retrieval system as shown on figure: 1. In our Information Retrieval system user gives the input query to IR system. That natural language query is processing into query processing module. In that query processing module query is passed from query handling and semantic analysis phase [17]. Domain specific semantic information retrieval system for retrieving appropriate data goes through following steps:

- User gives the natural language input query to IR system.
- Indentifying the concept, i.e., type of question is classified.
- Tokenizing the input query where key domain terms are extracted and prunes the irrelevant terms.
- Extraction of equivalent key terms from ontology.
- Expansions of extracted key terms with SPARQL query language. SPARQL is an ontology query language.
- Mapping of SPARQL query with RDF database and retrieving the conceptual terms and identifying the context relation.
- Retrieving the semantic answer

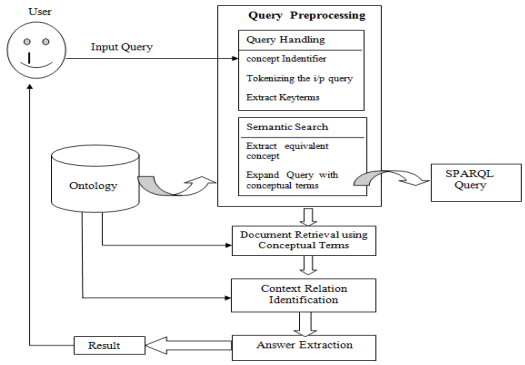

Figure 1: Architecture of Proposed IR System

3.1 QUERY HANDLING

The query given by the user through user interface is handled by query handling module. The meaningful concepts are identified means that the types of question are classified. After that meaningful concepts are extracted through the string tokenization concept from the input query and extracted the key terms. Then key terms are expanded through semantic query expansion by using WordNet [8]. WordNet is used for find the synonyms of the words. The natural language question given by the user is then semantically expanded to SPARQL query language by using a quepy tool. The quepy convert query to SPARQL semantically by using NLTK_DATA contains WordNet.

3.2 SEMANTIC ANALYSIS

In semantic search phase extract the meaningful concept from ontology. Ontology is in the form of RDF database. SPARQL is a query language for RDF database. Jena API is used for manipulating the SPARQL queries. SPARQL is analyzed the answer but search the data which is related to input query in knowledge base. SPARQL is a query language for retrieve the appropriate RDF data [14]. SPARQL match the prefix i.e. namespace with RDF namespace then search data from that URL properties. SPARQL query is map with RDF through JENA and retrieve the answer.

3.3 DOMAIN ONTOLOGY CREATION

Ontology has been created for a particular domain and is used to model the knowledge for this domain in terms of Concepts (various terms of a specific domain) Relationships between concepts . Ontology shows the hierarchical relationship between different

classes and their subclasses in graphical pattern as shown in figure 2.

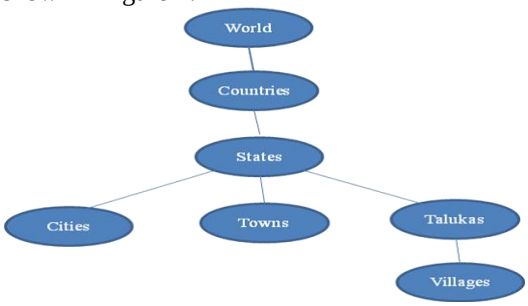

Figure 2: ontology showing 'World' hierarchy in the graph format

In the figure 2 class hierarchy has been shown. 'Village' is a sub class belonging to its super class 'Talukas' which is again a sub class to 'states'. Likewise 'Cities', 'Towns' are also subclasses of class 'State' which belongs to super class 'Countries'.

3.4 INFORMATION EXTRACTION

Answer extraction is a last module of the IR system, which is the understanding between semantic information retrieval system and the usual sense of text retrieval system. In answer extraction module, there is filtering of the metadata and generation of the answer. In answer extraction phase IR system identify the relevant concept, extract the concept then validate the relevant answer from extracted concept. The different stages of IE are:

- Identify: The answer candidates are identified within the filtered ordered paragraphs through parsing. We can use POS tagger for this after parsing the question. We also have heuristic measures as a good option
- Extract: The answer is extracted by choosing only the word or phrase that answers the submitted question through a set of heuristics. Researchers have presented miscellaneous heuristic measures to extract the correct answer from the answer candidates. Extraction can be based on measures of distance between keywords, numbers of keywords matched and other similar heuristic metrics.
- Validate: The answer is validated by providing confidence in the correctness of the answer. There are several ways to validate the final answer and it is always recommended to do so. One can use lexical resource like WorldNet to verify the correctness of final answer. Other source is specific knowledge resources. It can also be used to check questions belonging to specific domain. Even web search is a good option to validate the correctness for domain specific knowledge. The most attractive and easiest yet simplest technique is investigation using the redundancy of the web to validate answers based on frequency counts of question answer collocation.

IV. IMPLEMENTATION DETAILS

The proposed system is implemented with appropriate tools and the proposed techniques. The implementation details of our proposed system are described in short in the next sections:

1.1 SEMANTIC SEARCH USING QUEPY

In our proposed system semantic search of user input query is done by using SPARQL query. SPARQL is a query language for retrieving the semantic data from RDF database.

For creating the ontology, system use the database in RDF data format. RDF format shows the relationship between different object and their attributes. For RDF database language is SPARQL. SPARQL language is difficult to understand user as compared to natural language. So it is important to convert natural language query into SPARQL query [2]. Converting user input query, taken as natural language query into SPARQL query by using "Quepy" tool and interfacing between SPARQL query and RDF database by using Jena API.

Quepy is a tool for converting natural language input queries into SPARQL query. Quepy tool for generating database queries to be access through the dbpedia database. For accessing our own system database instead of dbpedia give the path of our own database. For accessing different more types of question we are adding the POS tagging in the different form of question templates of quepy program then accessing the our own database. Quepy is a python framework transfer natural language question into database language as follows.

Quepy is installed on ubuntu12.04. For checking installation successfully to show the quepy version

Figure 3: Installation of Quepy

As shown in figure 6 show the Input query is converted into SPARQL query. After giving the path of database input query converted to SPARQL query. In figure 7 shown that quepy is interfacing with UI.

```
onto@ubuntu: ~/Desktop/quepy-master
onto@ubuntu:~$ cd ~/Desktop/quepy-master
onto@ubuntu:~/Desktop/quepy-master$ python examples/dbpedia/main.py "who is  Tom
 cruise?"
who is  Tom cruise?
------------------

PREFIX owl: <http://www.w3.org/2002/07/owl#>
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX foaf: <http://xmlns.com/foaf/0.1/>
PREFIX skos: <http://www.w3.org/2004/02/skos/core#>
PREFIX quepy: <http://www.machinalis.com/quepy#>
PREFIX dbpedia: <http://dbpedia.org/ontology/>
PREFIX dbpprop: <http://dbpedia.org/property/>
PREFIX dbpedia-owl: <http://dbpedia.org/ontology/>

SELECT DISTINCT ?x1 WHERE {
  ?x0 rdf:type foaf:Person.
  ?x0 rdfs:label "Tom cruise"@en.
  ?x0 rdfs:comment ?x1.
}
```

Figure 4: NL query converted to SPARQL

1.2 ONTOLOGY CONSTRUCTION FOR ACADEMIC INSTITUTION

Protégé is a tool which creates data into RDF data format. We have used Protégé_3.4.8 tool to create ontology for Academic Institution.When we design Academic Institution Ontology, we go through the following steps as shown in figure 5. The first step is to collect all the details regarding Academic Institution. In second step, identify the classes and their subclasses. In third step, identify the properties of those classes and subclasses. In fourth step, create the instances of those properties. Then, develop and save the ontology. Finally, it is exported in RDF or OWL data format [18].

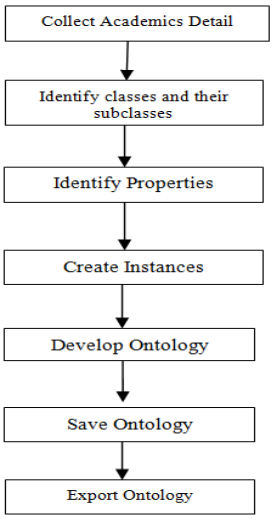

Figure 5: Steps for creating Academic Institution Ontology

4.2.1 CREATION OF CLASSES AND SUBCLASSES:

Academic Institution Ontology contains several classes and subclasses as shown in figure 6. The root node, Academic Institution, contains classes such as management group, departments, administrator and library. The class Department is divided into the branches viz. Computer Technology, Civil Engineering etc. Those subclasses are further divided into HOD, teaching staff, non-teaching staff and students.

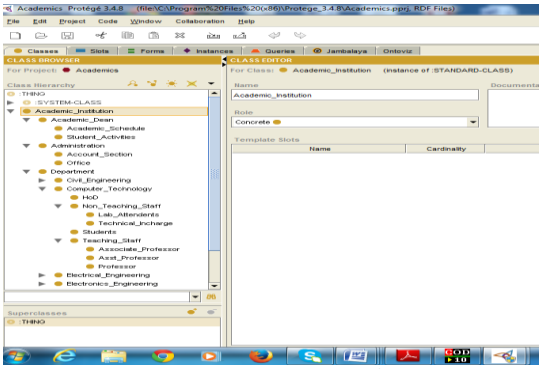

Figure 6: Classes and subclasses of Academic Institution Ontology

4.2.2 IDENTIFY PROPERTIES

Properties of different classes and their subclasses are as shown in figure 7. For example, the properties of subclass Professor are 'Name', 'Subject taught', 'Email ID' and 'Teaching experience' etc.

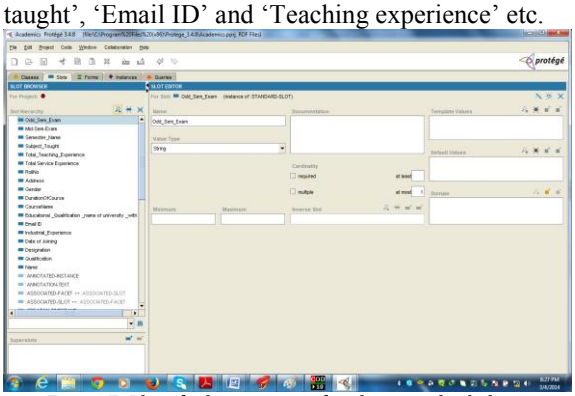

Figure 7: Identify the properties for classes and subclasses

4.2.3 CREATE INSTANCES

Create individual instances for different classes and subclasses as as shown in figure 8. Figure 8 shows an instance of subclass HOD with values of identified properties. For the property Name, give the value as 'A. R. Bhagat patil, Teaching experience is '22 years', qualification is 'Ph. D.', designation is 'professor'. Likewise give the values for properties of classes and subclasses.

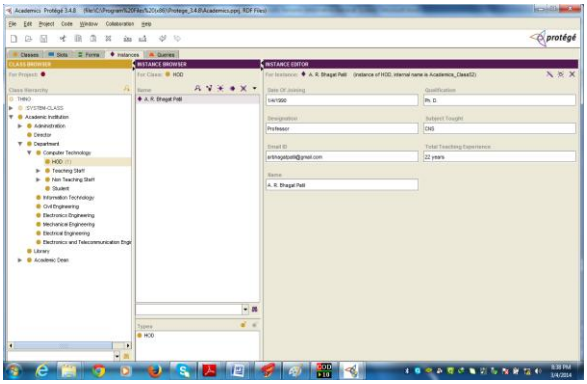

Figure 8: Create instances for HOD class

4.2.4 DEVELOPED ONTOLOGY

Hierarchical view of Academic Institution Ontology is shown in figure 9 and radial view of Academic Institution ontology is shown in figure 10. As shown in figure 9, blue line shows the relationships between classes and their subclasses, and red line shows the instances for classes and subclasses.

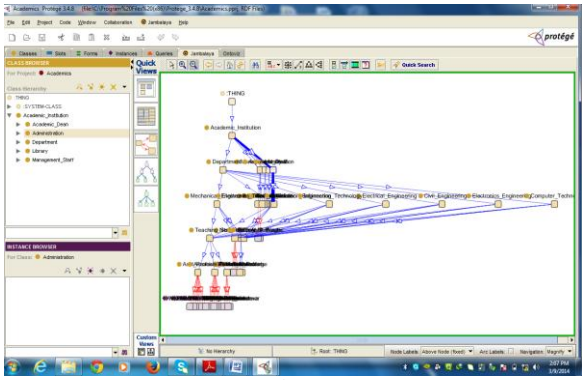

Figure 9: Heirarchical view of Academic Institution Ontology

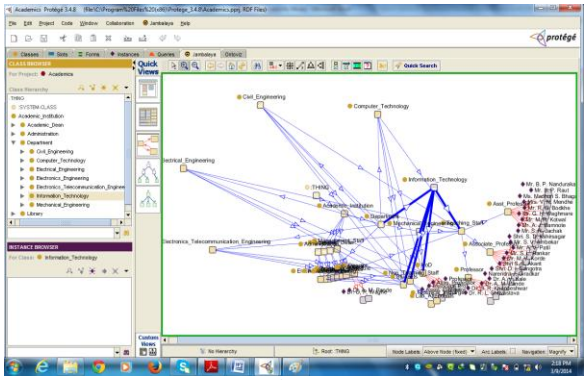

Figure 10: Radial view of Academic Institution Ontology

4.2.5 EXPORT ONTOLOGY IN RDF FORMAT

After export ontology in RDF format as shown in figure 11. As shown in figure 11 Ontology show the RDF schema. In KB take the values of properties identified for classes and subclasses.

```
</rdfs:Class>
<rdf:Property rdf:about="&kb;CourseName"
      a:maxCardinality="1"
      rdfs:label="CourseName">
     <rdfs:domain rdf:resource="&kb;Students"/>
     <rdfs:range rdf:resource="&rdfs;Literal"/>
</rdf:Property>
<rdf:Property rdf:about="&kb;Date_of_Joining"
      a:maxCardinality="1"
      rdfs:label="Date_of_Joining">
     <rdfs:domain rdf:resource="&kb;Associate_Professor"/>
     <rdfs:domain rdf:resource="&kb;Asst_Professor"/>
     <rdfs:domain rdf:resource="&kb;Director"/>
     <rdfs:domain rdf:resource="&kb;HoD"/>
     <rdfs:domain rdf:resource="&kb;Non_Teaching_Staff"/>
     <rdfs:domain rdf:resource="&kb;Principal"/>
     <rdfs:domain rdf:resource="&kb;Professor"/>
     <rdfs:range rdf:resource="&rdfs;Literal"/>
</rdf:Property>
<rdfs:Class rdf:about="&kb;Department"
      rdfs:label="Department">
     <rdfs:subClassOf rdf:resource="&kb;Academic_Institution"/>
</rdfs:Class>
<rdf:Property rdf:about="&kb;Designation"
      a:maxCardinality="1"
      rdfs:label="Designation">
     <rdfs:domain rdf:resource="&kb;Associate_Professor"/>
     <rdfs:domain rdf:resource="&kb;Asst_Professor"/>
     <rdfs:domain rdf:resource="&kb;Director"/>
     <rdfs:domain rdf:resource="&kb;HoD"/>
     <rdfs:domain rdf:resource="&kb;Non_Teaching_Staff"/>
     <rdfs:domain rdf:resource="&kb;Principal"/>
```

Figure 11: snapshot of RDF database

In this way by using protégé creating the ontological database in RDF data format. SPARQL query is mapped with RDF database by using jena API then SPARQL query is fired on RDF database then retrieve the relevent infotmation.

V. CONCLUSION

Thus, we have proposed the approach of Ontology based Information Retrieval system for overcoming the limitation of keyword-based information retrieval system. The proposed ontology based IR system extracts relevant information instead of giving list of all the documents containing related information. This system has been implemented using appropriate tools. Quepy tool is used for converting natural query to SPARQL query. SPARQL query is used to extract the RDF data with respect to user input query. Protégé tool is used for creating ontology in RDF data format. By using protégé tool, we have created ontology for Academic Institution. As of now, experimental result shows that our system retrieves answers to some specific types of queries. Future work will extend as IR system gives the answer to all the possible types of questions.

REFERENCES

Journal Papers:

[1]. Lakshmi Tulasi R., Goudar R.H., Sreenivasa Rao M., Desai P.D." Domain Ontology Based Knowledge Representation for Efficient Information Retrieval", *International Journal Of Information Systems And Communication*, 2012.

[2]. Miriam Fernández , Iván Cantador , Vanesa López , David Vallet , Pablo Castells , Enrico Motta ,"Semantically enhanced Information Retrieval: An ontology-based approach "*Web Semantics: Science, Services and Agents on the World Wide Web* (2011).

[3]. J.Uma Maheswari, Dr. G.R.Karpagam," A Conceptual Framework For Ontology Based Information Retrieval", Professor, Department of Computer Science and Engineering, PSG College of Technology, Coimbatore, Tamilnadu, *International Journal of Engineering Science and Technology* , 2010

[4]. Xiangsheng Yang, Ruoman Zhao, Chuan Zhang," An Ontology-based Framework for the Construction of Teaching Resource Library", Ningbo University.

[5]. SwathiRajasurya, Tamizhamudhu Muralidharan , Sandhiya Devi,Dr.S.Swamynathan," Semantic Information Retrieval Using Ontology In University Domain", Department of Information and Technology,College of Engineering,Guindy, Anna University,Chennai.

[6]. Christian Paz-Trillo , Renata Wassermann," An Information Retrieval application using Ontologies", Departament of Computer Science Institute of Mathematics and Statistics University of Paulo, Brazil, 3, 2005..

[7]. Tang Lijun, Chen Xu,"The Study of Semantic Retrieval Based on the Ontology of Teaching Management",*Advancedin Control Engineering and Information Science* CEIS 2011.

[8]. Sandhya Revuri, Sujatha R Upadhyaya , P Sreenivasa Kumar," Using Domain Ontologies for Efficient InformationRetrieval", Department of Computer Science and Engineering Indian Institute of Technology Madras Chennai - India.

[9]. Anatoly Gladun, Julia Rogushina, Victor Shtonda," Ontological Approach To Domain Knowledge Representation For Information Retrieval In Multiagent Systems", *International Journal "Information Theories & Applications" Vol.13*.

[10]. Zeng Dan, "Research on Semantic Information Retrieval Based on Ontology", Library of Wuhan University of Technology, Wuhan, P.R. China, 430070.

[11]. LIU Xiaoming, XU Jinzhong,LI Fangfang, "Domain-specific Ontology Construction from Hierarchy Web Documents", School of Computer Science and Technology, Beijing Institute of Technology, China..

[12]. F.Neches.R.;Finin.T.;Gruber.T;Enabling Technology for Knowledge Sharing;AI Magazine,1991,P36-56.

Proceedings Papers:

[13]. Rashmi Chauhan, Rayan Goudar, Robin Sharm, Atul Chauhan," Domain Ontology based Semantic Search for Efficient Information Retrievalthrough Automatic Query Expansion", Dept. of Computer Science and Engineering GEUDehradun, India. *International Conference on Intelligent Systems and Signal Processing (ISSP)* 2013.

[14]. Liu Xinhua, Zhang Xutang, Li zhongkai ," A Domain Ontology- based Information Retrieval ApproachforTechniquePreparation i*nternational Conference on Solid State Devices and Materials Science* 2012.

[15]. IT .Kanimozhi, Dr.A.Christy," Incorporating Ontology and SPARQL for Semantic Image Annotation" *Proceedings of 2013 IEEE Conference on Information and Communication Technologies (ICT)*, 2013.

[16]. Swathi Rajasurya , Tamizhamudhu Muralidharan , Sandhiya Devi, Nwe Ni Aung, Thinn Thu Naing," Sports Information Retrieval with Semantic Relationships of Ontology", University of Computer Studies, Yangon , *3rd International Conference on Information and Financial Engineering IPEDR* ,2011.

[17]. Thinn Mya Mya Swe," Intelligent Information Retrieval Within Digital Library Using Domain Ontology", Computer University, Mandalay, Myanmar. *International Conference on Applied Computer Science*.

[18]. Ayesha Ameen, Khaleel Ur Rahman Khan, B. Padmaja Rani,"Construction of University Ontology", *IEEE conference* 2012.

[19]. Guowei Chen,Pengzhou Zhang, "Keywords Retrieval Based On Ontology Inference", Communication University of China, *International Conference on Industrial Control and Electronics Engineering* 2012.

[20]. Rachid Ahmed-Ouamer, Arezki Hammache, "ontology-Based Information Retrieval for e-Learning of Computer Science", *IEEE conference* 2010.

[21]. B.Chandrasekaran;John R.Josephson; What Are Ontologies,and Why Do We Need Them? *IEEE Intelligent Systems*, [J], 1999. PP20-25