I J C T A, 9(17) 2016, pp. 8817-8825

Fuzzy Ontology Based Web Mining Approach for Extraction of Semantic Web Documents

Puneet Goswami* and Ramesh Chandra Sahoo**

ABSTRACT

Ontology represents relationships among set of terms and concepts in hierarchical fashion. Ontology plays crucial role in formulization of information related to given domain. Understanding these ontologies without having sufficient knowledge of ontology editors is like working on project without knowing its requirements. Traditional text mining methods and aero-text systems for extracting key phrases have been used but it needs to be improved to support large scale ontology constitution for real world applications. An ample amount of documents present on web puts the users in state of dilemma. Relevance means how closely the given query matches large number of documents.

The paper proposed fuzzy ontology based approach that retrieves information from web documents by using fuzzy relations and semantic context vectors. It discovers fuzzy ontology rather than textual descriptive ontology with crisp features only. The output membership fuzzy functions are produced by simulation tool named as MATLAB. The validation of proposed approach is done by evaluating information retrieval performance in two specific domains-weather domain (web pages containing information about weather forecasting and analysis) and Google TM collection (web pages containing news).

Keywords: Web Mining, Information Retrieval (IR), Ontology, Fuzzy Ontology Based Web Mining, and Semantic Web

I. INTRODUCTION

As the number of documents on web is increasing day by day, the methods of retrieving information from these documents are also growing massively. Various scientists and researchers are contributing towards the methods of information retrieval and machine learning. Online documents are composed of terms that are based on various extraction methods like vector approach, Bayesian, probabilistic approach etc. After evolution of ontology, we have gone through ontology methodology that analyses and classifies web documents. It was good but not best. It's representation of documents is not effective. To represent documents effectively, we have also viewed some probabilistic approaches like Bayesian Model. They are capable of finding probabilities among various terms and distinguish them as relevant or non relevant. This method does not tell about frequency of terms that are occurring in given document. So, there is need to use soft computing techniques to handle uncertainty caused by excessive number of documents on web. The techniques include fuzzy logic, neural networks, machine learning and many more. Ontology is abbreviated as FESC which means *Formal, Explicit, specification of shared conceptualization.* [7]. Formal specifies that it should be machine understandable. Explicit defines the type of constraints used in model. Shared defines that ontology is not for individual, it is for group. Conceptualization means model of some phenomenon that identifies relevant concept of that phenomenon. Building Ontology needs attention of domain expert that represents concepts and relations between them for a given domain. The proposed methodology builds fuzzy ontology for a given domain rather than generating standard ontology from textual databases. There are various uses of Ontology:

* Professor Computer Science & Engineering Department, SRM University Sonipat, India, *E-mail: goswamipuneet@gmail.com*

** Assistant Professor Computer Science & Engineering Department, NIET, Greater Noida, India, *E-mail- rsahoo22@gmail.com*

- Used for knowledge sharing and reuse.
- Can improve understanding between concepts.
- It is useful in Semantic Web that is information in machine form.
- Some search engines use ontology for finding relevant pages related to given query.

The paper is divided into following sections: Section 2 presents various literature studies conducted in context of fuzzy approach. Section 3 presents overview of semantic web and way of querying data in it. Section 4 depicts proposed fuzzy ontology approach and fuzzy output membership functions using MATLAB. Section 5 computes IR performance with/without fuzzy domain ontology. Section 6 concludes the given paper.

II. STATE OF ART

Various studies have been laid by researchers in context of generating fuzzy ontology. The FOGA framework has been proposed for generation of fuzzy ontology [20]. It deals with the fuzzy formal concept analysis (FCA) and clustering rather than textual formal concept analysis. FOGA method extends FCA approach that is being applied to extract ontologies with the help of fuzzy sets. The fuzzy sets are represented by membership functions. But the FOGA framework failed due to its small database size.

Cimano et.al [2] devised an automatic taxonomy learning algorithm that extracts hierarchical concepts from textual database. The learning algorithm used by them was formal concept analysis (FCA). It is method for deriving indirect relationships among set of objects holding set of attributes. FCA uses textual clustering techniques to generate lattice instead of fuzzy clustering techniques.

Chang Lee et.al [8] introduced the use of fuzzy ontology that includes some concepts related to domain. The attributes (classes, objects) used in designed ontology are predefined by experts. The taxonomy is generated on basis of these predefined concepts rather than discovering concepts automatically.

Yuefeng Li et al.[9] proposed ontology mining technique for extraction of patterns that satisfy user information needs. The technique has two components- top backbone and base backbone. The top backbone part us used to represent relations between different classes of ontology while base backbone is used to derive relationships between classes in top backbone. It is concluded that this work does not produce any fuzzy knowledge approach instead it leads to discovery of standard ontology according to user requirements.

Mohd. Abu et al.[1] extracts relationship between designed ontology on biological system. The approach saves the basic knowledge related to domain but it needs to be updated from time to time. The text documents are analyzed and the association between two biological entities is represented by fuzzy conjunction operator. It leads to generation of fuzzy relations that are used to retrieve information from medical document called GENIA.

III. SEMANTIC WEB AND ITS COMPONENTS

It is defined as collection of information linked in a way so that they can be easily processed by machines. From this statement, we conclude that SW is information in machine form. It is also known as framework for expressing information.

Architecture consists of following parts:

(i) URI and UNICODE: Semantic Web contains URI’s to represent data in triples based structures with the help of syntaxes designed for particular task.

- UNICODE supports intellectual text of style.

(ii) RDF and rdfschema: - RDF is Resource Description Framework. It processes metadata. It provides interoperation to work together between applications that exchange machine understandable information on web.

- rdfschema: - It is RDF vocabulary description language and represents relationship between groups of resources. There is RDF model designed for representing properties and their values.

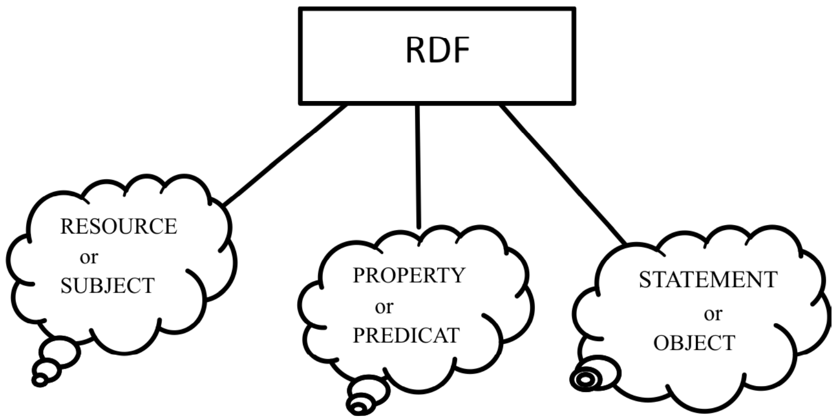

Figure 1: "RDF Model"

3.1 Querying in Semantic Web-

The basic query method takes a (subject, predicate and object) pattern and returns all triples that match the pattern. The triples determine type of index related to given subject.

```
Def triples (sub, pred, obj);
Try:
If sub ! = None:
If pred != None:
# sub pred obj:
If obj !=None:
If obj in self.spo [sub] [pred]:
Yield (sub, pred, obj)
# sub pred None
Else:
For retObj in self.spo[sub][pred]:
Yield (sub, pred, retObj)
Else:
# sub None obj
If obj != None:
For retPred in self.osp [obj][sub]:
Yield(sub, retPred, obj)
# sub None None
Else:
For retPred, objSet in self.spo[sub].items ():
For retObj in objSet:
```

```
Yield (sub, retPred, retObj)
Else:
If pred != None:
# None pred obj
If obj != None:
For retSub in self.pos[pred][obj]:
Yield (retSub, pred, obj)
Else:
None pred obj.
```

IV. FUZZY ONTOLOGY BASED APPROACH

The approach consists of following steps:

(a) The method is used in order to remove noisy/superfluous words from cluster of web documents stored in database. Standard document pre-processing, POS tagging and word stemming [17] are being applied on results produced by documents.

(b) After pre-processing, windowing process is performed over to reduce noisy words. It creates virtual window for each document that stores statistical information among similar terms used in documents called Tokens.

(c) [5, 15] proved that windows having number of terms from 5 to 10 is effective. If any word has weight lower than threshold values, it is discarded from window.

(d) Representation of terms in documents is done by statistical method named as Mutual Information (MI) and Balanced Mutual Information (BMI). The difference between them is that MI method is useful only when parameters are known while BMI can work even in absence of terms.

The relation between MI and terms is given by equation:

$$MI(t_1, t_2) = \log P\,(t_1, t_2)/P\,(t_1)\,P(t_2) \quad (1)$$

Where P (t_1, t_2) is probability that both terms are present in document. $P(t_{1)}$ is probability that term t_1 occurs in document. It is calculated as ratio of number of documents having term t_1 to total number of documents.

$$P(t_1) = |d_i| \,/\, |d| \quad (2)$$

The relation of BMI is given by:

$$\Pi_{c,t} = BMI\,(t_1, t_2)$$

$$= k[P(t_1, t_2) \log P\,(t_1, t_2) / P\,(t_1)\,P(t_2)] + [P(!\,t_1, !\,t_2) \log P(!\,t_1, !\,t_2) / P\,(!t_1)\,P(!\,t_2)] +$$

$$[P(!\,t_1, t_2) \log P(!\,t_1, t_2)/ P\,(!t_1)\,P(t_2)] - (1 - k)\,[P\,(t_1, !\,t_2) \log P(t_1, !\,t_2)/ P\,(t_1)\,P\,(!t_2)]$$

Where c refers to concepts, t is term used in those concepts.

(e) Concept Pruning takes place now. It states that same threshold value concept is used to discard noisy terms from concepts. After computing values, these values are scaled linearly to make them in range of membership function [0, 1].

Above figure generates fuzzy set that consists of objects drawn from a domain D and the membership of each object t_i € D in set is defined by membership function Π_f: T € [0, 1].

Some other estimation methods to find membership values are: Jaccard method [3], Conditional probability [4] and Kullback Divergence method [5].

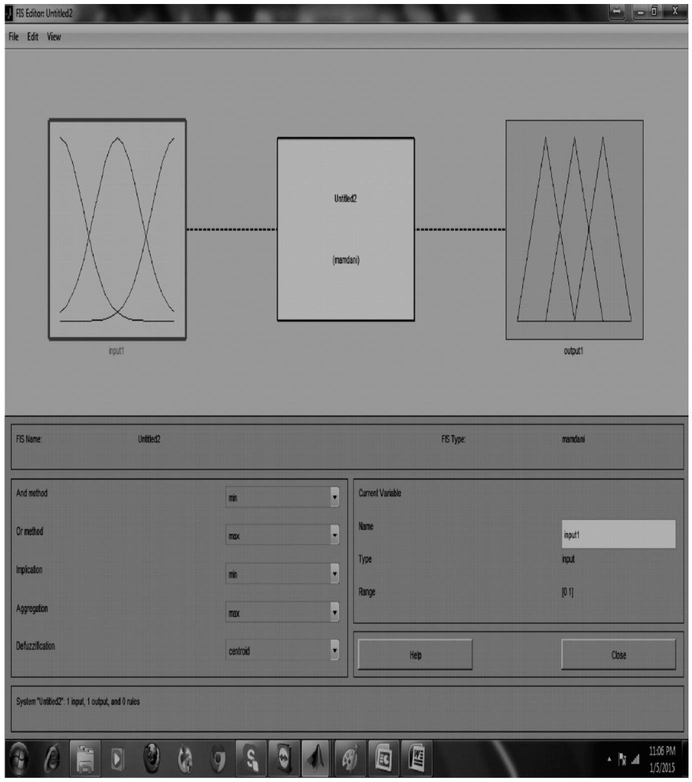

Figure 2: Fuzzy output membership function

Jaccard Method: $\Pi_{c,t}$ **= [P(c ^ t)/ P(c $_v$ t)]**

Conditional Probability: $\Pi_{c,t}$ **= P(c, t_1) / P(t_1)**

KL method**:** $\Pi_{c,t}$ **= [P(c, t) log P(c|t)/p(t)]**

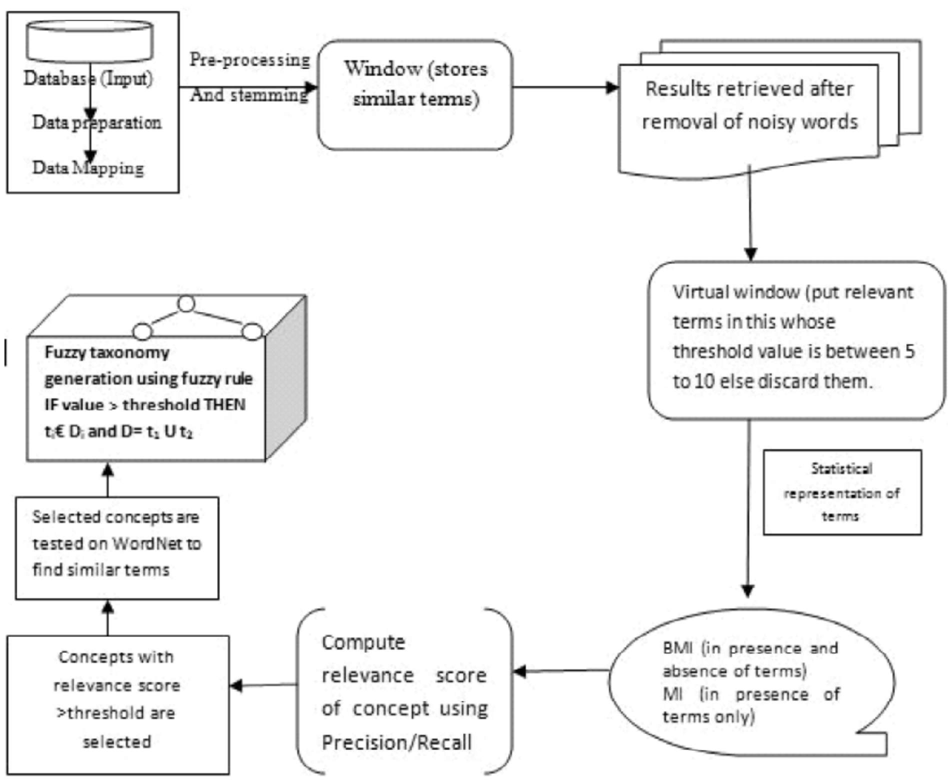

Figure 3: Fuzzy Ontology Based Approach

V. EVALUATION OF IR PERFORMANCE

There are two methods for evaluating performance as listed below:

It is evaluated on concept of Relevance. Relevance means that user should be satisfied with the results produced with respect to given query.

Factors affecting Relevance

- It depends not only on query data but also on context. It might happen that user is satisfied on some day and dissatisfied on another day.
- It depends on order of retrieval i.e. If first document satisfies user's needs then only user will move to second document.
- *Precision (P)* and *Recall (R)* are two measures to evaluate performance where Precision (P) = Relevant items retrieved / Total number of items retrieved.

  Recall (R) = Relevant items retrieved / Total relevant items in document.

The relevance formula for measuring Precision and Recall is given by

E = 1 – 1 / [$(1/P) + (1-$) 1/R] [22]

Where E = Effectiveness measure

P = Precision

R = Recall

$ = parameter that describes importance to P and R.

If $ = 0, then user has no importance to Precision

If $ = ½, then P = R

If $ = 1, then No Recall

On solving it, we have

E = 1 – 1 / [$/P + (1-$)/R]

Or

E = 1 – PR/ ($R + P-P$)

Or

E = 1 – PR / [$(R-P) + P]

Table 1
IR performance with/without ontology

Domain	*With fuzzy ontology Precision*	*Recall*	*Without ontology Precision*	*Recall*
weather (rain)	0.273	0.361	0.180	0.293
Google TM(news)	0.355	0.456	0.231	0.342
weather(food)	0.123	0.234	0.119	0.212
Google TM (stock)	0.234	0.289	0.121	0.232
weather (livestk)	0.345	0.478	0.237	0.432
Google TM (trade)	0.321	0.378	0.278	0.321
weather (humid)	0.456	0.675	0.311	0.564
weather (gauge)	0.347	0.543	0.245	0.459
Google TM(lives)	0.289	0.378	0.234	0.343

Precision is measured if set of users agree on relevance of retrieved documents. Measuring Recall is quite difficult because it depends on knowing the relevant documents which needs accessing of whole document. It is so difficult to access whole document.

5.1. Experiment

The experiment is conducted to compute IR performance with/without fuzzy domain ontology by taking two domains- Weather system and Google TM.

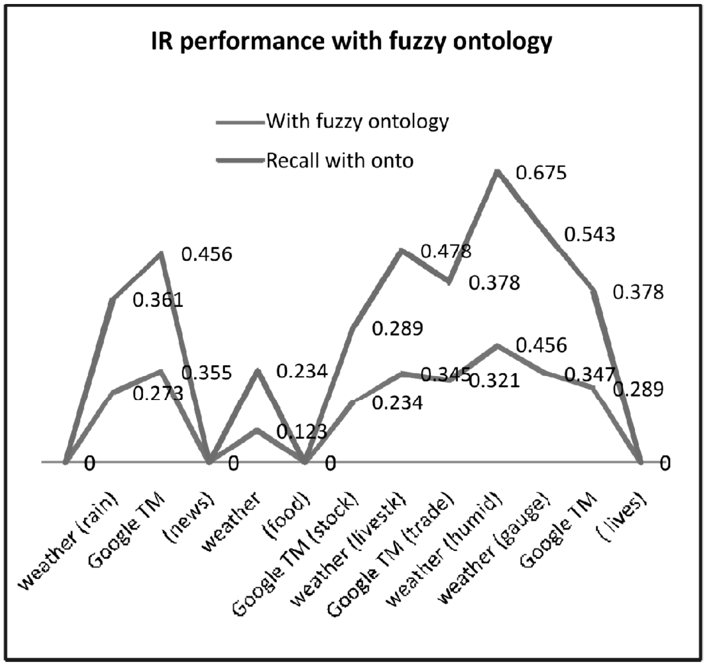

Figure 4: IR performance with use of fuzzy ontology

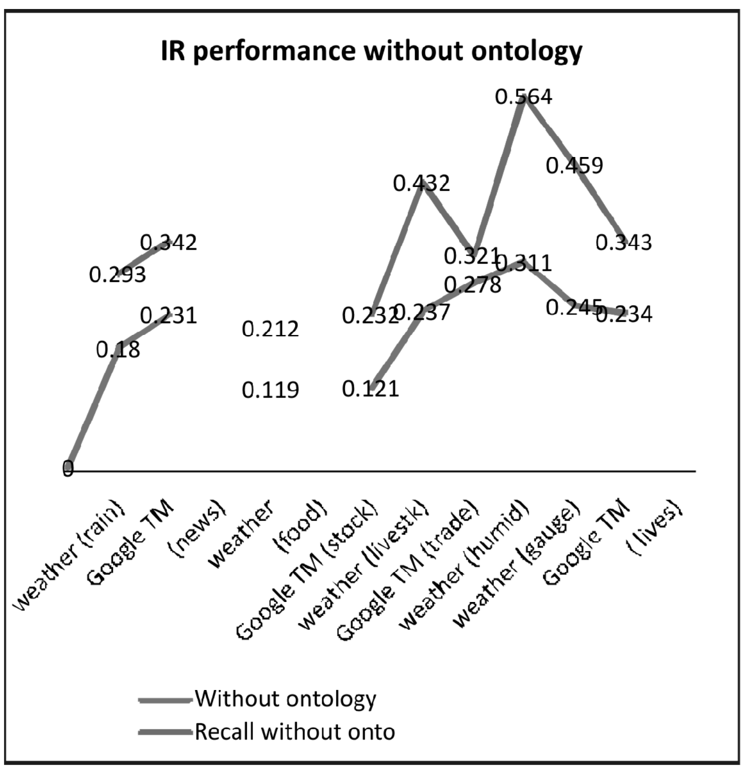

Figure 5: IR performance without ontology. (The results are scattered without ontology skipping some values)

VI. CONCLUSION AND FUTURE WORK

It is also possible to discover ontology from textual databases without involvement of any soft computing techniques. But this domain ontology with crisp concepts and relations is less likely to satisfy uncertain factors of real world applications. This paper proposes fuzzy ontology based web mining approach that uses fuzzy set and relations to discover fuzzy taxonomy. It performs concept pruning by putting selected concepts in virtual windows. And then statistical approaches like BMI, Kullback divergence are used to analyze them.

Our preliminary experiments show that the automatically generated fuzzy domain ontology can significantly improve the performance of information retrieval.

Future work involves comparing proposed fuzzy approach with other estimation membership approaches like Kullback, Conditional and Jaccard.

REFERENCES

[1] Muhammad Abulaish and Lipika Dey :Biological ontology enhancement with fuzzy relations A text mining framework. In Andrez Skowron, Rakesh Agrawal, Michael Luck and editors, Proceedings of the 2005 IEEE/WIC/ACM International Conference in Web Intelligence, pages 379-385, IEEE Computer Society.

[2] P. Cimiano; A.Hotho and S.Staab: Learning concept hierarchies from text corpos using formal concept analysis. Journal of Artificial Intelligence Research, 24:305-339, 2005.

[3] The World Wide Web Consortium. Web ontology language, 2004, Available from http://www.w3.org/2004/OWL/

[4] T.R.Gruber: A translation approach to portable ontology specifications. Knowledge Acquisition, 5(2):199-220, 1993

[5] Hongyan Jing and Evelyne Tzoukermann: Information retrieval based on context distance and morphology. In Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, language and analysis, pages 90-96, 1999.

[6] David Vallet; M.Fernandes: An Ontology-Based Information Retrieval Model, "European Semantic Web Symposium (ESWS)", 2006

[7] Vishal Jain; Gagandeep Singh Narula: Information Retrieval through Semantic Web - An Overview. In Proceedings of CONFLUENCE-2012 titled 'The Next Generation Information Technology Summit', pp 23-27, organized by the Amity School of Engineering and Technology, Noida (UP) during 27-28 September 2012.

[8] Chang-Shing Lee; Zhi Wei and Huang: A fuzzy ontology and its application to news summarization. IEEE Transactions on Systems, Man and Cybernetics, Part B, 35(5)"859-880, 2005.

[9] Yuefeng Li and Ning Zhang: Mining ontology for automatically acquiring web user information needs. IEEE Transactions on Knowledge and Data Engineering, 18(4): 554-568, 2006.

[10] Abd-Elraham Elsayed; Samhaa Ram; Mahmod Rafea: "Applying data mining for ontology building", "ACM Conference 26 (10)", 2003.

[11] John McCrae; Mauricio Espinoza: "Combining Statistical and semantic approaches to translation of ontologies and taxonomies", "In Proceedings of SSST-5, Fifth Workshop on Syntax, Semantics and Structure in Statistical Translation", 2011, pages 116–125.

[12] Alexander Maedche and Steffen Staab: Ontology learning. In handbook on Ontologies, pages 173-190, 2004.

[13] A. Doan; J. Madhavan; A. Halevy: "Ontology Matching: Machine Learning Approach". "Handbook on Ontologies in Information Systems, S. Staab and R. Studer", May 2004, pages 397-416.

[14] Roberto Navigli; pasola Velardi etal: Ontology learning and its application to automated terminology translation. IEEE Intelligent Systems, 18(1) : 22-31, 2003.

[15] Patrick perrin and Frederick Perrin: Extraction and replacement of contextual information for knowledge discovery in texts. Information Sciences, 151:125-152, 2003.

[16] Berners-Lee and Fischetti: "Weaving the Web: The Original Design of the World Wide Web by its inventor". "Scientific American", 2005.

[17] G Salton and MJ. McGill: Introduction to Modern Information Retrieval. McGraw Hill, New York, 1983.

[18] R. Tous and J. Delgado: A Vector Space Model for Semantic Similarity Calculation and OWL Ontology Alignment, "DEXA 2006" , pp. 307–316.

[19] Dayal U; Kuno H: Making the Semantic Web Real, “IEEE Data Engineering Bulletin, Vol.26, No.4”, pp 4-7, 2003.

[20] Quan Thanh; SIu Chang; Alvis Cheuk and Tru Hoang Cao: Automatic fuzzy ontology generations for semantic web. IEEE Transactions on Knowledge and Data Engineering, 18(6):842-856, 2006.

[21] Sukanya Ray; Nidhi Chandra: Domain Based Ontology and Automated Text Categorization Based on Improved Term Frequency and Inverse Document Frequency, IJMECS Vol.4, No. 4, May 2012.

[22] Gagandeep Singh Narula: Ontology Development Using Hozo and Semantic Analysis for Information Retrieval in Semantic Web in ‘ICIIP-2013 IEEE Second International Conference on Image Information Processing with IEEE Conference Record Number 31034 Jaypee University of Information Technology, Shimla, December 9-11, 2013’

[23] Xin Shi; Shurong Tong; Bo Li: Ontology Mapping of Design process Knowledge based on Classification: IJEME Vol.1, No. 5, November 2011.

[24] Avinash J Aggarwal; O.G. Kakde: Semantic Analysis of Natural Language Queries Using Domain Ontology for Information Access from Database, IJISA Vol. 5, No.12, November 2013

[25] Raymond Y.K. Lau and Yuefung Li: Mining fuzzy domain ontology from textual databases. IEEE/WIC/ACM International Conference on Web Intelligence, IEEE Computer Society, pages 156-162, 2007.