Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 50 (2015) 135 – 142

Procedia Computer Science

2nd International Symposium on Big Data and Cloud Computing (ISBCC'15)

Hybrid Fuzzy-Ontology Design using FCA based Clustering for Information Retrieval in Semantic Web

Balasubramaniam K[a] *

[a] *Research Scholar, Hindustan University, Chennai 603103, India*

Abstract

Ontology is a way to represent the domain knowledge into a human understandable and machine readable format. It is used as one of the major knowledge representation mechanism for semantic web. Introducing the ontology knowledge provides more relevant search results for the users information need. To deal with uncertain information, the mechanism supported by the regular ontology may not be adequate and the requirement for new technique arises. Fuzzy based methods are the proven methods to interpret the uncertain information. The combination of Fuzzy and Ontology based information retrieval provides better results as they mainly deal with the semantics and the uncertainty of information. Keyword matching is one another widely used method which matches the input keywords with the existing information domain to find the best match results. When the input queries are complex the fuzzy ontology based information retrieval which respects the user's keyword and the domain produces more accurate results. This work enlarges the fuzzy ontology knowledge results along with the input queries and keyword matching. The given algorithm is a hybrid technique based on matching extracted instances from the input queries and in information domain. Overall, compared to the existing query models supported by fuzzy ontology or keyword based models the hybrid ontology with keyword matching is sufficient and easy way to retrieve the documents in semantic web. The performance of the hybrid ontology approach is measured using improved precision, recall and f-measure values.

Peer-review under responsibility of scientific committee of 2nd International Symposium on Big Data and Cloud Computing (ISBCC'15)

Keywords: Fuzzy Ontology Generation framework (FOGA): Keyword Matching: Semantic Web: Hybrid FOGA with keyword: Information Retrieval Method.

* Balasubramaniam K. Tel.: +91 -9940671275; *E-mail address:*bala.kandasamy@gmail.com

Peer-review under responsibility of scientific committee of 2nd International Symposium on Big Data and Cloud Computing (ISBCC'15)
doi:10.1016/j.procs.2015.04.075

1. INTRODUCTION

Ontology is an effective conceptualism commonly used for the Semantic Web. The domain of ontology is beneficial in establishing a common vocabulary describing the domain of interest. This is important for the unification and the sharing of knowledge about the domain and connecting with other domains. The problem of these ontologies is that they are not constructed to describe the complete domain of data mining, but are simply made with a specific task in mind. For this purpose, Fuzzy logic can be incorporated to ontology to represent uncertainty information. Traditionally, fuzzy ontology is generated and used in text retrieval and search engines, in which membership values are used to evaluate the similarities between the concepts in a concept hierarchy. However, manual generation of fuzzy ontology from a predefined concept hierarchy is a difficult and tedious task that often requires expert interpretation. So, automatic generation of concept hierarchy and fuzzy ontology from uncertainty data of a domain is highly desirable. Clustering technique is applicable for producing the hierarchical grouping of documents. Clustering of document is very important for the purpose of document organization, summarization, topic extraction and information retrieval in an efficient way. The main objective of this research is to present a new automatic approach to extract ontology using clustering and FCA combined with a fuzzy rule-based language.

FOGA (Fuzzy Ontology Generation frAmework) is used for automatic creation of fuzzy ontology on insecurity information to tackle the difficulties in concept hierarchy is given by Jelsteen *et al* (2013). This FOGA has been very difficult to standardize ways of classifying information available on computer networks. The framework of semantic web is that intends to describe various strategies and technologies that can be used in order to start building broad and generally acceptable standards [1]. The fuzzy knowledge is an important task in huge amount of domains that face a many imprecise and vague knowledge and information, such as text mining, multimedia information system, medical informatics, machine learning, and human natural language processing. To incorporate fuzzy theory into ontology, one of the possible solutions is yielding a fuzzy ontology model for handling insecurity of information and knowledge [2]. Accordingly, fuzzy Ontologies contain fuzzy concepts and fuzzy memberships.

Finally, semantic web of information management would not suffer restrictions inherent to keyword matching; information would be classified and investigated on more abstract criteria such as concepts, validity, or quality. However, the semantic web is good but it performs function slowly and some concepts had to be tested behind closed doors first (i.e. on Intranets). Ontology is used to understand the meaning of domain dependent queries.

This paper proposes the hybrid system of FOGA by combining the FCA (Formal Concept Analysis) based clustering and keyword matching approach. Integrating FCA with FOGA is an effective technique for conceptual clustering, data analysis and for knowledge presentation. Many research issues to be addressed in this work; So far it is clear that content providers will be able to bring their contribution as well as benefit from a consensual classification scheme.

This research work is organized as follows: The related work is reviewed in Section 2. The proposed ontology-based information retrieval model is depicted in Section 3. The experiments and discussion on the results are described in Section 4. Finally, conclusion is made in Section 5.

2. LITERATURE SURVEY

Lee *et al* (2005) present, a fuzzy ontology and its application to news summarization. The fuzzy ontology with fuzzy concepts is an extension of the domain ontology with crisp concepts. It is more suitable to describe the domain knowledge than domain ontology for solving the uncertainty reasoning problems.

Inyaem *et al* (2010) performs fuzzy ontology is based on the concept that each index object is related to every other object in the ontology, with a degree of membership assigned to that relationship based on fuzzy set theory. This work proposes use cases based on the related process of the terrorism event extraction using fuzzy ontology, especially the terrorism fuzzy ontology construction methodology. Ferreira-Satler *et al* (2014) shows a fuzzy ontology based approach to automatically build user profiles from a collection of user interest documents. The ontological representation of the user profile enhances the performance in tasks such as filtering, categorization and information retrieval.

Cleophas *et al* (2010) discussed a new taxonomy of sub linear (multiple) keyword pattern matching algorithms

is presented. Based on an earlier taxonomy by the second and third authors, this new taxonomy includes not only suffix-based algorithms, but also factor- and factor-oracle-based algorithms.Ontology matching is a vital step whenever there is a need to integrate and reason about overlapping domains of knowledge are given by Albagli *et al* (2012).

Cao *et al* (2011) and Shvaiko *et al* (2013) review the state of the art of ontology matching and analyze the results of recent ontology matching evaluations. These results show a measurable improvement in the field, the speed of which is albeit slowing down.

Ontologies are frequently used in information retrieval being their main applications the expansion of queries, semantic indexing of documents and the organization of search results are shown by Jimeno-Yepes *et al* (2010). Dragoni *et al* (2012) and Kara *et al* (2012), present an ontology-based information extraction and retrieval system and its application in the soccer domain. In general, they deal with three issues in semantic search, namely, usability, scalability and retrieval performance. They propose a keyword-based semantic retrieval approach. A keyword retrieval system for document image is given by Wei *et al* (2014).

Web contains huge number of web pages and to find suitable information from them is very cumbersome task. Information Retrieval (IR) technology is major factor responsible for handling annotations in Semantic Web (SW) languages and in the present paper knowledgeable representation languages used for retrieving information are discussed by Singh *et al* (2014).

3. RESEARCH METHODOLOGY

Hybrid ontology is an effective theory commonly used for the Semantic Web. In this hybrid ontology, fuzzy logic can be incorporated for indicating uncertainty information. In general, creating of fuzzy ontology is done from predefined concept hierarchy. But, constructing a concept hierarchy for a specified domain is very complicated and deadly task. To overcome these issues, here Hybrid FOGA for automatic creation of fuzzy ontology on uncertainty information using FCA based clustering technique is intended. It also discusses resembling reasoning for increasing enrichment of the ontology with new upcoming data. The information can be retrieved in hybrid fuzzy ontology by using the keyword matching. At last, a method based on fuzzy concept for combining other attributes of database to the ontology is planned. The architecture for the Hybrid FOGA using FCA based clustering techniques for information retrievals are described in figure 1.

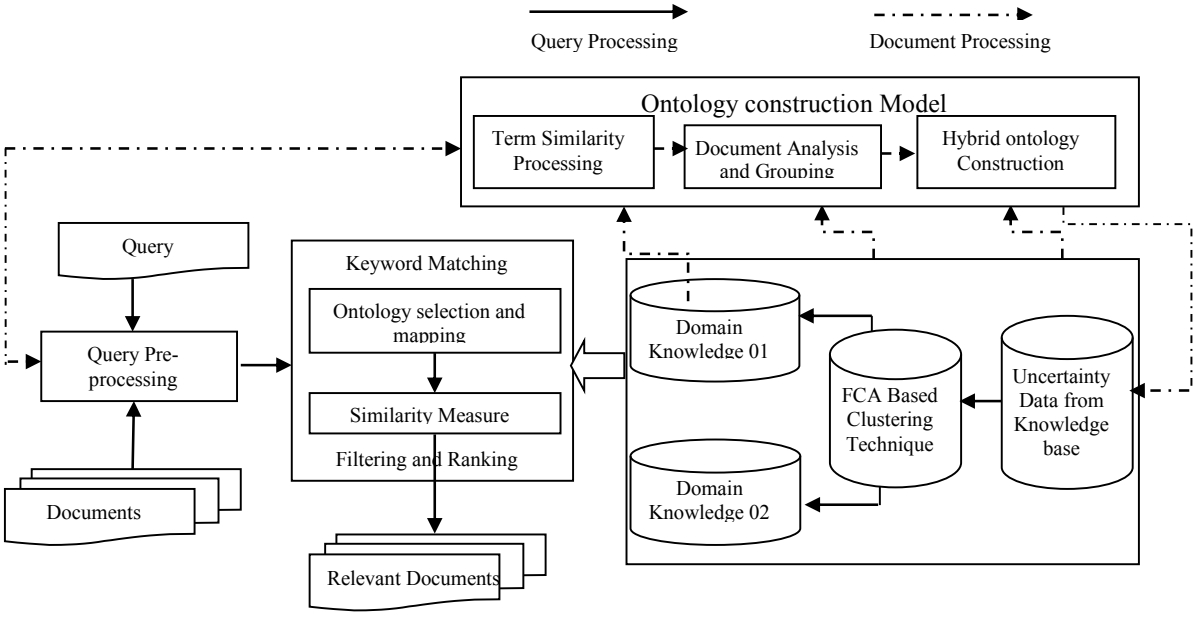

Fig.1. Architecture for Hybrid FOGA approach

3.1 FCA based clustering using FOGA framework

Ontology is widely used in Semantic webs for getting efficient information retrievals from the actual source system. In order to identify the improbability of the information ontology as well as web fuzzy logics can be associated together which is so called referred to as fuzzy ontology as the predefined hierarchy concept. Designing the hierarchy for the specific domains is not such a simple task and this proposed system will introduce a FOGA-fuzzy ontology generation framework [1] which is capable of producing the automatic fuzzy ontology for the purpose of identifying the uncertainty of the information.

Proposed FOGA which consists of the following components:

1. **Fuzzy Formal Concept Analysis:** It constructs a fuzzy formal context from a database containing uncertainty data. In addition, it also generates fuzzy formal concepts from the fuzzy formal context and organizes the generated concepts as a fuzzy concept lattice.
2. **Fuzzy Conceptual Clustering**: It clusters concepts on the fuzzy concept lattice and generates conceptual clusters. The clustering process is performed based on fuzzy information incorporated into the lattice using fuzzy logic.
3. **Hierarchical Relation Generation**: It generates hierarchical relations between conceptual clusters to construct a concept hierarchy.

The clustering techniques are incorporated into the FOGA to analyzing the formal concept for knowledge presentation. In order to prune the lattices generated for text mining, clustering is first performed on the data set to generate clusters of documents. Then, feature selection is used to extract frequent keywords (or terms) from documents in each cluster as attributes for the cluster. The clustering techniques have three properties to generate the FCA-based clustering.

Property 1: The number of clusters generated by a clustering algorithm is always lower than the number of starting objects to which one applies the clustering algorithm.

- All objects belonging to one same cluster have the same proprieties. These characteristics can be easily deduced knowing the centre and the distance from the cluster.
- The size of the lattice modelling the properties of the clusters is lower than the size of the lattice modelling the properties of the objects.
- The management of the lattice modelling the properties of the clusters is optimum than the management of the lattice modelling the properties of the objects.

Property 2: Let C1, C2 be two clusters, generated by a clustering algorithm and verifying the properties p1 and p2 respectively. Then the following properties are equivalent:

$C1 \Rightarrow C2$ (CR) $\Leftrightarrow$

- $\forall$ object O1 $\in$ C1 => O1 $\in$C2 (CR)
- $\forall$ object O1 $\in$C1, O1 checks the property p1 of C1 and the property p2 of C2. (CR)

Property 3: Let C1, C2 and C3 are three clusters generated by a classification algorithm and verifying the properties p1, p2 and p3 respectively. Then the following properties are equivalent:

C1, C2 = > C3 (CR) $\Leftrightarrow$

- $\forall$ object O1 $\in$C1 $\cap$ C2 = > O1 $\in$C3 (CR)
- $\forall$ object O1$\in$ C1 $\cap$ C2 then O1 checks the properties p1, p2 and p3 with (CR).

Validation of the two properties will be done due to the fact that all objects correspond to the unique cluster will ensure essentially the unique attributes as their clusters [6]. Therefore, this approach is highly appropriate to domains when the clustering algorithms automatically discover and generate the ontology.

In this work, a new technique called FOGA (Fuzzy Ontology Generation frAmework) is proposed which will automatically create a fuzzy ontology. This fuzzy ontology has the ability to deal with fuzzy knowledge and is

very effective in text and multimedia object representation and retrieval [2]. Ontology is an important section of the W3C standards for the Semantic Web which is used to identify standard conceptual vocabularies to replace data between systems, which will provide reusable knowledge and make easy interoperability across multiple heterogeneous systems and databases.

3.2 Keyword Optimization Techniques for Information Retrieval

The present Information Retrieval (IR) method provide a secure path for the user to state the information requires on the basis of keywords but it do not properly capture the clear meaning of a keyword queries [7]. Keyword optimization plays a major role in every field where a user needs a searching like aspect of internet marketing from content strategy. Generally, IR model is based on keyword search and its search makes queries with user's keyword and repossesses the information which using that model. A document or resources contains the keywords and the word knowledge is considered and summarized as the conception included in particular documents.

3.3 Integrating Hybrid FOGA framework and Keyword Optimization

The proposed model is a semantic document retrieval model that uses a FCA based clustering techniques and fuzzy ontology with keyword. It semantically retrieves a set of relevant documents according to a users query respecting the underlined field or domain. It can be used to retrieve any kind of documents in a specific domain written in any language. The proposed model aims to:

- As its hybrid method uses a fuzzy ontology with keyword for information retrieval. It uses a components set of concepts, relation between them, terms, relation between them, and a set of relations between concepts and terms.
- This is through using a fuzzy ontology with keyword during its expansion algorithm to expand each user keyword in a certain field or view.
- Rank the resulted semantically relevant documents according to some criteria, such as the document matching degree, its confidence degree, and its timeliness.

Algorithm 1: Hybrid ontology mapping with keyword
Input: Select fuzzy ontology and keyword
Output: ontology explained with a set of keywords in the specified view
Steps:
1. Divide the paper into different weighted zones
2. PKS (Paper Keyword Set) = expand all keywords in the keyword zone according to the given view using the fuzzy ontology
3. Arrange all PKS in decreasing order according to each keyword n-gram
4. annotate each zone with the PKS
5. for each zone, calculate the weight of each keyword in PKS
6. **For** each section
7. annotate it with its title
8. SKS (Section Keyword Set) = expand this annotation using the fuzzy ontology
9. arrange SKS in a descending order with respect to each keyword n-gram
10. calculate the value of each keyword in SKS
11. **End for**
12. Calculate the weight of each keyword in the keyword zone through summing its value from each zone and each section.

After applying the proposed term keyword algorithm, the resulted keywords are stored in a relational database. The hybrid FOGA are used to weight expanded the keywords to discover the concept and terms which related to keywords. This work applies the n-gram for listed keywords to discover the degree of match with a cluster and a particular domain. A method is constructed the keywords into two parts, where the first part is the ontology based queries and next part holds the keyword information for every properties. Finally, it's converted to a complete ontology query. A user fills the ontology form only if the front-end system is a graphical user interface and then

keywords are entered subsequent to the each property which enables the keyword matching. The proposed algorithm has assist to maintain these keyword lists in order to develop the classifications using three main objectives:

- Should be removed the uncover keywords
- Discover new keywords
- Assign a weighting system to reflect keywords' relative importance.

The keyword also included after the creation of new ontology. This new ontology is assembled through the initial matching of the keywords in *KeyWord* class. This process can completed through the users to select the list of keywords in openly from the *KeyWord* class or can enter haphazard keywords. In initial stage, the ontology must be issued sequentially for to transmit to the suitable server. Next, the lists of words from the *KeyWord* class are displayed within the server and it's returned these terms to the users to select for that matching the request process. For the second stage, the keywords are coordinated to the terms in *KeyWord* class by the text matching tool so the user can able to include the keyword between the ontology queries. To evaluate the weights of a specific keyword in a section by this proposed method:

- For only well written papers, considers the main sentence of each paragraph in this zone or in this section, otherwise, consider the whole paper paragraphs.
- Returns each pronoun in it to its referred noun.
- Removes all stop words.
- Stems each of the remaining keyword.
- Calculates the weight of the keyword that is belongs to PKS in a certain zone.

The following steps are describes the examples for hybrid FOGA using clustering techniques and keyword optimizations.

Constructing a multi-view query- When a certain user enters his query, he should specify the underlined field and domain. *"select all papers about bioinformatics according to the medical view"* where "*bioinformatics*" is the keyword that the user searches for. "*medical*" is the search point of view. This linguistic term is previously defined by the user according to his subjective view and stored in his account.

Applying the Query Operations- After user submits his query, the operations are performed on it. First the query is parsed, such that each searched keyword is extracted with its field search view. Each keyword is then expanded in its specified search point of view using the predefined fuzzy ontology.

Retrieving a set of relevant documents- the clustering method is used for similar keywords to retrieve the set of relevant documents semantically with respect to a certain user query through calculating document matching degree. A document matching degree is calculated as the max-min composition between the list of weighted keywords that annotate this document and the list of query's weighted expanded keywords. The result of this is a list of semantically relevant documents each associated with its matching degree.

4. EXPERIMENTAL RESULTS

The experimental result shows the integration of fuzzy ontology with keywords for information retrieval. The text documents collected from the IEEE web site and tested 1000 scientific text documents on the research area "Information Retrieval" that are used for experimentation. The journal collection is downloaded from the web which is related on data mining domain. The list of journals from IEEE considered here are Biomedical Engineering, Circuits and Systems, Communications and Computer Graphics and Application. Java application is used for evaluating the proposed clustering technique which could produce the enhanced OWL file using Protege tool. The HTML is used to design the journal abstract, that HTML pages are downloaded. By removing the HTML tag elements from the web documents for conversion of HTML page into text document. The text contents are maintained in separate text files. The classifier's performance using the selected keywords against the results obtained with the algorithm. These classes are used as a benchmark to evaluate the clustering results in terms of recall, precision, and F-measure. The results are compared with the ngram keyword search system. The extracted citation keywords of documents act as their attributes. Since these attributes are descriptors for the generated clusters, if more keywords are extracted and used, the more meaningful the cluster descriptors are constructed. To

verify this, vary the number of keywords N extracted from documents from 2 to 10, and the similarity threshold from 0.2 to 0.9 when performing FCA based clustering. The results also show that the F-measure tends to have the best performance when threshold is equal to 0.5. For calculation the precision and recall are as follows,

$$Precision = \frac{number\ of\ retrieved\ relevant}{number\ of\ retrieved}$$

$$Recall = \frac{number\ of\ retrieved\ relevent}{total\ number\ of\ retrieved\ in\ collection}$$

$$F - Measure = 2.\frac{precisiom.recall}{precision + recall}$$

Recall is the proportional of the correctly retrieved documents among the pertinent documents in the collection. Precision is the proportion of the correctly retrieved documents among the documents retrieved by the system.

Table 1 shows the precision and recall

Approaches	Precision	Recall
Standard FOGA	0.79	0.87
Keyword Matching	0.85	0.72
Hybrid FOGA with Keyword	0.89	0.65

Table 2 shows the F-Measure

Approaches	F-MEASURE
Standard FOGA	0.52
Keyword Matching	0.76
Hybrid FOGA with Keyword	0.85

In Table 1 and table 2 the precision, recall and F-measure for FOGA framework with keyword matching are presented.

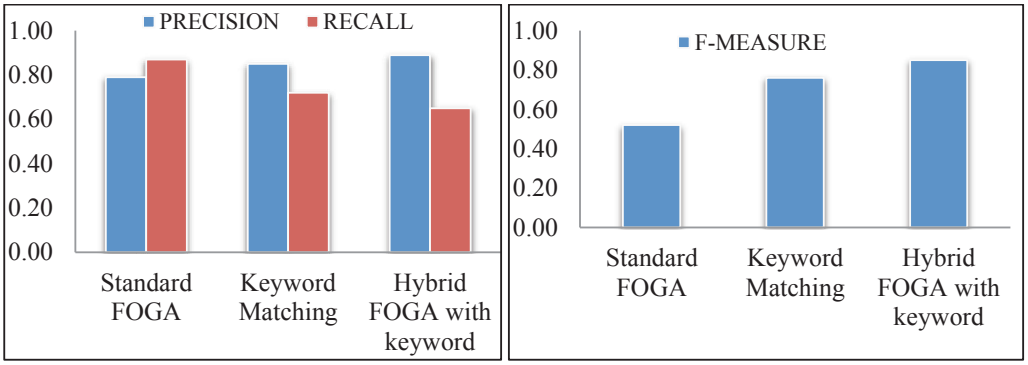

Fig.2. (a) Precision and Recall; (b) F-Measure.

Figure 2 (a) shows the precision and recall for proposed Hybrid FOGA with keyword. The proposed method have high precision and less recall rate when compare to other methods. Figure 2 (b) shows the F-Measure for proposed method.

5. CONCLUSION

In this work, the Hybrid FOGA framework using FCA for information retrieval based on clustering is proposed. The following steps are used for FOGA: Fuzzy Formal Concept Analysis, Fuzzy Conceptual Clustering and Fuzzy Ontology Generation. Fuzzy Formal Concept Analysis offers a conceptual framework used for analysing,

structuring and visualizing information for develop them more clearly. The proposed FOGA with Formal Concept Analysis is used to develop ontology from insecure data as it characterise the insecure information and develop a concept hierarchy by the uncertainty information automatically. The evaluation of the proposed FOGA technique use FCA is offer based on the ontology generation. Experimental result shows that the proposed approaches optimize the ontology definition, offers a good interpretation of the information and optimize the precision, recall and F-Measure for developing data. The future direction to work in this area would be to build a document annotation algorithm using our proposed fuzzy ontology tool.

REFERENCES

1. Tho, Quan Thanh, Siu Cheung Hui, Alvis Cheuk M. Fong, and Tru Hoang Cao. "Automatic fuzzy ontology generation for semantic web." *Knowledge and Data Engineering, IEEE Transactions on* 18, no. 6 (2006): 842-856.
2. Lee, Chang-Shing, Zhi-Wei Jian, and Lin-Kai Huang. "A fuzzy ontology and its application to news summarization." *Systems, Man, and Cybernetics, Part B: Cybernetics, IEEE Transactions on* 35, no. 5 (2005): 859-880.
3. Inyaem, Uraiwan, Phayung Meesad, Choochart Haruechaiyasak, and Dat Tran. "Construction of fuzzy ontology-based terrorism event extraction." In *Knowledge Discovery and Data Mining, 2010. WKDD'10. Third International Conference on*, pp. 391-394. IEEE, 2010.
4. Cleophas, Loek, Bruce W. Watson, and Gerard Zwaan. "A new taxonomy of sublinear right-to-left scanning keyword pattern matching algorithms." *Science of Computer Programming* 75, no. 11 (2010): 1095-1112.
5. Albagli, Sivan, Rachel Ben-Eliyahu-Zohary, and Solomon E. Shimony. "Markov network based ontology matching." *Journal of Computer and System Sciences*78, no. 1 (2012): 105-118.
6. Amel Grissa Touzi, Hela Ben Massoud and Alaya Ayadi, *"Automatic ontology generation for Data mining using FCA and clustering",* arxiv.org, no. 1311.1764.
7. Thanh Tran, Philipp Cimiano, Sebastian Rudolphand Rudi Studer, *"Ontology-Based Interpretation of Keywords for Semantic Search",* Springer The Semantic Web Lecture Notes in Computer Science Volume 4825, 2007, pp 523-536.
8. Shvaiko, Pavel, and Jérôme Euzenat. "Ontology matching: state of the art and future challenges." *Knowledge and Data Engineering, IEEE Transactions on* 25, no. 1 (2013): 158-176.
9. Cao, Dongxing, Zhanjun Li, and Karthik Ramani. "Ontology-based customer preference modeling for concept generation." *Advanced Engineering Informatics*25, no. 2 (2011): 162-176.
10. Jimeno-Yepes, Antonio, Rafael Berlanga-Llavori, and Dietrich Rebholz-Schuhmann. "Ontology refinement for improved information retrieval."*Information Processing & Management* 46, no. 4 (2010): 426-435.
11. Dragoni, Mauro, Célia da Costa Pereira, and Andrea GB Tettamanzi. "A conceptual representation of documents and queries for information retrieval systems by using light ontologies." *Expert Systems with applications* 39, no. 12 (2012): 10376-10388.
12. Kara, Soner, Özgür Alan, Orkunt Sabuncu, Samet Akpınar, Nihan K. Cicekli, and Ferda N. Alpaslan. "An ontology-based retrieval system using semantic indexing." *Information Systems* 37, no. 4 (2012): 294-305.
13. Ferreira-Satler, Mateus, Francisco P. Romero, Jose A. Olivas, and Jesus Serrano-Guerrero. "Fuzzy Ontology-Based Approach for Automatic Construction of User Profiles." In *Rough Sets and Current Trends in Soft Computing*, pp. 339-346. Springer International Publishing, 2014.
14. Singh, Gagandeep, and Vishal Jain. "Information Retrieval (IR) through Semantic Web (SW): An Overview." *arXiv preprint arXiv:1403.7162* (2014).
15. J.Jelsteen, D.Evangelin, J.Alice Pushparani, j.Nelson Samuel Jebastin. "Ontology Learning Process Using Fuzzy Formal Concept Analysis" International Journal of Engineering Trends and Technology (2013)