International Journal of Advance Research in Science and Engineering
Volume No.07, Special Issue No.01, March 2018
www.ijarse.com
IJARSE
ISSN: 2319-8354

A SURVEY ON DIFFERENT APPROACHES FORINFORMATION RETRIEVAL AND ONTOLOGY FRAMEWORK

Sharvali S. Sarnaik[1], Ajit S. Patil[2]

[1] *M.E. Student, Dept. of CSE,*

Kolhapur Institute of Technology College of Engineering,Kolhapur, (India)

[2] *Associate Professor, Head of Dept. CSE,*

Kolhapur Institute of Technology College of Engineering, Kolhapur, (India)

ABSTRACT

To retrieve the information and relevant document when the need arise. Fetching the relevant information is one of the challenging tasks as lot of documents are available now-a-days. In the past few years many researches have been going on information extraction. Ontologiesare one of the fields of information extraction. There are many algorithms which help to extract information with the help of ontology. Ontologies are used as a guide for these algorithms. The Term-frequency and inverse document frequency is also been used to weight the words from the document. This paper is a survey of different approaches of information retrieval and the ontology based information extraction.

Keywords:Information Extraction, Text Ranking, Schema Graph, Ontology Framework, TF-IDF

I.INTRODUCTION

With increase in development of technology and in the today's era every person saves their entire work in soft copy. Large amount of the information is generated in text format. The information is collected in the form of structured and semi-structured data. Organizations need useful information from the set of documents and from huge information set. It is very difficult to retrieve the useful and needed information or document from the set of documents. Finding the relevant document manually is really a very time consuming task as human will take lot of time searching the relevant information and getting the needed document. Hence retrieving the information is really one of the critical and complex tasks. The solution that can help in this scenario is called as information extraction. Information extraction will help to retrieve the documents as per user need. Information extraction is done with the help of different approaches. Clustering, text ranking, machine learning, sequence labeling, schema graph etc. are the approaches used for information extraction and for relevant document retrieval. Ontology is the field of information extraction. Many researches are being done on ontology based information extraction. Document mapping is being done to retrieve the matching document. There are many algorithmswhich help to divide the words and form in sequence manner which helps to extract the information. [3] Information extraction is defined as extraction of information from natural language processing (NLP). Weights are also assigned to words which are present in the document. [1] There are many methods to calculate

the weights for example methods like frequency function, Boolean function, entropy function and term frequency and inverse document frequency (TF-IDF). Term frequency-inverse document frequency helps in identifying the weights of the words that is how many times the particular term arise in the documents [14].

The main goal of this paper is to survey the different approaches used for information extraction and ontology based information extraction. There are many algorithms which help to extract information from the bunch of documents. Ontologies are used as a guide for these algorithms.

II.LITERATURE SURVEY

This paper contains study of research papers. The research papers which are included in this literature review are taken from the IEEEXplore Digital Library and Science Direct, Elsevier. The research papers were scrutinized on the basis of ontology framework, information extraction and term frequency and inverse document frequency. Initially the papers which are retrieved for the literature survey are 30 research papers. The papers which left for topic review were 22 papers. After analyzing the papers 17 papers were short listed for content review. This literature survey contains 15 research papers which are studied and analyzed.

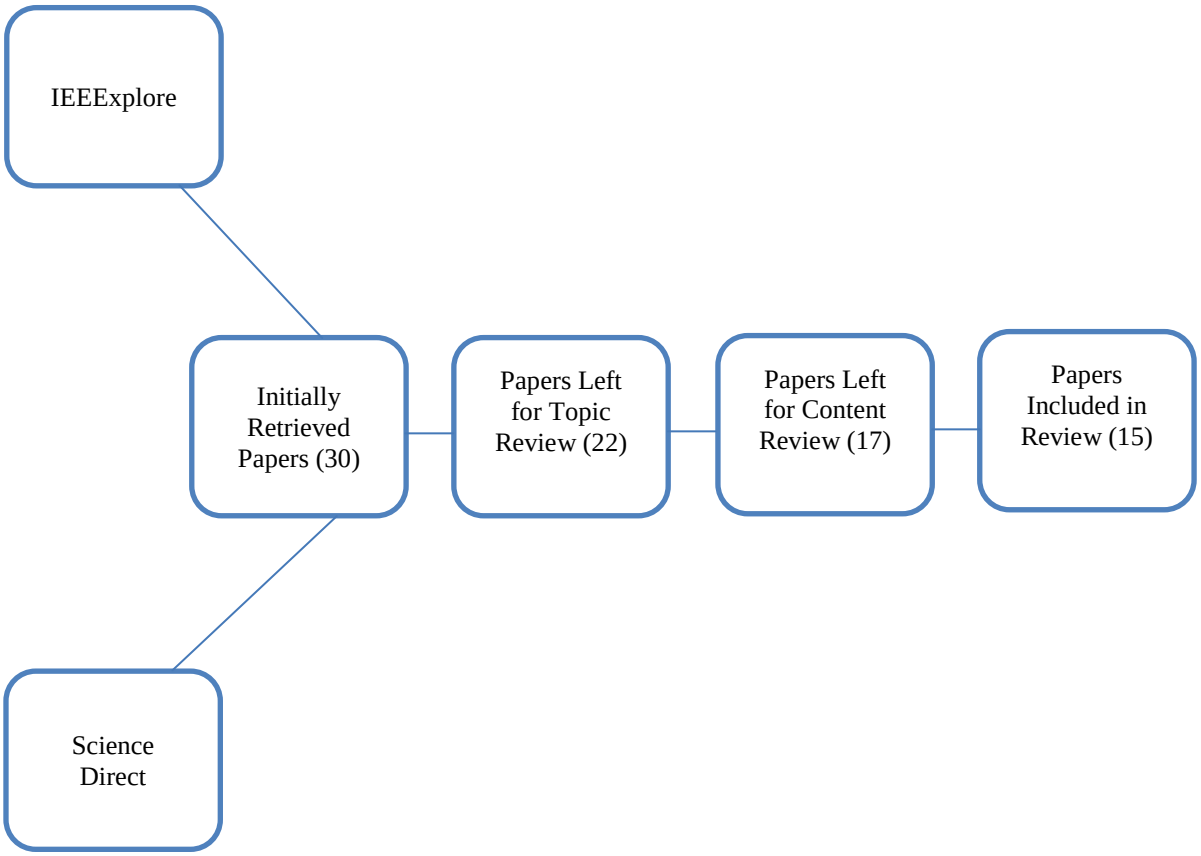

Figure 1: Procedure followed to retrieve the papers for literature survey

III.DIFFERENT APPROACHES FOR INFORMATION RETRIEVAL

3.1. INFORMATION EXTRACTION WITH CLUSTERING AND KEYWORD RANKING

One of the approaches for information extraction is through clustering and keyword ranking. Ying Qin [8] had implemented the framework for location information extraction and keyword extraction from the single document. The term frequency is applied on the Chinese document. Text ranking and unsupervised methods are used for comparison between the user document and corpus. Experimental results are shown by using 'AND' and 'OR' logic. Clustering algorithm can be used for graph based information extraction [15].

Bernardus AriKuncoro and Banbang Heru Iswanto [9] had done ranking keywords of Instagram user's image caption. Multiple words can also be weighted and retrieved with the help of TF-IDF [11][14]. The top twenty users' image captions are taken and the term frequency and inverse frequency factor of particular word is calculated. The top twenty user name has been taken as a input. According to the term frequency and inverse document frequency the ranking of top 10 keywords is done. Due to the TF-IDF the ranking of words is done easily. Prafulla Bafna, Dhanya Pramod, Anagha Vaidya [10] has used term frequency and inverse document frequency for document clustering. Clustering and ontology together helps to retrieve information [10][13]. The dataset used from different area such as News20, Reuters, emails, research papers etc.

3.2. INFORMATION EXTRACTION USING GRAPH TECHNIQUE

Another approach for information extraction is to construct the schema graph. The graph helps to categories the words. Information extraction has also done with the graph and text ranking. [15] Document is split into sentences to words accordingly the graph is constructed. Chaleerat Thamrongchote and wiwat vatanwood [6] have proposed business process schema graph for defining user story. The user stories are the small card which provides the requirements of the user. The card describes in the role-action-object format. The template of the user story is collected accordingly classes are defined and hierarchy of ontology is created. Schema graph of ontology has constructed. The relation between ontologies is described and the synonym is found to reduce nodes of ontology. The Chaleerat Thamrongchote and wiwat vatanwood had [6] defined two properties they are perform object and perform action in the ontology schema graph. Perform action links role class to action class and perform object links action class to object class. The subclass and hierarchy of ontology is also defined.

Jie Tao, Amit V. Deokar and Omar F. El-Gayar [4] have designed the framework for processing the textual format of initial public offering prospectus. The relationship between entities in the IPO prospectus is identified. Classes have defined from the prospectus. Three modules have implemented in this paper information extraction module, reasoning and learning module and analytics module. This proposed framework is useful for the average investors. The limitations of this framework are evaluation metrics are not more developed. The Kaijian Liu and Nora El-Gohary [3] had done validation using precision, recall and F-measure. Jie Tao, Amit V.

Deokar and Omar F. El-Gayar [4] had done validation against proposed framework using delta analysis and contextual analysis.

3.3INFORMATION EXTRACTION USINGTERM FREQUENCY AND INVERSE DOCUMENT FREQUENCY

Taizhong Guo and Tao Yang [1] hasanalyzed the weight of words which are used in unstructured data classification in big data. They have focused on the traditional term frequency-inverse document frequency algorithm and traditional weight calculation method. The TF-IDF is calculated as TF × IDF. This paper modifies the traditional TFIDF algorithm formula.

Mutual information and document frequency are used for domain ontology extraction [5]. The pre-processing is done from Chinese text. Mutual information is used to identify correlation two words in a set. N-gram algorithm is generated for two-word phrase. Term frequency is calculated between the words. Linguistic rules are used for screening. The accuracy of the results is not mentioned.

IV.ONTOLOGY BASED FRAMEWORK FOR INFORMATION EXTRACTION

Ontology is one of the fields of information extraction. Ontology framework can be used to extract information. Semantic web can be used with ontology. T. Muthamilselvan and B. Balmurugan [2] have focused on cloud automated framework which helps to retrieve the relevant documents. In the proposed framework they have worked on two ontologies. The semantic web is used as a tool to retrieve the documents. Semantic web is more widely used with ontology framework [2][12]. The dyadic deontic logic rule is used with graph derivation representation for semantic rich ontology. Similarity measures are calculated using cosine rule between two documents. The framework is used for e-health applications. In this paper [2] the solution is provided by constructing the ontology structure to handle polymorphism in ontology representation and it aims to estimate the degree of similarity. The accuracy of the retrieved document is not mentioned. Multiple ontologies has been used for performance evaluation with diabetes ontology. The GDR-DYDL has been computed with UML-GM, GDR-DL, GDR-DEOL for similarity measurements.

Kaijian Liu and Nora El-Gohary [3] have proposed information extraction framework. This framework automatically recognizes and extracts data. Multiple ontologies have been used. First ontology is used for sequence labeling with term identification and another for grammar for relationship association. The conditional random field is used to drive ontology based sequence labelling. To reduce the human work this paper uses machine learning. The [3] Kaijian Liu and Nora El-Gohary has derived four modules. First they had done pre-processing on the data which includes tokenization, sentence splitting and morphological analysis. Second module does the features extraction with the token associated. Next the ontology based sequence labelling training is done using condition random field with feature extraction. And then the performance is measured.

Tarek Helmey, Ahmed Al-Nazer, Saeed Al-Bukhitan, Ali Iqbal [7] have aimed to retrieve the information of users query. The queries are related to the health, food and nutrition. Multiple ontologies are used to retrieve the information from various domains. The ontologies from various domains are integrated to answer the user queries and for multi-lingual support.

Proposed Method	Used technique and tool	Used Text document and dataset	References
Cloud based automated framework for semantic rich ontology construction and similarity computation for E-health application	GDR, DDL rule, ONTOLOGY FRAMEWORK	Diabetes dataset from UCI repository	[2]
Proposed method for processing the textual content of IPO prospectus with ontology based Information Extraction	GATE, JAPE, APOLDA, SWRL, ONTOLOGY	SEC EDGAR database by SEC	[4]
Proposes the method to extract ontology concept from multiple text of same type	JIEBA WORD SEGMENTATION, TF-IDF, N-GRAM ALGORITHM	Corpus from network	[5]
Business process ontology for defining user story	SCHEMA GRAPH, ONTOLOGY	The user stories are collected from historical project data	[6]
Ontology based semantic information retrieval system and Jena semantic web framework	RDF, SPARQL, JENA API, ONTOLOGY	Free- text document	[12]

Figure 2: Example of some ontology based information retrieval methods

V.CONCLUSION

The purpose of this survey paper was to get the deep knowledge about the researches made in these recent years on the various approaches of information extraction, ontology based framework and TF-IDF factor etc. With this Survey we provide general overview on IE, ontology framework and TF-IDF factor. In this survey paper we achieved the two goals. Primarily we studied the different approaches of information extraction. On

comprehensive study about the research papers of various publications we came into conclusion that the information extraction can be done with the help of ontology. Secondary the research was done on ontology based information extraction. While performing the survey of information extraction, we found that the information can be extracted by constructing ontology framework with consideration of multiple approaches.

REFERENCES

[1] Aizhang Guo, Tao Yang, "Research and Improvement of feature words weight based on TFIDF Algorithm" IEEE 2016

[2] T.MuthamilSelvan, B.Balamurugan, "Cloud based automated framework for semantic rich ontology construction and similarity computation for E-health applications" 2352-9148, 2016 Elsevier Ltd

[3] Kaijian Liu and Nora El-Gohary, "Ontology-based sequence labelling for automated information extraction for supporting bridge data analytics" 1877-7058 Elsevier Ltd 2016

[4] Jie Tao, Amit V. Deokar and Omar F. El-Gayar, "An Ontology-based Information Extraction (OBIE) Framework for Analyzing Initial Public Offering (IPO) Prospectus", 978-1-4799-2504-9/14 IEEE 2014

[5] Yuefeng Liu and Minyoung Shi, Chunfang Li, "Domain Ontology Concept Extraction Method Based on Text" 978-1-5090-0806-3/16, 2016 IEEE, ICIS 2016

[6] Chaleerat Thamrongchote and wiwat vatanwood, "Business Process Ontology for Defining User Story" 978-1-5090-0806-3/16, IEEE 2016, ICIS 2016

[7] Tarek Helmey, Ahmed Al-Nazer, Saeed Al-Bukhitan, Ali Iqbal, "Health, Food and User's Profile Ontologies for Personalized Information Retrieval" Elsevier B.V 2015

[8] Ying Qin, "Applying Frequency and Location Information to Keyword Extraction In Single Document" 978-1-4673-1857-0/12 IEEE 2012

[9] Bernardus Ari Kuncoro and Banbang Heru Iswanto, "TF-IDF Method in Ranking Keywords of Instagram User's Image Caption" 978-1-4673-6664-9/15 IEEE 2015

[10] Prafulla Bafna, Dhanya Pramod, Anagha Vaidya, "Document Clustering: TF-IDF" 978-1-4673-9939-5 IEEE 2016

[11] Eko Darwiyanto, Ganang Arief Pratama, Sri Widowati, " Multi Words Quran and Hadith Searching Based on News Using TF-IDF" 978-1-4673-9879-4 IEEE 2016

[12] Amol N. Jamgade, Shivkumar J. Karale, "Ontology Based Information Retrieval System for Academic Library" 978-1-4799-6818-3/15 IEEE 2015

[13] Aradhana R Patil, Amrita A Manjrekar, "A Novel Method To Summarize and Retrieve Text Documents Using Text Feature Extraction Based on Ontology" 978-1-5090-0774-5/16 IEEE 2016

[14] Mohamed K. Elhadad, Khaled M. Badran, Gouda I. Salama, "A Novel Approach for Ontology-based Dimensionality Reduction for Web Text Document Classification" IEEE ICIS 2017, Wuhan, China

[15] Yan Ying, Tan Qingping, Xie Qinzheng, Zeng Ping, Li Panpan "A Graph-based Approach of Automatic Keyphrase Extraction" 1877-0509 ICICT 2017