ISSN(Online): 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Website: www.ijircce.com

Vol. 6, Issue 3, March 2018

A Survey on Enhanced Method for Retrieval of Information with the help of Ontology

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

M. E Student, Dept. of CSE, Kolhapur Institute of Technology College of Engineering, Kolhapur, India[1]

Head, Dept. of CSE, Kolhapur Institute of Technology College of Engineering, Kolhapur, India[2]

ABSTRACT: To search the documents as per user need and to retrieve the documents relevant to user document is a challenging and time-consuming task. This paper helps to fill this gap with the help of term frequency and inverse document frequency score. To recognize whether the retrieved information is really relevant to our document is a part of analysis which is an important factor. This paper provides the ontology based information retrieval system.

KEYWORDS: ontology, term frequency, inverse document frequency, information extraction, similarity measures

I. INTRODUCTION

Information extraction is one of the major needs of each and every other person. In day to day life as the world is getting more digitized and with increasing technology huge amount of data is generated. To extract the meaningful information from the huge amount of data is one of the time consuming work.

To manually go through the data & to extract the needed information is not that much possible for the human being and also that extracted information may be not accurate, there may be human errors. This problem is overcome by the development of the information extraction system which will automatically extract the needed information from the bunch of documents. Information extraction framework can be used many other application which will help to extract the information. To extract the information, the frequencies of the words available in the document need to be identified. The highest frequency of words will provide the priority ratio of the document. To find the frequency of the world can be done with term frequency & inverse document frequency.

Ontology Framework is also used for information extraction. Ontology based information extraction system works well to extract the information. Ontologies is a shared conceptualization. Ontology helps to derive the relationship between domain & entities. This paper helps to extract the needed information based on user input by constructing ontology based information extraction.

II. RELATED WORK

Aizhang Guo and Tao Yang [1] has analysed 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. This paper modifies the traditional TFIDF algorithm formula. The results are shown with the comparison between traditional TFIDF algorithm and modified TFIDF algorithm.

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. The dyadic deontic logic rule is used for graph derivation representation. 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 limitations of the paper are it is time-consuming to use graph derivation technique. The accuracy of the retrieved document is not mentioned.

DOI: 10.15680/IJIRCCE.2018.0603034

Kaijian Liu and Nora El-Gohary [3] have proposed information extraction framework. This framework automatically recognizes and extracts data. Ontology is used for sequence labeling with term identification. The conditional random field is used to drive ontology based sequence labelling. To reduce the human work this paper uses machine learning.

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 with the help of ontology based information extraction. 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 paper focused on single ontology but multiple ontologies can be used.

Yuefeng Liu and Minyoung Shi, Chunfang Li [5] have focused on extraction from multiple texts of the same type. Mutual information and document frequency are used for domain ontology extraction. 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 limitation of this paper is ontology is used for two-word phrase only. The accuracy of the results is not mentioned.

Chaleerat Thamrongchote and wiwat vatanwood [6] have proposed business process ontology 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.

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.

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.

III. EXISTING SYSTEM

There are lot many traditional methods to extract the information. The information extraction is done with the help of clustering, keyword ranking, machine learning etc. The clustering technique helps to cluster the group of similar words & a group of dissimilar words. On the basis of the cluster the information is retrieved. K means algorithm is also used by the traditional methods which help to find the needed information. Key word ranking is the method used which helps to rank the keywords according to the priority. The traditional method help to extract the information but the accuracy is the measure cause.

APPROACH 1: Pre-Processing Technique

The different techniques such as name entity recognition (NER), Conditional random fields and relation extraction are much more helpful for information extraction. NER helps to find the entity in the real word from sequence of words. The entity is recognized by string comparing and matching with the dictionary.

Conditional Random field (CRF) are generally used for sequence labelling [3] with the help of ontology. The sequence labelling is one of the probabilistic method. Complete sequence of words are used and not the single part of sentence or a single word.

Stop word removal and stemming is another method used for pre-processing where the words which are useless are the stop words. The words like "is, an, the" are stop words and that are eliminated from the sentences. Stemming is where the words contains "ing, ion" are separated and only the major part of the word is kept.

APPROACH 2: Information Extraction using Clustering & Keyword Ranking Technique

The extraction of information can be done with the help of clustering. Clustering is the major technique used. In the clustering the similar documents are grouped together and dissimilar documents are grouped together [9]. Cluster is nothing but the group of words or documents which having similar property. Clustering is one of the important traditional methods. Clustering is used for classification, visualization, document arrangement etc. There are many algorithms used which are used for clustering. The clustering can be agglomerative or divisive, that is bottom up & top down approach. K-Means algorithm is widely used to clustering. In k-means the documents are divided into partition of K cluster.

Keyword Ranking can be done with the calculation of term frequency and inverse document frequency factor[1]. The occurrence of the word in the document is calculated that is frequency of word, which is call term frequency. The TF-IDF helps to weight the words and to find the frequency of each world so that the word can be extracted priority wise preceding highest priority.

APPROACH 3: Information Extraction using Ontology

Ontology helps to extract the information in meaningful way. Ontology deals with what entities exist and how such entities are grouped. Ontology has the ability to reuse the data or information [2]. Ontology can be used for information retrieval, semantic web, Knowledge management and Recommendation system.There are ontology editors which help to create the ontology. Editors such as Protégé, Apollo, SWOOP etc., are available.

IV. PROPOSED SYSTEM

The proposed system defines a framework to information extraction with ontology repository. Input will be the user document & output will be the extracted information.

The proposed system of information extraction is shown in figure.1. The working of the proposed system is described with following blocks.

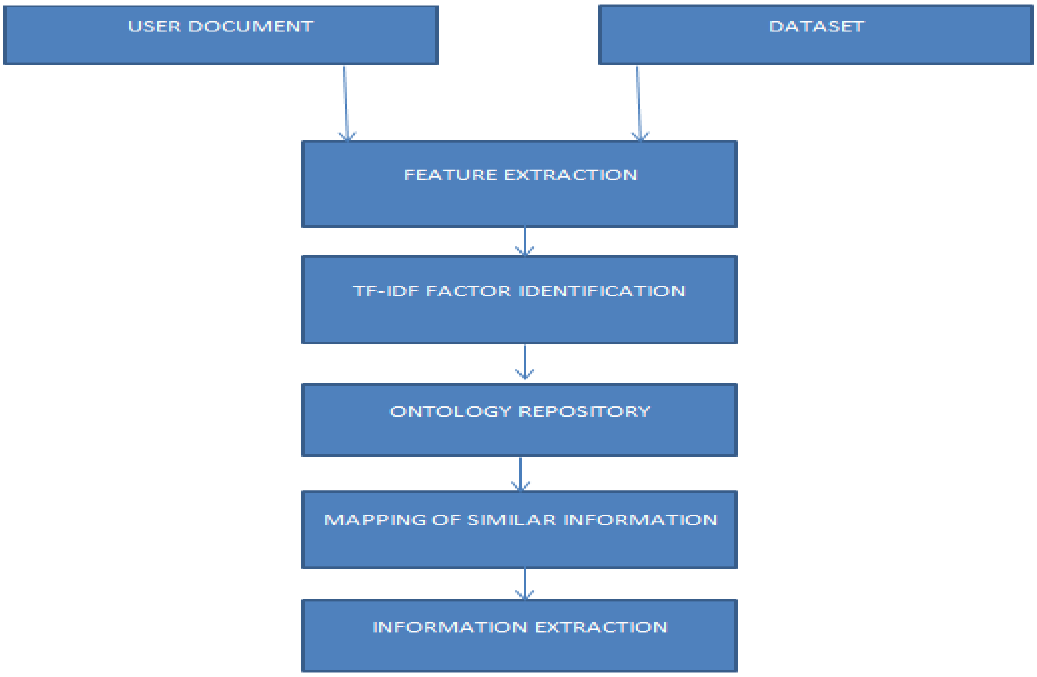

Figure 1: Block diagram of Information retrieval System

1 Feature Extraction: -

The feature extraction consists of Stopwords Removal and Stemming. Stop words in the document means the words which are useless, which don't provide any valuable information. Stop words removal is to remove words like "is, an, the". If such words are found in the document then they are deleted from the document. Stemming is another part of pre-processing. The word which ends with suffixes (for example: "es, ing") are found and then those are separated from the word and removed. The rest part of the word is kept as it is.

2 TF-IDF Factor Calculations: -

Term frequency inverse document frequency (TF-IDF) helps to calculate the frequency of words. Term frequency means how many times the words are present in the single document and the inverse document frequency means to calculate the frequency of the words from overall document i.e. from the dataset.

3 Ontology Repository: -

Ontology Repository consists of .owl file which is generated from the construction of ontology.

4 Mapping of similar information and Information Extraction: -

This will help to map the similar information from user input information and the dataset. On the basis of the mapping the threshold is set to extract the highest matched information.

VI. CONCLUSION

In this paper various methods for extraction of information are studied and analysed. We finalise enhanced approach to information extraction by using ontology. In this paper we have proposed framework consists of pre-processing of the user document it includes stop word removal and stemming etc. The same procedure is done on the set of documents (dataset). After the feature extraction done on the dataset and user document, term frequency and inverse document frequency factor are calculated. We have designed a modified approach to extraction of information.

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

[10]. 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

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

[12]. 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

DOI: 10.15680/IJIRCCE.2018.0603034