IJCSE International Journal of Computer Sciences and Engineering Open Access
Research Paper Vol.-7, Issue-5, May 2019 E-ISSN: 2347-2693

"Efficient Retrieval of Relevant Documents by Constructing Ontology Framework"

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

[1]Dept. of Computer Science and Engineering, Kolhapur Institute of Technology College of Engineering, Kolhapur, India
[2]Head of Dept. CSE, Kolhapur Institute of Technology College of Engineering, Kolhapur, India

Corresponding Author: sharvali_sarnaik@yahoo.co.in

DOI: https://doi.org/10.26438/ijcse/v7i5.17371740 | Available online at: www.ijcseonline.org

Accepted: 23/May/2019, Published: 31/May/2019

***Abstract*-**Information retrieval has a motive for obtaining the meaningful information on the basis of user demand. Information retrieval plays a major role in providing the information from huge amount of documents as per the requirements. Now days, the huge amount of data has been spread all over the world. We acquire data from various sources viz; internet, social media etc. some data is created by ourselves. In our system we have lot of documents stored but it is very difficult to address meaningful document or to find the information which relates our document. It is time consuming task to collect the needed information or document from the dataset available with us. In this paper, the focus is done over the information retrieval by constructing ontology framework. TF-IDF will help to find frequency of word present in document which will help to get the weightage of document. Input will be dataset & user document and the output will be documents matching the user document. The threshold is set to retrieve the accurate documents.

***Keywords*-** Information retrieval, Feature extraction, term frequency& inverse document frequency, ontology

I. INTRODUCTION

In the recent years the growth of population is increasing as well as the growth of technology is increasing rapidly. Due to the increase in technology, the bulk of information is generated and it is difficult to scrutinize the information. The information is increasing from the internet, social media etc. Large amount of the data is generated in text format. There are lots of documents present hence it is really hard for the user to search for the documents as per the demand. The text documents are in pdf, txt format. Large amount of documents sometimes does not have keywords to identify. Some people highlight the sentences from the document and important keywords in the document while creating the document, still when the person needs the relevant document it takes time to identify those keywords from the large document. To filter the information is the major and difficult task. The information retrieval system helps to retrieve the possible matched documents with the help of input query.

One method to retrieve the documents from the huge amount of document (dataset) is with the help of constructing ontology framework. Ontology helps to reuse the domain knowledge [2],[7]. Ontology helps to define the classes, set of instances and propertied [4]. Ontology can be used in the medical field, knowledge management, information retrieval, semantic search, web etc. Ontology framework can be used with any other application to retrieve the information. Since ontology can be reuse it is not necessary to construct new ontology every time[12]. In this proposed system the ontology framework for desktop application is constructed. The user document is taken as an input query. Feature extraction helps to reduce the noisy data from dataset.

With the help of term frequency-inverse document frequency (tf-idf) weight of words are identified, which help to find the highest priority document having related similar words to the user document. Input will be dataset & user document and the output will be documents matching the user document. The threshold is set to retrieve the accurate relating documents.

II. RELATED WORK

Aizhang Guo and Tao Yang [1] have analysis weights of feature word. They modifies the term frequency inverse document frequency 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.

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.

Yuefeng Liu and Minyoung Shi, Chunfang Li [5] have focused on text 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.

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.

Bernardus Ari Kuncoro 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]. Ying Qin [8] had implemented the framework for location information extraction and keyword extraction from the single document. 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].

III. METHODOLOGY

The proposed system designed with the original user document and the publicly available dataset. In this proposed system the documents which are relevant to the user input document is extracted hence there is need of a framework for information extraction. The set of documents are given as an input to the system. To find the useful words from the document the frequency of the words which occur in the document are checked. We can quickly search the efficiency of words from the documents. The ontology file is created and to store the ontology file the ontology repository need to be generated. To get the similar documents as per user input document, the similarity between the documents need to be checked. The main goal of the paper is to provide the relevant documents to the end user.

Following process takes place to reach the objective of paper:

1) Input: User document and dataset.
2) Output: Fetch data from user document and provide relevant files to user.
3) Validations: Identify similar words and its frequency of occurrence from input datasets.

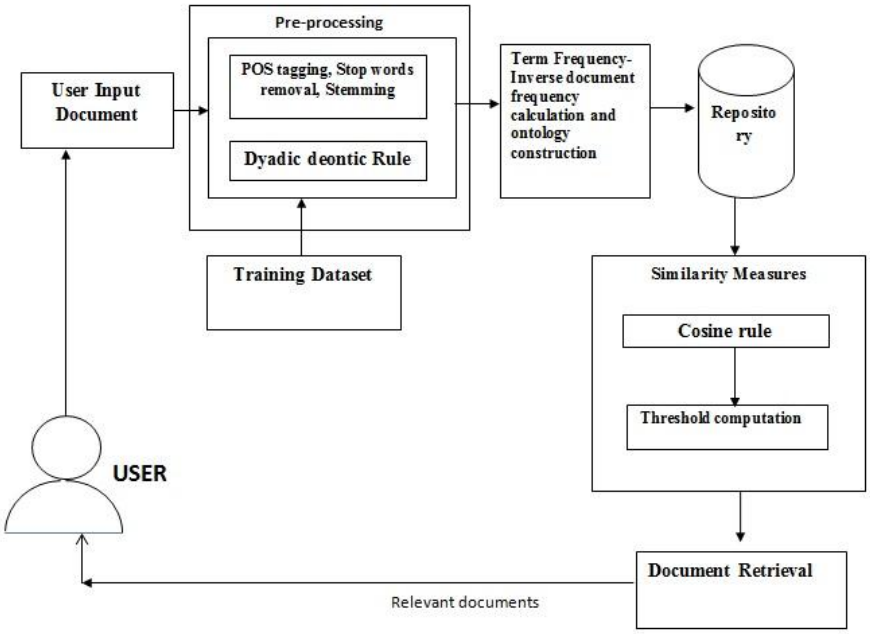

Figure 1: Block diagram of efficient retrieval of relevant documents

Stage 1: Pre-processing
In this stage cleaning of data is done. The dataset is taken from available valid dataset from internet. The pre-processing is performed on dataset in the training phase where the noisy data is removed from the documents. There are many missing values and inconsistent data in the documents which is not meaningful & it is of no use. Hence that should be removed. Stop words are removed from the uploaded data. 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. The porter stemming algorithm is used. After pre-processing, logical rules were created to form the sentences as per clauses, which will help to generate ontology.

Stage 2: Term Frequency-Inverse Document frequency factor calculation & ontology construction:
The module involves assigning weights to the words and calculating the weight of the words in the documents. After stemming & pre-processing, the sentences which we get in the pre-processing are now used to calculate the TF-IDF factor.

We calculate TF-IDF by using the traditional formula:

$$Weight = \frac{tf \times \log(\frac{N}{n} + 0.01)}{\sqrt{\sum_1^N tf^2 \log(\frac{N}{n} + 0.01)}}$$

After calculating the TF-IDF score with the help of formula, the .owl file is created and all .owl files are stored in ontology repository.

Stage 3: Similarity Measure
After pre-processing the user input document and finding the word frequency, the ontology of words is generated forming repository. The next step is to find the similarity between the words from documents. We use Cosine similarity measures to measure the similarity between the documents. The similarity is calculated basically for good and quick results. Here the similarity measure shows how many documents having similarity with the given input. Cosine similarity can be calculated as:

$$similarity(doc1,doc2) = cos(\theta) = \frac{doc1 doc2}{|doc1||doc2|}$$

Stage 4: Retrieval of documents:
The final step is to fetch the matching documents. The limit of document is computed, it helps to fetch the matched documents. After certain experimental results, we will set the limits which will help to retrieve relevant documents. Following is the algorithm for relevant document extraction:

Pseudo code: Relevant document extraction
Inputs:
Input A: User document (document provided by user called as base document)
Input B: Dataset (set of documents which are available publicly)
Output: set of relevant documents
Algorithm:
d= user document
D= documents from dataset
cnt= counter
x= index of relevant documents
y= index of irrelevant document
Input a= user document
Input b= documents from dataset
For loop d=1
    For loop cnt=1 to n

    If
    The limit of a & b is 0.5-1.00, input a and input b corresponding to each other hence index is incremented;
    x=x+1;
Else
    Irrelevant index is incremented
    y=y+1;
    Increment counter cnt=cnt+1;
End loop
End for loop

IV. RESULTS AND DISCUSSION

Results analysis is done through recall and precision. For training dataset the publically available dataset is taken. It contains 350 text files. The graph shows the performance evaluation with the input of dataset.

Table 1: Performance evaluation on Training dataset

Dataset(no. of documents)	Precision	Recall
150	0.6	0.3
200	0.7	0.4
250	0.8	0.5
300	0.9	0.6
350	1	0.7

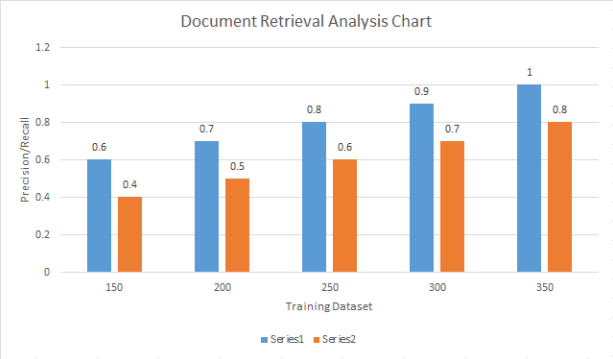

Figure 2: Document retrieval analysis graph

V. CONCLUSION AND FUTURE SCOPE

The paper represent a framework for information retrieval, in which graph is used to represent the relevance relations among the documents. The user gets the relevant documents that we say matching documents on user need. The term frequency and inverse document frequency helps to find out the weightage of the words, the frequency of the words present in document, which will help to get the most useful content document. The logical rule helps to form the statement, which helps to create ontology. The similar documents are measured with the help of cosine similarity and the threshold value helps to retrieve the document. In future, more rules can be generated and the accuracy of the document can be measured by time efficiency. More effective algorithm for retrieval can be generated.

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

[12] 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.

[13] 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.

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

[15] 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.

Authors Profile

Miss. Sharvali S. Sarnaik received her degree in Bachelor of Engineering in Information Technology from Kolhapur Institute of Technology, College of Engineering, Kolhapur, India in the year 2016. She is pursuing Master of Engineering in Computer Science and Engineering at Kolhapur Institute of Technology, College of Engineering, Kolhapur, India. Her Master of Engineering dissertation work based on information retrieval with the help of ontology construction. She is currently working as lecturer in department of Information Technology at Government Polytechnic Kolhapur, India. Her research interests includes Data mining, Web mining, Information retrieval and Ontology.

Mr. Ajit S. Patil received his degree in B.E Computer Science and Engineering from Government College of Engineering, Aurangabad, India in the year 2000. He has completed his M.Tech in Computer Engineering from Dr. Babasaheb Ambedkar Marathwada University, Raigad, India. His M.Tech dissertation work was based on comparisons of different distributed token circulation algorithm in Mobile Ad hoc Networks. His research interests include Computer networks, Mobile communication, and Delay tolerant networks. He is currently working as an Associates Professor in department of computer science and engineering at K.I.T's College of Engineering, Kolhapur, India and pursuing his PhD at Walchand College of Engineering, Sangli, India.

