Plagiarism Scan Report

Summary	
Report Genrated Date	22 Apr, 2018
Plagiarism Status	**100% Unique**
Total Words	892
Total Characters	5943
Any Ignore Url Used	

Content Checked For Plagiarism:

ABSTRACT

Many applications such as Text Summarization, Semantic Web, Search Engine Optimization, Sentiment Analysis, and such others make use of Document Classification as a key component in their realization. In order to aid the process of Document Classification, many of these applications rely on extracting domain specific keywords. The existing techniques used are pure NLP and extraction based on only term document frequency. However, these do not always guarantee accurate results. In our paper, we present an ontological approach to extraction of keywords which will give more precise results as they are based on the context of the search. This is done by creating domain-specific ontologies and using them to extract keywords present in the user's document.

Keywords: - Ontology, keyword extraction, entropy, domain analysis, contextual search

1. Introduction
There is a need for contextual data classification as many applications such as Search Engines,Text summarization depend on it.One important module of data classification is keyword extraction module which tells us what the document talks about. The existing keyword extraction softwares use methods that do not consider the domain of the document. Recognizing the domain helps us to extract only the relevant words pertinent to that document and not other frequently occuring words in that document thus decreasing the noise in the extracted keywords.
Upon research we found that an ontological approach might give us better results. An ontology is broadly, representation of knowledge. All the concepts related to a particular entity are identified and a relation is established between them.Ontology is created such that the machine understands these relation between words in terms of classes,subclasses and properties associated with the concepts.
In the paper we discuss a method to incorporate ontology for keyword extraction. The document entered by user along with the domain of the document is pre-processed then scoring parameters are applied to obtain candidate words and later these candidate words are mapped with words/concepts in the ontology for a particular domain.The later sections describe this process in further detail.

2. Literature Survey
Ontology is an important module in our proposed system.In order to appreciate it better

and to understand how it facilitates keyword extraction, we did a literature survey to understand how an ontology is created and its various applications.The following paragraph briefly discusses few papers that explains various algorithms for ontology creation.
In Paper [1] a keyword dictionary server is introduced that helps in keyword expansion using domain specific ontologies. it has been used to categorise web service keywords which have been classified on the basis of similarity calculation between two keywords.
Paper [2] talks about the skeleton of a semantic search engine that allows automatic query expansion. Firstly, a SPARQL query is built and later it is fired on the knowledge base to find appropriate RDF triples.Then, Web documents relevant which specified in the triples are fetched and ranked according to their relevance to the user's query and then are sent to the user. In Paper [3] the authors have found out synonyms using WordNet for user's query. A technique called ontological indexing is used which is based on calculating the context of the words in the provided document using ontology.
In paper [4], domain experts have created an ontology which is then supplied to the system. Here authors have discussed two algorithms : "semantic information extraction algorithm" and "semantic information re-recognizing algorithm". Text information is then extracted using created ontology and the two proposed algorithms. Paper [5] talks about using Ontology Based Information Extractors(OBIE) which is used for text grading.Authors highlight that the combination of OBIE which perform different functions provides a much better understanding of a graded text, and the ones with different functions can improve system performance.
In Paper [6] authors have used ontology to find relevant recent knowledge in the domain by exploiting their underlying knowledge as keywords. Using ontology-based and pattern-based information extraction technique it extracts instances and statements from the documents. Then a confidence value is used to maintain the stability of the ontology. Finally, the paper discusses a way to expand the ontology with the newly extracted keywords to validate the knowledge inside ontology.
Paper [8] reviews the concepts and methods related of ontology construction and extension, and also proposes an automatic ontology extension method based on supervised learning and text clustering. Paper [9] proposes a way to extract ontology directly from RDB in the form of OWL/RDF triples, for semantic web using direct mapping rules. Then, SPARQL queries are rewritten from SQL by translating the relational algebra.
In paper [7], authors have compared 11 ontology learning models. After proper analysis, five techniques for ontology learning and creation stood out in terms of accuracy, f-measure and precision.They are-

Ontolearn [10] which uses a text mining and statistical approach to learn concepts and build taxonomic relations.
The second method is Text2Onto [11] that makes use of a 'Probabilistic Ontology Model' and involves statistical and linguistic techniques to create an Ontology.
The CRCTOL [12] algorithm is also a statistical algorithm and extract concepts and relation using statistical approach
In OntoGain [14] which is an unsupervised algorithm a linguistic tool is used to preprocess text and extract relevant concepts.
HCHIRISM [13] is the other unsupervised algorithm which first crawls through a large number of websites to find relevant concepts for a given domain by using an initial keyword which is closely related to the domain.

Plagiarism Scan Report

Summary	
Report Genrated Date	22 Apr, 2018
Plagiarism Status	**100% Unique**
Total Words	783
Total Characters	4705
Any Ignore Url Used	

Content Checked For Plagiarism:

3. Proposed Model
The goal of the system is to extracted domain specific keywords based on ontological concepts. The proposed system consists of an ontology store initially. This ontology store can be created using web pages or existing downloadable ontologies can be collected and stored. The user inputs the text document along with the domain she seeks to find the keywords pertaining to. The system scans user's document to look for domain related keywords using the specific domain ontology from the ontology store. An elaborate procedure is given in Figure 1. The system consists of two main modules: Creating ontologies and Extracting keywords using these ontologies.

3.1 Ontology Store Creation
Ontology Store Creation can be done by accumulation of ontologies of various domains which are available on the internet. The system will be able to access these ontologies based on the domain the user wants. These will help speed up the process of ontology store creation. The results will be better if we consider the next approach.
The next approach to create the ontology store is by creation of ontologies using web pages. This system is semi-automatic since there is a need of expert intrusion when it comes to checking or verifying if a particular keyword belongs to the respective domain or not. This approach is time-consuming but can provide better results once the ontology store is created because of the availability of exhaustive ontologies. When scanned web pages are used, more keywords will be put into ontologies and thus, more keywords will be mapped from user's document while extraction.
The basic steps for this approach are:
Crawling of domain specific web pages.
Extracting terms from web pages by scraping.
Formulating triplets of subject-verb-object or noun-verb-noun.
Identifying the classes, relations and individuals.
Creating the OWL ontology.
Saving the ontology
3.2 Keyword Extraction
The keyword extraction module os the main module of the system. Here, we map the words in user's document along with the words in the respective domain ontology from the ontology store. The system, initially, pre-processes the data given by the user. The pre-processing consists of the following steps:
Converting the entire piece of text into lowercase.
Removing special characters from text.

Removing stop words from text.
Lemmatizing the text. (i.e. converting the word to its root form)
Once the data is pre-processed, we apply three scoring parameters from [15] to identify the relevance of the word to the topic. They are:
Entropy: Using frequency directly for calculation can sometimes misguide us; there could be some noisy words which may have very high frequency while relevant words may have less. Thus, instead of using the parameter frequency, we have used entropy. The formula for calculating entropy is as follows:
W1 = FNlog2(FN)
where W1 = Entropy of word in given document.
F = Occurrence frequency of word in document
N = Total Number of words in document.
Position of sentence: We consider the position of sentence where the word exists in the document. The idea behind this is that words in initial paragraphs carry more weightage than words in last. This is given by:
W2 = (St + 1Sf + 1)
where W2 = weight of given word, due to index position of the sentence in which the given word occurs first.
St = Total number of sentences in given document.
Sf = Sentence Index in which the given word occurs first.
Position related strength: Position related strength is calculated using two factors viz. Position of given word in the sentence and the length of that sentence. The idea behind this is that, a word has higher weightage when it comes in the initial part of the sentence than the rear.
Let, IK = Index position of Candidate Word "K" in given sentence "S".
LS = Length of sentence "S" in which the candidate word "K" is present.
P(K) = I(K) if (I(K)<(L(S)/2))
= 2 (L(S) - I(K)) else
where P(K) = Partial Position related strength of given distinct word
We combine strength due to length of sentence with the formula:
W3 = log2(L(S) + 1P(K) + 1
where W3 = Weight value of given distinct word, calculated by using position related strength of word in sentence and length of sentence in which it exist.
After applying the above three parameters, we multiply them to get a final score. A threshold is set which is the average weights of all words, and candidate words are filtered which are above average. These candidate words are them mapped to the ontology, to get the domain specific keywords. The arrangement of words according to their scores gives us a proper measure of relevance of keywords in the document to the given domain.

Plagiarism Scan Report

Summary	
Report Genrated Date	22 Apr, 2018
Plagiarism Status	**100% Unique**
Total Words	393
Total Characters	2540
Any Ignore Url Used	

Content Checked For Plagiarism:

4. Results
The above described approach is analysed on basis of accuracy,precision, recall and f-measure. Before discussing them in greater detail, it's important to realise that the accuracy of the approach depends heavily on the ontology. An exhaustive ontology will most definitely give much better result than an ontology with few classes and property.
The domain selected to test the system is "Library". An ontology was created for the same. The user document sample was taken from wikipedia article on library of roughly 700 words.
Accuracy is the ratio of correctly classified observation and total no of observations. Precision deals with correctly classified observation by total positive observations. Recall is ratio of correctly classified observation and all positively identified observation.F1 score is ratio of precision and recall. The corresponding values for our system are as follows:
Accuracy = TP+TN/(TP+FP+FN+TN) = 0.95
Precision = TP/(TP+FP) = 0.54
Recall = TP/(TP+FN) = 0.51
F1 Score = 2*(Recall * Precision) / (Recall + Precision) = 2
Where,
TP is true positive, which is correctly classified positive observation.
TN is true negative, which is correctly classified negative observation.
FP is false positive, which is incorrectly classified positive observation.
FN is false negative, which is incorrectly classified negative observation.

For our document the values are:
True Positive(TP) : 18 True Negative(TN) : 668
False Positive(FP) : 20 False Negative(FN) : 15

5. Conclusion
We realized the need for a contextual extraction for keywords and found that an ontological approach gives accurate results.We have created a system which can be broadly classified into two modules- keyword extraction and mapping candidate words with those present in the ontology.This approach can be generalized for any domain and for multi domain systems also. The accuracy of the system lies in the ontology used for this purpose. Using gold standard ontology is expected to give best results but because only few are available we can create our own as defined briefly in this paper and make it as exhaustive as possible.

6. ACKNOWLEDGEMENT

We thank our college Vivekanand Education Society's Institute of Technology for considering our project and extending help at all stages throughout the implementation of the project. We would like to extend our gratitude to Head of the Computer Department Dr. (Mrs.) Nupur Giri and our Principal Dr. (Mrs.) J.M. Nair, for giving us the valuable opportunity to do this project.