Plagiarism Scan Report

Summary	
Report Genrated Date	03 Nov, 2017
Plagiarism Status	**100% Unique**
Total Words	543
Total Characters	3586
Any Ignore Url Used	

Content Checked For Plagiarism:

Abstract— In Semantic Web, Text Summarization, Search Engine Optimization and many more such technologies, document classi□ication is a key aspect in getting the results, Extracting domain keywords □rom documents helps to optimize the task o□ document classi□ication involved in In□ormation Retrieval. The existing state o□ the art techniques extensively depend on keyword extraction based on term document □requency. Also, the existing technologies rank words based on the title o□ the document, which, in some cases, is imprecise because sometimes the title o□ the document doesn't have words relevant to the context in mind. To overcome such problems, we propose the idea o□ ontology based keyword extraction □or increasing the accuracy o□ document classi□ication and in turn its applications. The objective o□ this paper is to extract domain speci□ic keywords □rom the given text document with the help o□ Domain Dictionary created using Ontology. This approach can be □urther extended towards revitalizing text summarization techniques.

Keywords— ontology; domain dictionary; keyword extraction; in□ormation retrieval; term document □requency

I. INTRODUCTION

Document classi□ication is required □or many applications such as Search Engine Optimization, Semantic Web, Metadata Tagging, etc. Existing applications or APIs □or doing so □or a given domain could not be □ound. Furthermore, the ambiguity in existing applications □or applications o□ documents classi□ication is high, which results in incorrect data and thus, incorrect results. Moreover, Term Document Classi□ication is used extensively in applications o□ In□ormation Retrieval. In some cases, this results in data which is not in the context. This method needs to be re□ined.

Thus, there is a need □or coming up with a new methodology □or implementation o□ the same, and we are proposing it through Ontology. Ontology will help us improve the accuracy o□ these applications to a great extent by giving

results based on the context of the query.
Ontology is a way of knowledge representation in which all the concepts related to a particular object are explored and relations are established between these concepts. Ontology should be such that it is machine readable which enables it to comprehend it and it should be expressed in some manner. Ontology is formal description of concepts, properties of these concepts which are basically the features of the concepts and restriction imposed on these properties. To develop an ontology we need to first define all the classes (concepts) in the domain under consideration, then identifying classes as superclass or subclass and the relationship between these classes. Afterwards, we need to define the slots (properties) and the restriction imposed on them that is permissible values or each slot. All this constitutes the knowledge base.
When two parties are communicating, both should have already established the context of the communication. For e.g., if party 'A' talks about 'Jaguar' with party 'B' , then B must know whether A is talking about Jaguar the animal or car or operating system. Ontology establishes this context between both the users.
In this paper we propose a system where the user provides the document(text) and the domain of concern, the system extract keywords with help of pre-created ontology and gives all keywords along with important parameters such as requency of word, how strong is the association between the word and domain, etc. which are discussed in greater detail in the paper.

Plagiarism Scan Report

Summary	
Report Genrated Date	03 Nov, 2017
Plagiarism Status	**100% Unique**
Total Words	536
Total Characters	3643
Any Ignore Url Used	

Content Checked For Plagiarism:

LITERATURE SURVEY
International scholars have studied this □ield □or many years now and explored various methods. Paper [1] proposes a keyword dictionary server that provides keyword expansion using domain speci□ic ontologies. This has been achieved using the □unctional metadata o□ services like service name, category, provider and description. Paper [2] proposes the skeleton o□ a semantic search engine that □ollows automatic query expansion. For all the terms, SPARQL query is built and then it is □ired on the knowledge base that □inds appropriate RDF triples in knowledge Base. Web documents relevant to the requested concepts and individuals speci□ied in these triples are then retrieved and ranked according to their relevance to the user's query and then are sent to the user. Paper [3] uses WordNet as a dictionary □or □inding synonyms o□ user's query. This paper explores a technique called ontological indexing which is based on calculating the context o□ the words using ontology. In paper [4], ontology is created by domain experts and is supplied to the system. Here 2 algorithms are proposed □or extraction : "semantic in□ormation extracting algorithm" and "semantic in□ormation re-recognizing algorithm". Text in□ormation is extracted using ontology and the 2 proposed algorithms.
Paper [5] talks about using Ontology Based In□ormation Extractors(OBIE) □or text grading. They argue that the combination o□ in□ormation extractors that per□orm di□ferent unctions can provide a better understanding o□ a graded text, and the combination o□ in□ormation extractors that have di□ferent implementations can improve the per□ormance o□ the extraction process. Paper [6] enables the ontology to □ind related recent knowledge in the domain □rom communities, by exploiting their underlying knowledge as keywords. It extracts instances and statements □rom the documents using the ontology-based and pattern-based in□ormation extraction technique. A con□idence value is used in order to maintain the stability o□ the ontology. Finally, the proposed system enriches

the ontology with the new extracted instances and statements and validates the knowledge inside the ontology. Paper [8] reviewed the related concepts and methods o□ ontology construction and extension, proposed an automatic ontology extension method based on supervised learning and text clustering.Paper [9] proposes an approach to extract ontology directly □rom RDB in the □orm o□ OWL/RDF triples, to ensure its availability at semantic web. Then it automatically constructs an OWL ontology □rom RDB schema using direct mapping rules.Later, rewriting SPARQL query □rom SQL by translating SQL relational algebra into an equivalent SPARQL. In paper [7], authors have developed a □ramework or comparing 11 ontology learning systems. A□ter analyzing these methodologies, 3 are selected and then applied to inance domain. Out o□ the 11 ontology learning systems, we have selected 5 systems : Ontolearn [10] uses text mining and statistical techniques to learn concepts and build taxonomic relations; Text2Onto [11] uses Probabilistic Ontology Model and involves statistical, linguistic techniques to create ontology; CRCTOL [12] algorithm is an statistical algorith[illegible] which extracts concepts and relations; OntoGain [14] is an unsupervised algorithm which uses linguistic tools to preprocess text and extract concepts; HCHIRISM [illegible] is also an unsupervised algorithm which recursively [illegible]a[illegible]ze[illegible] a large number o□ web sites in order to □ind impo[illegible]a[illegible] concepts □or a domain by introducing an initial keyword. A comparative study o□ all these methods is put in App[illegible]nd[illegible].

Plagiarism Scan Report

Summary	
Report Genrated Date	03 Nov, 2017
Plagiarism Status	**100% Unique**
Total Words	467
Total Characters	2935
Any Ignore Url Used	

Content Checked For Plagiarism:

PROPOSED MODEL
The two main components o the the proposed model are:
i) Domain Ontology Creation
ii) Parameter application on text
The Domain Ontology creation module generates an
Ontology based on the [13] HCHIRSIM model.
The second module shortlists the important keywords
rom the text using a traditional eature-based se on
approach.
Figure 1. demonstrates the system design
incorporating the two modules to accom h the goal o
Domain speci ic keywords extraction.
The two main inputs to the system a) User's
document rom which words a to xtracted and ii) The
desired domain. O these the n is given as input to the
Ontology Creation Module.
The proposed model work as ollows:
1. The Domain Ontolo used is created.
2. The user's documen s pre-processed and cleaned.
The two pre-proce s steps are Cleaning and
Lemmatization
a. Cleanin involves the removal o special
character stopwords rom the text.
b. Lemmatiz ion is used to derive the root
orm o the word rom its in lection version.
3. The words rom this preprocessed text are then
mapped to concepts and instances in the Ontology
created in step 1. Mapping involves looking up the
existence o the words in preprocessed text in the
Ontology.
4. Parallely the keywords rom the original document
are shortlisted based on the ollowing parameters:
i. Frequency o the word in the
English language.
ii. Frequency o occurrence in the
text.

iii. Position weightage.
iv. Part o Speech o the word.
v. Number o times the word is used as a either a Subject, Object or a Predicate.
vi. Distance between current and previous occurrence.

5. The Keywords obtained rom the Ontology Mapping and Parameter Application are then ranked.
6. Finally the highest ranked keywords are output as the Domain speci ic keywords.

IV. CONCLUSION

In this paper, we have explained the need or Ontology in identi ying keywords in a given document and have understood how it is use ul in various disciplines. We highlighted that the existing systems use NLP and o ten extract keywords based on title. The delta change in the proposed system is to extract keywords based on domain. Our proposed system has two main modules which are Ontology creation module and Keyword extraction module. By employing Ontology we can overcome the ambiguity that the existing system su fers rom.

Major application o proposed system is Classi ication o documents, which urther can be applied to Search engine optimization, Semantic web et

ACKNOWLEDGMENT

We are thank ul to our college Vivekanand Education Society's Institute o Technology or considering our project and extending help at all stages needed during our work o collecting in ormation regarding the project. We are deeply indebted to Head o the Computer Department Dr.(Mrs.) Nupur Giri and our Princip l Dr. (Mrs.) J.M. Nair , or giving us this valuable opportunity to do this project.