VIVEKANAND EDUCATION SOCIETY'S

INSTITUTE OF TECHNOLOGY

Department of Computer Engineering

Project Report on

Ontology based Domain Dictionary

In partial fulfillment of the Fourth Year (Semester–VIII), Bachelor of Engineering (B.E.) Degree in Computer Engineering at the University of Mumbai

Academic Year 2017-2018

Submitted by

Snehal Bhagat, D17A - 06

Padmaja Kolle, D17A - 40

Shruti Zade D17A - 77

Bhavik Dand, D17A - 20

Project Mentor

Mrs. Lifna C.S.

(2017-18)

VIVEKANAND EDUCATION SOCIETY'S INSTITUTE OF TECHNOLOGY
Department of Computer Engineering

Certificate

This is to certify that ***Snehal Bhagat, Padmaja Kolle, Shruti Zade, Bhavik Dand*** of Fourth Year Computer Engineering studying under the University of Mumbai have satisfactorily completed the project on "***Ontology based Domain Dictionary***" as a part of their coursework of PROJECT-II for Semester-VIII under the guidance of their mentor ***Asst. Prof. Lifna C.S*** in the year 2017-2018 .

This project report entitled ***Ontology based Domain Dictionary*** by ***Snehal Bhagat, Padmaja Kolle, Shruti Zade, Bhavik Dand*** is approved for the degree of Bachelor of Engineering (Computer Engineering)

Programme Outcomes	Grade
PO1, PO2, PO3, PO4, PO5, PO6, PO7, PO8, PO9, PO10, PO11, PO12 PSO1, PSO2	

Date:

Project Guide:

Project Report Approval
For
B. E (Computer Engineering)

This project report entitled ***Ontology based Domain Dictionary*** by ***Snehal Bhagat, Padmaja Kolle, Shruti Zade, Bhavik Dand*** is approved for the degree of Bachelor of Engineering (Computer Engineering)

Internal Examiner

External Examiner

Head of the Department

Principal

Date:

Place:

Declaration

We declare that this written submission represents our ideas in our own words and where others' ideas or words have been included, we have adequately cited and referenced the original sources. We also declare that we have adhered to all principles of academic honesty and integrity and have not misrepresented or fabricated or falsified any idea/data/fact/source in our submission. We understand that any violation of the above will be cause for disciplinary action by the Institute and can also evoke penal action from the sources which have thus not been properly cited or from whom proper permission has not been taken when needed.

---	---
(Signature)	(Signature)
Ms. Snehal Bhagat (06)	Ms. Padmaja Kolle (40)
---	---
(Signature)	(Signature)
Ms. Shruti Zade (77)	Mr. Bhavik Dand (20)

Date:

Acknowledgement

We are thankful to our college Vivekanand Education Society's Institute of Technology for considering our project and extending help at all stages needed during our work of collecting information regarding the project.

It gives us immense pleasure to express our deep and sincere gratitude to Assistant Professor **Mrs. Lifna C. S.** (Project Guide) for her kind help and valuable advice during the development of project synopsis and for her guidance and suggestions.

We are deeply indebted to Head of the Computer Department **Dr.(Mrs.) Nupur Giri** and our Principal **Dr. (Mrs.) J.M. Nair ,** for giving us this valuable opportunity to do this project.

We express our hearty thanks to them for their assistance without which it would have been difficult in finishing this project synopsis and project review successfully.

We convey our deep sense of gratitude to all teaching and non-teaching staff for their constant encouragement, support and selfless help throughout the project work. It is great pleasure to acknowledge the help and suggestion, which we received from the Department of Computer Engineering.

We wish to express our profound thanks to all those who helped us in gathering information about the project. Our families too have provided moral support and encouragement at several times.

Computer Engineering Department

COURSE OUTCOMES FOR B.E. PROJECT

Learners will be to:-

Course Outcome	Description of the Course Outcome
CO 1	Able to apply the relevant engineering concepts, knowledge and skills towards the project.
CO2	Able to identify, formulate and interpret the various relevant research papers and to determine the problem.
CO 3	Able to apply the engineering concepts towards designing solution for the problem.
CO 4	Able to interpret the data and datasets to be utilized.
CO 5	Able to create, select and apply appropriate technologies, techniques, resources and tools for the project.
CO 6	Able to apply ethical, professional policies and principles towards societal, environmental, safety and cultural benefit.
CO 7	Able to function effectively as an individual, and as a member of a team, allocating roles with clear lines of responsibility and accountability.
CO 8	Able to write effective reports, design documents and make effective presentations.
CO 9	Able to apply engineering and management principles to the project as a team member.
CO 10	Able to apply the project domain knowledge to sharpen one's competency.
CO 11	Able to develop professional, presentational, balanced and structured approach towards project development.
CO 12	Able to adopt skills, languages, environment and platforms for creating innovative solutions for the project.

Abstract

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

Two main modules exist in our project, namely, the Ontology Creation module and the Keyword Extraction module. The Ontology Creation module makes use of data from web pages and the WordNet API. WordNet is a platform which comprises of thousands of words, which proves to be really useful for data population in an ontology. The Keyword Extraction module initially pre-processes the text to give words in their pure form. Then, it uses the ontologies created in the previous module and maps the words in the text to those in the ontologies. It then chooses those words which lie in the ontology and assigns ranking to those words based on certain parameters.

Finally, the keywords with the highest ranking are chosen to be displayed. The results are compared to that of IBM Alchemy API and that of an expert who does the process impartially and ideally.

Index

Chapter 1 : Introduction

This chapter is an introduction to the project. Section 1.1 gives a brief introduction to the project along with the explanation of the concept of Ontology. Section 1.2 explains the motivation and need. Section 1.3 tells us the problem definition. Section 1.4 explains the relevance to the current scenario. This chapter ends with Section 1.5 giving a brief of the methodology used.

1.1 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, it was 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 project a method is proposed 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

Artificial-Intelligence literature contains many definitions of an ontology; many of these contradict one another. An **ontology** is a formal explicit description of concepts in a domain of

discourse (**classes** (sometimes called **concepts**)), properties of each concept describing various features and attributes of the concept (**slots** (sometimes called **roles** or **properties**)), and restrictions on slots (**facets** (sometimes called **role restrictions**)). An ontology together with a set of individual **instances** of classes constitutes a **knowledge base**. In practical terms, developing an ontology includes:

- defining classes in the ontology,
- arranging the classes in a taxonomic (subclass–superclass) hierarchy,
- defining slots and describing allowed values for these slots,
- filling in the values for slots for instances.

Then one can create a knowledge base by defining individual instances of these classes filling in specific slot value information and additional slot restrictions.

- An ontology is an explicit, formal specification of a shared conceptualization.
- The term ontology is borrowed from philosophy where ontology is the 'account of existence'.
- For AI systems, 'what exists' is what that can be represented.

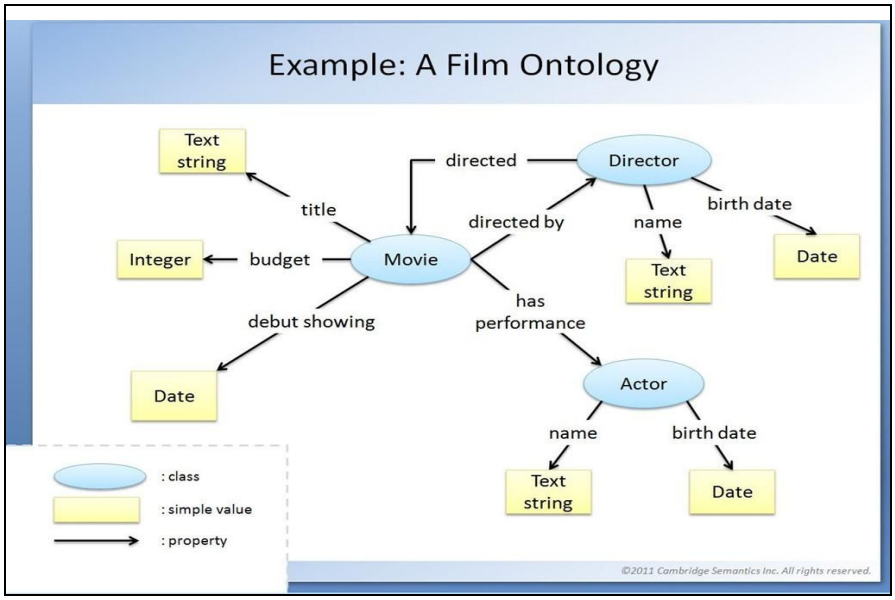


Fig. 1: Example: A Film Ontology which illustrates classes, their attributes and relations between them.

Breakdown of formal definition:

- explicit: explore all related concepts
- formal: knowledge representation should be machine readable and mathematically described.
- conceptualization: we create a model of a domain & in that domain we identify concepts and relationships between them

1.2. Motivation

Document Classification is required for many applications such as Search Engine Optimization, Semantic Web, Metadata Tagging, etc. Existing applications or APIs for doing so for a given domain could not be found. Moreover, the ambiguity in existing applications for applications of documents classification is high, which results in incorrect data and thus, incorrect results. Thus, there is a need for coming up with a new methodology for implementation of the same, and we are proposing it through Ontology. Ontology will help us improve the accuracy of these applications to a great extent by giving results based on the context of the query.

1.3. Problem Definition

Term Document Classification is used extensively in applications of Information Retrieval. In some cases, this results in ambiguous data retrieval; data which is not in the context. This method needs to be refined and thus, we are proposing Ontology as a way to do it. Extracting domain specific keywords from the given text document and domain with the help of Domain Dictionary created using Ontology. The input of the system is a document or a piece of text and the output of the system are domain specific keywords.

1.4 Relevance of the Project

Most of the Information Retrieval systems today are based on Pattern matching search algorithms and Natural language processing. But these techniques tend to lose the context of the

initial retrieval query. Using Ontologies can help improve Information Retrieval as these can aid effective query expansion such that the targeted Information is retrieved easily. Additionally, existing text classification tools are more likely to misinterpret homonyms and hyponyms in a given text. Using ontologies on the other hand, a single word can be mapped to multiple concepts across different ontologies, which can increases the accuracy of the classification. Apart from Information Retrieval, classification, keyword extraction etc., the use of Ontologies can help fuel the progress of the Semantic Web as its primary focus is the semantic interpretation of the user queries.

1.5 Methodology used

Enterprise Knowledge defines an ontology as "a defined model that organizes structured and unstructured information through entities, their properties, and the way they relate to one another." Think of an ontology as another way to classify content (like a taxonomy) that allows you to relate content based on the information in it as opposed to a term describing it.

For example, 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. Thus ontology establishes context between both the users.

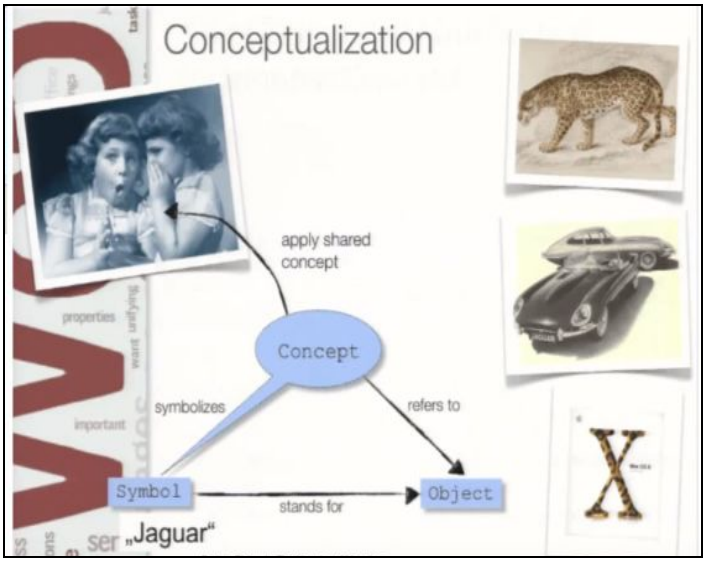


Fig. 2: Conceptualization. Resolving the ambiguity by context.

There are many reasons why this is valuable for your organization. Ontologies can allow an organization to:

- Manage content more effectively;
- Maximize findability and discoverability of information;
- Increase the reuse of “hidden” and unknown information; and
- Elevate SEO on external search engines.

Chapter 2: Literature Survey

This chapter comprises of the literature review done. Section 2.1 comprehends the papers researched. The links to the paper can be found in the References Chapter. Section 2.2 gives the summarized inference drawn from the survey. Section 2.3 is a collection of the drawbacks found in the existing systems. Ontology is an important module in our proposed system. In order to appreciate it better and to understand how it facilitates keyword extraction, a literature survey was done in order 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.

2.1 Research Papers

Paper [1] proposes a keyword dictionary server that provides keyword expansion using domain specific ontologies. The ontology modelling has been used to categorise web service keywords. This has been achieved using the functional metadata of services like service name, category, provider and description. Words have been classified on the basis of similarity calculation between two keywords.

The major modules involved are ontology management module, keyword ontology, keyword expansion module, similarity calculation module and interface module. Repositories such as Programmableweb (provides listing of web services), Wordnet and Wikipedia have been used for ontology creation. The defined keyword ontologies have 770 keywords and 23 categories. When a query is fired the keyword expansion module is fired to speed the retrieval of information. The keyword expansion module on receiving a query keyword, searches and returns the set of synonyms, hyponyms, hypernyms and components of query keywords (using OWL Reasoner).

This words are selected on the basis of similarity with the query keyword. If a word has a relation with the query keyword in the ontology, similarity calculation module calculates

similarity value based on shortest path distance between two keywords.

Datasets used are Programmableweb, Wordnet and Wikipedia pages while tools and technologies used are Java, OWL and Apache Server.

Paper [2] proposes the skeleton of a semantic search engine that follows automatic query expansion. For all the terms (expanded and initial query terms), SPARQL query is built and then it is fired on the knowledge base that finds appropriate RDF triples in knowledge Base. Web documents relevant to the requested concepts and individuals specified in these triples are then retrieved. Finally, the retrieved documents are ranked according to their relevance to the user's query and then are sent to the user. If a user wants to find specific information; can search with another module of our system that works without query expansion.

The existing search engines provide a huge amount of information for any specific query but there are several problems with current searching systems as:

- Identifying hyponyms and synonyms for query keywords
- A lot of information retrieved i.e. information excess
- Poor precision and poor recall

The authors have presented a new technique for ontology based query expansion that has been integrated with the help of specific tools, to our search scheme. The overall architecture is described below:

1. Domain Ontology Construction - Ontology has been created for a particular domain
2. User Interface - The user enters the query
3. Query Handling - The meaningful concepts are extracted from the query entered by the user through concept identifier and expanded through semantic query expansion
4. Semantic Search Engine
5. Semantic Query Expansion
6. Semantic Similarity Computation
7. Query Expansion Algorithm

Dataset is prepared from Wordnet while the tools and technology used are Jena API with

Java Eclipse, Protégé tool and SPARQL.

Paper [3] uses WordNet as a dictionary for finding synonyms of user's query. The technique is based on context of word. Using context of words helps improvise the search results. The technique is called ontological indexing. This technique is compared with text based search. The words on the web pages are mapped to concepts in ontology. Mapping score is generated for each word. Results of search depend on value of mapping score. Moreover the issue of word sense disambiguation is solved up to some extent using parts of speech tagger.

Ontological indexing uses context of the query to find the results from the database. Keyword search retrieves results based on the keyword match but does not consider the context of the word. XML data is converted into RDF using Piazza. Ambiguity is removed and Instantiated ontology is created based on WordNet ontology. Ontology is created in Protégé and is stored in OWL format which is again converted in RDF. Web pages are crawled for text based search. Ontological indexing is used to find mapping score. Page rank is then compared. WordNet is used to generate data. Web pages are stored in xml database. Xml is then converted to RDF format. Tools and Technology used are Piazza, OWL, WordNet, Protégé, Java and JENA API.

In the system in Paper [4], each input text is viewed as a non-formalized instance of a domain ontology. And the outputted semantic elements extracted from the text are formalized elements composed of classes, individuals and properties. To perform the extraction, two extraction algorithms are proposed and used in the system. They are "semantic information extraction algorithm" and "semantic information re-recognizing algorithm".

The system is composed of four parts. Because the extracting algorithms are generated based on the structure of domain ontologies, the system can extract texts of any domain as long as the OWL domain ontology is offered.

A domain ontology is built by domain experts. Once input into the system, the domain ontology will be passed to ontology parsing module which will process it and store the semantic information it contains into a database, and after that, the algorithms in semantic extraction

module can extract information from texts of this domain according to the data in the database. A text is divided into paragraphs, each paragraph is split into sentences and each sentence is segmented into a series of single words by some existing software for word segmentation. Ontology parsing module's main functions are to parse the inputted domain ontology and get all information about classes, individuals and properties from it. Based on the information of T-Class table, together with the keywords of the classes stored in the vocabulary base, the algorithm extracts the classes and their corresponding keywords in the sentence and stores them in the database. The sentence is divided into sections by explicit properties that have been extracted. In each section, the algorithm finds out the classes and individuals and lists all the occurrences of combination. It is found that the precision and recall ratios are affected by the ontology inputted into the system. The system can perform better with a better-defined ontology.

Dataset used are OWL Ontologies created from web pages while tools and technology used are JENA API and OWL.

Paper [5] has the purpose of text grading with a feedback system, with answers to why a piece of information is wrong or right. They have identified that the information extractors, which are the OBIE components that do the extraction process, can have multiple dimensions. These dimensions allow us to combine different information extractors in one hybrid OBIE system, letting the system have multiple configurations. They argue that the combination of information extractors that perform different functions can provide a better understanding of a graded text, and the combination of information extractors that have different implementations can improve the performance of the extraction process.

Given a domain, there may be a well-developed ontology. Pre-processing is done and based on regular expression, extraction rules capture information by identifying specific elements in text. On the other hand, with machine learning methods such as Support Vector Machine, Naive Bayes, or Conditional Random Fields, the information extraction task is transformed into a labelling and supervised learning task, where classification methods and probabilistic models try to identify which elements from a sentence are part of the sought information. In the case of

extracting correct statements, they use axioms from the ontology to design the information extractors.

In more detail, they have found that the combination of information extractors that have different implementations can obtain a higher precision and recall than using only one type of implementation. They also found that the extraction of incorrect statements is more complex than the extraction of correct statements, which leads to high variability in the performance of information extractors. The experiment results show that this variability can be reduced through the use of a hybrid configuration.

The first set of experiments use data collected from an undergraduate biology class (real dataset), while the second set of experiments use a synthetic data set generated from correct and incorrect statements to test the scalability of the system.

Paper [6] enables the ontology to find related recent knowledge in the domain from communities, by exploiting their underlying knowledge as keywords. They determine a confidence value during the enrichment and validation process to ensure the stability of the enriched ontology. It extracts instances and statements from the documents using the ontology-based and pattern-based information extraction technique. We extend existing lexico-syntactic pattern in order to extract the knowledge. A confidence value is used in order to maintain the stability of the ontology. Finally, the proposed system enriches the ontology with the new extracted instances and statements and validates the knowledge inside the ontology.

The model consists of three main phases. (1) Selection of Knowledge Source, this phase finds and selects recent knowledge sources that align with the original ontology domain. (2) Knowledge Extraction, this phase extracts and identifies relevant instances, then finds related statements from knowledge source documents. (3) Knowledge Enrichment and Validation, using the result from the second phase, this phase enriches the ontology by adding new knowledge. It uses predefined Named Entity Recognition (NER) to identify generic named entities such as person, location, and organization. The enrichment and validation process considers the confidence value inside the ontology.

Dataset used are Google News documents while tools & technology used are GATE1 Processing Resources and NER.

In Paper [7], authors have developed a framework for comparing ontology learning systems and place a number of the more prominent ones into it. They have selected 11 specific projects for this study, which are:

1. Asium (1998)
2. OntoLearn (2002)
3. Sanchez & Moreno (2004)
4. OntoLT (2004)
5. Text2Onto (2005)
6. CRCTOL (2005)
7. OntoGen (2006)
8. OntoGain (2010)
9. OntoPlus (2011)
10. HChirSim (2011)
11. PARNT (2013)

Later, the comparison framework is discussed which includes general characteristics, such as the purpose of ontology, its coverage (general or domain specific), and the formalism used. It also includes the design process used in creating ontology and the methods used to evaluate it. Finally, they have selected three methodologies for ontology construction and apply them in Finance domain ontology constructed from the web.

This paper presents a framework for comparing ontology learning systems and gives an overview of eleven prominent ontology-learning systems according to this framework. The comparison includes features of the input, the methods of learning and knowledge acquisition, the elements learned, the resulting ontology and also the evaluation process. Then, from this comparison, three algorithms, which illustrate the greatest differences, are selected and applied in Finance domain ontology constructed from the web.

Datasets are prepared from web pages. Technologies used are Wordnet, SymOntoX, Semcor, Google Search API, SCHUG, Gate, Berkeley parser, OpenNLP, OHSUMED, Lucene and Genia. Tools used are Asium, OntoLearn, Sanchez & Moreno, OntoLT, Text2Onto, CRCTOL, OntoGen, OntoGain, OntoPlus, HCHIRSIM and PARNT.

Paper [8] reviewed the related concepts and methods of ontology construction and extension, proposed an automatic ontology extension method based on supervised learning and text clustering. In order to achieve information sharing and interaction better, the authors proposed the need for professional ontology to guide the search.

The main content of this article is the study on automatic ontology extension method, and raise an automatic ontology extension method based on text clustering and supervised learning. This paper studies the existing methods of semi-automatic ontology extension. On this basis, the authors have proposed an automatic ontology extension method based on supervised learning and text clustering.This algorithm is designed to have a feedback mechanism, in which "positive feedback" is to get the classifications of web pages, filter some effective keywords, add them to the training set and update the stopwords dictionary with the words which are not related to the field. "Negative feedback" is a process by using Bayesian classifier to do supervised learning from candidate vocabulary set according to training set and classify them one by one. In addition, they have set a threshold. If the score of word is lower than the threshold, the word will be added to stopwords dictionary.

The methods extract web pages from the web database, take web pages as texts and do the text clustering algorithm. As datasets, the proposed method uses web pages. This method uses the K-means clustering algorithm to separate the domain knowledge, and to guide the creation of training set for Naïve Bayes classifier.The classification result that is obtained is displayed by ontology visualization tool, OntoGraf, Chinese word splitting component, IKAnalyzer.

Paper [9] proposes an approach to extract ontology directly from RDB in the form of OWL/RDF triples, to ensure its availability at semantic web. Then it automatically constructs an OWL ontology from RDB schema using direct mapping rules.Later, rewriting SPARQL query

from SQL by translating SQL relational algebra into an equivalent SPARQL. To identify and discuss the problem of direct mapping RDB to ontology including querying relational data and its ontology using semantic query (SPARQL), they have identified some of the primary obstacles in integrating RDBs with semantic web are that, how an ontology can be constructed automatically from RDBs as RDF/OWL. Being an important step towards realizing benefits of semantic web research, and how the user can be assisted to formulate queries in order to retrieve more accurate information. There is a lot of difficulties exist in generating ontology from RDB or querying, including unclear generation approaches, query formulation, manage and query data stored in RDF files, determination of how to retrieve the transformation data, analysis of RDB and ontology, and their similarities, and dealing with relationships and null value.

The goal of this paper is to propose a novel approach for automatically building ontology (RDF graph with OWL vocabulary) from RDBs (Schema and data) and manage querying semantically on generated ontology. In order to accomplish an alternative for common query (SQL) on RDB data, the combination of ontology (OWL/RDF graphs) and semantic query language (SPARQL). As datasets, the proposed method uses data from Relational databases stored in web servers. Technologies used are OWL/RDF and SPARQL.

2.2. Inference

Paper	Tools	Domain	Performance Measure (%)		
			Precision	Recall	F Measure
Integrated Approach to Web Ontology Learning and Engineering	OntoLearn	Tourism	84	52.74	64.8
A Framework for Ontology Learning and Data-driven Change Discovery	Text2Onto	Tourism	17.38	29.95	22
Mining Ontological Knowledge from Domain-Specific Text Documents	CRCTOL	Terrorism	86	57.1	68.63
A Domain Ontology Learning from Web Documents	HCHIRSIM	Medical-Cancer	**89.02**	**93.87**	**91.38**
Unsupervised ontology acquisition from plain texts: The OntoGain system	OntoGain	Computer Science	73	62	66.8

Table 1: Tool Comparison

2.3 Patent Search

After extensive patent search, it was found that for “Ontology based Domain Dictionary”, “Ontological approach to keyword extraction” and similar topics, no patents were filed. However a brief summary of projects which are somewhat similar to our system and or algorithm is presented.

In [15] the present invention relates to a method and apparatus for extracting a user keyword and constructing a synonym ontology database using a web log. A customized information providing system according to the present invention includes an SMS server, a customized information DB, a service server and similar sources.

In [16] the inventor talk about how to exploit the concept that terms in one article are related to other terms found in other.In this a first set of first articles is selected from the knowledge base for a domain corpus. A second set of second articles related to the first set of first articles is identified. The second set of second articles is selected from the knowledge base for the domain corpus. The domain corpus is made available to access.

In [17] Systems and methods are disclosed for obtaining a structured listing of attributes, such as a product description in a product record. Relevant words are identified in the document and corresponding candidate attributes are identified in a taxonomy. Attribute-value pairs are then evaluated with respect to a plurality of rules. Attribute-value pairs and outputs of the one or more rules are evaluated using a machine-learning algorithm.Final attribute-value pairs are stored and used to respond to search queries and facilitate comparison of products.

Chapter 3: Requirement Gathering

This chapter mentions the requirements needed for the system.Section 3.1 comprises of functional requirements without which the system is incomplete. Section 3.2 comprises of non-functional requirements which help the system to attain perfection. Section 3.3 and 3.4 highlight the constraints, hardware & software requirements, techniques and tools respectively. Section 3.5 talks justifies the choice of the selected tools in the previous section.

3.1 Functional Requirements

A functional requirement defines a function of a system or its component. A function is described as a set of inputs, the behavior, and outputs. Following are the functional requirements of the proposed system:

1. Application of optimal parameters for scoring words in the document.
2. Keyword retrieval from a given document based on the desired domain.
3. Ontology modelling for the relevant scope.
4. Retrieval of highest relevance keywords from the classified set.
5. Classification of words based on domain relevance.
6. Reduction in ambiguity pertaining to synonyms and homographs.
7. Analysis of results with respect to existing systems.

3.2 Non-Functional Requirements

A non-functional requirement is a requirement that specifies criteria that can be used to judge the operation of a system, rather than specific behaviors. Following are the non-functional requirements of the proposed system:

1. Performance: All data access and update operations will be executed by the system within a short period of time.

2. Reliability: The system will be able to handle large amount of text as input and ensure tolerance.
3. Accessibility: The user will have access to all the functionalities and data visualization for a given input.

3.3 Constraints

Following constraints need to be kept in mind while using the proposed system:

1. Input should be in the form of text only.
2. Images/ Mathematical equations are not supported.
3. Text should be in English Language only.
4. Currently system will be limited to specific domains only(viz. 'library').

3.4 Hardware & Software Requirements

Following are the hardware and software requirements needed for the proposed system:

1. Python Libraries: For implementing the algorithm of ontology creation and keyword extraction.
    a. nltk, re
    b. owlready2
    c. word_tokenize, sent_tokenize
    d. pos_tag
    e. ne_chunk
    f. bs4- BeautifulSoup
    g. urllib
    h. string
2. WordNet: Data dictionary for populating the ontology.
3. Protege: Creation of ontology will be done in Protege.
4. HTML, CSS, Javascript, Django framework, chart.js library
5. RDF/XML: Data will be converted into RDF format which is a form of XML.

3.5 Selection of the Hardware, Software, Technology and Tools

The two main programming languages under consideration for the project were Java and Python. Of these, Java had support for the Jena API that allowed us to create and manipulate Ontologies in RDF format. On the other hand Python gave us access to multiple Natural Language Processing Libraries that would aid in keyword extraction and document processing. Python also has support for the library "owlready2" which would allow us to deal with Ontologies in the '.owl' format. Also, the Django web framework is written in Python which one could leverage to design a functional UI that would work in cohesion with the Python code.

In summary, the use cases of Python outnumbered that of Java and hence Python was chosen as the most suitable programming language for the Project.

For text-processing various libraries available for text-processing were tested namely rake, spacey, nltk and the google nlp library among others. Of these the 'nltk' library best fit the requirements as it allowed near accurate stemming, lemmatization, chunking and POS-tagging of words.

Chapter 4: Proposed Design

Chapter 4 comprises of all diagrams and system designs which provide an overall view as well as a detailed view of the functioning of the proposed system. Section 4.1 shows the block diagram. Section 4.2 comprises of modular diagram of our system. Section 4.3 gives a detailed design of the system with the data flow diagrams, flowchart. Section 4.4 comprehends the timeline of the project.

4.1 Block Diagram of the system

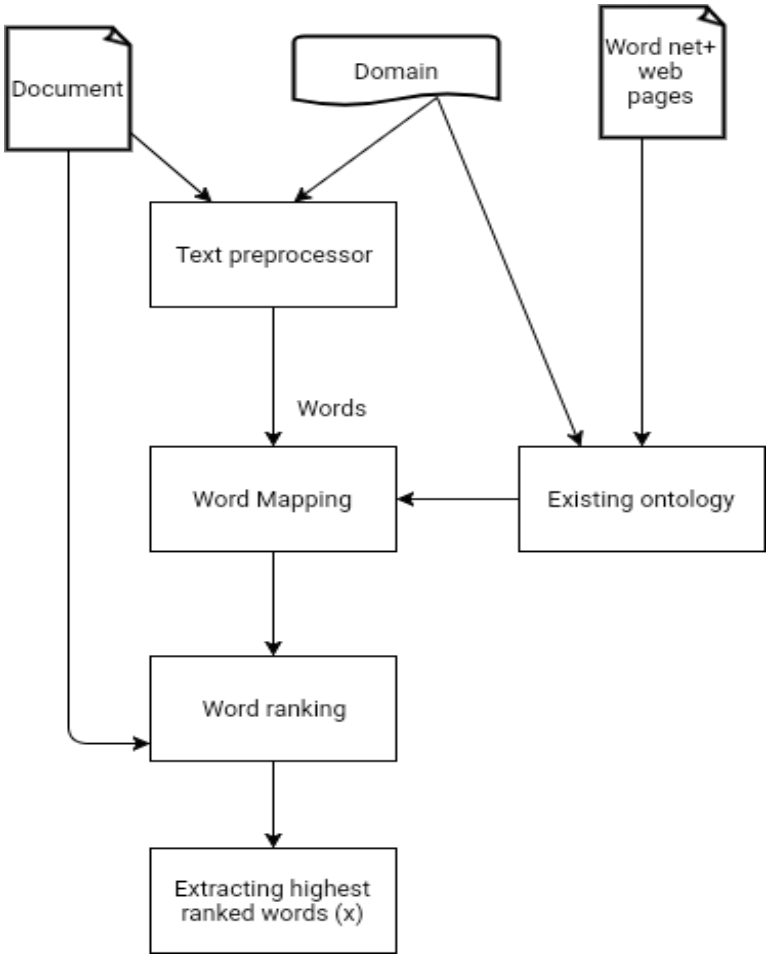


Fig 3 : Block Diagram

The block diagram shows the basic approach taken. As represented, the system can be viewed to be split up in 2 parallel phases; one for keyword extraction from user document based on domain and the other for domain related keyword extraction from ontology. After this, the phases merge and it maps out common words obtained from both phases. Later they are displayed as output with scores based on their relevance to the domain.

4.2. Modular diagram representation of the system

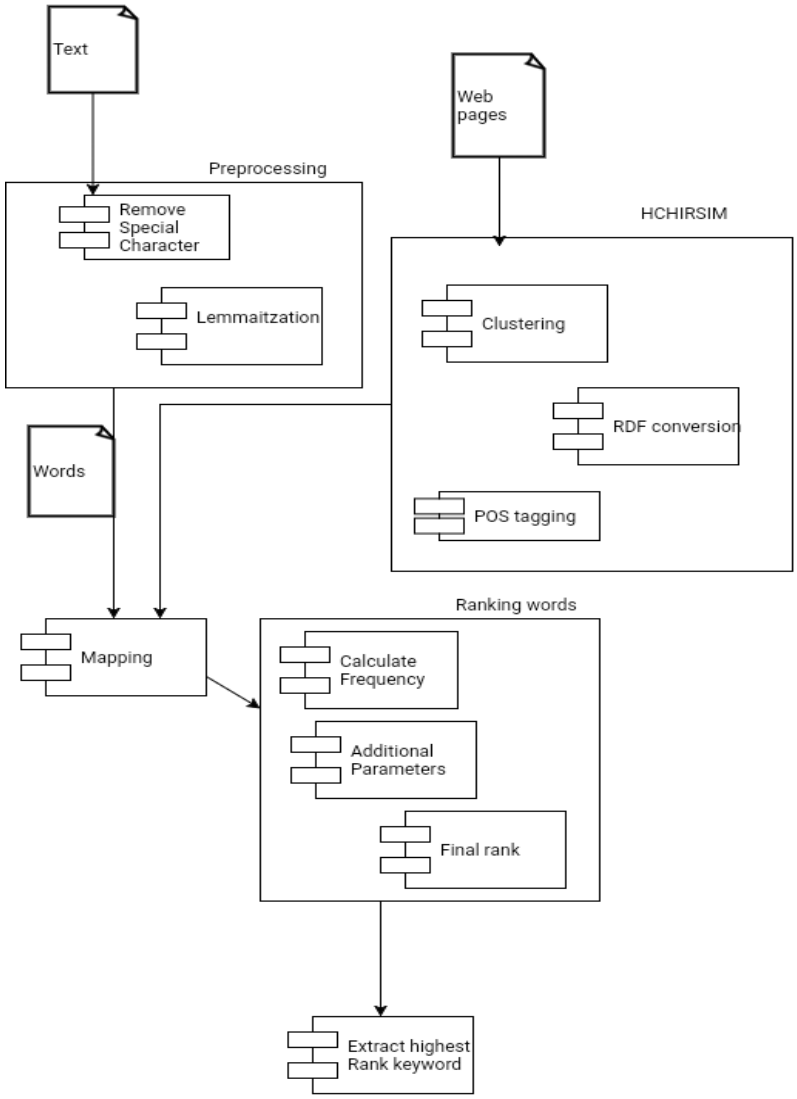


Fig 4 : Modular Diagram

The above block diagram comprises of several modules which result in accomplishing the output of the system. Initially, text document entered, is given to the Text Pre-Processing Module, where stemming or lemmatization is done along with removal of stopwords. Then, in the Class Selection Module, the algorithm classifies words into candidate classes and words which are not relevant to the domain are removed. The OWL ontology is created with the defined classes with the help of WordNet which is a data dictionary. The OWL ontology is then converted into a three tuple RDF format. Words are ranked according to predefined parameters and words with high ranking are extracted using extraction algorithm.

4.3 Detailed design of the system

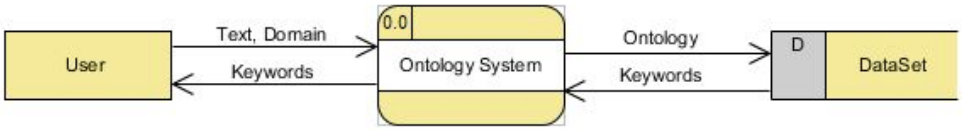


Fig 5 : Level 0 DFD

The input to the system is a text document which can either be uploaded, posted or could be a URL of a document. The system processes the document and gives domain specific keywords as output.

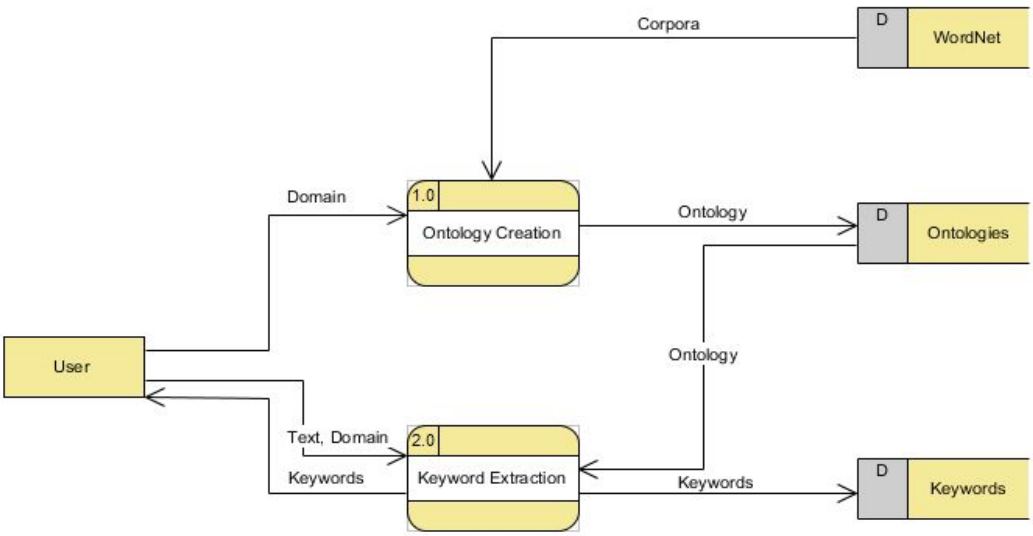


Figure 6 : Level 1 DFD

In level 1 DFD it can see that there are 2 important modules Ontology creation and keyword extraction. The Domain is relevant to both these modules.The Domain Ontology to be used is created using the algorithm described in chapter 5. The user's document is pre-processed and cleaned. The extracted keywords are stored later for mapping. The ontology created also has to be stored.

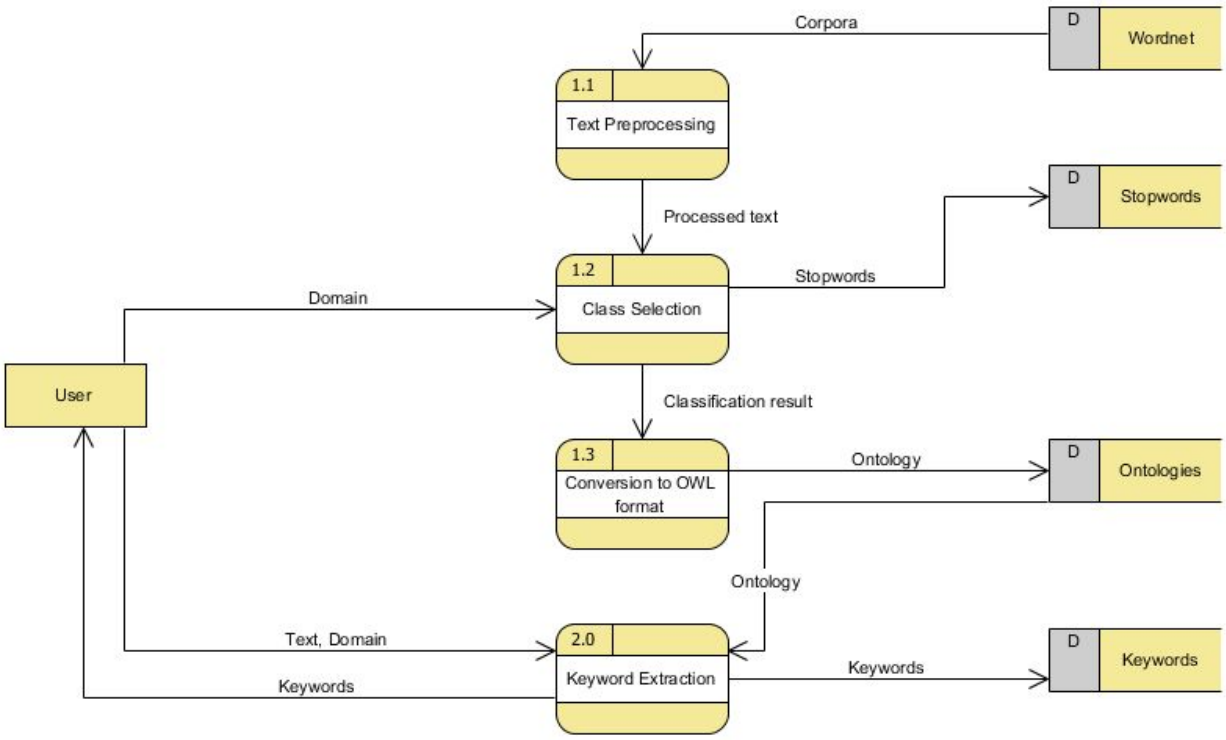


Figure 7 : Level 2 DFD

Level 2 DFD shows all sub modules of ontology creation and keyword extraction.They work as follow:

1. The Domain Ontology is created and stored beforehand.The general algorithm is to identify the important class and convert them to OWL format.
2. The user's document is pre-processed and cleaned. The two pre-processing steps are Cleaning and Lemmatization.
    a. Cleaning involves the removal of special characters and stopwords from the text.
    b. Lemmatization is used to derive the root form of the word from its inflection version.

3. The words from this preprocessed text are then mapped to concepts and instances in the Ontology created in step 1. Mapping involves looking up the existence of the words in preprocessed text in the Ontology.
4. Parallely the keywords from the original document are shortlisted.
5. The Keywords obtained from the Ontology Mapping and Parameter Application are then ranked.
6. Finally the highest ranked keywords are output as the Domain specific keywords.

4.4. Project Scheduling & Tracking using Timeline

The following screenshots displays timeline of the project. Smartsheet software was used to create this timeline.

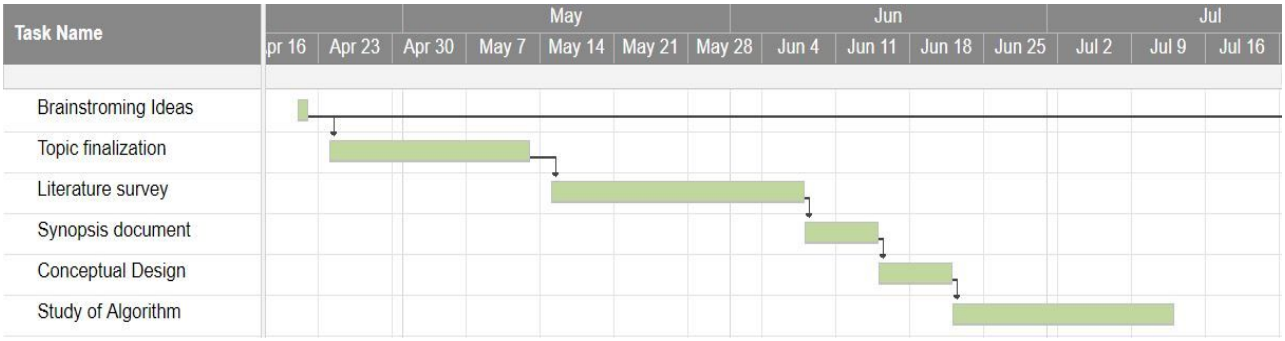


Fig. 8 : Project Scheduling and Timeline (A)

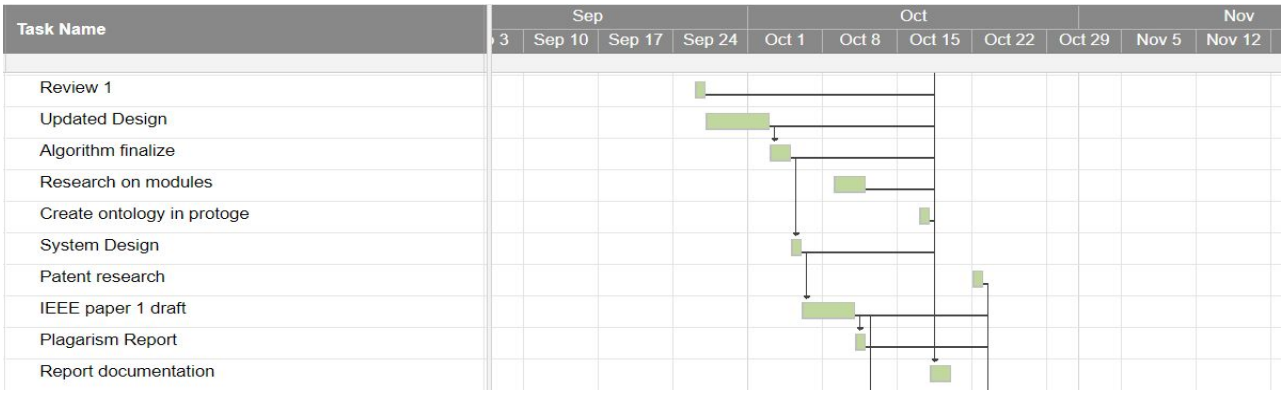


Fig. 9 : Project Scheduling and Timeline (B)

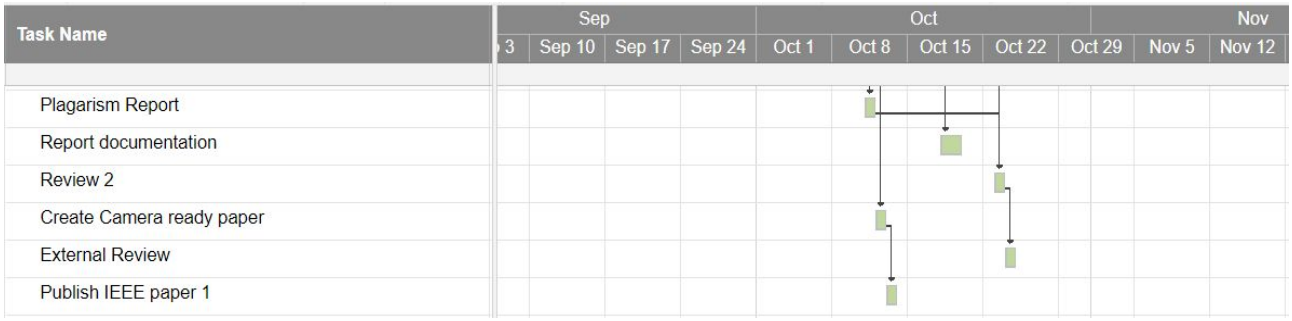


Fig. 10 : Project Scheduling and Timeline (C)

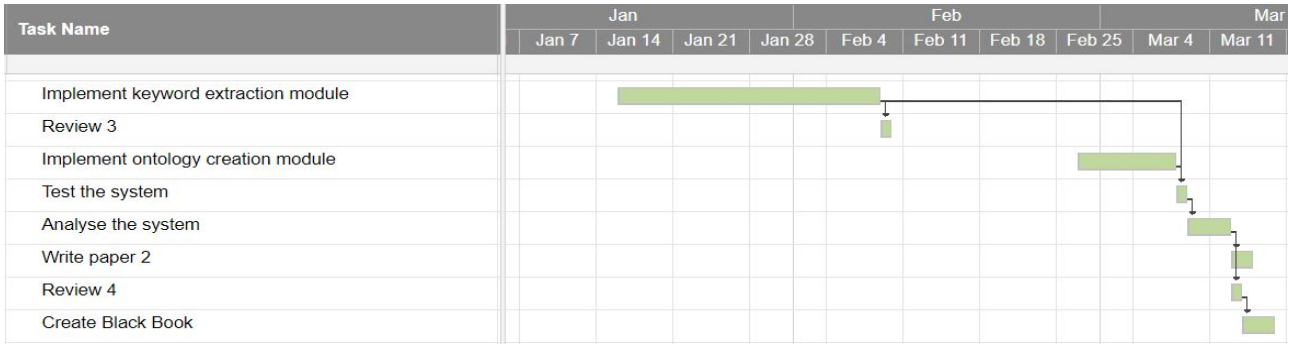


Fig. 11 : Project Scheduling and Timeline (D)

Task Name	Duration	Start	Finish	Predecessors
Brainstroming Ideas	1d	04/21/17	04/21/17	
Topic finalization	15d	04/24/17	05/12/17	1
Literature survey	18d	05/15/17	06/07/17	2
Synopsis document	5d	06/08/17	06/14/17	3
Conceptual Design	5d	06/15/17	06/21/17	4
Study of Algorithm	15d	06/22/17	07/12/17	5
Review 1	1d	09/26/17	09/26/17	
Updated Design	4d	09/27/17	10/02/17	
Algorithm finalize	2d	10/03/17	10/04/17	8
Research on modules	3d	10/09/17	10/11/17	
Create ontology in protoge	1d	10/17/17	10/17/17	
System Design	1d	10/05/17	10/05/17	9
Patent research	1d	10/22/17	10/22/17	
IEEE paper 1 draft	3d	10/06/17	10/10/17	12
Plagarism Report	1d	10/11/17	10/11/17	14
Report documentation	2d	10/18/17	10/19/17	1, 7, 8, 9, 10, 11, 12
Review 2	1d	10/23/17	10/23/17	13, 14, 15
Create Camera ready paper	1d	10/12/17	10/12/17	14, 15
External Review	1d	10/24/17	10/24/17	17
Publish IEEE paper 1	1d	10/13/17	10/13/17	18
Implement keyword extraction module	18d	01/16/18	02/08/18	
Review 3	1d	02/09/18	02/09/18	21
Implement ontology creation module	7d	02/27/18	03/07/18	
Test the system	1d	03/08/18	03/08/18	21, 23
Analyse the system	2d	03/09/18	03/12/18	24
Write paper 2	2d	03/13/18	03/14/18	25
Review 4	1d	03/13/18	03/13/18	25
Create Black Book	3d	03/14/18	03/16/18	27
External review 2	1d	04/24/18	04/24/18	

Figure 12 : Project Scheduling with dates

Chapter 5: Implementation Details

Chapter 5 talks about the algorithm behind the system in greater detail. All the steps for implementation are also explained in this chapter. Section 5.1 explains the algorithm, section 5.2 explores existing system and identifies the drawbacks, lastly section 5.3 discusses evaluation measures of the developed system.

5.1 Algorithms implemented

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 this 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.

5.1.1 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 one considers the next approach.

The next approach to creating 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 and Saving the ontology.

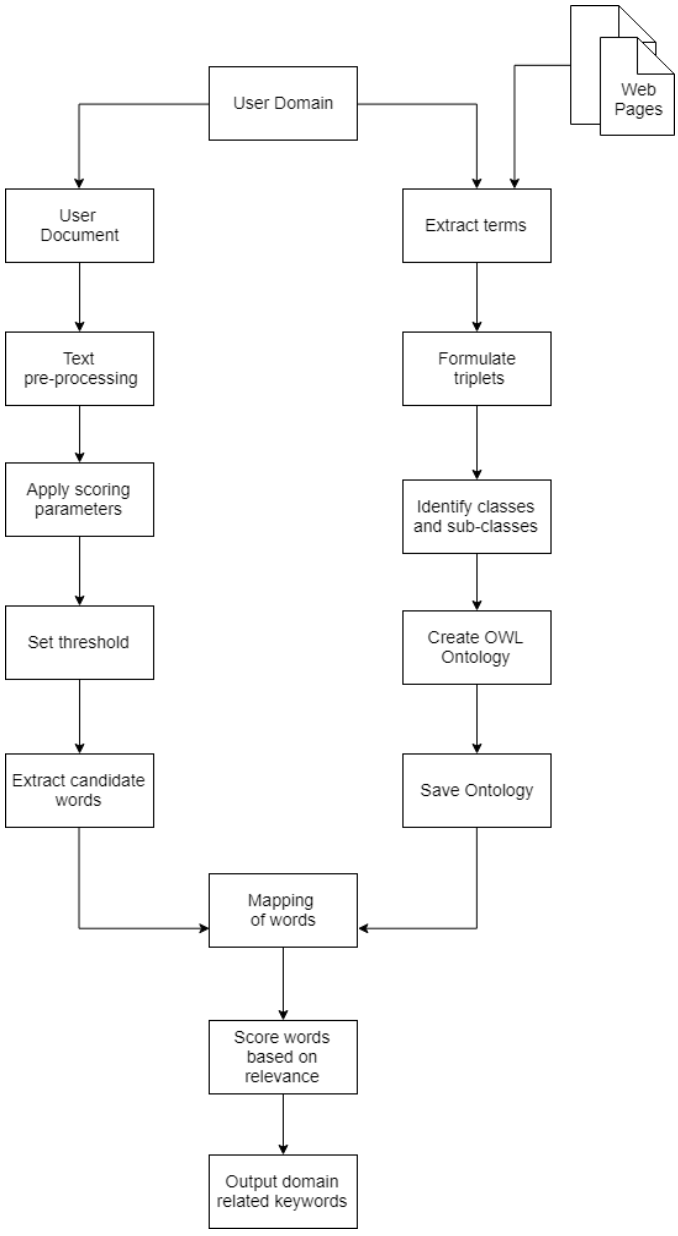


Fig. 13 System flow diagram

5.1.2 Keyword Extraction

The keyword extraction module os the main module of the system. Here, 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, 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, entropy has been used. The formula for calculating entropy is as follows:

$$W1 = F/N\log2(F/N)$$

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: 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:

$$W2 = (Stotal + 1/Sf + 1)$$

where W2 = weight of given word, due to index position of the sentence in which the given word occurs first.

Stotal = 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.

$$P(K) = I(K) \quad \text{if } (I(K)<(L(S)/2))$$
$$= 2\,(L(S) - I(K)) \quad \text{else}$$

where P(K) = Partial Position related strength of given distinct word

Combine strength due to length of sentence with the formula:

$$W3 = \log2(\square\ L(S) + 1/P(K) + 1)$$

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, 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.

5.2 Comparative Analysis with the existing algorithms

This section explores system that has some resemblance to the system- Ontology based Domain Dictionary to perform a comparative study.

Sketch Engine: Sketch Engine is the ultimate tool to explore how language works. Its algorithms analyze authentic texts of billions of words (text corpora) to identify instantly what is typical in language and what is rare, unusual or emerging usage.

IBM Alchemy: Used to analyze text to extract metadata from content such as concepts, entities, keywords, categories, relations and semantic roles.

Aylien: Aylien is an artificial intelligence startup that focuses on creating technologies that help machines understand humans better. The firm provides text analysis and news API's that allow users to make sense of human-generated content at scale. They also provide a range of content analysis solutions to developers, data scientists, marketers and academics. Their core offerings include packages of information retrieval, machine learning, natural language processing and image recognition API's.

Drawbacks:

- Existing systems use pure NLP for keyword extraction, which is more time consuming and tedious methodology.
- Keywords are extracted from titles, which result in ambiguity in content where literal words in the title does not match the context of the document.
- Cannot differentiate between homonyms

5.3 Evaluation of the developed system

The above described approach is analysed on basis of accuracy,precision, recall and f measure. Before discussing them in greater details, its important to realise that the accuracy of the approach depends heavily on the ontology. 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. Ontology for the same domain was created. The user document sample was taken from wikipedia’s article on library of roughly 700 words.

Accuracy is the ratio of correctly classified observation and total no of observation. 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.

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

Chapter 6: Testing

This chapter compares and contrasts the extraction results obtained by applying three different extraction methods on the same test-document. The Document to be tested contains standard text scraped from Wikipedia. The test compares the 'Domain relevance' of the extracted keywords to the domain of choice 'library'. The three methods used are the implemented system, the IBM Alchemy API and manual selection of keywords.

6.1 Implemented System

- **Shortlisted Keywords**

  Student, public, similar, form, community, institution, sources, academic, include, both, building, reference, book, made, space, borrowing, physical, room, source, database, provides, material, accessible, may, digital, format, information, access, provide, collection, resource, library

- **Shortlisted domain specific keywords**

  Library, resource, collection, information, access, material, space, format, room, source, database, book, form.

- **Domain Relevance**

  Domain Relevance to Library Domain = 67.37%

6.2 Alchemy API

- **Shortlisted Keywords**

  academic libraries, academic institutions, library search databases, Specific course-related resources, group study, institutional electronic resources, electronic journals databases, post-secondary educational institutions, electronic citation software, effective search techniques, group study space, quiet study space, institutional scholarly research, great libraries, Modern libraries, mean library, academic library, digital access,

similar resources, students, virtual space, academic knowledge, physical building, Blu-ray Disc, French bibliothèque, modern languages, public body, unrestricted access, academic careers, information needs, cuneiform script, public facilities, public institutions, extensive collection, clay tablets, classical Greece, Classical period, earliest form, various resource, century BC, electronic means, general public, private individual, Internet access, professional assistance, common areas, institutional collections, Mediterranean world, digital tools, quiet areas.

- **Shortlisted domain specific keywords**

  academic libraries, academic institutions, library search databases, Specific course-related resources, institutional electronic resources, electronic journals databases, post-secondary educational institutions, group study space, quiet study space, institutional scholarly research, digital access, similar resources, students, academic knowledge, physical building, public body, information needs, public facilities, public institutions, extensive collection, various resource, Internet access, common areas, digital tools, quiet areas.

- **Domain Relevance**

  Domain Relevance to Library Domain = 49.01%

6.3 Manual Selection

- **Shortlisted Keywords**

  Library, collection of sources, information, resources, reference, borrowing, physical or digital access, material, physical building, room, books, periodicals, newspapers, manuscripts, films, maps, prints, documents, microform, CDs, cassettes, videotapes, DVDs, Blu-ray Discs, e-books, audiobooks, databases, shelves of books, Latin, Greek, bookcase, Bibliotheca, Bibliotheke, modern languages, French bibliothèque, archives of the earliest form, clay tablets, cuneiform script, Sumer, written books, Classical period, Mediterranean world, Constantinople, Alexandria, organized, public body, institution, corporation, afford, purchase, professional assistance, research, services of librarians, experts at finding and organizing information, interpreting

information needs, quiet areas for studying, common areas, facilitate group study, collaboration, electronic resources, Internet, unrestricted access to information, formats, digital tools, academic libraries, college, university campuses, students,faculty, academic institutions, accessible to members of the general public, post-secondary educational institutions, main function, support in research, resource linkage, Specific course-related resources, textbooks, article readings, reserve, ability to check out laptop computers, web cameras, scientific calculators, workshops, courses, graded coursework, citations, effective search techniques, journal databases, electronic citation software, skills, achieve success, careers, meeting rooms, digitally oriented, print/physical, digital, scholarly writing software, computer workstations, computer labs, journals, library search databases, portals, institutional electronic resources, electronic repository, curation of digital copies, theses, dissertations.

- **Shortlisted domain specific keywords**

  Library, collection of sources, information, resources, reference, borrowing, physical or digital access, material, physical building, books, periodicals, newspapers, manuscripts, films, maps, documents, e-books, audiobooks, databases, shelves of books, bookcase, Bibliotheca, Bibliotheke, archives of the earliest form, written books, public body, institution, research, services of librarians, experts at finding and organizing information, interpreting information needs, quiet areas for studying, common areas, group study, electronic resources, Internet, unrestricted access to information, digital tools, college, university campuses, students, faculty, academic institutions, general public, support in research, resource linkage, specific course-related resources, textbooks, article readings, reserve, ability to check out laptop computers, web cameras, scientific calculators, workshops, citations, journal databases, electronic citation software, print/physical, digital, scholarly writing software, computer workstations, library search databases, institutional electronic resources, theses, dissertations.

- **Domain Relevance**

  Domain Relevance to Library Domain = 60.7%

Chapter 7: Result Analysis

This chapter highlights all the parameters considered in the system along with parameters proposed to evaluate the accuracy and success of our system. Later sections deal with the UI of system and graphical and textual output.

7.1. Parameters Considered

7.1.1. Keyword Extraction

The following listed parameters are used in the keyword extraction module. Their understanding will help to interpret the result more accurately. They are:

1. Entropy:weight directly does not depend on frequency,but on the probability of word. w1=F/N(log(F/N))
   where W1 = Entropy of word in given document.
   F = Occurrence frequency of word in document
   N = Total Number of words in document.
2. Position of sentence in document in which the given distinct word exists: Giving more weight to words that come in the beginning of a sentence as they are more important. w2=(S_total+1)/(sj+1),
   where Stotal = Total number of sentences in given document.
   Sf = Sentence Index in which the given word occurs first.
   W2 = weight of given word, due to index position of the sentence in which the given word occurs first.
3. Position of distinct word in sentence and length of sentence and Strength due to combined effect of length of sentence and position related strength.
   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

Combine strength due to length of sentence with the formula:

W3 = log2(□ L(S) + 1/P(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.

7.1.2. Result Evaluation

The measures for evaluation are explained in brief below:

1. Accuracy : Accuracy is perhaps the most intuitive performance measure. It is simply the ratio of correctly predicted observations.
   Accuracy = (TP + TN) / (TP + TN + FN + FP)
2. Precision : Precision looks at the ratio of correct positive observations.
   Precision = TP / (TP + FP)
3. Recall: It's the ratio of correctly predicted positive events.
   Recall = TP / (TP + FN)
4. F1 Score : The F1 Score is the weighted average of Precision and Recall.
   F1 Score = 2*(Recall * Precision)/(Recall + Precision)

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.

7.2. Screenshots of User Interface (UI)

The User Interface of the system is displayed in form of snapshots below.

Fig. 14 : Welcome screen

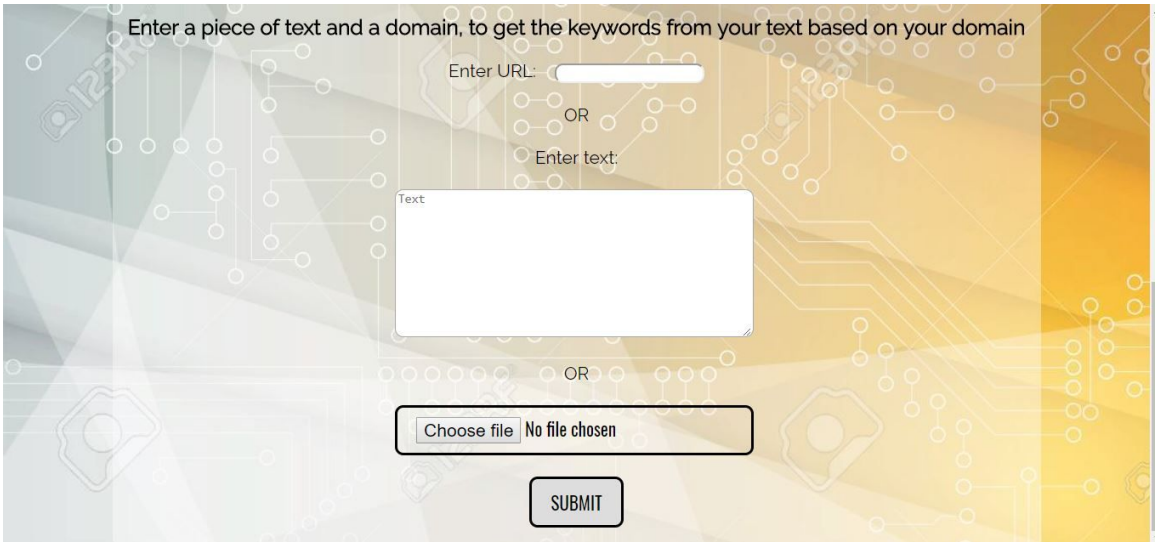


Fig. 15 : On this screen user can submit data in 3 ways: via URL, via Uploading a file or via entering text in textbox.

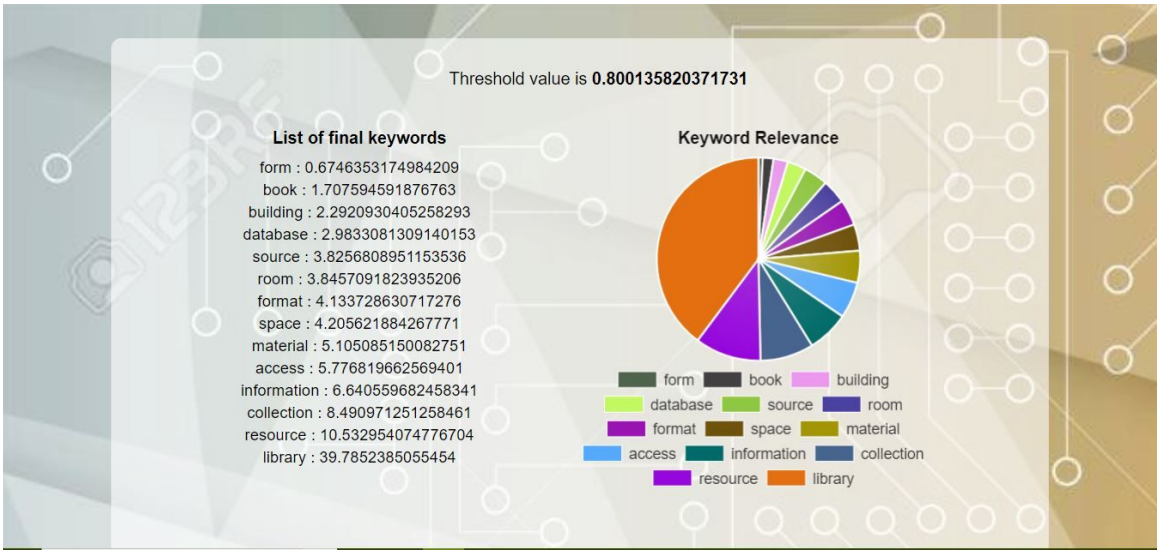


Fig. 16 : Displays the extracted keywords, threshold value and displays a pie chart.

Fig. 17 : Displays domain relevance of the text submitted and list displaying Named Entity Relationship.

7.3. Graphical Outputs

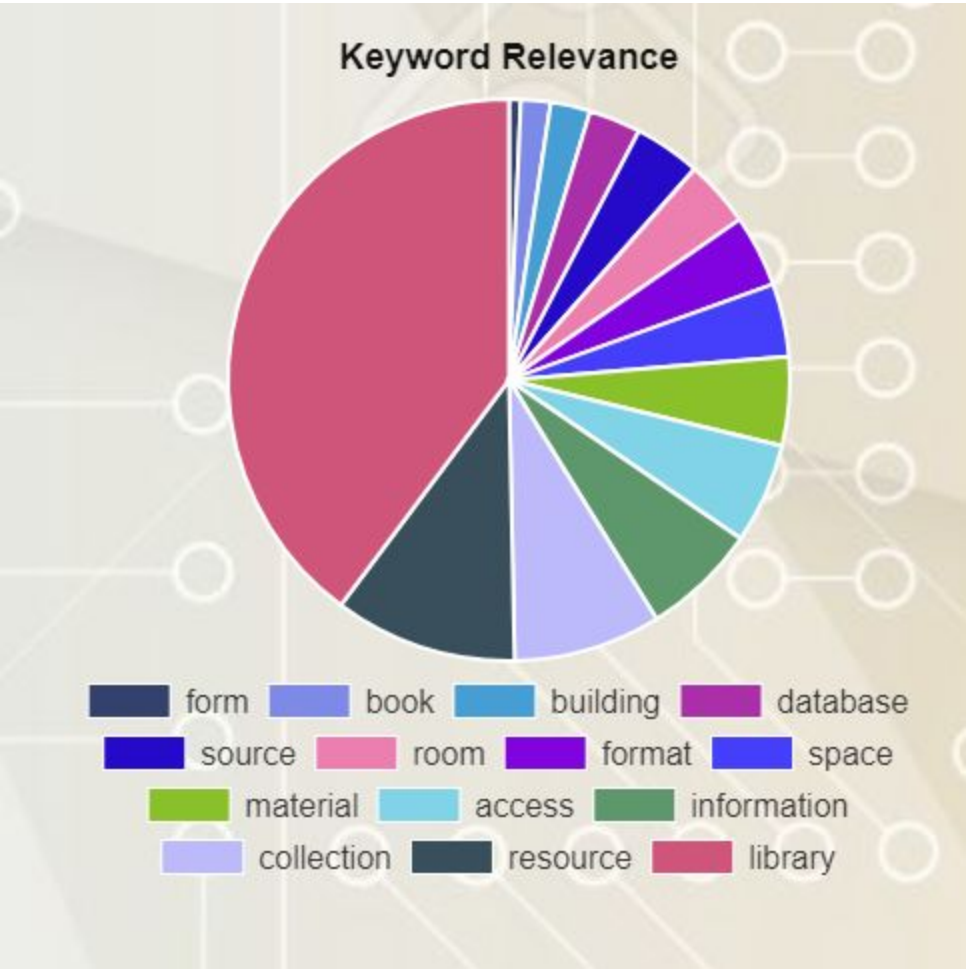


Fig. 18 : Pie chart denoting domain relevance of the extracted keywords

Chapter 8: Conclusions and Future Scope

The last chapter briefly recaps the project methodology and highlights the application. Later, the future scope of the project is also discussed.

The existing systems use NLP and often extract keywords based on title. The delta change in the proposed system is to extract keywords based on domain. Proposed system has two main modules which are Ontology creation module and Keyword extraction module. By employing Ontology the system overcame the ambiguity that the existing system suffers from. Major application of proposed system is Classification of documents, which further can be applied to Search engine optimization, Semantic web etc.

The system is semi-automated because of the need of an expert for Ontology creation. It can be made fully automated by replacing the semi-automated ontology creation module with an automatically created ontology by eliminating the need for an expert. Moreover, the system can be extended for multi-domain keyword extraction from a particular text by incorporating ontologies from several domains.

References

1. HunKyung Yoo, YooMi Park, TaeDong Lee "Ontology based Keyword Dictionary Server for Semantic Service Discovery" (2013)
2. Rashmi Chauhan, Rayan Goudar, Robin Sharma, Atul Chauhan "Domain Ontology based Semantic Search for Efficient Information Retrieval through Automatic Query Expansion" (2013)
3. Komal Mule and Arti Waghmare. “Context Based Information Retrieval Based On Ontological Concepts” (2015)
4. Hongsheng Wang, Lu Yuan, Hong Shao. "Text Information Extraction Based on OWL Ontologies" (2008)
5. Fernando Gutierrez, Dejing Dou, Adam Martini, Stephen Fickas and Hui Zong. "Hybrid Ontology-based Information Extraction for Automated Text Grading" (2013)
6. Dhomas Hatta Fudholi, Wenny Rahayu and Eric Pardede "Ontology-based Information Extraction for Knowledge Enrichment and Validation" (2016)
7. Omar Ismaïl, Bouchra Frikh,Brahim Ouhbi - “Building Ontologies: a State of the Art, and an Application to Finance Domain” (2014)
8. Qiuxia Song, Jin Liu, Xiaofeng Wang, Jin Wang "A Novel Automatic Ontology Construction Method Based on Web Data" (2014)
9. Mohamed A.G. Hazber, Ruixuan , Xiwu , Guandong, Yuhua Li-“Semantic SPARQL query in a relational database based on ontology construction” (2015)
10. Philipp Cimiano, Johanna Volker "A Framework for Ontology Learning and Data-driven Change Discovery" (2005)
11. Michele Missikoff, Roberto Navigli, Paola Velardi "Integrated Approach to Web Ontology Learning and Engineering" (2002)
12. Xing Jiang, Ah-Hwee Tan "Mining Ontological Knowledge from Domain-Specific Text Documents" (2005)

13. Euthymios Drymonas, Kalliopi Zervanou, Euripides G.M. Petrakis "Unsupervised Ontology Acquisition from Plain Texts: The OntoGain System" (2010)
14. B. Frikh, A. S. Djaanfar and B. Ouhbi "A hybrid Method for Domain Ontology Construction from the Web" (2010)
15. Jeon Shin “Method and apparatus for keyword extraction and synonym ontology database based on web log analyzing”,European Patent KR20130131770, December 04, 2013
16. Labrou, Stergios “Generating a domain corpus and a dictionary for an automated ontology”,United States Patent 8,560,485,October 15, 2013
17. Garera; Nikesh Lucky , Rampalli, Ravikant, Subramaniam Sun, Yalin “Ontology-based attribute extraction from product descriptions”,United States Patent 9,208,442,December 8, 2015

Project Evaluation Sheet 2017 - 18

Class: D17A/B/C
Group No.: G41

Title of Project: Ontology based Domain Dictionary

Group Members: Snehal Bhagat, Padmaja Kolle, Bhavik Daud, Shruti Zade

	Engineering Concepts & Knowledge (5)	Interpretation of Problem & Analysis (5)	Design / Prototype (5)	Interpretation of Data & Dataset (3)	Modern Tool Usage (5)	Societal Benefit, Safety Consideration (2)	Environment Friendly (2)	Ethics (2)	Team work (2)	Presentation Skills (3)	Applied Engg &Mgmt principles (3)	Life - long learning (3)	Professional Skills (5)	Innovative Approach (5)	Total Marks (50)
Review of Project Stage 1	3	3	3	2	4	1	1	1	1	2	3	2	5	5	35
Comments:	Deep study needed in concepts of Semantic Techniques														

Name & Signature Reviewer1

	Engineering Concepts & Knowledge (5)	Interpretation of Problem & Analysis (5)	Design / Prototype (5)	Interpretation of Data & Dataset (3)	Modern Tool Usage (5)	Societal Benefit, Safety Consideration (2)	Environment Friendly (2)	Ethics (2)	Team work (2)	Presentation Skills (3)	Applied Engg &Mgmt principles (3)	Life - long learning (3)	Professional Skills (5)	Innovative Approach (5)	Total Marks (50)
Review of Project Stage 1	3	3	2	3	4	1	2	2	2	2	3	2	4	3	36
Comments:															

Date: 26th September 2017

24/10/17.

Name & Signature Reviewer2
Sheetal Balasaat

Inhouse/ Industry: Inhouse

Class: D17 A/B/C

Group No.: 41

Project Evaluation Sheet 2017 - 18

Title of Project: Ontology Based Domain Dictionary

Group Members: Padmaja Kolle, Snehal Bhagat, Shruti Zade, Bhavik Dand

	Engineering Concepts & Knowledge (5)	Interpretation of Problem & Analysis (5)	Design / Prototype (5)	Interpretation of Data & Dataset (3)	Modern Tool Usage (5)	Societal Benefit, Safety Consideration (2)	Environment Friendly (2)	Ethics (2)	Team work (2)	Presentation Skills (3)	Applied Engg &Mgmt principles (3)	Life - long learning (3)	Professional Skills (5)	Innovative Approach (5)	Total Marks (50)
Review of Project Stage 1	04	04	04	03	04	02	02	02	02	03	03	02	04	04	43
Comments:	Evaluation Parameters are close to threshold, need to improve it														

Pallavi Saindane

Name & Signature Reviewer1

	Engineering Concepts & Knowledge (5)	Interpretation of Problem & Analysis (5)	Design / Prototype (5)	Interpretation of Data & Dataset (3)	Modern Tool Usage (5)	Societal Benefit, Safety Consideration (2)	Environment Friendly (2)	Ethics (2)	Team work (2)	Presentation Skills (3)	Applied Engg &Mgmt principles (3)	Life - long learning (3)	Professional Skills (5)	Innovative Approach (5)	Total Marks (50)
Review of Project Stage 1	3	4	3	3	5	2	2	2	2	3	2	2	3	4	40
Comments:	Usage of existing resources for compatibility. Hence scope affected														

Date: 15th March,2018

Name & Signature Reviewer2

Sheetal B.

Appendix

The following chapter gives additional information on the topic explored in the contents of this report. The chapter presents list of figures, tables and information related to paper publication such as IEEE paper, plagiarism report.

A. List of Figures

B. List of Tables

C. Papers Published

Paper 1: Ontology Based Domain Dictionary

Ontology based Domain Dictionary

Padmaja Kolle	Snehal Bhagat	Shruti Zade	Bhavik Dand	Lifna C. S.
B.E.	B.E.	B.E.	B.E.	Asst. Prof.
Department of Computer Engineering V.E.S.I.T, Mumbai India padmaja.kolle@ves.ac.in	Department of Computer Engineering V.E.S.I.T, Mumbai India snehal.bhagat@ves.ac.in	Department of Computer Engineering V.E.S.I.T, Mumbai India shruti.zade@ves.ac.in	Department of Computer Engineering V.E.S.I.T, Mumbai India bhavik.dand@ves.ac.in	Department of Computer Engineering V.E.S.I.T, Mumbai India lifna.cs@ves.ac.in

Abstract* — *Document classification is a key component in the realisation of many applications, including Text Summarization, Semantic Web, Search Engine Optimization, Sentiment Analysis among many others._Extracting domain keywords from documents helps to optimize the task of document classification involved in Information Retrieval. The existing state of the art techniques extensively depend on keyword extraction based on term document frequency. Also, these techniques rank words based on the title of the document, which, in some cases, is imprecise as the title of the document may not have words relevant to the context of the document. To overcome such problems, we propose a model for ontology based keyword extraction to increase the accuracy of document classification and in turn its applications. The objective of our paper is to extract domain specific keywords from the given text document with the help of a Domain Dictionary created using Ontology. This approach can further be extended towards revitalizing text summarization techniques.

Keywords* — *ontology; domain dictionary; keyword extraction; information retrieval; TDF

I. INTRODUCTION

Document Classification is required for many applications such as Search Engine Optimization, Semantic Web, Metadata Tagging, etc. There doesn't exist a generic application/API to classify documents according to domains. Furthermore, the ambiguity in existing applications for documents classification is high, which eventually results in incorrect data and findings. Moreover, Term Document Classification is used extensively in applications of Information Retrieval. In some cases, this results in data which is out of context. This method needs to be refined.

Thus, there is a need for coming up with a new methodology for implementation of the same, and we are proposing it through Ontology. Ontology will help us improve the accuracy of 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 machines to comprehend it and express it in some manner. Ontology is a formal description of concepts, properties of these concepts which are basically the features of the concepts and restrictions imposed on these properties. To develop an ontology we need to first define all the classes (concepts) in the domain under consideration, then differentiate classes as superclass or subclass and identify the relationship between these classes. Afterwards, we need to define the slots (properties) and the restriction imposed on them that is permissible values for 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 the car or the

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 then extracts keywords aided by the pre-created ontology and gives all keywords belonging to the domain based on important parameters such as frequency of the word, strength of association between the word and the domain, etc. which are discussed in greater detail in the paper.

II. Literature Survey

International scholars have studied this field for many years now and explored various methods. Paper [1] proposes a keyword dictionary server that provides keyword expansion using domain specific ontologies. This has been achieved using the functional metadata of services like service name, category, provider and description. Paper [2] proposes the skeleton of a semantic search engine that follows automatic query expansion. For all the terms, SPARQL query is built and then it is fired on the knowledge base that finds appropriate RDF triples in knowledge Base. Web documents relevant to the requested concepts and individuals specified 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 for finding synonyms of user's query. This paper explores a technique called ontological indexing which is based on calculating the context of the words using ontology. In paper [4], ontology is created by domain experts and is supplied to the system. Here two algorithms have been proposed for extraction : "semantic information extracting algorithm" and "semantic information re-recognizing algorithm". Text information is extracted using ontology and the two proposed algorithms.

Paper [5] talks about using Ontology Based Information Extractors(OBIE) for text grading. They argue that the combination of information extractors that perform different functions can provide a better understanding of a graded text, and the combination of information extractors that have different implementations can improve the performance of the extraction process. Paper [6] enables the ontology to find relevant recent knowledge in the domain from communities, by exploiting their underlying knowledge as keywords. It extracts instances and statements from the documents using the ontology-based and pattern-based information extraction technique. A confidence value is used in order to maintain the stability of 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 of 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 from RDB in the form of OWL/RDF triples, to ensure its availability for semantic web. Their system then automatically constructs an OWL ontology from RDB schema using direct mapping rules. Later, SPARQL queries are rewritten from SQL by translating SQL relational algebra into an equivalent SPARQL. In paper [7], authors have developed a framework for comparing 11 ontology learning systems. After analyzing these methodologies, three were selected and then applied to finance domain. Out of the 11 ontology learning systems, we have shortlisted and compared the following five : Ontolearn [10] uses text mining and statistical techniques to learn concepts and build taxonomic relations; Text2Onto [11] uses Probabilistic Ontology Model and involves statistical and linguistic techniques to create an ontology; the CRCTOL [12] algorithm is a statistical algorithm which extracts concepts and relations; OntoGain [14] is an unsupervised algorithm which uses linguistic tools to preprocess text and extract concepts; HCHIRISM [13] is also an unsupervised algorithm which recursively analyzes a large number of websites in order to find important concepts for a domain by introducing an initial keyword. A comparative study of these methods can be found in the Appendix.

III. Proposed Model

The two main components of 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 from the text using a traditional feature-based selection approach.

Figure 1. demonstrates the system design incorporating the two modules to accomplish the goal of Domain specific keywords extraction.

The two main inputs to the system are i) User's document from which words are to be extracted and ii) The desired domain. Of these the domain is given as input to the Ontology Creation Module.

The proposed model works as follows:

1. The Domain Ontology to be used is created.
2. The user's document is pre-processed and cleaned. The two pre-processing steps are Cleaning and Lemmatization.
    a. Cleaning involves the removal of special characters and stop-words from the text.
    b. Lemmatization is used to derive the root form of the word from its inflected version.

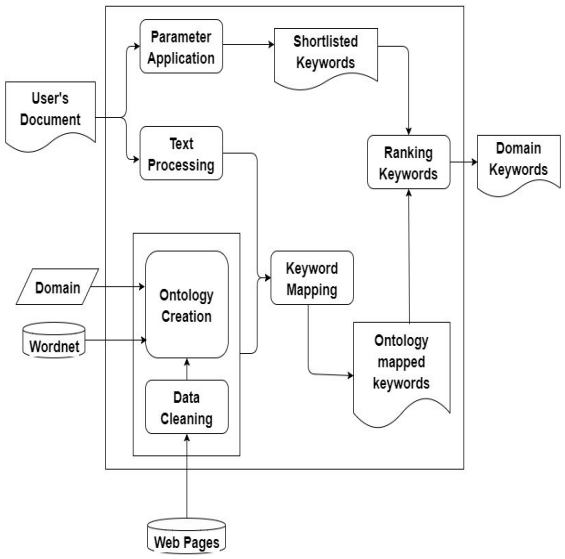


Figure 1. System Design

3. The words from this preprocessed text are then mapped to concepts and instances in the Ontology created in step 1. Mapping involves looking up the existence of the words from the preprocessed text in the Ontology.
4. Parallely the keywords from the original document are shortlisted and scored on the basis of the following parameters:
    i. Frequency of the word in the English language.
    ii. Frequency of occurrence in the text.
    iii. Position weightage.
    iv. Part of Speech of the word.
    v. Number of 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 from the Ontology Mapping and Parameter Application are then ranked.
6. Finally the highest ranked keywords are output as the Domain specific keywords.

IV. Results

The results are based on three major parameters viz. f-measure, precision and recall. According to the literature survey, the HCHIRSIM model resulted in the precision of 89.02%. The expected outcome of the project is of a precision ranging from 80 to 90%. The recall and f-measure values are expected to increase by a margin of approximately 5%.

V. Conclusion

In this paper, we have explained the need for Ontology in identifying keywords in a given document and have understood how it is useful in various disciplines. We highlighted that the existing systems use NLP and often extract keywords based on title. The delta change in the proposed system is to extract keywords based on domain.

The proposed system has two main modules which are Ontology creation module and Keyword extraction module. By employing Ontology we can overcome the ambiguity that existing system suffers from. Major application of proposed system is Classification of documents, which further can be applied to Search engine optimization, Semantic web etc.

Appendix

Paper	Tools	Domain	Performance Measure (%)		
			Precision	Recall	F Measure
[10]	OntoLearn	Tourism	84	52.74	64.8
[11]	Text2Onto	Tourism	17.38	29.95	22
[12]	CRCTOL	Terrorism	86	57.1	68.63
[13]	HCHIRSIM	Medical-Cancer	89.02	93.87	91.38
[14]	OntoGain	Computer Science	73	62	66.8

Table 1: Comparison of Ontology Creation Algorithms

Acknowledgment

We are thankful to our college Vivekanand Education Society's Institute of Technology for considering our project and extending help at all stages needed during our work of collecting information regarding the project. We are deeply indebted 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.

References

[1] HunKyung Yoo, YooMi Park, TaeDong Lee "Ontology based Keyword Dictionary Server for Semantic Service Discovery" (2013)

[2] Rashmi Chauhan, Rayan Goudar, Robin Sharma, Atul Chauhan "Domain Ontology based Semantic Search for Efficient Information Retrieval through Automatic Query Expansion" (2013)

[3] Komal Mule and Arti Waghmare. "Context Based Information Retrieval Based On Ontological Concepts" (2015)

[4] Hongsheng Wang, Lu Yuan, Hong Shao. "Text Information Extraction Based on OWL Ontologies" (2008)

[5] Fernando Gutierrez, Dejing Dou, Adam Martini, Stephen Fickas and Hui Zong. "Hybrid Ontology-based Information Extraction for Automated Text Grading" (2013)

[6] Dhomas Hatta Fudholi, Wenny Rahayu and Eric Pardede "Ontology-based Information Extraction for Knowledge Enrichment and Validation" (2016)

[7] Omar Ismaïl, Bouchra Frikh,Brahim Ouhbi - "Building Ontologies: a State of the Art, and an Application to Finance Domain" (2014)

[8] Qiuxia Song, Jin Liu, Xiaofeng Wang, Jin Wang "A Novel Automatic Ontology Construction Method Based on Web Data" (2014)

[9] Mohamed A.G. Hazber, Ruixuan , Xiwu , Guandong, Yuhua Li-"Semantic SPARQL query in a relational database based on ontology construction" (2015)

[10] Michele Missikoff, Roberto Navigli, Paola Velardi "Integrated Approach to Web Ontology Learning and Engineering" (2002)

[11] Philipp Cimiano, Johanna Volker "A Framework for Ontology Learning and Data-driven Change Discovery" (2005)

[12] Xing Jiang, Ah-Hwee Tan "Mining Ontological Knowledge from Domain-Specific Text Documents" (2005)

[13] B. Frikh, A. S. Djaanfar and B. Ouhbi "A hybrid Method for Domain Ontology Construction from the Web"

[14] Euthymios Drymonas, Kalliopi Zervanou, Euripides G.M. Petrakis "Unsupervised Ontology Acquisition from Plain Texts: The OntoGain System" (2010)

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 accomp 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 , invo s 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 Principal Dr. (Mrs.) J.M. Nair , □or giving us this valuable opportunity to do this project.

IEEE IEEE BOMBAY SECTION

International Conference on Smart City and Emerging Technologies (ICSCET)

organized by

Universal College of Engineering

(Gujarati Linguistic Minority Institution)

(Approved by AICTE, DTE & Affiliated to University of Mumbai)

Near Bhajansons Dairy & Punyadham, Kaman Bhiwandi Road, Vasai East, Mumbai – 401212, Maharashtra, India.

in association with

Institute of Electrical and Electronics Engineers (IEEE)

This is to certify that Dr./Ms./Mr. **SNEHAL BHAGAT**

has presented a paper on

Ontology Based Domain Dictionary

in the "International Conference on Smart City and Emerging Technologies" (ICSCET-2018)

organized on 5th January 2018 at Universal College of Engineering

in association with ***IEEE Bombay Section.***

Prof. Asir Khan	Dr. Ajoy Kumar	Dr. Jitendra Patil
Conference Chair	*Principal*	*Campus Director*

IEEE IEEE BOMBAY SECTION

International Conference on Smart City and Emerging Technologies (ICSCET)

organized by

Universal College of Engineering

(Gujarati Linguistic Minority Institution)

(Approved by AICTE, DTE & Affiliated to University of Mumbai)

Near Bhajansons Dairy & Punyadham, Kaman Bhiwandi Road, Vasai East, Mumbai – 401212, Maharashtra, India.

in association with

Institute of Electrical and Electronics Engineers (IEEE)

This is to certify that Dr./Ms./Mr. PADMAJA KOLLE

has presented a paper on

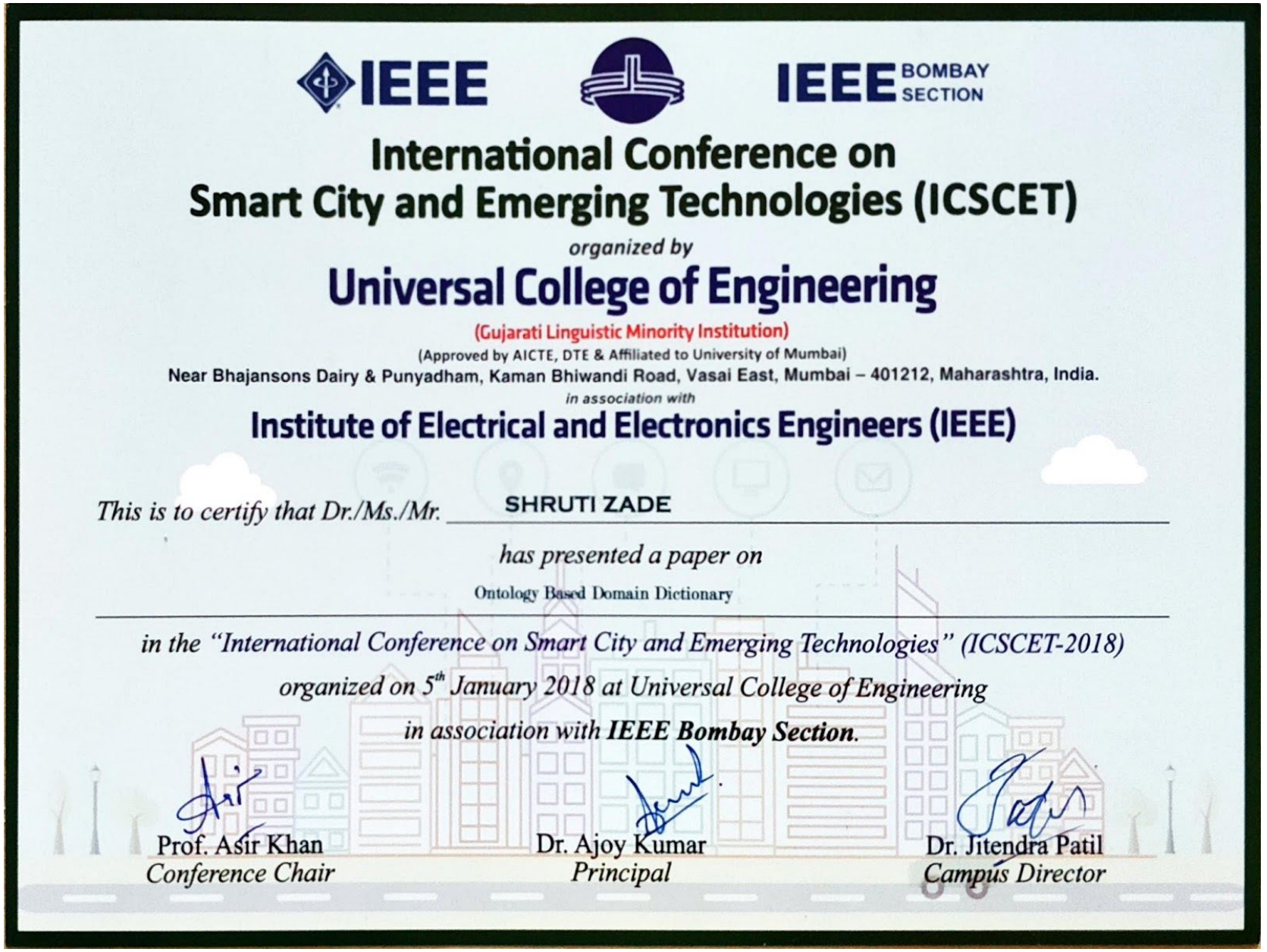
Ontology Based Domain Dictionary

in the "International Conference on Smart City and Emerging Technologies" (ICSCET-2018)

organized on 5th January 2018 at Universal College of Engineering

in association with ***IEEE Bombay Section.***

Prof. Asir Khan	Dr. Ajoy Kumar	Dr. Jitendra Patil
Conference Chair	*Principal*	*Campus Director*



IEEE IEEE BOMBAY SECTION

International Conference on Smart City and Emerging Technologies (ICSCET)

organized by

Universal College of Engineering

(Gujarati Linguistic Minority Institution)

(Approved by AICTE, DTE & Affiliated to University of Mumbai)

Near Bhajansons Dairy & Punyadham, Kaman Bhiwandi Road, Vasai East, Mumbai – 401212, Maharashtra, India.

in association with

Institute of Electrical and Electronics Engineers (IEEE)

This is to certify that Dr./Ms./Mr. SHRUTI ZADE

has presented a paper on

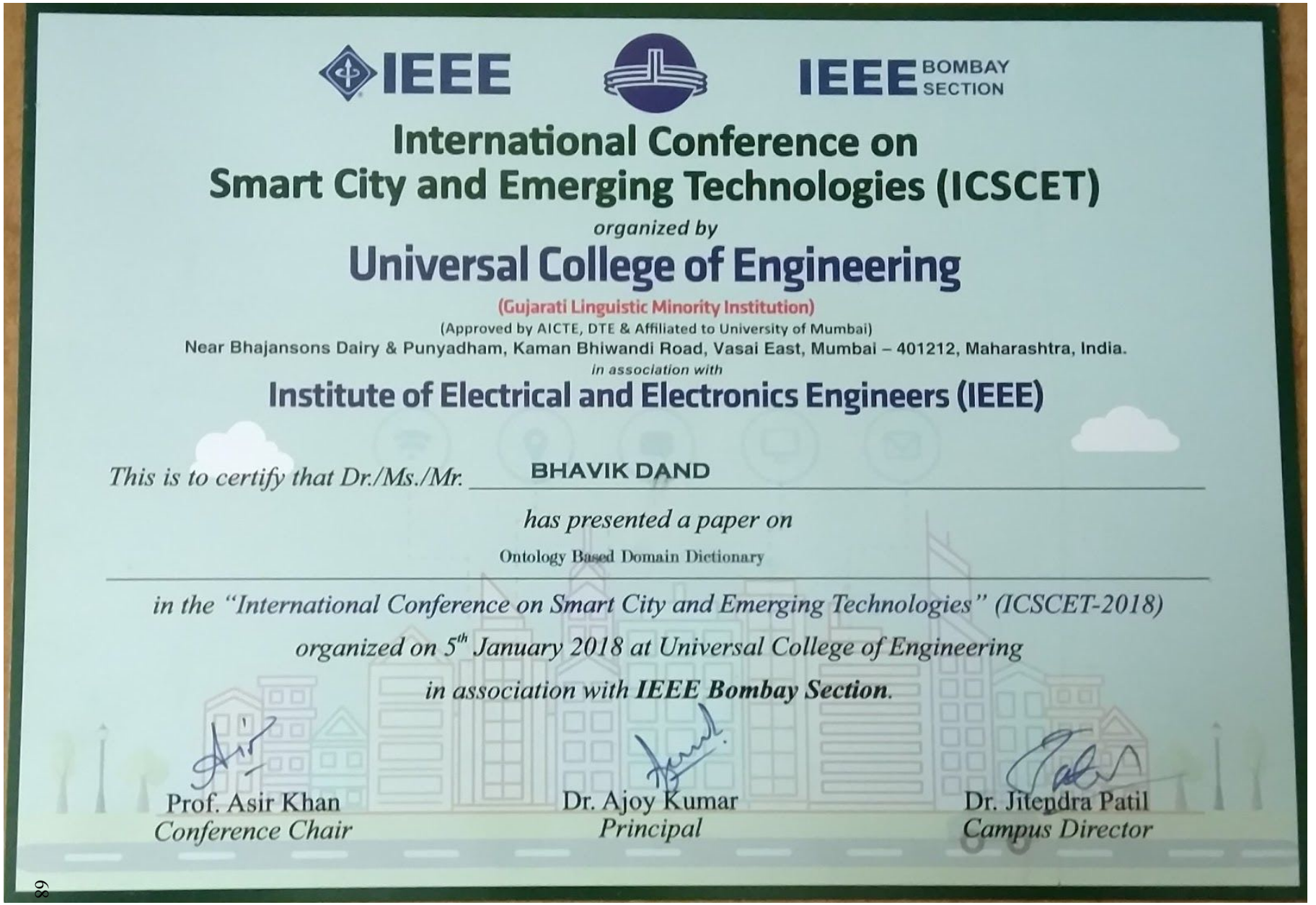
Ontology Based Domain Dictionary

in the "International Conference on Smart City and Emerging Technologies" (ICSCET-2018)

organized on 5th January 2018 at Universal College of Engineering

in association with ***IEEE Bombay Section.***

Prof. Asir Khan	Dr. Ajoy Kumar	Dr. Jitendra Patil
Conference Chair	*Principal*	*Campus Director*



IEEE

IEEE BOMBAY SECTION

International Conference on Smart City and Emerging Technologies (ICSCET)

organized by

Universal College of Engineering

(Gujarati Linguistic Minority Institution)

(Approved by AICTE, DTE & Affiliated to University of Mumbai)

Near Bhajansons Dairy & Punyadham, Kaman Bhiwandi Road, Vasai East, Mumbai – 401212, Maharashtra, India.

in association with

Institute of Electrical and Electronics Engineers (IEEE)

This is to certify that Dr./Ms./Mr. BHAVIK DAND

has presented a paper on

Ontology Based Domain Dictionary

in the "International Conference on Smart City and Emerging Technologies" (ICSCET-2018)

organized on 5th January 2018 at Universal College of Engineering

in association with ***IEEE Bombay Section.***

Prof. Asir Khan	Dr. Ajoy Kumar	Dr. Jitendra Patil
Conference Chair	*Principal*	*Campus Director*

Paper 2: An Ontological Approach for Keyword Extraction

An Ontological Approach for Keyword Extraction

Shruti Zade [#], Padmaja Kolle [#], Snehal Bhagat [#], Bhavik Dand [#], Lifna C.S [$]
[#] B. E Student, Department of Computer Engineering
[$] Assistant Professor, Department of Computer Engineering
Vivekanand Education Society's Institute of Technology, Mumbai, India
Email : shruti.zade@ves.ac.in, padmaja.kolle@ves.ac.in, snehal.bhagat@ves.ac.in, bhavik.dand@ves.ac.in, lifna.cs@ves.ac.in

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 only on term-document frequency. However, these do not always guarantee accurate results. In this 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, a 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.

This paper discusses 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 the proposed system. In order to appreciate it better and to understand how it facilitates keyword extraction, a literature survey to understand how an ontology is created and its various applications has been presented. 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, relevant Web documents which are 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 to 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.

3. Proposed Model

The goal of the system is to extract 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 this 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 the next approach is considered. The next approach to creating the ontology store is by creating 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 is the main module of the system. Here, the words in user's document are mapped 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)

After data pre-processing, three scoring parameters discussed in [15] were applied to identify word relevance. They are :

- **Entropy:** Using frequency directly for calculation can sometimes give misguided results; 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, entropy has been used as a parameter. The formula for calculating entropy is as follows:

$$W1 = \frac{F}{N} log2(\frac{F}{N})$$

  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:** The position of sentence where the word exists in the document has been considered. The idea behind this is that words in initial paragraphs carry more weightage than words in last. This is given by:

$$W2 = (\frac{St+1}{Sf+1})$$

  Where W2 = weight of given word, due to index position of the sentence in which it occurs first.

  St = Total number of sentences in given document.

  Sf = Sentence Index in which the given word occurs first.

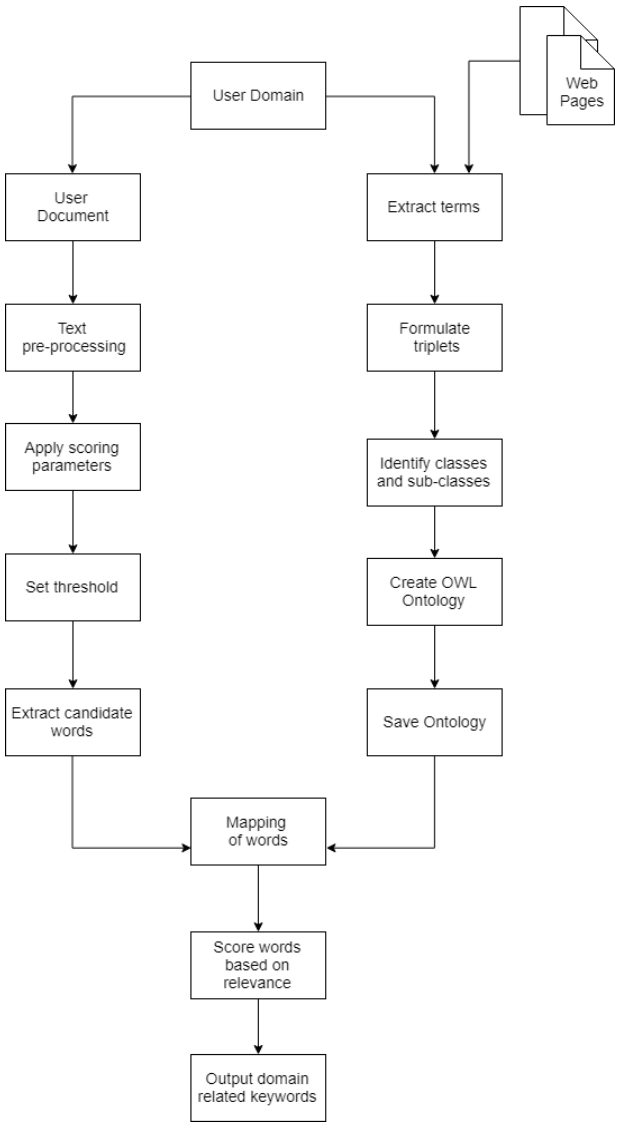


Fig. -1: System Flow Diagram

- **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.

$$P(K) = I(K) \qquad if\ (I(K) < (L(S)/2))$$

$$= 2 \times (L(S) - I(K)) \qquad else$$

  where P(K) = Partial Position related strength of given distinct word

  We combine strength due to length of sentence with the formula:

$$W3 = log2(\Sigma \frac{L(S)+1}{P(K)+1})$$

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.

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. Exhaustive ontology will most definitely give much better result than an ontology with few classes and properties. The domain selected to test the system is "Library". An ontology was created for the same. The user document sample was taken from wikipedia's 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

This paper realizes the need for a contextual extraction for keywords and has found that an ontological approach gives accurate results. A system was created 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, as also for multi domain systems. 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 this paper creates a custom ontology as defined briefly in this paper and make it as exhaustive as possible.

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.

REFERENCES

[1] HunKyung Yoo, YooMi Park, TaeDong Lee "*Ontology based Keyword Dictionary Server for Semantic Service Discovery*". 2013

[2] Rashmi Chauhan, Rayan Goudar, Robin Sharma, Atul Chauhan "*Domain Ontology based Semantic Search for Efficient Information Retrieval through Automatic Query Expansion*" 2013

[3] Komal Mule and Arti Waghmare. "*Context Based Information Retrieval Based On Ontological Concepts*". 2015

[4] Hongsheng Wang, Lu Yuan, Hong Shao. "*Text Information Extraction Based on OWL Ontologies*". 2008

[5] Fernando Gutierrez, Dejing Dou, Adam Martini, Stephen Fickas and Hui Zong. "*Hybrid Ontology-based Information Extraction for Automated Text Grading*". 2013

[6] Dhomas Hatta Fudholi, Wenny Rahayu and Eric Pardede "*Ontology-based Information Extraction for Knowledge Enrichment and Validation*". 2016

[7] Omar Ismaïl, Bouchra Frikh,Brahim Ouhbi - "*Building Ontologies: a State of the Art, and an Application to Finance Domain*". 2014

[8] Qiuxia Song, Jin Liu, Xiaofeng Wang, Jin Wang *"A Novel Automatic Ontology Construction Method Based on Web Data"*. 2014

[9] Mohamed A.G. Hazber, Ruixuan , Xiwu , Guandong, Yuhua Li-"*Semantic SPARQL query in a relational database based on ontology construction*". 2015

[10] Michele Missikoff, Roberto Navigli, Paola Velardi "*Integrated Approach to Web Ontology Learning and Engineering*". 2002

[11] Philipp Cimiano, Johanna Volker "*A Framework for Ontology Learning and Data-driven Change Discovery"*. 2005

[12] Xing Jiang, Ah-Hwee Tan *"Mining Ontological Knowledge from Domain-Specific Text Documents*". 2005

[13] B. Frikh, A. S. Djaanfar and B. Ouhbi "*A hybrid Method for Domain Ontology Construction from the Web"*

[14] Euthymios Drymonas, Kalliopi Zervanou, Euripides G.M. Petrakis "*Unsupervised Ontology Acquisition from Plain Texts: The OntoGain System*". 2010

[15] Niraj Kumar,Kannan Srinathan,Vasudeva Varma"*Evaluating Information Coverage in Machine Generated Summary and Variable Length Documents*".2010

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. 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.

An Ontological Approach for Keyword Extraction

Submitted by: Shruti Zade, Lifna C.S., Padmaja Kolle, Snehal Bhagat, Bhavik Dand

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.

Detail	Value
Current Status	accepted
Copyright Form	Received on 21 Apr 2018 (View or Save)
Registration Fee	View Applicable Fee or Pay Online Now
Digital Certificate(s)	Provided after publishing
Print Certificate	For hardcopy of certificate, please proceed to pay. Click Here

First Author Details	
Author Name	Shruti Zade
Designation	Student
Country	India
Area of Research	Knowledge Representation
Organisation	Vivekanand Education Society's Institute Of Technology
Co-author(s)	Lifna C.S., Padmaja Kolle, Snehal Bhagat, Bhavik Dand