Ontology based Domain Dictionary

Padmaja Kolle
B.E.
Department of Computer Engineering
V.E.S.I.T, Mumbai
India
padmaja.kolle@ves.ac.in

Snehal Bhagat
B.E.
Department of Computer Engineering
V.E.S.I.T, Mumbai
India
snehal.bhagat@ves.ac.in

Shruti Zade
B.E.
Department of Computer Engineering
V.E.S.I.T, Mumbai
India
shruti.zade@ves.ac.in

Bhavik Dand
B.E.
Department of Computer Engineering
V.E.S.I.T, Mumbai
India
bhavik.dand@ves.ac.in

Lifna C. S.
Asst. Prof.
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.

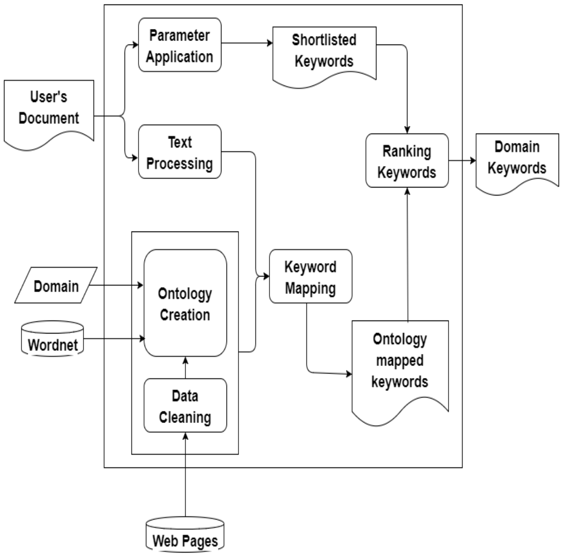

Figure 1. System Design

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

			Performance Measure (%)		
Paper	**Tools**	**Domain**	**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)