INTERNATIONAL JOURNAL OF ADVANCE RESEARCH, IDEAS AND INNOVATIONS IN TECHNOLOGY

IJARIIT

ISSN: 2454-132X

Impact factor: 4.295

(Volume 4, Issue 2)

Available online at: www.ijariit.com

An ontological approach for keyword extraction

Shruti Zade
shruti.zade@ves.ac.in
Vivekanand Education Society's Institute of Technology,
Mumbai, Maharashtra

Lifna C. S
lifna.cs@ves.ac.in
Vivekanand Education Society's Institute of Technology,
Mumbai, Maharashtra

Padmaja Kolle
padmaja.kolle@ves.ac.in
Vivekanand Education Society's Institute of Technology,
Mumbai, Maharashtra

Snehal Bhagat
snehal.bhagat@ves.ac.in
Vivekanand Education Society's Institute of Technology,
Mumbai, Maharashtra

Bhavik Dand
bhavik.dand@ves.ac.in
Vivekanand Education Society's Institute of Technology,
Mumbai, Maharashtra

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 depends on it. One important module of data classification is keyword extraction module which tells us what the document talks about. The existing keyword extraction software 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 no other frequently occurring 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 broad, 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 this 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 the 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 have been presented. The following paragraph briefly discusses few papers that explain 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 categorize 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. The 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 an 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 of 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 the 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:

- On to learn [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 extracts concepts and relation using a 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 the 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 for 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 of 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 the 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 the word in the document

N = Total Number of words in the document.

- **The position of the sentence:** The position of the 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.

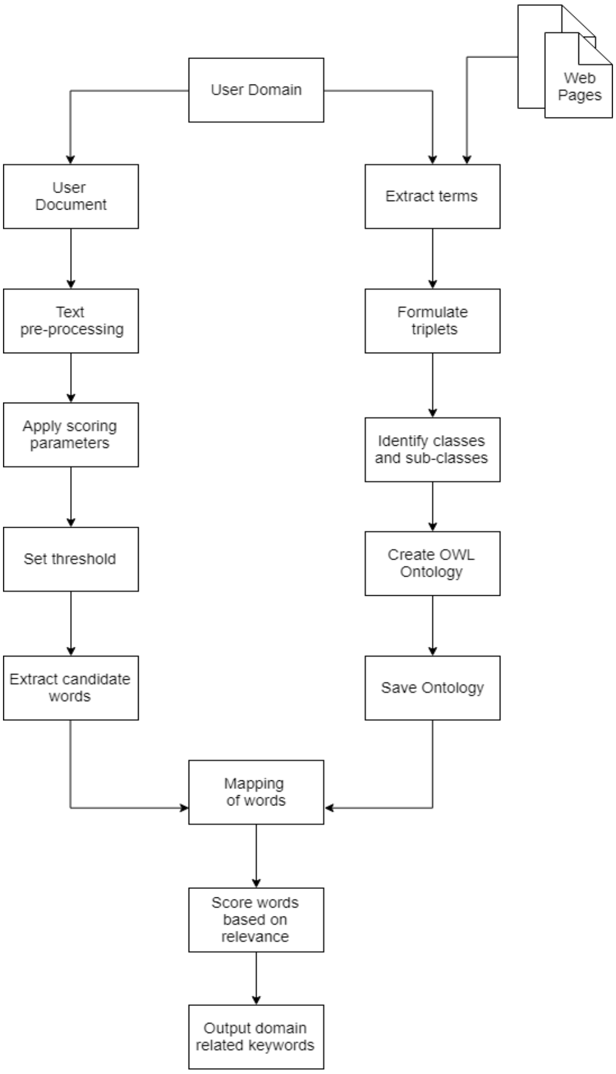

Fig. -1: System Flow Diagram

- **Position related strength:** Position related strength is calculated using two factors viz. The 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.

$$P(K) = I(K) \qquad if\ (I(K) < (L(S)/2))$$
$$= 2 \times (L(S) - I(K)) \quad else$$

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

We combine strength due to the length of a 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 exists.

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 analyzed on basis of accuracy, precision, recall, and f-measure. Before discussing them in greater detail, it's important to realize that the accuracy of the approach depends heavily on the ontology. The 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 a library of roughly 700 words. Accuracy is the ratio of correctly classified observation and total no of observations. Precision deals with the correctly classified observation by total positive observations. Recall is the ratio of correctly classified observation and all positively identified observation. The F1 score is the 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 truly positive, which is correctly classified positive observation.
TN is a true negative, which is correctly classified negative observation.
FP is false positive, which is incorrectly classified positive observation.
FN is a 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 a few are available this paper creates a custom ontology 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.

7. 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, and 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.