Automatic Generation of Domain Specific Keywords

Tabraiz Anwer, Adeel Ahmed

Department of Computer Science, SZABIST
Karachi, Pakistan
tabraizanwer@yahoo.com
adeel.ahmed@szabist.edu.pk

Abstract— **Identifying and understanding the domain specific terminologies from the noisy text is an important task. Classifying the document into some domain and finding domain related Technical terms or keywords may have comprised of one or more keywords. Therefore finding the starting and ending boundaries of technical keywords is very important. Finding such words help us link them to specific knowledge domain or target web pages which describes the knowledge or information in the topic. It helps users to understand the topic at hand better. Research Study would focus on finding domain specific keywords for multiple domains such as algorithms, database or networking.**

***Keywords*— Information Retrieval, Domain Specific Terms Extraction**

I. Introduction

Domain specific terms have some semantic to the respected domain. For Example, terms such as Database, Database Systems and Database Management System are associated with the domain database. Novice user or reader reading advance topic or research paper related to the domain “database” who did not identify these terms as unique or separated terms can have lot of trouble to understand that research paper. Extracting domain specific terminologies from the noisy text is a classification task where it is classified into predefined domains. Mostly it was employed with key phrase extraction [1], word sense disambiguation [2] and query expansion and cross-lingual text categorization [3].

II. LITERATURE REVIEW

A. Information Retrieval

Information Retrieval is finding knowledge / Information from the noisy text (Unstructured text) within large collection. Mostly Information Retrieval is done by all of us in our everyday life to facilitate business, education and Entertainment. Web search engines is the most popular and heavily used Information Retrieval Service, whether it is accessing latest information, news and events, finding and comparing online product for shopping. For Business, they need searching in their documents, email and memos.

B. Applications of Information Retrieval

Majorly Information Retrieval’s central task is searching. This field also covers inter-related problems, such as manipulation, storage and retrieval of language data.

1. *Document filtering:* In this typical search, the Application identifies the possible interest of the users from the given search queries in advance and behalf of these statements application filter the documents according to it. Such as users interest is only in business section of the news. Spam filter in an email system to block unwanted emails.

2. *Text Categorization:* Categorization of unlabelled articles. At the beginning some sample data will be provided with some classes such as “Sports”, “Technology” and “gadgets” to a categorization system and then it will automate itself according to the previous categorized articles submitted earlier at the stage.

3. *Summarization:* This system summarizes the articles or document to few paragraphs or phrases describing their content.

4. *Information Extraction:* This System identifies the entities such as name, places and describes relationship between these entities or link them to some dictionary already maintained previously.

   In this paper, we present the approach for extracting multi-word term from the document and classifying/label it into the domain. For domain specificity we use statistical method of term frequency (TF) and inverse document frequency (IDF) over document as well as on domain to capture domain related terms.

III. DISCUSSION AND ANALYSIS OF RELATED WORK

Many works has been done since late 80s in the field information extraction, where so many optimized solutions have been proposed. Most of these solutions are based on supervised methods of extracting domain specific terminologies. But the unsupervised approaches for extracting domain specific terms are still open. Here I have discussed and analysed some of these most appropriate proposed solutions. However, to date, the research to approach the task in an un-supervised manner is that of Park et al. [6] and Su Nam Kim et al. [7]. In Unsupervised methods, there have the obvious advantages but they need laborious manual classification of training instances and they are applicable to arbitrary sets of domains, tasks and languages. Su Nam Kim et al. [7], proposed method for unsupervised approach using inverse document frequency method. She used the Reuters sample data to work on it and generated the TF-IDF of those documents with related to domain and compared the results with Park et al. [6].

IV. AUTOMATIC GENERATION OF DOMAIN SPECIFIC TERM

In this section we elaborate (the method for classifying documents) the process of automatic generation of domain specific keywords. Firstly I have built the dictionary of domain specific terms through parsing the document and generated the frequency of term related to domain through TF-IDF method specific to domain secondly we generated the frequency of term of document and then compare the term of documents with the dictionary built with the previously generated domain specific terms. The higher the terms related to the domain. That domain must relate to that domain.

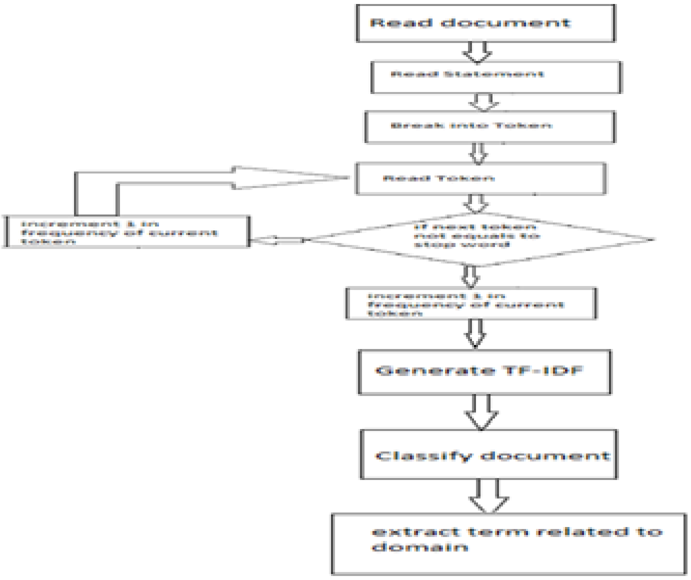

Read document
Read Statement
Break into Token
Read Token
increment 1 in frequency of current token
if next token not equals to stop word
increment 1 in frequency of current token
Generate TF-IDF
Classify document
extract term related to domain

Fig 1: Working of Token generation algorithm

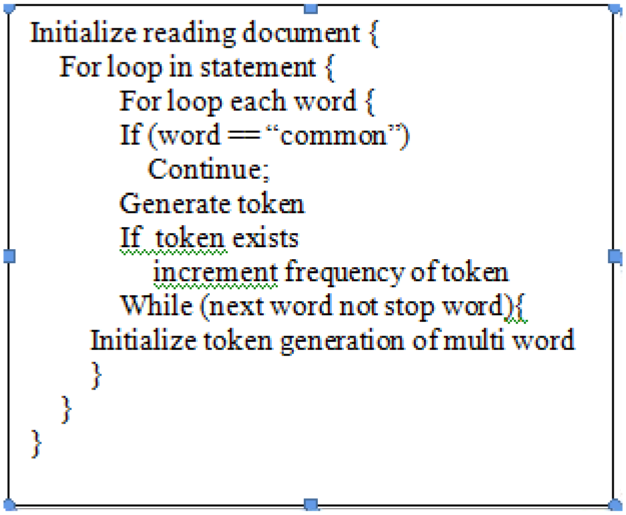
A. Proposed Method

1) Building Dictionary of Domain Specific Term through TF-IDF method: In this step we parsed the Reuters document sample data 21578 document sample data. I have applied my process to 20578 documents related to 123 domains through which I get that 1,853,214 words not including the common or stop words and the terms are multi-words. Reuters sample data is in xml format with some attributes are defined with that documents. For e.g.;

- <TOPIC></TOPIC>, //topic of the document
- <PLACES></PLACES>, //to which place it related
- <BODY> </BODY>, // content of the document.

In this topic section, there are textual categorizations that the document belongs to which document and place tag define to which country or region it is related to and in the body tag there is main text of the document.

2) Tokenization: In generating the frequency we have started the tokenization process. Tokenization algorithm generates token that is not only single word but generates the multi-word token. it is started from first word and start making token with second word and then third word until it get “comma”, “semi-colon”, “fill-stop” or any other common words that is listed as stop word in our file. These stop words are listed on majorly information retrieval sites [13] that are working on term extraction. Belong with the process of token generation it is also start making the frequency of the generated token if the token is not generated it save the token and if the token is already generated it will increment the frequency of that token.



```
Initialize reading document {
   For loop in statement {
         For loop each word {
         If (word == “common”)
            Continue;
         Generate token
         If  token exists
             increment frequency of token
         While (next word not stop word){
      Initialize token generation of multi word
      }
   }
}
```

Fig 2: Multi-Word Token Generation Algorithm

3) *Flow Chart of the Algorithm:* The flow chart diagram in Figure 1 represents that how the algorithm of token generation works.

4) *Algorithm for Multi-Word Token Generation:* Algorithm for Multi-Word Token Generation is presented in Figure 2.

5) *TF-IDF Method:* After generation of frequency we implement TF-IDF method. Term frequency is the term occurs in the document TF (t,d). Inverse document Frequency is obtained by dividing the number of document by the number of document that term appears in using formula given below:

$$IDF(t) = \log \frac{|D|}{|\{d \in D : t_i \in d\}|}$$

Here t_i is the term related to the document and d is the no. of documents in which t_i exists. |D| is the total number of documents.

TF-IDF is calculated by:

TF-IDF = TF product with IDF

With this TF-IDF, I have generated the TF-IDF of the terms related to domains and once the task has been completed, we have selected the term having higher threshold values. Specifically choosing the point at which there is significant drop in TF-IDF. Table 1 show the terms related to the domain extracted by this method. Through this step we have successful identified the term related to the domain.

6) *Generating TF-IDF of document specific term:* Now, in the second step, we have to classify the domain of the document. At this stage dictionary is maintained and TF-IDF method is also apply on this document we can now have the term of this document. Now those terms that have higher threshold values are compared to the domain dictionary. If terms of a document are mostly related to any one of domains, we can say that the parsed document is related to that specific domain. That is available through same method by generating TF-IDF of that document and but in this step domain is not specified.

TABLE I
SOME TERMS FROM DICTIONARY BUILT THROUGH REUTERS DOCUMENTS

Terms	Domains
Oil	Crude-oil
Barrels	Crude-oil
Share	Earn
Average share	Earn
Net loss	Earn
Interest rates	Interest
Current account	Bop
Exchange rates	Money-fx

7) *Comparing the document specific term with domain built dictionary*: When the TF-IDF of document is generated, the selected terms of the document is compared with domain built dictionary so that the document can be classified into the pre-defined domain. When we parsed the Reuters documents the major domains were “crude-oil”, “trade”, “earn” and “interest” now when the TF-IDF of the document is generated I have got the results shown in Table II.

TABLE II
SOME TERMS FROM DOCUMENT AND TF-IDF OF THAT TERMS

Terms	TF-IDF
Oil	13.86
Prices	7.7
Crude	8.1
Crude oil	8.9
Barrel	9.1

After generating the TF-IDF, I have selected those terms and compare TF-IDF of those terms with the built dictionary of the terms that I have done in my first step and I have got the following result of the above terms as shown in Table III.

TABLE IVII
TF-IDF OF THE DOMAIN HAVING TERM “OIL” IN THAT DOMAIN

Domain where term “oil” exists	TF-IDF
Crude-oil	866.342
Natural gas	141.677
Earn	97.84
Ship	87.77

In Table III, we can see the term “oil” has higher TF-IDF in domain Crude

TABLE VV
TF-IDF OF THE DOMAIN HAVING TERM “BARREL” IN THAT DOMAIN

Domain where term “barrel” exists	TF-IDF
Crude-oil	555.928
Natural gas	49.49

Fuel	44.74
Earn	24.74
Trade	12.048
Money-fx	8.83

In Table IV, we can see term that the term “barrel” has higher TF-IDF in domain Crude-oil

TABLE V
TF-IDF OF THE DOMAIN HAVING TERM “PRICES” IN THAT DOMAIN

Domain where term “prices” exists	**TF-IDF**
Crude-oil	165.818
Earn	43.73
Trade	37.46
Wheat	28.45

So through the Table V, we can easily decide that the document that was parsed was specifically pointing to the domain crude-oil. And hence the term having higher TF-IDF of the document majorly relates to the domain.

V. COMPARISON BETWEEN MY PROPOSED ARCHITECTURE AND PRE WORK

Results of optimized parameters of my proposed architecture and architecture proposed by Su Nam Kim et al. [6] work was for the extraction of domain specific terms while this approach was inspired by the paper of Su Nam Kim. But the author only extracted the domain specific keywords. But in my research I have improved her algorithm by including multi-word term and generated their frequency also. And instead of defining the document related to the domain. In my work I have built dictionary through the above approach and parse the document of unknown domain and find that the document relates to which domain and what will be the terms related to that domain in the documents.

VI. FUTURE WORK

This paper shows that how it will be implemented unsupervised approach for document categorization and term detection in the current document. In Our Approach we had built dictionary and the classified the document in semi-unsupervised method and term recognition is built on behalf of that dictionary. In future work, this paper work can guide in the knowledge based application and give the idea of to how to implement unsupervised method of document classification and term extraction. In future it will be helpful in research based organizations where term and domain of the documents are unknown. Through this document can be parsed and domain will be identified and term related to that domain will be highlighted so that it can refer to some knowledge based dictionary.

VII. CONCLUSION

Even though, we presented approach to identify the relation of document with the domain and extract the domain specific terminologies from the document through unsupervised manner, using simple TF-IDF approach. The TF-IDF approach of Su Nam Kim et al (2009) extract the single word terms and we try to improve to get multi-word terminologies. Through our method terms frequency are high due to combination of word for extracting multi-word terms. Qualities of term are high when distributed over all domains. Major problem may be arises when the document may be consist of terms related to two or more domains. We verified that this approach will be helpful in extracting multi-word domains although we could not prove the utilization of these terms. But there is indication that it needs further analysis. Further it can be used in applications such like highlighting the domain specific terms for e.g. in helping students reading the re-search papers and un-aware of terms related to its domain. We conclude it can be helpful in text categorization.

REFERENCES

[1] S. M. Metev and V. P. Veiko, *Laser Assisted Microtechnology*, 2nd ed., R. M. Osgood, Jr., Ed. Berlin, Germany: Springer-Verlag, 1998.

[2] J. Breckling, Ed., *The Analysis of Directional Time Series: Applications to Wind Speed and Direction*, ser. Lecture Notes in Statistics. Berlin, Germany: Springer, 1989, vol. 61.

[3] S. Zhang, C. Zhu, J. K. O. Sin, and P. K. T. Mok, “A novel ultrathin elevated channel low-temperature poly-Si TFT,” *IEEE Electron Device Lett.*, vol. 20, pp. 569–571, Nov. 1999.

[4] M. Wegmuller, J. P. von der Weid, P. Oberson, and N. Gisin, “High resolution fiber distributed measurements with coherent OFDR,” in *Proc. ECOC’00*, 2000, paper 11.3.4, p. 109.

[5] R. E. Sorace, V. S. Reinhardt, and S. A. Vaughn, “High-speed digital-to-RF converter,” U.S. Patent 5 668 842, Sept. 16, 1997.

[6] (2002) The IEEE website. [Online]. Available: http://www.ieee.org/

[7] M. Shell. (2002) IEEEtran homepage on CTAN. [Online]. Available: http://www.ctan.org/tex-archive/macros/latex/contrib/supported/IEEEtran/

[8] *FLEXChip Signal Processor (MC68175/D)*, Motorola, 1996.

[9] “PDCA12-70 data sheet,” Opto Speed SA, Mezzovico, Switzerland.

[10] A. Karnik, “Performance of TCP congestion control with rate feedback: TCP/ABR and rate adaptive TCP/IP,” M. Eng. thesis, Indian Institute of Science, Bangalore, India, Jan. 1999.

[11] J. Padhye, V. Firoiu, and D. Towsley, “A stochastic model of TCP Reno congestion avoidance and control,” Univ. of Massachusetts, Amherst, MA, CMPSCI Tech. Rep. 99-02, 1999.

[12] *Wireless LAN Medium Access Control (MAC) and Physical Layer (PHY) Specification*, IEEE Std. 802.11, 199