Journal of Computer Science 7 (9): 1363-1367, 2011
ISSN 1549-3636
© 2011 Science Publications

An Automatic Topic Identification Algorithm

Hossein Shahsavand Baghdadi and Bali Ranaivo-Malançon
Multimedia University, Faculty of Information Technology
Cyberjaya, Malaysia

Abstract: Problem statement: Topic is a stream of words which stands for the content of a text. Knowing the topic of a document can help people to be aware from its content and facilitate their searching process. **Approach:** This paper proposes an automatic algorithm to identify the topic for a textual document based on the chunks corresponding to each sentences in the document. **Results and conclusion:** We achieved 86% matching for both total and partial matching in our experimental data sample.

Key words: Web document, text-document topic, partial matching, experimental data, Identification algorithm, chen's algorithm, syntactic parser, Adaptive Resonance Theory (ART), Maximum Ratio Balance (MRB)

INTRODUCTION

Understanding the content of a document can be a time-consuming procedure if it has been done manually. However, in some cases people are looking for some documents in a specific area and it is not possible for them to read all the documents to identify the relevant materials. Even by reading the summary of documents people have to spend lots of time for searching process. On the other hand, determining the area of documents by their summaries is a critical issue for both human and machine since the summary is normally one or more paragraphs. Knowing the topic of documents can address this issue by reducing the amount of text which should be read and consequently the required time to identify the domain of documents.

In this study, we propose an automatic algorithm to identify the topic for a textual document. This algorithm consists of five steps and it is capable to be run by the machine. During this algorithm, we first split the sentences inside the text and then we pars them by a syntactic parser to determine the chunks correspond to each sentences. After determining the chunks we select the most important chunks in each sentence and consider them as the topic for associated sentence. Then we try to calculate the weight for the sentences according to their selected chunks. Finally we identify the most weighted sentence's topics as the topic for whole document. This algorithm has some similarities to Chen's Algorithm (Chen, 1995) in terms of steps, however, some basic parts like selected parts in each sentence and calculation formula for topics of sentences have been modified which are described in next sections. This study is organized as following:

In next section we clarify the concept of topic and we distinguish it from title. Then we investigate some similar study and after that we explain the details about our algorithm. After that we describe an experiment which has been conducted base on proposed topic identification algorithm and provide the associated results. At the end we conclude this study in conclusion section.

Topic and Title: In this study, we define the term "topic" as a stream of terms which represent the content of text. A topic is different from a title, which is also a sequence of terms but rather represent the name of a study and does not necessary represent the content of this study. Most of the documents are embossed by their titles; however, the title is not necessarily stands for the content of documents and it is not possible to judge about the content of documents by only their titles. The automatic identification of the topic of a given document is not an easy task as a document may contain multiple topics.

Related works: Many research studys have been conducted for topic identification. Majority of approaches to detect the topic for a document are based on clustering algorithms. Cluster analysis or simply clustering is the assignment of a set of observations into subsets (called clusters) so that observations in the same cluster are similar in some sense. They try to extract or generate a stream of terms have been fallen in a most prior cluster by their algorithms.

Corresponding Author: Hossein Shahsavand Baghdadi, Multimedia University, Faculty of Information Technology, Cyberjaya, Malaysia

Anaya-Sánchez *et al*. (2010), the authors proposed a new method based on hierarchical clustering to generate the document's topic with most frequent terms and select some sentences to create a description for generated topic. Although the algorithm is capable to study with only one document, its accuracy depends on number of documents it has been studying with. They first tokenize the text and remove the prepositions, this leads to have a bag of words. Then they make the word list in order with most frequent at the top. They make all possible couples from the list and select the couple with most probability to have a semantic relation between the words. They also define the probability of generating a pair of terms for a collection of documents. At this moment they have a list of couples has been ordered by their probability and they can select as many couple as they need in respect to the accuracy needed. By using that approach, they attain 71% accuracy in document topic detection (Anaya-Sánchez *et al*., 2010).

(Ayad and Kamel, 2002) have proposed another algorithm purely based on clustering techniques. They exploit hierarchical, partitional and incremental clustering as following order to extract the topic from a set of documents. They used vector space model and Term Frequency-Inverse Document Frequency, or TF-IDF to determine the similarity between clusters and define a topic for each cluster. To generate the cluster's topic, the most common approach is to use the cluster's representative words. It can be done by truncating by those words regarding to a predefined length threshold. Those terms can be selected against a weight threshold. The authors in this study chose a close method by re-computing the term's weight in respect of revealed cluster structure. Here authors considered the most frequent terms in each cluster which are rare in other clusters as the best representative terms in that cluster. The overall topic accuracy that they achieved is 78%.

(Rajaraman and Tan, 2001) proposed a method to discover a text-document topic based on self-organizing neural netstudys. They have exploited Adaptive Resonance Theory (ART) netstudys which are a class of self organizing neural netstudys. Fuzzy ART incorporates computations from fuzzy set theory into ART net-studys.

Tiun *et al*. (2001) proposed a three-step algorithm to identify a Web document topic. They first extract the text part form web document based on predefined tags. Then they run a mapping module to map the extracted keywords on the words of ontology concepts that have been stemmed and sense-tagged. This mapping module exploits Yahoo ontology and Word Net as extended ontology database. The final module is optimization module which is responsible to shrink the ontology tree into an optimized tree where only active concepts and the intermediate active concepts are chosen. To determine the most suspicious nodes to be the topic, they created an algorithm which can find the node with greatest accumulated mixture distribution among the optimized tree. The algorithm which entitled "Ratio Balance Algorithm" is able to determine the Maximum Ratio Balance (MRB) of a single path node using subtraction of actual accumulated mixture weight with supposed accumulated mixture weight. By that way they have succeeded to obtain maximum 69.8% accuracy in topic identification (Tiun *et al*., 2001).

(Chen, 1995) Presented his study on topic identification based on two kinds of grammatical pairs: noun-noun and noun-verb. To select this pair, he first determined the importance of each noun and verb by Inverse Document Frequency (IDF).

$$IDF(W) = \log((P - O(W)) / O(W)) + c \quad (1)$$

Where, P is the number of documents in Corpus, i.e., 500, O (W) is the number of documents with word W and c is a threshold value. The threshold value for nouns is 0.77 and for verbs are 2.46. These values are used to represent the unimportant words, whose IDF values are negative. That is, their IDF values are reset to zero. Then he calculates the strength of each pair. The strength of one occurrence of a verb-noun pair or a noun-noun pair is computed by the importance of the words and their distances. (2) and (3) demonstrate how these values are calculated.

$$SNV (N,V) = IDF(N).IDF(V) / D (N,V) \quad (2)$$

$$SNV (N,N) = IDF(N).IDF(N) / D(N,N) \quad (3)$$

D is the distance and it is measured by the difference between cardinal numbers of two words. He assigns a cardinal number to each verb and noun in sentences. The cardinal numbers are kept continuous across sentences in the same paragraph. The strongest pair can be considered as a topic for each paragraph or entire document. Chen achieved to obtain around 80% accuracy (both total and partial matching) in identifying the discourse topic (Chen, 1995).

The last research which has been investigated in this study is by (Coursey and Mihalcea, 2009), in which Wikipedia is used to determine the document's topic. The method consists of two main steps. Firstly, they created a conceptual knowledge graph from Wikipedia where the nodes are entities of categories in Wikipedia. The edges are the proximity relation between articles inside this encyclopedia. The graph is created and will be used for later calculations. Then in second step, they first identified the encyclopedic conceptual weight in a text and then built the connections between the content of the document and the graph that they created in first

part. Then they perform a graph centrality algorithm on entire graph. Therefore all the nodes are ranked including new input document.

Automatic topic identification approach: As it has been investigated in literature review, there are many approaches to identify a document's main topic. They use different methods to address this issue and they obtain various results based on their techniques. Automatic Topic Identification Algorithm is closely similar to Chen's algorithm (Chen, 1995) in terms of study steps and general concepts. Chen tried to calculate IDF which is a weight for each noun and verb in a text by using (1) and then created all possible couples of noun-noun or noun-verb in each sentences. Then he calculated the weight for each couple based on (3) and (4) and considered the most weighted pair as the topic for the sentences. He then selected the most weighted sentences topics as the identified topic for whole document. Automatic Topic Identification follows some steps of Chen's algorithm, however, we reorganized his steps and also we modified his token selection method. Knowing that a document is made of many sentences and each sentence will output candidate topic, we intend to apply a technique for weighting these topics and then select the topic with the highest weigh. The weighting technique that will be used in this study has not been defined yet. Our algorithm consists of five different steps:

Split the text into sentences: The first step in our algorithm is splitting the sentences on the given text. In fact, the proposed algorithm is considered as a "divide and conquer" approach; therefore, the first step should be dividing the problem until it cannot be divided more. A sentence is a smallest text part which is capable to have a topic. Hence, we split the document into corresponding sentences. During this research we widely exploit Proxem Antelope (Proxem, 2009) which provides an open-source plenty of NLP tool. One of these tools is Text Splitter which splits a text into sentences. By performing this tool we would have a set of sentences.

Pars the sentences: In this time, Chen's algorithm tries to calculate the weight for each noun and verb and then creates all possible pairs. That may cause some overhead due to calculate the weight for some unimportant terms. Our proposed algorithm intends to pars the sentences and determines the candidate terms first to avoid any useless calculation. We believe that syntactic parts like Noun Phrase (NP) and Verb Phrase (VP) are playing most important roles to present the meaning of the sentence and therefore we should consider them instead of grammatical roles like noun and verb to identify the candidate topic for each sentence.

These syntactic parts are accessible through a dependency syntactic parser. In this study, we use the Stanford dependency parser (The Stanford Parser) which is an open-source tool available in Proxem Antelope package. For example, in sentence "My dog also likes eating bananas", the parser has recognized "my dog" as an NP subject and "likes eating bananas" as the VP. Figure 1 illustrates the syntactic analysis for above sentence which has been done by Stanford Parser.

Select the candidate parts: We select noun phrase (NP) and the head of a Verb Phrase (VP) instead of just pairs of nouns and noun-verb. We assume that the most important parts from a sentence are the NP's that function as subject or complement and the head of the VP. To illustrate it, in sentence "My dog also likes eating bananas", the phrase "my dog" is selected as the NP and "likes" is selected as the head of the VP and "bananas" as an NP complement. The combination of these three segments will be considered as candidate topic. Hence, the topic for this sentence is identified as "My dog likes bananas". At the end of this step, we have a set of candidate topics.

Calculate the weight for each candidate topic: At this moment we can calculate the IDF and SNV for only required syntactic parts. By this way, there is no need to calculate these amounts for irrelevant parts and in fact, we avoid any calculation overhead. Regarding to our modification in selected part of sentence, the calculation formula is changed to (4).

SNV (NP, head (VP)) = IDF (NP) . IDF (head (VP)) / D (NP, head (VP)) (4)

Select the final topic: When we determine the candidate topic and its associated weight for each sentence, we select the most weighted one and consider it as the main topic for the whole document. In case there are more than one candidate topics with greatest weight, we consider all of them as the main topic.

Experiment: As it is mentioned in previous chapters, "topic" stands for stream of terms which carry the semantic and meaning of text inside the document. However, it is not necessarily as same as the title which embossed on the top of document. Therefore, one proper method to evaluate the accuracy of topic identifier could be the comparing the identified topic by Automatic Topic Identification Algorithm for a random documents with the real topic which is determined manually for that document. In fact, this method is human result against machine result.

Your query

My dog also likes eating bananas.

Tagging

```
My/PRP$  dog/NN  also/RB  likes/VBZ  eating/VBG  bananas/NNS  ./.
```

Parse

```
(ROOT
  (S
    (NP (PRP$ My) (NN dog))
    (ADVP (RB also))
    (VP (VBZ likes)
      (S
        (VP (VBG eating)
          (S
            (ADJP (NNS bananas))))))
    (. .)))
```

Fig. 1: Dependency parsing with stanford dependency parser

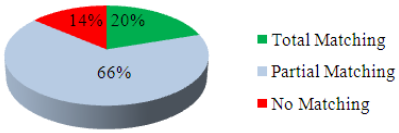

Fig. 2: Percentage of different results for automatic topic identification algorithm

	Total matching	Partial matching	No matching
Occurred in 200	40	132	28
Percentage	20 %	66 %	14 %

Fig. 3: Detail of automatic topic identification algorithm experiment

Wikipedia is an online encyclopedia which their pages are entitled exactly by their topics. Therefore, the Wikipedia page's titles can be considered as their topics. Due to this, a set of random pages from Wikipedia with their topics could be a suit dataset for this evaluation. To achieve this purpose, a set of 200 random pages with their topics have been selected.

By conducting such comparison, three different results would be considered for each case; two topics are totally matched, or partially matched (have some words in common) or totally different. The percentage of each group can demonstrate the accuracy of Automatic Topic Identification Algorithm. The result of this experiment is drawn as a pie chart in Fig. 2 and Fig. 3 illustrates the full details.

According to the pie chart (Fig. 2), Automatic Topic Identification Algorithm is able to recognize the exact correct topic in 20% of cases and in 66% of cases; it is able to identify a similar topic. Moreover, in 14% of cases, there were no common word between real topic and identified topic by this algorithm. Therefore, we reached the matching of 86% for both total and partial matching in this experiment.

Future study: Although Automatic Topic Identification is an algorithm and it is independent from any implementation, the developed version of this algorithm to conduct the experiment is able to process only English pages.

This limitation is emerged from NLP tools which are able to process only English texts. To address this issue, two approached are considered. In the first approach, we can use the modules which are able to process in other languages. To implement it, we need to add the new tools in our library in Utilities layer and determine the language of text before using the proper library. There are many tools to determining the language of text. One of them is Google Language Detector which is accessible by API (Application Programming Interface) technology. This tool is also available online (Fig. 6.1). In this figure, the language for the term "Daneshgah" (in Persian "دانشگاه") which means "university" is correctly detected Persian.

CONCLUSION

Identifying the topic for documents can reduce the required time for read and facilitate the searching process for those who are looking to find the relevant documents in a specific domain. The proposed method is an automatic algorithm to identify the main topic for any typical textual document. The main idea in this algorithm is dived the problem and conquer the simpler problem until addressing the main issue. In this way, we split the text into sentences and try to identify the topic for each sentence by selecting its appropriate syntactic parts. We also calculate a weight for each sentence's topic and consider the most weighted topics as the main topic of the whole text. The idea of mentioned algorithm is based on Chen's topic identification algorithm (Chen, 1995). We reorganized its steps and modified its selection policy. We select NP and VP instead of noun-noun and noun-verb pairs from each sentence. By this modification, we achieved 86% of matching for both total and partial matching among 200 random documents from the Wikipedia.

REFERENCES

Anaya-Sánchez, H., A. Pons-Porrata and R. Berlanga-Llavori, 2010. A document clustering algorithm for discovering and describing topics. Pattern Recognit. Lett., 31: 502-510. DOI: 10.1016/J.PATREC.2009.11.013

Ayad, H., and M. Kamel, 2002. Topic discovery from text using aggregation of different clustering methods. Proceedings of the 15th Conference of the Canadian Society for Computational Studies of Intelligence on Advances in Artificial Intelligence, (CACSIAAI'02), Springer-Verlag London, UK., pp: 161-175.

Chen, K.H., 1995. Topic identification in discourse. Proceedings of the 7th Conference on Eurpean Chapter of the Assication for Computational Linguistics,(ECACL'95), Morgan Kaufmann Publishers Inc. San Francisco, CA, USA., pp: 267-271 DOI: 10.3115/976973.977012

Coursey, K. and R. Mihalcea, 2009. Topic identification using Wikipedia graph centrality Proceedings of Human Language Technologies Annual Conference of the North American Chapter of the Association for Computational Linguistics, Companion, (NACACLC'09), Association for Computational Linguistics Stroudsburg, PA, USA., pp: 117-120.

Rajaraman, K. and A.H. Tan, 2001. Topic detection, tracking, and trend analysis using self-organizing neural networks. Adv. Knowl. Discovery Data Mining, 2035: 102-107. DOI: 10.1007/3-540-45357-1_13

SNLP, 0000. The Stanford Parser: A statistical parser. (n.d.). The Standfor National Langucage Processing Group.

Tiun, S., R. Abdullah and T.E. Kong, 2010. Automatic topic identification using ontology hierarchy. Comput. Linguistic Intell. Text Process., 2004: 444-453. DOI: 10.1007/3-540-44686-9_43