Computer and Information Science; Vol. 7, No. 4; 2014
ISSN 1913-8989 E-ISSN 1913-8997
Published by Canadian Center of Science and Education

Ontology Based Data Mining Approach on Web Documents

Hamideh Hajiabadi[1]

[1] Birjand University of Technology, Iran

Correspondence: Hamideh Hajiabadi, Birjand University of Technology, Iran. E-mail: hajiabadi@birjandut.ac.ir

Received: September 3, 2014 Accepted: September 14, 2014 Online Published: October 17, 2014

doi:10.5539/cis.v7n4p123 URL: http://dx.doi.org/10.5539/cis.v7n4p123

Abstract

Internet which is included plenty of huge data source is now rapidly increasing in all domains. It is considered as valuable data sources if the data can be processed that results in information. Data mining techniques are widely utilized in web documents in order to extract information. In this paper a data mining approach based on Ontology is proposed to classify web documents in order to facilitate applications based on classified text documents like search engines. The proposed approach is implemented and applied on several web documents. The experimental results show considerable progress.

Keywords: key phrase, data mining, ontology

1. Introduction

There are main resources on the internet containing information in any field of human activities that the availability of information technology (IT) make growth of digital data volumes exponentially. Although the size of data resource are massive and unstructured, there are a lot of application mining on the web and help people to find relative information [11]. Readers can easily recognize the main topic of the documents by frequent key phrases which contain one or more key words. Indeed key phrases point to major concept of the document. Many text-based applications such as search engines, text summarization and text clustering are based on key phrases. [7] [12]

Because of the massive information included on the web, it is impossible to manually extract key phrase. Then an effective method should mine and classify the text documents. The process cannot be performed manually because of the huge volumes of data. Consequently automated techniques are developed in order to classified documents. Although these kind of techniques produce several errors, but these are more preferable in comparison to manual techniques. [10] [13]

An ontology based approach is proposed in this paper such that automatically extract key phrases and classify web documents [6]. For this reason a qualified ontology which explain categories and sub categories precisely is needed. In order to build such ontology Wikipedia which is a "multilingual, web-based, free-content encyclopedia project supported by the Wikimedia Foundation and based on an openly editable model" is used. Wikipedia is expanded by collaborative attempt, and consequently all of the fields are approximately contained in. in order to build such a qualified ontology the approach proposed by C. Schonberg is used [5]. The ontology extracted from Wikipedia is named to WikiOnt. Because of the huge amount of categories included in the Wikipedia the ontology extracted is qualified.

Remaining of the paper is organized as follows: section 2 explains a number of related works which are carefully explained with each benefits and drawbacks. The algorithm generating ontology from Wikipedia is explained comprehensively in section 3. Section 4 explains how to extract key phrases. Section 5 contains evaluation results and the conclusion section is followed.

2. Related Works

Researches concentrates on extracting key phrases from text documents which are divided into supervised techniques and unsupervised ones. Unsupervised approaches are concentrated on this section. These approaches are conducted by two steps. Initially using several policies variety of key phrases are candidate, then variety of ranking algorithms are utilized in order to select most essential key phrases.

Brace Well in 2005 proposed an approach initially select some cluster from document, then the selected clusters are ranked. Ranking process is performed by calculating frequencies of terms occurred in the document.

Eventually top ranked categories are selected as key phrases. [2]

There are other kind of researches which are used graph-based ranking algorithm in order to extract key phrases. For this reason a graph with terms as nodes and their associations as edges are generated. Then a graph based ranking algorithm is applied to the graph and top ranked terms are selected as key phrases. A graph based ranking approach to extract key phrases are proposed by Litvak in 2008 and Haung in 2004. [3][4]

In [8] a full-text documents are represented by a graph which its nodes and edges are determined by entry recognition and UMLS semantic network, then the concepts are ranked by using relative important algorithm.

Sarkar in 2011 proposed an approach which extract key phrases from Bengali documents using two important features, the first occurrence and the frequency. [9]

Yu presents an approach which uses CRF in order to extract key-phrases, and then a classifier like SVM is used to classify txt documents. [10]

In this paper a new approach based on ontology is proposed which extract key phrases in order to classify text documents.

3. Ontology

The ontology used for this reason should be contained all the fields with hierarchy structure. Consequently it cannot build manually. As briefly explained earlier the ontology is generated by mining Wikipedia. Wikipedia is a free encyclopedia contains numerous articles written in hundreds of languages. Because of the collaborative efforts of lots of users in adding new articles, Wikipedia expand enormously and consequently it contains almost all of the fields and sub fields.

Schonberg [5] used wikis in order to develop ontology. in this paper Schonberg's algorithm is used to generate a large qualified ontology with all categories and sub categories exist in Wikipedia.

The technique proposed by Schonberg follows three main steps:

1. Categories are extracted from Wikipedia
2. A hierarchy category graph is developed.
3. An Ontology is created from category graph.

Conducting those steps an ontologies is constructed which is well qualified for this reason. In order to facilitate the ontology is called WikiOnt. Figure 1 shows a view of WikiOnt.

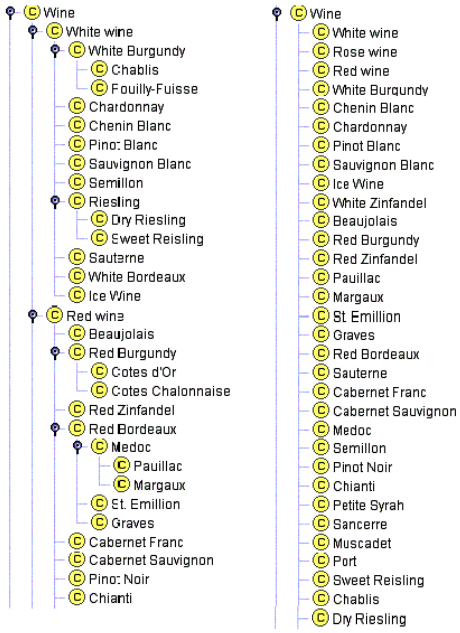

Figure 1. View of WikiOnt

4. Proposed Approach

There are three main steps included which are described as follows:

1. The key phrases are extracted and grouped.
2. The effect ratio of each key phrases are calculated.
3. The category of text document is specified by utilizing WikiOnt.

4.1 Extracting Key Phrases

In order to extract key phrase 2 steps are performed:

1. A set of candidate key phrases are selected.
2. Candidate key phrases are ranked using a ranking algorithms.
3. Ten top ranked key phrases are selected.

The input of the proposed approach is plaintext documents. Consequently the documents should be converted to plaintext one, it can be easily done by document converter.

Stanford POS Tagger (Note 1) system is applied to plaintext documents in order to assign pos tag to each sense. After that key phrases included less than 3 words are selected. Number of key phrases can be reduced enormously by eliminating key phrases with invalid pos tag information. For example key phrases which are not labeled as adjectives, verbs, or nouns, are removed.

4.2 Calculating Effect Ratio for Each Key Phrase

It is clear that key phrases only included one keyword occurs more than those containing more than. Key phrases are divided into 3 list: the ones containing one keyword, key phrases included 2 keywords and the last list contains key phrases having 3 keywords. So frequency of each key phrase is compared with those of the same keyword number.

Always key phrases appeared in the beginning of the documents are more important. For each key phrases 4 parameters are evaluated: *importance ratio*, *first appearance ratio*, *key phrases frequency* and *effect ratio*. *importance ratio* is evaluated by dividing *first appearance ratio* into *total* where *First appearance ratio* is the number of key phrases appeared in the document earlier and *Total* is the total number of key phrases included in the document.

If a key phrase occurs in bold style in the document, it is more important than the key phrases with the same frequency ratio which are not bold anywhere. Consequently its *importance ratio* is multiplied by 1.2. *Key phrase Frequency* is evaluated by counting key phrase 's occurrence in the document.

effect ratio is evaluated using equation (1):

$$effect\ ratio = important\ rate \times key\ phrase\ ferequency \quad (1)$$

When effect ratio is calculated for each key phrase, Wordnet is utilized to clarify synonym key phrases. The synonym key phrases are grouped. And the *total effect ratio* is assigned to each group by total sum of *effect ratio* of each key phrase included in the group.

Eventually the groups are sorted by their total effect ratios and ten top ranked groups are selected which are most frequented in the text document.

4.3 Classifying Text Document Using WikiOnt Ontology

After selecting more important key phrases, it is amied to map each group to a source included in WikiOnt. Wordnet plays key role in the mapping process.

For this reason each group g_i of key phrases should be mapped to a resource r_i of WikiOnt using mapping algorithm. In order to classify text document, the process is conducted. At first the resource set $\{r_1,r_2,\ldots r_{10}\}$ is partitioned into 5 subset $\{r_1, r_2\}, \{r_3, r_5\}, \cdots, \{r_9, r_{10}\}$

The following algorithm is performed for each subset $\{r_i, r_j\}$:

Step 1: if r_i and r_j are the same, then the result is r_i or r_j. There is no difference.

Step 2: if r_i is ancestor of r_j the result will be r_j.

Step 3: if none of above conditions fires, then a closest resource r_x is discovered such that it must be ancestor of both resource r_i and r_j.

After executing the algorithm, result in 5 resources, one for each set$\{r_i, r_j\}$. Proposed algorithm is conducted recursively until only one resource is remained. The remaining resource is the nearest class to the document.

As explained earlier most of the unsupervised proposed algorithm is performed automatically. The proposed algorithm is implemented and performed on a variety of text documents. The results are obtained and explained carefully in the following section.

5. Results

The proposed technique is applied on several text documents. Three articles are selected from http://whatis.techtarget.com/ and the results for each are demonstrated in the following. Table 1 illustrates the results.

Table 1. Correct class and result obtained

	correct class	**result obtained**
Document1	Data masking	Data masking
Document2	Role mining	Mining
Document3	Authentication	Authentication

The results indicate that result is to some extend more accurate for documents which are more general. If the document is in a specific filed, the proposed approach always returns one ancestor of its exact class.

In the following a text document is given to the proposed approach and the steps are explained in details. Figure 2 demonstrate an input text which is given to a POS tagger and the results are shown.

Input: "In computer science and information science, an ontology formally represents knowledge as a set of concepts within a domain, and the relationships between those concepts."
Output: IN In NN computer NN science CC and NN information NN science, DT an NN ontology RB formally VBZ represents NN knowledge IN as DT a NN set IN of NNS concepts IN within DT a NN domain, CC and DT the NNS relationships IN between DT those NNS concepts .
IN - Preposition
NN - Singular noun
CC - Coordinating conjunction
DT - Determiner
RB - Adverb
BZ - Verb, 3rd ps. sing. Present

Figure 2. PoS tagger

To facilitate the process only the key phrases containing one word are considered. The effect ratio for each is calculated in the table 2.

Table 1. The Effect ratio of key phrases

	Key phrase	**Effect ratio**
1	Computer	1
2	Science	3.4
3	information	1.3
4	Ontology	1.5

5	represents	1.6
6	knowledge	1.7
7	set	1.4
8	Concepts	1.9
9	Domain	1.10
10	Relationships	1.11

The synonym key phrases are grouped and each grouped is mapped to a resource in the WikiOnt ontology. Figure 3 demonstrate the mapping process.

6. Conclusion and Future Work

In this paper an automated ontology-based approach to extract categories of text documents is proposed. At first key phrase are extracted and some top key phrases are selected, ranking process is performed based on calculating frequency of each key phrases. A rich qualified ontology is built from Wikipedia. It is named to WikiOnt. Each key phrase of the document is mapped to a resource included in WikiOnt. Wordnet is a lexicon dictionary which is severally used in this process.

It is better that future works focus on the way the key phrases are extracted and selected. Better selection of key phrases result in better classification. Furthermore future works should focus on the number of key phrases which are nominated, it significantly effects on result.

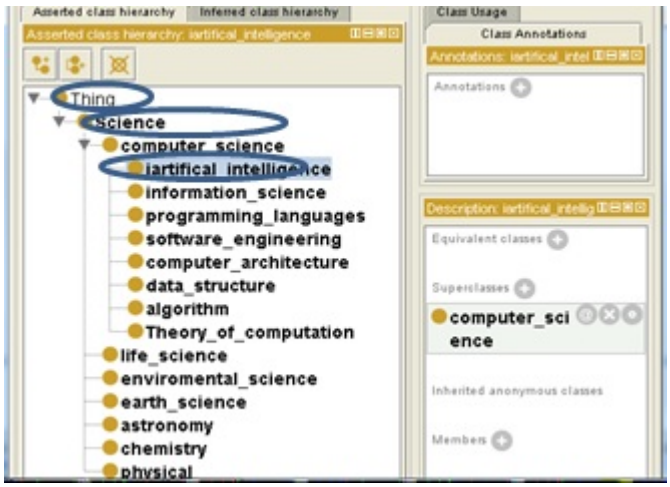

Figure 3. Mapping proccess. The result is information sciense

References

Berger, A. L., & Mittal, V. O. (2000). OCELOT: A system for summarizing Web pages. *In Proceedings Of the 23rd annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, 144–151.

Bracewell, D. B., Ren, F., & Kuroiwa, S. (2005). Multilingual single document keyword extraction for information retrieval. *In Proceedings of the 2005 IEEE International Conference on Natural Language Processing and Knowledge Engineering*, 517–522.

Christian, S., Helmuth, P., & Burkhard, F. R. (2010). Ontology Extraction and Wikipedia Expansion Using Language Resources, 11th international conference on Web Age Information Management (WAIM 2010) china, 2010 LNCS volume 6186.

Denny, J. C., & Smithers, J. D. (2002). A new tool to identify key biomedical concepts in text documents, with special application to curriculum content. *In Proceedings Of the AMIA Symposium*, 1007.

Ercan, G., & Cicekli, I. (2007). Using lexical chains for keyword extraction. *Inf. Process. Manage, 43*(6), 1705-1714.

Gutwin, C., Paynter, G., Witten, I., Nevill-Manning, C., & Frank, E. (1999). Improving browsing in digital libraries with keyphrase indexes. *Decision Support Systems, 27*, 81–104.

Huang, C., Tian, Y., Zhou, Z., Ling, C. X., & Huang, T. (2006). Keyphrase extraction using semantic networks structure analysis. *In ICDM '06: Proceedings of the Sixth International Conference on Data Mining*. Washington, DC, USA: IEEE Computer Society, 275–284

Jones, S., & Mahoui, M. (2000). Hierarchical document clustering using automatically extracted keyphrases.

Litvak, M., & Last, M. (2008). Graph-based keyword extraction for single-document summarization. In MMIES '08: Proceedings of the Workshop on Multi-source Multilingual Information Extraction and Summarization. Morristown, NJ, USA: Association for Computational Linguistics, 17–24.

Pudota, N., Dattolo, A., Baruzzo, A., Ferrara, F., & Tasso, C. (2010). Automatic keyphrase extraction and ontology mining for content-based tag recommendation. *International Journal of Intelligent Systems - New Trends for Ontology-Based Knowledge Discovery archive, 25*(12).

Sarkar, K. (2011). Automatic Keyphrase Extraction from Bengali Documents: A Preliminary Study. *2011 Second International Conference on Emerging Applications of Information Technology*, 125-129.

Schönberg, C., Pree, H., & Freitag, B. (2010). Rich Ontology Extraction and Wikipedia Expansion Using Language Resources. WAIM, 151-156.

Song, M., Bleik, S., Yu, H., & Hon, W. (2011). Extracting Biomedical Concepts from Fulltext by Relative Importance in a Graph Model. *2011 IEEE International Conference on Bioinformatics and Biomedicine Workshops*, 586-593.

Yu, F. Xuan, H., & Zheng, D. (2012). Key-Phrase Extraction Based on a Combination of CRF Model with Document Structure. *2012 Eighth International Conference on Computational Intelligence and Security,* 406-410.

Note

Note 1. http://nlp.stanford.edu/software/tagger.shtml.