© International Journal of Combinatorial Optimization Problems and Informatics, Vol. 5, No. 1, Jan-April 2014, pp. 21-25. ISSN: 2007-1558.

Ontology Based Data Mining Approach on Web Documents

Hamideh Hajiabadi
Birjand University of Technology
hajiabadi@birjandut.ac.ir

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 an ontology based data mining approach proposed to classify web documents in order to facilitate applications based on classified text documents like search engines. The proposed approach is implemented an applied on several web documents and the results are demonstrated.

Keywords—**key phrase; ontology; data mining.**

I. INTRODUCTION

Nowadays Word Wide Web is a main resource which everyone uses to extract required information. The information on the Web is huge and unstructured. Search engines are examined to facilitate in formation extraction. Key phrases which include one or more key words are descriptors that help reader to recognize topics and main concepts included in the documents. Informally we can say that key phrases indicate major idea of document. Key phrases are used in many text-based applications such as search engines. They are also useful in text clustering and text summarization.[7]

Manual key phrase extractions are costly and time consuming because of numerous documents exist on the web. Consequently it is essential to use techniques which automatically extract key phrases to reduce costs and time. In a manual process the supervisor should manage the process, and of course the results obtained this way are more effective and accurate than the dynamic ones. In this paper an ontology based approach proposed to extract key phrases dynamically and hence document topic.[6]

For our technique we need a rich ontology with high quality and a large number of relationships that accurately explain categories and sub categories. It Is better if ontology be more comprehensive. Then we try to build large qualified ontology.

Wikipedia is a good choice for that reason. Using the approach proposed by (Christian Schonberg, Helmut Pree, and Burkhard Freitag Rich 2010) an ontology is extracted and we named it to WikiOntfor further needs.

The paper is structured as follows: quite a number of related works is studied in section 2, for each its benefits and weaknesses are explained. Section 3 explains ontology built from Wikipedia. In section 4 a technique to extract key phrases is clarified. The evaluation results are demonstrated in section 5 and finally the conclusion section is presented.

Received Nov 20, 2013 / Accepted Dec 19, 2013

II. RELATED WORKS

Method existed to extract key phrases are divided into supervised and unsupervised. This section only concentrates on unsupervised ones. Unsupervised approaches firstly selects a variety of key phrases using some strategies; next by making use of some ranking techniques the most essential key phrases are selected.

Brace well in 2005 selected and ranked some cluster from the given document; ranking process is done by calculating frequency occurring each term in the document. Finally top ranked categories are selected as key phrases. [2]

Some works made use of graph-based ranking method to extract key phrases. In this method a graph with terms and its relatedness is build. Next using graph-based ranking algorithm, the terms are ranked. Finally top ranked terms are selected as key phrases. Litvak in 2008 and Haung in 2004 proposed a graph based ranking approach to extract key phrases. [3][4]

In this paper an ontology-based approach to extract key phrases and categories of given text documents is proposed.

III. ONTOLOGY

Ontology used for this job should be large and high qualified. It's expensive and time-consuming to build such ontology manually. Hence we should generate it automatically.

Wikipedia is a free encyclopedia contains millions of articles written in hundreds of languages. Because of the collaborative efforts of users in adding new articles, its growth is great and of course it is high qualified. So Wikipedia is perfect for our reason.

Schonberg in 2010 proposed an approach to develop ontology generated from wikis [5].We use the method presented in that paper 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. Extracting Category graph from Wikipedia
2. Developing and Extracting category graph.
3. Creating Ontology from category graph.

Following those steps an ontologies constructed which is perfect for this job. As told earlier we name it to WikiOnt.

IV. PROPOSED APPROACH

The approach is done in 3 steps as follows:

1- Extracting and grouping key phrases
2- CALCULATING EFFECT RATE OF KEY PHRASES
3- CATEGORIZING TEXT DOCUMENT USING WIKIONT

A.EXTRACTING KEY PHRASES

Key phrase extraction approach is always based on 2 stages:

1- A set of candidate key phrases are selected.
2- Candidate key phrases are ranked using various number of employed ranking algorithms.
3- Ten top ranked key phrases are selected.

Our proposed approach only uses plain text documents. So first we made use of document converter to extract plain texts from given document. Next a quite number of delimiters are used to separate sentences.

The previous work output is used as the input for Stanford POS Tagger[1] system to assign pos tag to each sense. Then key phrases up to 3 words are selected. To reduce number of key phrases they are removed by their pos tag information. For example key phrases which are not labeled as adjectives, verbs, or nouns, are removed.

b. Calculating effect rate of key phrases

Key phrases which contain one word have higher frequency than those which contain 2 or more words. So frequency of each key phrase is calculated in the list with those of the same word number.

Phrases which are appeared in the beginning of the documents are more important than the others ones. So we employ *importance rate* by dividing *first appearance rate* into *total* such that:

First appearance rate is the number of key phrases appeared earlier.

Total is total number of key phrases.

If a key phrase is bold in some parts of the document it means it is more important than the key phrases with the same frequency rate that are not bold in other parts. Consequently we multiply its *importance rate* by 1.2.

Key phrase Frequency: number of times a key phrase appears.

In order to find the effect of each key phrase, we employ effect rate as follow:

$$effect\ rate = important\ rate \times key\ phrase\ ferequency$$

After calculating above features for each separate key phrase, we make use of Wordnet to realize synonym key phrases and collect them into one group. We assign *total effect rate* to each group by total sum of effect rate of each key phrase included in the group.

Finally we rank the groups by their total effect rates and ten top ranked groups are selected. It is important to note that selected groups have the most impact on the text document.

c. Categorizing Text Document Using WikiOnt

[1]http://nlp.stanford.edu/software/tagger.shtml.

In this section each group of key phrases are mapped to are source in WikiOnt. Wordnet has a key role in the mapping process.

It means each group g_i of key phrases maps to a resource r_i exist in WikiOnt. In order to obtain most specific category for text document, we partition the so-called set $\{r_1, r_2, \ldots r_{10}\}$ to 5 subset $\{r_1, r_2\}, \{r_3, r_5\}, \cdots, \{r_9, r_{10}\}$

Following algorithm is done for each separate subset$\{r_i, r_j\}$:

Step 1: if r_i and r_j are similar, then the result could be each of them.

Step 2: if r_i is ancestor of r_j then result is r_j.

Step 3: if none of above happens then a resource r_x is found which is ancestor of both r_i and r_j.

After algorithm is done, each subset$\{r_i, r_j\}$converts to a resource. It means we have 5 resources instead of 5 subsets. Afterward two resources are selected and so-called algorithm is applied to selected resources. This algorithm runs recursively until only one resource remains. This resource is the nearest category to text documents.

Most of this process is done automatically and unsupervised. Consequently it saves a much costs in both time and effort. The proposed technique is done on some text documents and the results are explained in the following section.

V. RESULTS

The proposed technique is practice on some text documents. Bellow the results on three articles selected from http://whatis.techtarget.com/are demonstrated and practiced by proposed approach. The results are displayed in table I.

Table I specific category and retrieve category

	specific category	retrieval category
Document1	Data masking	Data masking
Document2	Role mining	Mining
Document3	Authentication	Authentication

The results indicate that the retrieval categories are more accurate for the text documents which are less specific. If topic of text document be general, the approach always returns correct category but if it be specific it always returns one ancestor of it.

VI. CONCLUSION AND FUTURE WORK

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

It is better future works focus on extracting and selecting key phrases. If the key phrases are selected best possible, the results will be more accurate.

Additionally future works should concentrate on the number of key phrases selected, it greatly effects on result obtained.

REFERENCES

[1] Christian Schonberg, Helmuth Pree, and Burkhard Freitag Rich, Ontology Extraction and Wikipedia Expansion Using Language Resources, 11th international conference on Web Age Information Management (WAIM 2010) china, 2010 LNCS volume 6186

[2] D. B. Bracewell, F. Ren, and S. Kuroiwa, "Multilingual single document keyword extraction for information retrieval," in Proceedings of the 2005 IEEE International Conference on Natural Language Processing and Knowledge Engineering, 30 October - 1 November 2005, pp. 517–522.

[3] M. Litvak and M.Last, "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, 2008, pp. 17–24.

[4] C. Huang, Y. Tian, Z. Zhou, C. X. Ling, and T. Huang, "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, 2006, pp. 275–284

[5] C.Schönberg, H.Pree, B.Freitag: Rich Ontology Extraction and Wikipedia Expansion Using Language Resources. WAIM 2010: 151-156

[6] N. Pudota, A. Dattolo, A. Baruzzo, F. Ferrara, C. Tasso, "Automatic keyphrase extraction and ontology mining for content-based tag recommendation" , Published in: International Journal of Intelligent Systems - New Trends for Ontology-Based Knowledge Discovery archive Volume 25 Issue 12, December 2010

[7] G. Ercan, I. Cicekli: Using lexical chains for keyword extraction. Inf. Process. Manage. 43(6): 1705-1714 2007