One Method of Keyword Extraction for Tibetan News Webpage

Tao Xu, Xiangzhen He*, Fucheng Wan and Xianghe Meng

ABSTRACT

Keywords of Tibetan text is important in the area of text text clustering/categorization, automatic abstracting, IR and so on. However, there are no keywords in the Tibet news WebPages. Besides, many algorithm for keywords extraction need the manually annotated corpus, so it is poor augment ability. Because Keywords can be considered as a set of words which are important and subject correlated cohesively in a document, this paper improved the CHI-Squared Statistic, use the idea of recommendation to extract keywords. Experiments from Tibetan news webpages demonstrate that this method is better than the method of TFIDF integrating with location information.

KEYWORDS

Tibetan information processing, CHI-Squared Statistic, keywords extraction

INTRODUCTION

Tibetan belongs to Sino-Tibetan family. It is an old and it is widely used in contemporary Tibetan text. With the development of Tibetan information, especially the popularity of Tibetan web, Tibetan text increases at high speed. Among so many news webs, if information is searched just by headlines, users may miss valuable
information or that they are interested in. If news webs are labeled with keywords, the above problem can be easily solved.

The keywords extraction research abroad has been developing quickly, and some practical systems have been set up. The earliest one is KEA[1] system, which is famous for its good extendibility. This system trains keyword extraction model by using the discrete Bayesian Method. GenEx[2] system extracts keywords by using genetic algorithm and C4.5 decision tree induction algorithm. There are also methods that extract keywords of a single text directly. Literature [3] is a good example. This method extracts representative words directly by marking the words based on the distribution of the subjects and words in LDA (latent dirichlet allocation). They build keyword extraction models according to the following three characteristics: the length of the noun phrases, the frequency and the frequency of the first word. Although the keyword extraction research in our country is not so developed as abroad, yet lots of research and achievements has been made. A mixed method is synthesizing linguistic analysis method and statistical learning method or adding heuristic knowledge (such as the location of the words, the length of the words, HTML tab, etc. In 2011, Yang Chunming proposed a fast automatic keyword extraction algorithm which was based on areas of the documents and extracting keywords by a simple lexical item statistical algorithm. Liu Zhiyuan in Tsinghua University proposed a keyword extraction method based on subjects which was first used in micro log for keyword extraction. Yin Qing and Hu Xuegang proposed a keyword extraction method based on a density clustering model for Chinese news web pages. In conclusion, methods of keywords extraction can be divided into two parts: the supervised learning method and the unsupervised learning method. The first needs lots of labeled corpus training classification models, aims at a certain field and has subjects that change with time. Therefore, this method is lack of universality. What’s more, since lots of training corpus is needed and there is no corpus labeling standard in the area of keywords extraction nowadays, this method is by no means time consuming. The second is faster for it needs no manual work of labeling. And its subjects are visual and practical. Since news web pages cover a wide range, if the supervised training method is used, the keywords extracted will varies, and lots of labeled corpus by manual work will be needed. Therefore, the second method is used in this thesis.

KEYWORD EXTRACTION FOR TIBETAN NEWS WEB PAGES

Basic Thinking

There are two steps in the Tibetan keyword extraction technology proposed by this thesis. The first step is the processing of the Tibetan news pages and building of the candidate keyword set. The processing of news web pages includes filtering the contents, extracting the title as well as saving some necessary labels; the processing of the text includes segmentation of the words, labeling of the part of speech, the
statistics of word frequency, processing of location information, phrasing of Tibetan sentences, processing of stop words and the forming of the candidate keyword set.

The second step is improving the keyword extraction on chi-squared statistic. With the enlightenment of chi-squared statistic, the correlation between each candidate keyword and each word in the word groups for comparing (chi-squared statistic) will be computed. That is extracting keywords by computing the correlation between the candidate keywords and each important word of the text according to the text.

**Candidate Keywords**

Nouns are considered as keywords only in many documents. Other words with other parts of speech work differently in keyword extraction for news web pages. For example, verbs have positive impact on the quality of the keywords extracted. But adjectives and adverbs work just the opposite. News web pages are different from other texts for they focus on the people and things of lately social events that drew great attention. So verbs of the text can express the subjects of the news more appropriately. Compared with nouns and verbs, descriptive words like adjectives and adverbs can not express the subject appropriately.

**Keyword Extraction Based on Chi-squared Statistic**

Chi-squared statistic was brought into keyword extraction for Tibetan texts in this thesis. Firstly, a group of lexical item H (whose tf-idf value is analyzed by means of corpus) is chosen from a single text as the initial generalization of the text (lexical item groups for comparison). Each sentence in the text is regarded as a topic sentence. And the co-occurrence distribution rate of candidate keywords ti (including the lexical items in group H) and the lexical item h in group H is calculated. The Degree of Bias between t and h is calculated through Chi-squared statistic. NiP* is the theoretical value (expected number). Ni is the number of co-occurrence distribution of ti and the lexical items in group H. And Phi is the tf-idf value of lexical item h)

Thus the chi-squared statistic of ti can be defined as following:

\[ \chi^2(t_i) = \sum_{h \in H} \frac{(f_{o(t_i,h)} - N_iP_h)^2}{N_iP_h} \]  

(1)

The lexical items in group H have the higher tf-idf value than any other lexical item in the text. Though they are calculated based on word frequency, they are of certain generality in some way. Tentatively H is regarded as the generalization of the text. In this case, the degree of bias of each ti and the lexical items in group H is calculated. The degree of bias is expressed by chi-squared statistic. And the sum between the lexical item ti and the chi-squared statistic in group H can be expressed
as $\chi^2(t_i)$ (as formula 1). Thus $\chi^2(t_i)$ is regarded as the degree of bias. That is, the bigger the value of $\chi^2(t_i)$ is, the higher of the degree of bias (between the lexical item $t_i$ and the text) will be, the better the lexical item can be qualified as the key item of the text.

**EXPERIMENT ANALYSIS**

**Preparation of Experiment Data**

The experiment data in the text is divided into two groups. The first group is used to obtain lexical item group $H$. The institute of national languages information technology in Northwest University for nationalities has set up a Tibetan text balanced corpus which covers a wide range like military affairs, politics, physical education, amusement and so on in the forms of newspaper and online books. Base on the corpus, the TFIDF value in the texts to be processed will be calculated, and the lexical item $H$ is hence obtained.

The second group is used in the experiment of keyword extraction for Tibetan news web pages. As there is no test dataset in Tibetan keyword extraction, to keep fair, the dataset in the text is from Tibetan website (www.tibentcm.com). 50 articles are extracted from three kinds of news web pages. And the standard of annotating keywords manually is as follow:

1. 3 to 8 keywords for each article;
2. The keyword phrases of each article must cover all the topics of the article;
3. Each keyword must appear more than once in the article;
4. The number of the Tibetan words must be no less than 200 for each article to be tested

**The Index of Evaluation**

Precision, Recall and F-measure which are in common use in information retrieval are used to evaluate the experiment result. The Precision and Recall of each article of each article is calculated separately. Then the Avg_Precision and Avg_Recall of all the articles are calculated. And lastly, based on the Avg_Precision and Avg_Recall, F-measure is calculated.

$$\text{Avg}_{\text{Precision}} = \text{Average(Precision}_1, \text{Precision}_2, ..., \text{Precision}_n) \quad (2)$$

$$\text{Avg}_{\text{Recall}} = \text{Average(Recall}_1, \text{Recall}_2, ..., \text{Recall}_n) \quad (3)$$

$$\text{F-measure} = \frac{2 \times \text{Avg}_{\text{Precision}} \times \text{Avg}_{\text{Recall}}}{\text{Avg}_{\text{Precision}} + \text{Avg}_{\text{Recall}}} \quad (4)$$
The Parameters Setting Experiment

In the method introduced in the text, one thing very important is the setting of the number of the lexical item H for comparing in the texts to be extracted. The proportion of the number of the lexical item H for comparing to the numbers of the article is set as Hp. In order to set the best Hp value, experiments are carried out when different Hp values are set. The best results are in table 1.

TABLE 1. THE INFLUENCE OF THE Hp VALUE ON KEYWORD EXTRACTION.

<table>
<thead>
<tr>
<th>Hp value</th>
<th>Avg_Precision</th>
<th>Avg_Recall</th>
<th>F-measure</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.2</td>
<td>0.621</td>
<td>0.328</td>
<td>0.429</td>
</tr>
<tr>
<td>0.3</td>
<td>0.782</td>
<td>0.371</td>
<td>0.503</td>
</tr>
<tr>
<td>0.35</td>
<td>0.810</td>
<td>0.523</td>
<td>0.636</td>
</tr>
<tr>
<td>0.4</td>
<td>0.632</td>
<td>0.451</td>
<td>0.526</td>
</tr>
</tbody>
</table>

From the above, we can see that the influence of Hp on the keyword extraction result is obvious. The words in H are the words of high TFIDF value in the texts to be extracted, so these words are important from the point of view of information retrieval. Thus these words can reflect the theme of the article. That is, the lexical items which closely associate with those words can be extracted from theory and experiment (calculated by chi-squared statistic). These words can reflect the theme better. Thus they are the keywords extracted in this text. However, the number of the words in H should be appropriate. Just as illustrated in table 1, Hp=35% is the best value.

Comparison of the Method in This Paper and the TFIDF Method

TABLE 2. COMPARISON OF THE RESULTS OF VARIOUS METHODS.

<table>
<thead>
<tr>
<th>methods</th>
<th>Avg_Precision</th>
<th>Avg_Recall</th>
<th>F-measure</th>
</tr>
</thead>
<tbody>
<tr>
<td>TFIDF</td>
<td>0.554</td>
<td>0.499</td>
<td>0.472</td>
</tr>
<tr>
<td>TFIDF+location</td>
<td>0.627</td>
<td>0.547</td>
<td>0.584</td>
</tr>
<tr>
<td>The method in</td>
<td>0.810</td>
<td>0.523</td>
<td>0.636</td>
</tr>
<tr>
<td>this paper</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

In experiment 2, there are three methods: 1) TFIDF method; 2) the method of TFIDF and location dataset; 3) the method in this paper. Table 2 shows the results of the second data group based on the above three methods.

From table 2, compared with TFIDF method, the Avg_Precision and the Avg_Recall of the method that integrates TFIDF and location dataset obviously increase. TFIDF method mainly focuses on the contribution of the frequently used words to the theme of the article. And the location dataset added mainly focuses on the importance of the words in special locations. And the method in this paper
perfectly explains the differences between “important words” and “keywords”, which enables the keywords to better reflect the theme of the article.

CONCLUSIONS

Based on chi-squared statistic, this paper calculates the tightness of the lexical items and the theme in the article to be extracted, which is verified in Tibetan news web pages. The experimental results show that the method in this article is effective. And the best value of the lexical item Hp for comparing is also calculated from the experiment. Nevertheless, the standard of the data for Tibetan tests is to be improved. And the number of the data needs increasing. What’s more, the reflection ability in the method in this text needs more experiments and methods, like word clustering method.

ACKNOWLEDGMENT

This research is supported by National Science-technology Support Plan Projects (NO. 2015BAD29B01). This research is supported by the Fundamental Research Funds for the Central University(NO.31920150031)

REFERENCES