Advances in Social Science, Education and Humanities Research, volume 448
Proceedings of the 2020 International Conference on Advanced Education, Management and Social Science (AEMSS2020)

Text Mining of Network Public Opinion Based on Link Template

Tianzhi Wang*

Department of Mathematics, Yunnan Normal University, Kunming, P.R. China
**Corresponding author*

ABSTRACT
Collecting information is the basis of network public opinion analysis, judgment and developing countermeasure. How to improve the efficiency and accuracy of retrieval is an important problem. This paper expounds the selection of search words from the forms of synonym, antonym, hypernym and hyponym, fallible form of retrieval words; the web link filtering by analyzing from the structure of web pages; extraction the webpage text mainly from the template learning analysis; the text part filtering by analyzing the frequency of search words, the relevancy of web page theme and relevancy of URL theme. The research results improve the efficiency and accuracy of text mining.
Keywords: *Subject term, Link template, Network public opinion, Text mining*

1. INTRODUCTION

With the development of society and technology, the Internet has been deeply into daily life and work. Network public opinion has formed a strong force, has been more and more attention. How to accurately identify and monitor network public opinion information and conduct effective guidance has important practical significance for maintaining social stability and creating a good public opinion atmosphere for national social development. Research on network public opinion mining is also very active. Web page text is an important form of public opinion, web crawler is an important technical means and tools of text mining. Many web crawlers use theme-based link filtering algorithm [1-2], ignoring the internal structure characteristics of website links. Domain name link filtering algorithm [3] is mainly based on the structural characteristics of web address, supplemented by the link filtering algorithm of topic, to clear pages that do not belong to the website or are not related to the topic. Theme acquisition tools oriented to intelligence acquisition analyze and extract urls, construct web page learning templates, and then extract and process the text [4]. The main work of network public opinion mining includes selecting search words, filtering URL, constructing web page information extraction template, extracting web page body and so on. The process of web text mining includes web document collection, preprocessing, constructing and reducing feature set, knowledge extraction and evaluation.

2. SEARCH TERM SELECTION

In information retrieval, the information of the same subject is often scattered under multiple words. There are four types of search words: synonym, antonym, anagram and error-prone form [5]. To improve recall rate in information retrieval, it is necessary to consider the full name, abbreviation, old name, etc. Synonyms mainly include: scientific name and common name; Full names and abbreviations, such as "National Southwest Associated University" being referred to as "Southwest Associated University"; New and old names; foreign languages and their abbreviations; different translations. In addition, different regions have different titles for the same thing, and the same person may have many different titles, such as Li Bai whose courtesy name was Taibai, was also known as the Hermit of Green Lotus. Antonym phenomenon, some antonyms present the same problem from the opposite side. The phenomenon of hypernym and hyponym, for the current subject to analyze its position in the subject knowledge, as well as the relationship between concepts, if not in-depth analysis will greatly affect the recall rate of information retrieval. The phenomenon of error-prone form of search words, no matter in publications or electronic resources, is widespread, but in network information resources, error-prone words are emerging in an endless stream. Even if *Chinese Science and Technology Journal Database* is a database retrieval system, there are a lot of typos. The existence of wrong words is an obstacle to improve the recall rate of information retrieval.

Since then, classified retrieval language has been widely used as a normative reference book in academic circles, especially *Chinese Library Classification* (the fifth edition, *National Library Press*, August 2010). Aiming at the theme of higher education, reference is made to "Chinese library classification code >> culture, science, education, sports >>education >> higher education". We select its catalogue as the collection of subject words. Interested readers may contact the author. From the perspective of the subject words of the classification of pictures in the middle, there are still some imperfections that need to be supplemented, such as colleges and universities in Yunnan Province, Chenggong University Town, holidays and other

subject words closely related to the public opinion of Yunnan colleges and universities. Take higher education of Yunnan Province as an example, the subject words are added: (1) names, abbreviations and nicknames of 86 colleges and universities in Yunnan Province; (2) the place names of colleges and universities and their abbreviations and nicknames; (3) other words related to colleges and universities in Yunnan Province.

3. WEBSITE LINK FILTERING

Web pages often includs information such as navigation, text, ads, posters, lace news, copyright and related links. In the process of fetching if directly save the entire web page, it will reduce the accuracy of the retrieval results, increase the storage space and running time web crawling. In order to make the web information as accurate as possible, first filter was carried out on the web link.

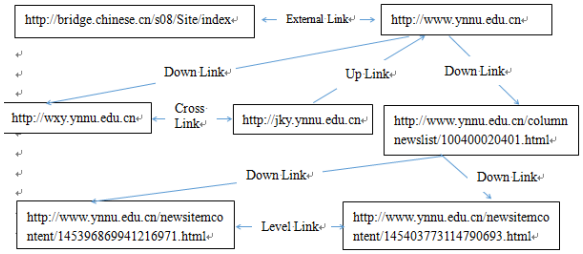

Figure 1 Website link structure

There is no uniform way to generate website URL, and the form of URL provided by each website is not the same, but most portal websites usually create pages according to a certain directory structure, supplemented by fixed format URL. Website links can be divided into up link, down link, Level link, external link and cross link from the structure, see figure 1. The parent page of the current page is called the up link: the URL is the parent directory of the current page or the parent domain of the current domain name. A subordinate page of the current page is called a downlink: the URL is a subdirectory or subdomain of the current page, a page that provides a more detailed list or body. Pages in the same directory as the current page are called level links. With the current page in the site is not the same as the page of the site called external link, and the user needs the content is not associated, with advertising or friendship links in the majority. A page that is not in the directory where the current page is located, but has the same path depth as the current page, is called a cross link, and if the current page is a content page, then the cross link is most likely also a content page. In addition, there are frame chains and dark web chains in web links. Frame chain generally exists in the form of <iframe></iframe>, which is easy to identify. News dynamic websites seldom use such links. The dark network chain often appears in the form <form></form>, which requires the retrieval of the search term set provided by the user to determine whether to collect it or not.

The goal of information collection is to grab pages with body content and remove list or index pages. Users usually specify the main site address of the collection target site. Based on the above analysis, down link, level link and cross link are the main retrieval targets of web pages. The home page usually contains several types of links, such as navigation, content, advertising and friendship links, etc. The link of the column page and the link of the body page are both down links relative to the home page, which need to be retrieved. Column page usually contains navigation, body, advertisement and other contents. The link belongs to level link or cross link, while the link of the body belongs to down link, which needs to be retained after URL re-judgment. In the body page, in addition to the up links to the home page and column pages, there are also level and cross links to other content pages. These level and cross links need to be retrieved, while the up links need to be re-examined. The accuracy of extracting from the link rules is generally not too high, the text in the link still needs to be judged, to determine whether it is the body page or the list page, to determine whether to extract.

4. PAGE BODY EXTRACTION BASED ON TEMPLATE

After filtering and extracting the URL, the body page can be downloaded directly. However, in addition to the body information, these pages also contain a large number of advertisements, column links, scripts, page style and other irrelevant content, which will lead to low retrieval efficiency and inaccurate results, which should be removed. Most web sites use one or more sets of page templates to generate dynamic or static web pages, where the body content is pulled from a database and displayed differently. The body page usually has four sections: the top of the menu including the login area, the side of the page including links to ads and other recommended information, the body of the page that needs to be extracted, the bottom of the page including the link to the site map and the copyright notice. Through the learning of templates, it generates one or a series of templates for extracting the body of target websites, and improves the efficiency and accuracy of webpage retrieval.

4.1. Web Page Extraction Preprocessing

The preprocessing of web page extraction includes two processes: coding conversion and noise information filtering. Common webpage character encoding methods include: Unicode, ASCII, extended ASCII, GBK/GB2312, GB18030, utf-8, etc. When visiting a webpage, it needs to be converted into the default encoding, otherwise there will be garble codes when reading. Tag content unrelated to body content in web pages is noise in information retrieval, which needs to be filtered and filtered to improve the efficiency of template learning. Noise information mainly includes <!---> <meta><link>< style><hr><form><img> <bgsound><:hover><:visited> and other tags and content.

4.2. Template Learning

HtmlAgilityPack and other toolkits are used to convert the pre-processed HTML source code into a DOM tree. Through traversing the DOM tree, the sample webpage is compared to get the extraction template of the text. These information nodes are relatively stable in the HTML presentation, forming templates, while the body content varies from page to page. There are static nodes, dynamic nodes and mixed nodes. The same part of the two sample web pages is the static node content, the dynamic node content contains the body content, and the mixed node needs further analysis. As shown in figure 2.

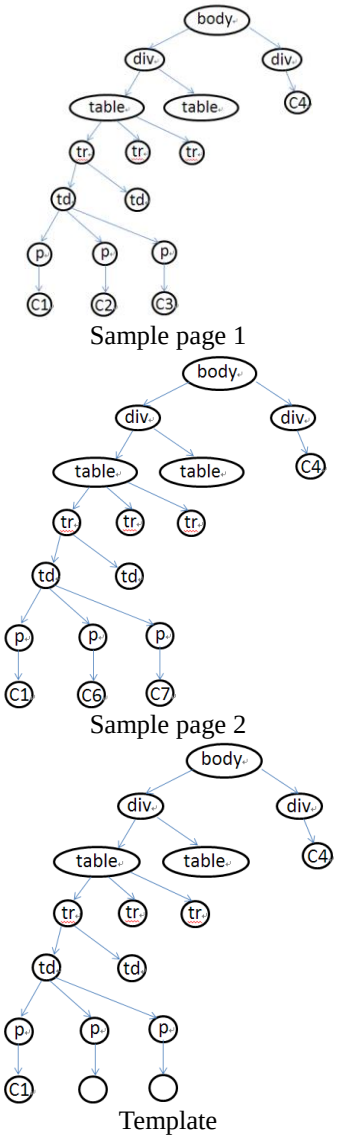

Figure 2 Page template learning

Note: the first div of Sample page 1 is denoted as div_{11}, and the second one is denoted as div_{12}. The first div of Sample page 2 is div_{21} and the second div_{22}.The remaining labels (table, tr, td, P) are marked as follows.

The body node in Sample page 1 and Sample page 2 in the figure is the parent node and of type mix. The (div_{11}, div_{21}) node is the parent, of type mix, and (div_{12}, div_{22}) has the same leaf C_4, so it is static. ($table_{11}$, $table_{21}$) is the parent node, designated as mix; ($table_{12}$, $table_{22}$) node is leaf node, and is static type node. Similarly, (tr_{11}, tr_{21}) (td_{11}, td_{21}) are all mixed nodes. (P_{11}, P_{21}) is static node with the same tag name, attribute and text content. Nodes (P_{12}, P_{22}) and (P_{13}, P_{23}) have the same tags and attributes, but different text contents, and are of dynamic type. It is not accurate to label all parent nodes with children as mix-type nodes. Traverse the children of a node of type mix, which is static if they and their children are of static type. This results in an extraction template based on the content of the target site.

4.3. Main Body Extraction

The process of text extraction is similar to that of template generation. The steps are: generate DOM tree according to HTML; Extract the body content from a static type node by matching the DOM tree of the template with the DOM tree to be processed. In order to improve the differentiation of different types of pages, a comprehensive industry portal usually uses multiple sets of templates to publish different types of information. The page to be processed needs to use a new template to extract the body, and the page needs to learn from the template and extract the body after generating the new template. The text extraction method based on natural language processing has higher accuracy but higher running time and computer resource consumption. In the template-based learning method, only the peer nodes in the DOM tree are compared. If the height of the DOM tree is H and the maximum number of nodes in each layer is N, then the maximum number of computation times is only H^N times. If different nodes are encountered in the process of comparison, the node will be marked as static type node directly, without the need to compare its children. Therefore, the actual calculation times are much less than H^N times, and the speed is relatively faster than the natural language-based method.

5. TEXT FILTERING

In URL extraction, all urls that match the user's retrieval terms are put into the queue to be collected. The matching rule is: if the URL points to the text in the page (including body text, anchor text, advertising text, etc.) and matches any word in the retrieval word set, it means that the webpage meets the collection requirements and can be collected. In addition to the research from the website address template, we also need to focus on the search term. The search term may appear in anchor text or advertising text instead of the body, and even if it appears in the body, it may be a skip, which not only increases the search

burden, but also reduces the efficiency. Therefore, the collected web pages need to be filtered again before being stored and indexed to ensure the accuracy of the collected information.

5.1. Text Filtering Based on Word Frequency Statistics

Generally speaking, the more frequently a word appears in the text, the closer it is to the theme of the article. But the length of the text is not the same, the short web page information may only be dozens of words, and the longer may have thousands of words, the use of word frequency statistics is very likely to filter out the more important pages. Therefore, the matching degree should consider the proportion of the total number of search terms in the body. In addition, the user assigns weight to each search term, representing the degree of importance. The process is as follows:

i Remove stop words. ICTCLAS Chinese Word Segmentation tool [6] is used for processing, to remove meaningless function words such as "of" and "about", form the text vocabulary set T, and calculate the total number of text vocabulary as Count(T).

ii Count frequency. Count(t_i) is denoted as the frequency of search words appearing in the paper.

iii Look up. If the user-provided search term t_i does not appear in the body vocabulary set T, then: t_i is not in the thesaurus of the word segmentation tool, so it cannot be recognized. There is no t_i in the text itself. For the first case, match again by string alignment. If Count(t_i) cannot be obtained, it means there is no t_i in the body, and t_i scores 0.

iv Filtering the body. Web page scores are calculated, and pages above a certain threshold are allowed for download and further processing.

$$Score(p) = \sum_{i=1}^{n} \frac{Count(t_i)}{Count(T)} * r_{ti}. \quad (1)$$

Where, Count(t_i) represents the word frequency of the i-th retrieval word, Count(T) represents the total number of words in the page body, and r_{ti} represents the weight set by the user of the i-th retrieval word.

The results of body filtering will be saved together with the downloaded page and page URL and other information for retrieval and browsing to complete the theme information collection process.

5.2. Filtering Based on Web Topic Relevance

Filtering based on web topic relevance is used to assess the degree to which web pages, urls, and keywords are relevant to the topic. The larger the value, the more likely the content of the page or URL is to be relevant to the user's desired topic. The relevancy of web theme is expressed by the function γ_ω:

$$\gamma_\omega(d) = \begin{cases} 0 & \text{if } \exists_{x\in E}\ x \in d \vee \forall_{x\in I} x \notin d \\ \frac{|\{x|x\in I, x\in d\}|}{|I|} \times \alpha + \cos(C, \vec{d}) \times (1-\alpha) & \text{else} \end{cases} \quad (2)$$

Where, d is a web page, $x \in d$ indicates that x belongs to the keyword set d; α is a user-configured coefficient used to adjust the importance of set I and vector C in the calculation of topic relevancy, and the value range is (0,1]; $\vec{d}$ for page d keyword set characteristic vector and the characteristic value of each keywords for the TF - IDF in d value; the cosine function represents the similarity between two eigenvectors.

5.3. Filtering Based on URL Topic Relevancy

If a web page is theme-related, the more similar the URL, anchor text and the content of the web page body are, the more likely it is to be theme-related. The content similarity between URL anchor text and web page is introduced to calculate the theme relevance of URL, so as to make up for imperfect theme knowledge and small information content of anchor text. Because the URL contains less information, some theme-related keywords in the URL anchor text have not been extended into the theme knowledge when the theme knowledge is not perfect, and the theme relevance value of the URL anchor text is very low. In other words, if the above formula is used to calculate the topic relevancy of URL anchor text, most of the relevancy obtained is 0. Therefore, using the theme relevancy of anchor text as the theme relevancy of URL to filter and sort URL will cause a lot of omissions. URL theme relevance function is:

$$\gamma_a(u) = \gamma_\omega\big(d(u)\big) \times \{\gamma_\omega\big(t(u)\big) + \cos(\overrightarrow{t(u)}, \overrightarrow{d(u)})\}. \quad (3)$$

Among them, u is a URL, $d(u)$ is the web page where u is located, $t(u)$ is the anchor text of u, $\overrightarrow{t(u)}$ is the eigenvector of anchor text of u, $\overrightarrow{d(u)}$ is the feature vector of the body of the page where u is located.

Through the search term selection, website link filtering, template-based learning web body extraction, text filtering analysis and research, it greatly improves the efficiency and accuracy of network public opinion text mining.

6. CONCLUSIONS

Through the analysis and study of synonyms, antonyms, upper and lower bits and the error-prone forms of search words, it enlarges the extension of the search words, reduces the omissions, and improves the reliability and accuracy of search. Through the analysis and research of the web page structure of the website links, such as the downlink chain, the uplink chain, the horizontal chain and the cross chain, reducing the noise of the complex web information, the "template learning" of web page extraction is constructed, which reduces the scope of web

page retrieval, makes it accurate, and improves the accuracy and efficiency of web page retrieval. Through the above process, the text of the web page is segmented into words, the relevance of the topic of the web page is calculated and filtered to complete the retrieval task.

REFERENCES

[1] Yuzhao Bai, Jiuzhen Liang, Research and implementation for focused crawler based on probabilistic model, J. Computer Engineering and Science, 35(1)(2013) 160-165.

[2] Alberti B, Anklesaria F, Lindner P, et a1. The internet gopherprotocol: a distributed document search and retrieval protocol, J. The Journal of Universal Computer Science, 24(2) (1991) 235-246.

[3] Yang Wen, Weny Chen u, Ye Yuan, et al, Domain link filtering algorithm for crawlers, J. Library and information work, 58(20) (2014) 125-127.

[4] Jun Gu, Jia Weng, Xin Xu, Design and implementation of theme acquisition tool for information acquisition, J. Library and information work, 58(20) (2014) 91-99.

[5] Yanhong Shen, Influence of search term selection on recall rate in information retrieval, J. Information exploration, 11 (2006) 73-74.

[6] NLPIR Chinese Word Segmentation System (aka ICTCLAS2016) [EB/OL]. [2014-06-18]. http://ictclas.nlpir.org/.