ISSN: 1813-4890

CLTM: An Innovative Method of Chinese Legal Text Mining

Xiao Li[1,*], Dong Sui[2], Yongtang Bao[3], Yangwu Zhang[1], Jing Li[1]

[1]School of Information Management for Law, China University of Political Science and Law, Beijing 102249, China

[2]School of Electrical and Information Engineering, Beijing University of Civil Engineering and Architecture, Beijing, China

[3]College of Computer Science and Engineering, Shandong University of Science and Technology, Qingdao China, 266590, China

*phd_li@yeah.net

Abstract

This paper puts forward an innovative method of Chinese legal text mining, which is abbreviated as CLTM. This method contains two main functional components. The first part is Chinese legal text segmentation based on innovative mode, the second part is Chinese legal text mining based on innovative TF-IDF. Part I includes word segmentation method and UDT (User defined Thesaurus) of CLTM. The first part uses an innovative and improved method to segment Chinese legal texts, which greatly improves the accuracy of Chinese legal text segmentation. The second part uses innovative TF-IDF to implement Chinese legal text mining, which greatly improves the accuracy and robustness of Chinese legal text mining.

Keywords

Chinese Legal; Text Segmentation; Text Mining; User Defined Thesaurus; Innovative TF-IDF.

1. Introduction

Text mainly refers to the information structure composed of certain symbols or codes, which can be expressed in different forms, such as language, text, image and so on. The text is made by a specific person, and its semantics will inevitably reflect the ideological content of people's specific positions, views, values and interests.

Text analysis[1][2][3] means to go deep into the depth of the text from the surface of the text, so as to find those deep meanings that can not be grasped by ordinary reading. As a powerful research method to explore the nature of information content, text analysis is also one of the common methods used by cultural researchers. In the research, the text analysis method studies the text more from the aspects of rhetoric and narration, so as to grasp the deep meaning of the text from the outside to the inside.

The theoretical resources of text mining[4][5][6] come from Hermeneutics and humanism, and there are several different research orientations, such as the "New Criticism" method originated from British and American literary criticism, the semiotic analysis method represented by Roland Barthes, the narrative analysis method[7] focusing on story analysis and narrative perspective analysis[8], intertextuality, dialogue theory analysis method taking into account the macro social environment and micro text deconstruction[9][10], Derrida's deconstruction The research methods of Text Sociology and the study of British culture.

ISSN: 1813-4890

Text analysis is like "paoding jieniu". Read the article word by word, and carefully interpret its meaning. From its theoretical resources, it can be seen that text analysis is more suitable for the study of literature and narrative.

2. CLTM Method

2.1 Chinese Legal Text Segmentation based on Innovative Mode

Heidegger said that 'where words are broken, nothing can exist'. Unlike English, Chinese sentences do not have spaces between words, which blurs the boundaries of words and phrases. In order to make the computer understand the text more easily, the first step of Chinese information processing is Chinese word segmentation[11]. Chinese word segmentation is to add boundary marks between words in Chinese sentences. This paper uses innovative approach to segment Chinese legal text based on innovative mode.

2.1.1 Word Segmentation Method

The segment component of CLTM provides four word segmentation modes. It can initialize the word segmentation engine and use the method for word segmentation. When loading the word segmentation engine, you can customize the thesaurus path and start different engines.

The maximum probability method (mpsegment), which is responsible for constructing directed acyclic graph and dynamic programming algorithm according to trie tree, is the core of word segmentation algorithm.

The hidden Markov model (hmm-segment) is based on the HMM model built on the people's daily and other corpora to segment words. The main algorithm idea is to represent the hidden state of each word according to the four states (B, e, m, s). HMM model consists of dict/hmm_ model Utf8. The word segmentation algorithm is Viterbi algorithm.

Mixsegment is one of the four word segmentation engines with better word segmentation results. It combines the maximum probability method and the implicit Markov model.

Querysegment first uses a hybrid model to segment words, and then enumerates all possible words in the sentence for the longer words to find out the words in the thesaurus.

The word frequency of words in the user dictionary, which defaults to "Max", the maximum value in the system dictionary. You can also select the minimum value of "Min" or the median value of "median". Stop thesaurus used for keyword extraction. It can also be used in word segmentation, but the corresponding path used in word segmentation cannot be the default stoppath.

Whether to write the file word segmentation results to the file. The default value is No. This parameter is used only when the input content is a file path. This parameter is only valid for word segmentation and part of speech tagging. The maximum index length is the maximum number of possible words in the index model. Whether to check the code of the input file. The method checks by default.

2.1.2 UDT (User Defined Thesaurus) of CLTM

This method customizes the user thesaurus. It builds a word segmentation thesaurus based on the transformation of the deep blue thesaurus. It can quickly convert the Sogou cell thesaurus and other input method thesaurus into the thesaurus format in this method.

UDT is an input method thesaurus conversion software, which supports more than 20 input method tools and Thesaurus. It includes Chinese pyim (Linux), fit input method (MAC), libpinyin (Linux), MacOS comes with simplified Pinyin, QQ Pinyin (text thesaurus and qpyd format classification Thesaurus), QQ Wubi, rime input method (Linux zhongzhouyun, windows xiaolanghao, Mac OS moustache), win10 Microsoft Pinyin, win10 Microsoft Wubi, baidu Pinyin PC (text thesaurus, bdict format), Bing input method, Cangjie platform, Google Pinyin, pole

ISSN: 1813-4890

Wubi, pole zhengma, Lingus thesaurus LD2, Pinyin plus, palm input method, Sogou Pinyin (text thesaurus, bin format backup thesaurus and scel format cell Thesaurus), Sogou Wubi, Microsoft Pinyin 2010, Xiaoxiao input method (pinyin, Wubi, Zheng code and Erpi), Xiaoya Wubi, Sina Pinyin, Yahoo Qimo input method (phonetic alphabet), Ziguang Pinyin (text thesaurus and UWL format classification Thesaurus), custom format. This greatly improves the word segmentation accuracy of the system

2.2 Chinese Legal Text Mining based on Innovative TF-IDF

TF-IDF (term frequency – inverse document frequency) is a commonly used weighting technique for information retrieval and data mining. It is often used to mine keywords in articles. The algorithm is simple and efficient. It is often used in the industry for the first text data cleaning. TF-IDF has two meanings. One is term frequency (abbreviated as TF) and the other is inverse document frequency (abbreviated as IDF). When there are TF (word frequency) and IDF (inverse document frequency), multiply the two words to get the TF-IDF value of a word. The larger the TF-IDF of a word in the article, the higher the importance of the word in the article. Therefore, by calculating the TF-IDF of each word in the article, the top words are the keywords of the article. For word w of document d, calculate its TF value:

$$tf_{w,d} = \frac{n_{w,d}}{\Sigma_k n_{k,d}}. \tag{1}$$

$n_{w,d}$ indicates the number of times a word appears in the passage. $\Sigma_k n_{k,d}$ represents the total number of articles. At this time, a corpus is needed to simulate the language environment. Idfi is as follows:

$$idf_w = log\frac{N_1}{N_2+1}. \tag{2}$$

N_1 represents the total number of documents in the corpus. N_2 represents the number of documents containing the word w. If a word is more common, the denominator is larger, and the frequency of inverse document is smaller and closer to 0. The reason for adding 1 to the denominator is to avoid the denominator being 0 (that is, all documents do not contain this word). Log means to take the logarithm of the obtained value. It can be seen that TF-IDF is directly proportional to the number of occurrences of a word in the document and inversely proportional to the number of occurrences of the word in the whole language. Therefore, the algorithm for automatically extracting keywords is very clear. It is to calculate the TF-IDF value of each word of the document, and then arrange it in descending order, taking the first few words.

IDF is calculated in the document set, but the distribution of different classes in the document set is uneven. For example, 100 food articles and 1000 lipstick articles form a document collection. According to the definition of traditional IDF, "food" IDF will be larger than "lipstick" because there are many fewer food articles. But in fact, it is no more important than "lipstick", but there are too many articles containing "lipstick". Based on this, this paper proposes an innovative scheme:

$$idf_w = log\frac{N_1}{N_2+1}\ log\frac{p_w+0.01}{P_w+0.01}. \tag{3}$$

p_w indicates the frequency of the feature word w in the current category, P_w indicates the frequency of the feature word w in other categories. The IDF is calculated as it can also be seen

that the ratio of the frequency of occurrence in the current class to that in other classes is used to measure the importance of a word. 0.01 is to prevent it from P_w being 0.

3. Conclusion

Based on the characteristics of Chinese legal text, this paper implements an innovative Chinese legal text mining method based on this scheme. Based on the weakness of TF-IDF, according to the characteristics of Chinese legal text, an innovative algorithm is proposed. The specific improvement strategies are as follows. This article is looking for category keywords. Obviously, a word appears a lot in this category and less in other categories. This article wants to find the keywords of different contents under the same category. Obviously, it should focus on the less words under the category. For example, "eating" is very important in distinguishing between food and makeup, but if it is all about food, "eating" is meaningless. Its core idea is as follows. Specify the basis for dividing the content set, and then select the most important factors to weight the keywords. The unsupervised neural network method will be tried later. IDF is also suitable for exposing keywords when extracting content keywords unsupervised. Tagging is done based on a given keyword. Basically, when a keyword appears, you can consider tagging. As for accuracy, it depends on judging the context in which the keyword appears and whether the context is also very consistent. It is very consistent with the description that this content is attached with enough keyword related descriptions, rather than just mentioning a word.

Acknowledgments

This work was supported by National Key R&D Program of China (No. 1171-23318102) and Ministry of Education Cooperative Education Project (No. 202002237006).

References

[1] Tingting Bi, Peng Liang, Antony Tang, Chen Yang. A systematic mapping study on text analysis techniques in software architecture, Journal of Systems and Software, Volume 144, 2018, Pages 533-558, ISSN 0164-1212, https://doi.org/10.1016/j.jss.2018.07.055.

[2] Manyu Li, Application of sentence-level text analysis: The role of emotion in an experimental learning intervention, Journal of Experimental Social Psychology, Volume 99, 2022, 104278, ISSN 0022-1031, https://doi.org/10.1016/j.jesp.2021.104278.

[3] Yasmine Elkhayat, Mohamed Marzouk, Selecting feasible standard form of construction contracts using text analysis, Advanced Engineering Informatics, Volume 52, 2022, 101569, ISSN 1474-0346, https://doi.org/10.1016/j.aei.2022.101569.

[4] Miriam Alzate, Marta Arce-Urriza, Javier Cebollada, Mining the text of online consumer reviews to analyze brand image and brand positioning, Journal of Retailing and Consumer Services, Volume 67, 2022, 102989, ISSN 0969-6989, https://doi.org/10.1016/j.jretconser.2022.102989.

[5] Dominic D. Martinelli, Evolution of Alzheimer's disease research from a health-tech perspective: Insights from text mining, International Journal of Information Management Data Insights, Volume 2, Issue 2, 2022, 100089, ISSN 2667-0968, https://doi.org/10.1016/j.jjimei.2022.100089.

[6] Jonathan Benchimol, Sophia Kazinnik, Yossi Saadon, Text mining methodologies with R: An application to central bank texts, Machine Learning with Applications, Volume 8, 2022, 100286, ISSN 2666-8270, https://doi.org/10.1016/j.mlwa.2022.100286.

[7] Li X, Li S, Qin H, et al. Spatiotemporal consistency-based adaptive hand-held video stabilization. Sci China Inf Sci, 2020, 63(1): 114101, https://doi.org/10.1007/s11432-018-9764-0.

[8] Sal Consoli, Uncovering the hidden face of narrative analysis: A reflexive perspective through MAXQDA, System, Volume 102, 2021, 102611, ISSN 0346-251X, https:// doi.org/ 10.1016/ j.system. 2021. 102611.

[9] Shuheng Du, Profound connotations of parameters on the geometric anisotropy of pores in which oil store and flow: A new detailed case study which aimed to dissect, conclude and improve the theoretical meaning and practicability of "Umbrella Deconstruction" method furtherly, Energy, Volume 211, 2020, 118630, ISSN 0360-5442, https://doi.org/10.1016/j.energy.2020.118630.

[10] Dong-shin Shin, Tony Cimasko, Youngjoo Yi, Development of metalanguage for multimodal composing: A case study of an L2 writer's design of multimedia texts, Journal of Second Language Writing, Volume 47, 2020, 100714, ISSN 1060-3743, https://doi.org/10.1016/j.jslw.2020.100714.

[11] Ling Zhao, Ailian Zhang, Ying Liu, Hao Fei. Encoding multi-granularity structural information for joint Chinese word segmentation and POS tagging, Pattern Recognition Letters, Volume 138, 2020, Pages 163-169, ISSN 0167-8655, https://doi.org/10.1016/j.patrec.2020.07.017.