Legal Knowledge and Information Systems
M. Palmirani (Ed.)

doi:10.3233/978-1-61499-935-5-215

K-Means Clustering for Controversial Issues Merging in Chinese Legal Texts

Xin Tian [a], Yin Fang [a], Yang Weng [a], Yawen Luo [b], Huifang Cheng [c] and Zhu Wang [b,1]
[a] *College of Mathematics, Sichuan University*
[b] *Law School, Sichuan University*
[c] *China Justice Big Data Institute*

Abstract. In the fact of growing number of cases, Chinese courts have gradually formed a trial mode to improve the efficiency of trials by conducting trials around the controversial issues. However, identifying the controversy issue in specific cases is not only affected by the uncertainty of facts and laws, but also by the discretion of the judges and extra-case factors, and cannot be expressed as a standard format, which lead to the controversial issues based case retrieval a challenge problem. In this paper, we propose a controversial issues merging algorithm based on K-means clustering for Chinese legal texts. The proposed algorithm can determine the number of clusters of the given cause of action automatically and merge the controversial issues semantically, which makes the case information retrieval more accurate and effective.

Keywords. information retrieval, K-means clustering, controversial issues

1. Introduction

Over the past 20 years, the number of cases accepted by the people's courts has increased rapidly. From 2013 to 2017, more than 89 million cases were accepted by Chinese people's courts at all levels. In order to cope with the high-speed growth of cases, Chinese courts have gradually formed a trial mode of launching trials around controversial issues to improve the efficiency of trials, and as embodied in judgments, it is mainly to develop reasoning around controversial issues in the reasoning part of the courts [1]. In Chinese adjudicative documents, controversial issues mainly exist in civil and commercial cases, and are also common in administrative cases, but seldom exist in criminal cases.

According to the content of disputes involved, controversial issue can be generally divided into factual controversial issue and legal controversial issue. In terms of trial organization, factual controversial issue can help to focus on fact investigation, while legal controversial issue can help to organize court debate. Both play a role in improving the efficiency of court trial. The controversial issues sorted out, organized to investigate, and debated by the courts during trials will be embodied in the judgment, and become the contents which could mostly restore the scene of court trials and the judgment thoughts of the judges in the judgment.

[1] Corresponding Author: wangzhu@scu.edu.cn.

It should also be noted that judges' summary about controversial issues is unformatted. As a verbal interaction aiming at multiple parties, the summary of controversial issues reflects the judges' skill in using laws and court trial rules to ascertain the facts of a case. The determination of controversial issues in individual cases is influenced not only by the uncertainty of facts and laws, but also by the discretion of administrative judges and extrajudicial factors, so it is impossible to express controversial issues in a precise format.

Due to the limited access to cases by judges, it is very difficult for the judges to draw lessons from the experience of other judges in summarizing and discussing controversial issues, except from the cases tried or discussed by themselves, and this has greatly hindered the accumulation of legal knowledge and the dissemination of judges' experience. It is difficult to retrieve unformatted controversial issues with key words, so the combination of homogeneous controversial issues has become the basis for judges to retrieve similar controversial issues [2].

As a matter of fact, the controversial issues are finite for a case in same cause of action, but it is difficult to distinguish the similar controversial issues due to huge corpus and different expressions. Therefore, we need machine learning algorithm for identifying the similar controversial issues in a large legal corpus. Most legal information is expressed in text, such as facts of the case, laws and rules, etc. Firstly, we should transform this semantic information into vector space. Several approaches are utilized for the semantic vectorization, such as *term frequency-inverse document frequency* (TF-IDF) [3], *Latent semantic analysis* (LSA) [4], Word2Vec and Doc2Vec. Unsupervised learning is one of the machine learning task of inferring a function that describes the structure of unlabeled data, and the clustering is a form of unsupervised learning that classifies data into different classes or clusters automatically.

In this paper, we extract controversial issues from the Chinese adjudicative documents of personality right disputes in 2017 and introduce four classes of controversial issues including repeated cause of action, general procedure law, general substantive laws and non-general substantive laws controversial issue or factual controversial issue group. The first three controversial issues were extracted by regular expression, then the last one was extracted by machine learning. Experiments show that semantic-based methods capture the semantic information in the text, whose clustering precision is higher than others.

2. Controversial Issues Overview

The courts divide controversial issues into factual controversial issues and legal controversial issues during court trials, mainly aiming to ascertain the facts first, and then proceed to legal reasoning. But from the perspective of referential property to other judges, some legal controversial issues may not have reference significance, and they are of limited types and general-purpose, so they may be sorted out first artificially. However, the combination of controversial issues based on machine learning mainly aims at the non-general substantive laws controversial issues which are not sorted out in advance and the factual controversial issues. We divide controversial issues in judgments into four types:

First type: Controversial issue group of repeated cause of action (G1). Such controversial issues are featured by that, upon the request of the parties concerned, the judges

consider that the issues with the nature of controversial issue are actually the causes of action involved in the cases. For example, in the disputes over portrait right, the expression Does the defendant infringe upon the plaintiffs portrait right? is clearly directed at the cause of action. Similar controversial issues may be combined directly.

Second type: Controversial issue group of general procedure law (G2). Such controversial issues are featured by that, in different causes of action, there will probably be similar procedural controversial issue group. For example, for such controversial issues of Is the plaintiff a competent subject?, there must be clear regulations in the Civil Procedure Law, so it is available to sort out the controversial issues by type first before technical combination.

Third type: Controversial issue group of general substantive laws (G3). Such controversial issues are featured by that, the judges make value judgment on whether minor premise (facts of a case) meets major premise (legal provisions) according to the clearly expressed provisions of law. For example, for such cause of action, like Is the amount of compensation claimed by the plaintiff reasonable?, there must be clear regulations in substantive laws, which widely exist in different causes of action, so it is worth sorting out the causes of action before technical combination.

Fourth type: non-general substantive laws controversial issue and factual controversial issue group (G4). Different causes of action involve different non-general substantive laws controversial issue and factual controversial issue groups. Wherein, the level-four causes of actions under the same level-three cause of action are possibly repeated with the factual controversial issue group of the level-three cause of action, and have relatively great retrieval value for similar cases. Non-general substantive laws controversial issues have relatively great reference significance. Such causes of action are mainly sorted out by relying on machine learning.

3. Feature Extraction and K-means

Feature extraction aims to transfer text to a vector which represents the feature of the text. There are many methods which can capture different features of text, such as frequency-based (TF-IDF, LSA) and network-based (Word2Vec, Doc2Vec). In fact, the network-based methods are more appropriate because the combination of controversial issues are based on semantics and the network-based methods can capture semantic feature exactly. For instance, Word2Vec expresses the semantic information of words by learning huge corpus and makes the embedding vectors of similar words more compact as well [5]. Doc2Vec is an extension of Word2Vec that is designed to get an embedding vector of documents [6]. The difference between them is that Doc2Vec adds the identification information of the document, and it can be regarded as the topic of the document.

K-means is the most important hard clustering algorithm [7]. It aims to assign the data set into K clusters, where the value of K is already known. As we know, K-means is a simple and efficient clustering method, but a big disadvantage of this method is that it needs to determine the number of clusters first. In reality, often K is nothing more than a good guess based on experience or domain knowledge. It is similar in merging controversial issues. In cases of small data sets, we can give an ideal K with the help of some legal experts, but we can not do so in huge data sets. In this part, we will focus on a method for determining an appropriate K.

Schtze et al. introduced a heuristic method in the "*Introduction of Information Retrieval*", which can capture some possible values of K [8]. In more details, we first perform i (e.g. $i = 10$) clusters with a fixed K (Note that you should initialize each one differently) and compute the objective functions of each cluster. Then the minimum of these objective function values can be denoted by $J_{min}(K)$. Now, we implement this process and compute $J_{min}(K)$ with the increase of K. Finally we can capture a series of $J_{min}(K)$ with different K, and find the "knee" in the curve - the point where the successive decreases in J_{min} become noticeably smaller.

4. Experiments

In this part, the experiments are implemented on the adjudicative documents of personality right disputes in 2017 to demonstrate the performance of our approach. We attempt to prove that the feature model can capture semantic information, and meanwhile the clustering algorithm can also assign controversial issues according to our expectations, especially, an appropriate K can be given according to the heuristic method.

Throughout the experiments, we utilize two indicators to measure the degree of conformity between ground truth clusters and our evaluation from algorithm outputs: *adjusted mutual information* (AMI) and *V-measure* [9]. Another measure is V-measure [10]. V-measure is an entropy-based measure which explicitly measures how successfully the criteria of homogeneity and completeness have been satisfied.

The proposed method is implemented on the cause of action of portrait right disputes belongs to disputes over personality right. There are 87 controversial issues in the cause of action of portrait right disputes that were extracted by regular expression and then clustered. In more details, the component we set will be retained at 50 in LSA, which means the dimension of the text feature vector from LSA is 50. We utilized huge corpus to train Word2Vec vector embedding including two parts: Controversial issues of disputes over private lending, disputes over traffic accident liability for motor vehicles and disputes over personality right, on the other hand, adjudicative documents of personality right disputes which includes the facts affirmed by the court, the reasoning of the courts and the result of judgement.

We set the number of K separately according to "*manual*" and "*heuristic*". When deciding about the "*manual*" number of K, we rely on expert opinion. The ideal value of K we choose is 33, which equals to groundtruth. The "*heuristic*" number of K comes from former method are 31, 29, 27, 40 and 36 respectively. The experimental results are shown in Table 1, we see that, as expected, the Word2vec and Doc2vec based on semantic feature get better results, meanwhile the "*heuristic*" results are comparable with the "*manual*". It further proves that the semantic-based method we used can capture the semantic information inside the document, and the heuristic search of K in K-means is applicable to this problem.

5. Conclusions and Future Works

In this paper, the controversial issues in judgments are divided into four types. The non-general substantive laws controversial issues and the factual controversial issues are

Table 1. Comparison of several methods

Method	*AMI*	*V − measure*
LSA_manual	0.4263	0.7759
LSA_heuristic	0.4137	0.7689
Word2Vec (dim=50)_manual	0.4318	0.7926
Word2Vec (dim=50)_heuristic	0.4627	0.7864
Word2Vec (dim=100)_manual	0.4603	0.8097
Word2Vec (dim=100)_heuristic	0.4334	0.7727
Doc2Vec (dim=50)_manual	**0.4936**	**0.8229**
Doc2Vec (dim=50)_heuristic	0.4720	**0.8390**
Doc2Vec (dim=100)_manual	**0.5300**	0.8202
Doc2Vec (dim=100)_heuristic	0.4902	0.8184

merged by the proposed machine learning algorithm. More precisely, we utilize network-based word embedding models such as Word2Vec and Doc2Vec to extract text feature and cluster data with K-means. Finally, the results of experiments demonstrate that the proposed algorithm is more effective than baseline. However, we found that there is a hierarchical structure between different cause of action. Therefore, it will be better to introduce some hierarchical clustering algorithms for the intrinsic hierarchical structure of controversial issues in Chinese legal texts.

Acknowledgements

This work was supported by National Key R&D Program of China (No. 2018YFC0830300).

References

[1] CHEN Gui-ming. (2004). Issues Concerning Several Relations in the Design of Pretrial Preliminary Procedure. *The Political Science and Law Tribure*, 23(4), 10-15.

[2] Hu Ya-qiu. (2012). Arrangement Procedure of Controversial Points in the Evolution of Civil Litigation System. *Journal of Soochow University* , 12(3), 58-67.

[3] Salton, G., & McGill, M. J. (1986). Introduction to modern information retrieval.

[4] Deerwester, S., Dumais, S. T., Furnas, G. W., Landauer, T. K., & Harshman, R. (1990). Indexing by latent semantic analysis. *Journal of the American society for information science*, 41(6), 391-407.

[5] Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

[6] Le, Q., & Mikolov, T. (2014, January). Distributed representations of sentences and documents. *In International Conference on Machine Learning* (pp. 1188-1196).

[7] Lloyd, S. (1982). Least squares quantization in PCM. *IEEE transactions on information theory*, 28(2), 129-137.

[8] Schtze, H., Manning, C. D., & Raghavan, P. (2008). *Introduction to information retrieval* (Vol. 39). Cambridge University Press.

[9] Vinh, N. X., Epps, J., & Bailey, J. (2010). Information theoretic measures for clusterings comparison: Variants, properties, normalization and correction for chance. *Journal of Machine Learning Research*, 11(Oct), 2837-2854.

[10] Rosenberg, A., & Hirschberg, J. (2007). V-measure: A conditional entropy-based external cluster evaluation measure. *In Proceedings of the 2007 joint conference on empirical methods in natural language processing and computational natural language learning (EMNLP-CoNLL).*