Research on Hybrid Automatic Classification of Multi-type Documents Based on CNKI: Post-print

Authors: Li Xiangdong, Liu Kang, Ding Cong, Gao Fan

Date: 2017-10-11T00:00:00+00:00

Abstract

[Purpose] To address issues such as feature mismatch arising from different document types and improve the classification performance of texts to be classified. [Method] Texts from document types different from those of the target texts were used as the training set for the corpus, and a third-party resource, "HowNet," was introduced for semantic feature expansion. [Results] Using this method, classification experiments were conducted on four types of documents: web pages, books, non-academic journals, and academic journals. Compared with classification methods without expansion, the classification accuracy improved by 1.2% to 11.0%. [Limitations] Public corpora were not used for testing for each document type; therefore, the generalizability of the proposed method and the objectivity of the experimental results require further verification. [Conclusion] Experimental results demonstrate that the method possesses certain feasibility and practicality, can eliminate semantic differences between different document types to varying degrees, and enhances the effectiveness of automatic text classification through two approaches: corpus construction and feature expansion.

Full Text

Preamble

ChinaXiv Cooperative Journal, Issue 267, 2016, No. 2

Research on Mixed Automatic Classification of Multiple Document Types Based on HowNet

Li Xiangdong[1],[2], Liu Kang[1], Ding Cong[1], Gao Fan[1]
[1](School of Information Management, Wuhan University, Wuhan 430072, China)

[2](Center for the Studies of Information Resources, Wuhan University, Wuhan 430072, China)

Abstract

[**Objective**] This study addresses the feature mismatch problem caused by different document types and improves the classification effectiveness of target texts. [**Methods**] We use texts from document types different from the target classification texts as the training corpus and introduce the third-party resource HowNet for semantic feature extension. [**Results**] Experiments conducted on four document types—webpages, books, non-academic journals, and academic journals—show that compared with classification methods without feature extension, the classification accuracy improves by 1.2% to 11.0%. [**Limitations**] Not every document type was tested using publicly available corpora, so the generalizability of the proposed method and the objectivity of the experimental results require further verification. [**Conclusions**] The experimental results demonstrate that the proposed method is feasible and practical. It can eliminate semantic differences between different document types to varying degrees and improve text automatic classification performance through both corpus construction and feature extension.

Keywords: Third-party resource; HowNet; Feature extension; Semantic difference
Classification Codes: TP393; G35

Introduction

With the rapid development of the Internet, online information resources are growing exponentially, enabling people to continuously acquire information in various forms such as text, images, audio, and video. Textual information can originate from numerous document types including webpages, books, and academic journal articles, allowing access to information on the same topic with different connotations, quality levels, and publication speeds. Therefore, research on using text classification technology to organize these texts systematically for more efficient categorization and retrieval holds significant practical and applied value.

Automatic text classification involves multiple stages including training set construction, feature selection, and classification algorithms. Traditional automatic classification research typically uses the same document type for both training and target texts. However, studies in information resource management have shown that using different document types for training sets can also improve classification effectiveness for target texts [1-3], making mixed-type classification a promising approach. Nevertheless, these studies have overlooked the fact that different document types exhibit distinct characteristics in word usage and writing style, causing features from the training set and target texts that express the same concept to fail to match properly. Consequently, when using

different document types as training sets, appropriate methods are needed to overcome lexical and semantic differences between training and target texts, thereby increasing shared features and improving classification performance.

This study employs third-party resources to perform feature extension on training sets and target texts from different document types. By expanding the quantity and semantics of features, we increase the likelihood of matching features that express the same concept in both training and target texts, achieving the goal of creating more common features between them. Our experimental data includes scientifically classified, long-accumulated academic literature such as books and journal articles, as well as news-oriented, non-academic texts with frequent updates like webpages and newsletters. We conduct automatic classification experiments using different document types as training sets and test sets respectively, propose a semantic feature extension method based on HowNet [4], and demonstrate through experiments that it can improve classification effectiveness across multiple document types.

2.1 Research Status and Development Trends

Machine learning-based text automatic classification requires algorithms to learn from training sets and apply the acquired knowledge to test set classification. Traditional machine learning classification typically uses the same document type for training and test sets, whereas mixed-type classification can leverage existing or easily obtainable training sets to classify test sets of different document types. This approach draws from transfer learning' s cross-domain classification concept [5]. Cross-domain classification is a frontier topic in machine learning research, with the fundamental premise of classifying training and test sets from different domains. These domains can differ in subject content, product reviews, or even language. Studies [6-7] have used third-party resources like Wikipedia as intermediaries to correlate features between training and test sets from different thematic domains, reducing semantic feature differences caused by thematic variations and constructing feature spaces with more common features to enhance classification effectiveness. This paper applies this cross-domain concept, using HowNet as a third-party resource to increase feature matching possibilities between training and test sets of different document types, representing a cross-document-type or cross-source classification problem.

Short text feature extension has also become a hot topic in recent text classification research, with the core idea of improving classification effectiveness by expanding the number of common features or semantic information between training and test sets. For instance, study [8] used Wikipedia' s related concept sets as feature extension word sets, leveraging conceptual links and category relationships in Wikipedia to extend features in both training and test short texts, thereby improving performance through feature quantity expansion. Similarly, study [9] approached from the perspective of enriching feature semantics, extracting domain high-frequency words as features and extending them into concepts and sememes based on HowNet, then calculating feature similarity

using the information content of shared sememes across different concepts to achieve classification, which also improved effectiveness.

This study applies short text classification feature extension methods to training and test sets composed of different document types, expanding feature quantity and semantics to create more common features between different document types, thereby helping to improve classification effectiveness.

2.2 Research Significance

Machine learning-based text automatic classification is the mainstream approach, with basic processes including corpus construction, text modeling, feature selection, feature extension, and classification algorithm implementation. In artificial intelligence research, the main focus is on all stages except corpus construction. In information management, however, documents are the primary research and application objects, with numerous research achievements on document classification, content features (such as subjects), types, and characteristics. Therefore, when conducting automatic classification research, there is natural emphasis on the textual characteristics of training and test sets as documents in the corpus construction stage, attempting to apply these characteristics to improve classification efficiency. This paper draws on information management research that uses different document types as training and test sets to improve performance, and attempts to further enhance classification effectiveness from the corpus perspective by narrowing semantic differences arising from document type variations.

3.1 Classification Framework Based on HowNet Semantic Feature Extension

To address feature mismatch between training and test sets caused by document type differences, this paper proposes a text classification method with feature extension for different document types, following the basic feature extension principles in literature [10]. The specific classification framework is shown in Figure 1 [Figure 1: see original paper].

Figure 1. Text Classification Framework Based on HowNet Semantic Feature Extension

(1) **Preprocessing**: Perform tokenization, stop-word removal, and other preprocessing on training and test set texts from different document types to obtain initial feature sets for each document.

(2) **Semantic Core Word Set Extraction**: Calculate TF-IDF weights for feature words in the training set and extract words above a certain threshold to form the semantic core word set.

(3) **Feature Extension**: For each preprocessed target text d, calculate semantic similarity between each feature word in d and feature words in the

training set' s semantic core word set using HowNet' s lexical dictionary. Extend feature words with similarity values above a threshold into text d to obtain the extended target text. This enables features with similar semantics in the test set to be extended and matched with features in the training set through HowNet.

(4) **Classification**: Use the KNN algorithm to calculate similarity between the target text and the training set' s semantic core word set, assigning the category with the highest similarity to the target text.

3.2 Acquisition of Training Set Semantic Core Word Set

TF-IDF weighting is widely used for feature weighting in text classification. The main idea is that if a feature term appears frequently in a document but rarely in other documents, it has strong category discrimination capability [11]. Therefore, this paper uses feature words with high TF-IDF values in the training set as semantic core words for feature extension. We calculate TF-IDF values for each feature word in each category of the training set and select words above the threshold as semantic core words. The specific process is as follows:

Input: Training set D, TF-IDF threshold *weight*
Output: Training set semantic core words

1. Perform part-of-speech filtering on the training set, retaining only nouns, verbs, and adjectives that significantly impact classification.
2. Calculate TF-IDF weights for each feature word in all documents.
3. Normalize feature words in each document. Let w_i be the TF-IDF value of feature word i in a document, and normalize it using formula (1).
4. Select feature words with proportions greater than threshold *weight* in each category as high-frequency words of the training set.

3.3 Semantic Similarity Calculation Based on HowNet

This paper uses HowNet to calculate semantic similarity between feature words, thereby 挖掘 ing relationships between training and test sets composed of different document types. Thus, HowNet-based semantic similarity calculation forms the foundation of our feature extension method. In HowNet' s structure, words are represented by senses (义项), meaning a word can have multiple senses, and each sense is represented by sememes (义原). Therefore, word similarity can be calculated through sememe similarity.

(1) Sememe Similarity Calculation

HowNet' s sememe tree is a hierarchical system constructed from hyponymy relationships between sememes. This paper calculates sememe similarity using the shortest path distance between sememes in the tree. Let the shortest path distance between two sememes be d; the similarity between these two sememes is calculated as follows [12]:

$$sim(p_1, p_2) = \frac{\alpha}{d + \alpha}$$

where p , p represent two sememes, d is the shortest path distance between p and p in the sememe tree, and is an adjustable parameter.

Literature [13] argues that considering only the shortest path distance cannot accurately calculate sememe similarity and proposes a method incorporating sememe hierarchical depth. The main idea is that for two sememes with the same shortest path distance, the deeper the level, the more specific the meaning, and thus greater similarity weight should be assigned. The calculation formula is:

$$sim(p_1, p_2) = \frac{\min(depth_1, depth_2)}{\min(depth_1, depth_2) + d} \times \frac{\alpha}{d + \alpha}$$

where *depth* , *depth* are the depths of sememes p and p in the sememe tree respectively, d is the shortest path distance, and is an adjustable parameter. This paper uses formula (3) to calculate sememe similarity.

(2) Sense Similarity Calculation

In HowNet, sense descriptions are generally divided into four categories [14]: primary features, secondary features, relational sememe features, and relational symbol features. Sense similarity is the weighted sum of similarities between corresponding components of the semantic descriptions. The similarity between sense s and sense s is calculated as follows [15]:

$$sim(s_1, s_2) = \sum_{i=1}^{4} \beta_i \times sim_i(s_1, s_2)$$

where *sim_i(s , s)* represents the similarity between corresponding components, + + + = *1*, and _*i* *0*. Literature [12] provides methods for calculating similarities between components in sense descriptions, which ultimately reduces to sememe similarity calculation.

(3) Word Similarity Calculation

Literature [15] defines word similarity as the likelihood that two words can replace each other in different contexts without changing the syntactic-semantic structure of the text. Assuming words w and w have m and n senses respectively: w = *{s 1, s 2, ···, s ^m}* and w = *{s 1, s 2, ···, s ^n}*, the similarity between w and w is the maximum similarity among all sense combinations, using a maximum matching approach. The calculation formula is [15]:

$$sim(w_1, w_2) = \max_{\substack{1 \le i \le m \\ 1 \le j \le n}} sim(s_1^i, s_2^j)$$

4.1 Experimental Materials

This study collected three types of documents—webpages, books, and journals—from sources including the Sogou corpus [16], library catalogs, and electronic journal databases. Journals were further divided into academic and non-academic categories. Webpage documents were selected from the sports, IT, and military categories of the Sogou corpus. Book documents were obtained from a university library' s OPAC catalog, extracting titles and abstracts from books in the sports, computer technology, and military categories of the Chinese Library Classification system. Journal documents were selected from CNKI in the same three categories according to the Chinese Library Classification.

We established multiple experimental datasets for each of the four document types and repeated experiments. Each dataset included a training set and test set of one document type, all comprising three major categories: sports, computer technology, and military. Each document type contained 600 texts, totaling 2,400 documents.

4.2 Experimental and Evaluation Methods

To eliminate the impact of imbalanced data on experimental results, all corpora used in this study were balanced, with each category containing approximately the same number of texts and similar text lengths, and with no overlap between training and test sets. We employed five-fold cross-validation for training and classification, using the macro-averaged F1-score as the evaluation metric [18].

4.3 Experimental Results

(1) Individual Document Type Classification Results

Individual document type classification refers to experiments where training and test sets are from the same document type. We conducted five-fold cross-validation experiments on the four document types separately, ensuring no text overlap between training and test sets. The classification results are shown in Figure 2 [Figure 2: see original paper].

Different classification algorithms significantly affect results. This study selected the classic KNN algorithm for classifier construction. Theoretically, the Naive Bayes algorithm can also achieve good classification results, but its feature independence assumption is often violated, making it a common benchmark for comparison. Our experiments showed that Naive Bayes produced results almost identical to KNN, validating KNN' s effectiveness. Therefore, we selected KNN as our classification algorithm. Figure 2 shows that individual classification of each document type achieved good results, with accuracies above 70%.

(2) Mixed Document Type Classification Results

Mixed document type classification refers to experiments where training and test sets are from different document types. We conducted five-fold cross-validation

experiments on all combinations of the four document types. The results are shown in Figure 3 [Figure 3: see original paper].

Mixed classification experiments showed that classification between webpages and non-academic journals performed well, both above 80%, with webpage-to-non-academic-journal cross-classification reaching 83.9%—even better than non-academic journals' individual classification. Book-to-academic-journal classification also performed well, above 70%, with academic-journal-to-book classification achieving 78.4%, higher than books' individual classification accuracy of 71.8%. This proves the validity of mixed-type classification.

However, classification between webpages and books, webpages and academic journals, books and webpages, and books and non-academic journals performed poorly, all below 60%. This indicates that the selection of training and test document types significantly impacts classification effectiveness. Well-matched document types can even outperform individual classification, while poorly matched types yield low effectiveness, demonstrating that document type combination substantially influences results.

(3) Mixed Classification with HowNet-Based Semantic Feature Extension

These experiments used all combinations of the four document types as training and test sets (including same-type combinations), applying our proposed feature extension method with HowNet as the third-party resource to extend test set features before conducting five-fold cross-validation. Results are shown in Figure 4 [Figure 4: see original paper].

The results show that after HowNet-based semantic feature extension, individual classification effectiveness improved across all four document types to varying degrees: book-only classification increased from 71.8% to 74.1%, and webpage-only classification from 91.4% to 91.7%. Cross-classification effectiveness also improved noticeably. Well-matched document types showed smaller improvements (e.g., non-academic-journal-to-webpage increased from 86.7% to 88.9%), while poorly matched types improved significantly (e.g., non-academic-journal-to-academic-journal increased from 25.4% to 36.4%), indicating greater improvement potential for low-match combinations.

Experiment (1) compared KNN and Naive Bayes for individual classification, finding nearly identical results and validating our selection of KNN. Experiment (2) conducted mixed classification without feature extension, with results in Figure 3 showing high matching between webpages and non-academic journals and between books and academic journals—ideal for validating our method. Experiment (3) compared our HowNet-based semantic feature extension method with non-extension results. Individual classification improvements were 0.3%, 2.3%, 1.5%, and 0.8%, proving the method enhances same-type classification. Cross-classification between webpages and non-academic journals improved by 1.2% and 2.2% respectively, demonstrating effectiveness for both same-type and cross-type classification. Comparison of extended features revealed that while seman-

tic differences were eliminated, some "noise words" —features with low discriminative power due to frequent appearance across multiple categories—were introduced, interfering with classification and limiting improvement magnitude. Although webpage-to-non-academic-journal cross-classification was slightly lower than webpage-only classification, using non-academic journals as training sets avoids the cumbersome real-time updates required when using webpages themselves, making our method practically significant.

5 Conclusion and Future Work

This study investigated automatic classification of multiple document types using HowNet as a third-party resource for feature extension. Experimental results prove that cross-classification between well-matched document types can achieve equal or better results than single-type classification. From the perspectives of corpus construction and feature extension, we proposed a HowNet-based semantic feature extension method that uses HowNet' s semantic structure to eliminate differences in word usage and writing style across document types, further improving mixed-type classification effectiveness. Experiments demonstrate that this method effectively enhances current classification performance. It can leverage scientifically classified, long-accumulated literature to efficiently classify rapidly growing, frequently updated document types with better results, thus offering high practical value.

This research was conducted based on the relatively mature Vector Space Model (VSM) for text representation, essentially achieving semantic extension through external resources to narrow differences between document types and enable cross-type classification. Future work will further investigate document type differences, explore methods to eliminate interference from "noise words"to improve classification effectiveness, experiment with probabilistic topic models (LDA) for text representation, evaluate other third-party resources like Wikipedia for cross-type classification, and assess the adaptability of various classic classification algorithms such as Support Vector Machines (SVM) for cross-document-type classification.

References

[1] Xue Chunxiang, Xia Zuqi, Hou Hanqing. A Comparison of Automatic Classification Between Corpus-based Model and Experiences-based Model [J]. Journal of Nanjing Agricultural University: Social Sciences Edition, 2005, 5(4): 85-91.
[2] Pong J Y H, Kwok R C W, Lau R Y K, et al. A Comparative Study of Two Automatic Document Classification Methods in a Library Setting [J]. Journal of Information Science, 2008, 34(2): 213-230.
[3] Li Xiangdong, Hu Yiquan, Ba Zhichao, et al. The Study of Mixed Automatic Categorization on Digital Library Collections [J]. Library Journal, 2014, 33(11): 42-48.

[4] HowNet Knowledge Database [DB/OL]. [2015-06-15]. http://www.keenage.com/.
[5] Pan S J, Yang Q. A Survey on Transfer Learning [J]. IEEE Transactions on Knowledge and Data Engineering, 2010, 22(10): 1345-1359.
[6] Wang P, Domeniconi C, Hu J. Using Wikipedia for Co-Clustering Based Cross-Domain Text Classification [C]. In: Proceedings of the 8th IEEE International Conference on Data Mining. IEEE, 2008.
[7] Lu Z, Zhu Y, Pan S J, et al. Source Free Transfer Learning for Text Classification [C]. In: Proceedings of the 28th Association for the Advancement of Artificial Intelligence Conference on Artificial Intelligence. 2014.
[8] Zhao Hui, Liu Huailiang. Classification Algorithm of Chinese Short Texts Based on Wikipedia [J]. Library and Information Service, 2013, 57(11): 120-124.
[9] Ning Yahui, Fan Xinghua, Wu Yu. Short Text Classification Based on Domain Word Ontology [J]. Computer Science, 2009, 36(3): 142-145.
[10] Li Xiangdong, Cao Huan, Ding Cong, et al. Short-text Classification Based on HowNet and Domain Keyword Set Extension [J]. New Technology of Library and Information Service, 2015(2): 31-38.
[11] Shi Congying, Xu Chaojun, Yang Xiaojiang. Study of TFIDF Algorithm [J]. Journal of Computer Applications, 2009, 29(S1): 167-170.
[12] Liu Qun, Li Sujian. Word Similarity Computing Based on How-net [J]. Computational Linguistics and Chinese Language Processing, 2002, 7(2): 59-76.
[13] Wu Jian, Wu Zhaohui, Li Ying, et al. Web Service Discovery Based on Ontology and Similarity of Words [J]. Chinese Journal of Computers, 2005, 28(4): 595-602.
[14] Li Shengqi, Tian Qiaoyan, Tang Cheng. Disambiguating Method for Computing Relevancy Based on HowNet Semantic Knowledge [J]. Journal of the China Society for Scientific and Technical Information, 2009, 28(5): 706-711.
[15] Sun Jianwang, Lv Xueqiang, Zhang Leihan. Short Text Classification Based on Semantics and Maximum Matching Degree [J]. Computer Engineering and Design, 2013, 34(10): 3613-3618.
[16] SogouT [DB/OL]. [2015-06-03]. http://www.sogou.com/labs/dl/t.html.
[17] Tan S. An Effective Refinement Strategy for KNN Text Classifier [J]. Expert Systems with Applications, 2006, 30(2): 290-298.
[18] Feng Guohe. Review of Performance Evaluation of Text Classification [J]. Journal of Intelligence, 2011, 30(8): 66-70.

Author Contributions:
Li Xiangdong: Conceived research ideas and design, reviewed and revised the final manuscript;
Liu Kang: Implemented the system, conducted experiments, wrote the manuscript;
Ding Cong: Conducted experiments, performed literature review;
Gao Fan: Collected data, performed literature review.

Conflict of Interest Statement: All authors declare no conflict of interest.

Supporting Data:
The supporting data can be found in the online version at http://www.infotech.ac.cn:
[1] Li Xiangdong, Liu Kang, Ding Cong, Gao Fan. book.txt. Book-type documents.
[2] Li Xiangdong, Liu Kang, Ding Cong, Gao Fan. web.txt. Webpage-type documents.
[3] Li Xiangdong, Liu Kang, Ding Cong, Gao Fan. acd.txt. Academic journal-type documents.
[4] Li Xiangdong, Liu Kang, Ding Cong, Gao Fan. nonacd.txt. Non-academic journal-type documents.

Received: 2015-08-12
Revised: 2015-10-19

Note: Figure translations are in progress. See original paper for figures.