Research on Automatic Discovery and Annotation Methods for Citation Metadata: A Case Study of Foreign Language Citations (Postprint)

Authors: Jiang Lin, Wang Dongbo

Date: 2017-11-08T00:00:00+00:00

Abstract

【Purpose】 Building upon a summary of current citation metadata extraction methods, this study explores automatic extraction methods for citation metadata by combining semantic knowledge and machine learning methods. 【Method】 In the experiments, a neural network model was employed to train word vectors on a manually segmented corpus. Leveraging the phenomenon that metadata of the same type tends to be relatively concentrated in a certain region of the vector space, automatic classification and labeling of metadata was achieved through a Support Vector Machine (SVM) classification algorithm. 【Result】 In experiments using foreign language citation data as the test set, the proposed method achieved high precision and recall, particularly demonstrating strong processing capability for citations containing multiple languages and abbreviations. 【Limitation】 There are certain limitations in the fine-grained extraction of temporal content from citation metadata. 【Conclusion】 Experimental results demonstrate that this method achieves good effectiveness in the automatic discovery and labeling of citation metadata, and can significantly enhance the applicability and fault tolerance of the approach.

Full Text

Automatically Detecting and Tagging Foreign Language Citation Metadata

Jiang Lin[1],[2], **Wang Dongbo**[3]
[1](School of Information Management, Nanjing University, Nanjing 210023, China)
[2](Jiangsu Key Laboratory of Data Engineering and Knowledge Service, Nanjing 210023, China)

[3](College of Information Science and Technology, Nanjing Agricultural University, Nanjing 210095, China)

Abstract: [Objective] Building upon existing citation metadata extraction methods, this study explores automatic extraction techniques by integrating semantic knowledge with machine learning approaches. [Methods] The experiments employed neural network models to train word vectors on manually segmented corpora. Leveraging the phenomenon that similar metadata types tend to cluster in specific regions of vector space, we implemented a Support Vector Machine classification algorithm to automatically categorize and annotate metadata. [Results] Experiments on foreign language citation datasets achieved high precision and recall rates, demonstrating particularly strong capability in handling citations containing multiple languages and abbreviations. [Limitations] The approach exhibits certain limitations in fine-grained extraction of temporal content within citation metadata. [Conclusions] Experimental results indicate that this method effectively automates the discovery and tagging of citation metadata while significantly improving adaptability and fault tolerance.

Keywords: Bibliographic Metadata; Metadata Extraction; Machine Learning; Neural Network
Classification Number: G254

Citation data appears extensively in scientific literature, particularly in technical documents. Such data not only reflects the continuity of scientific development but also demonstrates respect for and protection of others' intellectual property rights. Bibliographic references typically comprise numerous descriptive elements including titles, authors, publishers, publication dates, and other bibliographic fields. In most document metadata standards, citation data is considered an important metadata type with various applications in digital library and semantic web construction. Traditional libraries often required manual extraction or entry of bibliographic metadata, but the current explosion in literature volume has made manual processing impractical. Additionally, large volumes of legacy paper documents require automated metadata extraction during digitization. Citation metadata extraction forms the foundation for domain retrieval, citation network analysis, article contribution evaluation, and topic discovery. However, due to inconsistent standards, citation metadata often exhibits diverse styles—different languages, subjects, and publication types (books, journals, conferences) employ distinct citation formats. In terms of content structure, different citations contain varying numbers of metadata elements arranged in different sequences. In English scientific literature alone, common styles include APA, MLA, Chicago, AMA, IEEE, and ACM, among six major formats [1]. Given both the importance of citations and their stylistic diversity, mining information contained within citation data has become an important yet challenging task in information extraction. This paper proposes a machine learning-based approach for automatic citation metadata extraction and tagging that can circumvent inconsistencies in manually compiled citation templates and demonstrates effective cross-language applicability.

As a subtask of metadata extraction, citation metadata extraction holds significant research value in computer science and library science, with multiple methodological approaches having evolved. Overall, these methods fall into three categories: rule-based, template-based, and machine learning-based approaches.

Corresponding Author: Jiang Lin, ORCID: 0000-0003-3211-7783, E-mail: 18205185622@163.com

Rule-based methods have been widely applied in practical citation extraction systems. For instance, Wei et al. [2] utilized a Layer-upon-Layer Tagging approach to extract citation metadata, employing progressive annotation through format attribute layers and dictionary semantic layers to achieve automatic metadata tagging. Besagni et al. [3] proposed combining part-of-speech tagging with rule refinement for citation metadata extraction and annotation. Li et al. [4] suggested using regular expressions to extract paper metadata.

Template-based methods represent another commonly adopted approach. These methods typically construct a template database first, then complete extraction by searching and matching templates. Day et al. [5] identified six primary reference formats in computer science literature and developed the multi-layer knowledge representation framework INFOMAP, upon which they built a knowledge-based citation metadata extraction system—essentially a multi-layer template-based metadata extraction method. Cortez et al. [6] proposed an unsupervised citation metadata extraction approach that automatically generates templates using existing domain metadata as training data. Huang et al. [7] and Chen et al. [8] represented citation strings as protein sequences, storing citation template sequence representations in a DNA database. They then employed BLAST (Basic Local Alignment Search Tool), a similarity comparison tool used in DNA databases, to find similar DNA sequences—i.e., citation templates—for analyzed citations, finally parsing citation data according to matched templates.

These rule-based or template-based methods generally offer high analysis efficiency, particularly achieving high recognition rates for citation styles covered by their rules or templates. However, researchers recognize inherent limitations: when new citation styles emerge, additional rules or templates become necessary. As citation styles proliferate, the burden of rule/template creation increases, leading to higher system redundancy and reduced applicability.

In contrast to rule and template-based approaches, many researchers have turned to machine learning methods for automatic metadata discovery and indexing. In natural language processing, classification algorithms have been widely applied to sequence labeling problems. For example, Han et al. [9] treated metadata extraction as a classification problem and introduced Support Vector Machines (SVM) to metadata extraction tasks, improving upon HMM' s independence assumption limitations while achieving satisfactory results, though at the cost of losing the close relationship between state transitions and observation sequences. Additionally, Conditional Random Fields (CRF)

represent a currently popular approach. Peng et al. [10] applied CRF to automatic citation metadata extraction, achieving excellent results on the Cora dataset, a public test set for paper metadata extraction. Yu et al. [11] tested CRF for extracting paper header and citation metadata from Chinese scientific paper datasets, likewise obtaining favorable outcomes.

In summary, machine learning-based methods achieve excellent results in metadata extraction but introduce additional overhead such as manual annotation and lengthy training times. Moreover, given the diversity of citation styles and languages in reality, it is impossible to exhaustively cover all citation formats. Particularly when authors manually add citation data, template misuse inevitably occurs, significantly reducing automatic recognition accuracy. In this sense, neither manually crafted rules/templates nor machine learning-trained templates exhibit strong adaptability. Therefore, we aim to improve machine learning algorithms to enhance cross-language adaptability and break template limitations, thereby increasing automatic annotation accuracy and universality.

Key Techniques for Automatic Discovery, Extraction, and Tagging of Citation Data

Addressing existing challenges in citation metadata extraction, this paper proposes an improved feature analysis-based extraction method that eliminates traditional template dependencies while offering cross-language platform advantages. The technical implementation roadmap is illustrated in [Figure 1: see original paper].

The experimental citation data primarily originated from the Chinese Social Sciences Citation Index (CSSCI) citation database. Since the experiments required constructing word vector space models, Chinese citations would have necessitated word segmentation, whose quality significantly impacts results. Therefore, the experiments primarily used foreign language citation data for effectiveness testing, obtaining foreign citation data through regular expression filtering.

Observation of extensive foreign language citation data revealed that punctuation marks such as periods, commas, and colons commonly serve as metadata separators. However, periods also frequently appear in name abbreviations, version numbers, etc. To enhance separator recognition, the experiments established the following preprocessing rules: (1) **Separator Replacement Rule:** Since citation data often exhibits mixed Chinese and English punctuation usage, increasing recognition difficulty, all punctuation was standardized to English punctuation. (2) **Period Replacement Rule:** When a period follows an uppercase letter and precedes an English letter and punctuation, it typically indicates an English name; when surrounding single digits (e.g., “Windows 3.0”), it usually denotes software version numbers; periods also combine with adjacent words as abbreviations (e.g., “St.” , “Vol.” , “Aug.”). In these cases, periods were replaced with asterisks and no longer treated as metadata separators.

Training of Metadata Classification Feature Values

Current neural network-based word vector computation has achieved excellent results. For example, Mikolov et al. from Google developed an automatic generation technology for dictionaries and terminology that can transform one language into another with remarkable effectiveness.

Consider English and Spanish as examples. Through training, we obtain their respective word vector spaces E(English) and S(Spanish). Selecting five English words—one, two, three, four, five—with corresponding word vectors u1, u2, u3, u4, u5 in space E (left side of [Figure 2: see original paper]), we apply Principal Component Analysis (PCA) for dimensionality reduction to obtain two-dimensional vectors v1, v2, v3, v4, v5. Similarly, we extract their Spanish counterparts (uno, dos, tres, cuatro, cinco) and apply PCA, as shown in the right side (S) of [Figure 2: see original paper].

[Figure 2: see original paper] reveals that the five words occupy similar relative positions in both vector spaces, demonstrating structural similarity between vector spaces of different languages. This validates the reasonableness of using distance to measure word similarity in vector space and shows that functionally similar words tend to cluster in the same region. Based on this phenomenon, we constructed word vector space models for words in metadata using neural network models. By the same principle, words frequently appearing in the same metadata type will cluster in specific regions of the vector space. Consequently, different citation metadata types—such as author names, titles, journal names, and dates—will respectively aggregate in distinct regions of the vector space model, enabling effective automatic indexing of citation metadata while reducing language-related impacts on classification effectiveness. Since Chinese citations lack obvious word boundaries and require segmentation software whose errors introduce significant interference, the experiments primarily used foreign language citations as examples.

In the experiments, preprocessed training data underwent manual identification and annotation, serving two purposes: providing training sets for word vector construction and for SVM feature analysis classification. Annotated sample data are shown in [Figure 3: see original paper].

In [Figure 3: see original paper], metadata was annotated by type, with each line representing one metadata type. The annotated metadata categories and classification information are detailed in .

(1) Word Vector Training This study employed the CBOW model [13-14] for word vector construction, using segmented metadata as training data. The model' s core idea predicts the current word Wt given its context Wt-2, Wt-1, Wt+1, Wt+2. [Figure 4: see original paper] illustrates the CBOW model network structure, which resembles a neural network comprising three layers: input, projection, and output.

Input Layer: Contains word vectors v(Context(w)1), v(Context(w)2)··· v(Context(w)2c) R^m for 2c words in Context(w), where m denotes word vector length and c represents the number of words taken before and after word w.

Projection Layer: Performs summation of the 2c vectors as shown in formula (1).

Output Layer: Corresponds to a binary tree using corpus vocabulary as leaf nodes, weighted by word frequency to construct a Huffman tree. The tree contains N leaf nodes (N=|D|), each corresponding to a word in dictionary D.

Neural network-based word vector construction offers two main advantages: First, word similarity is reflected through word vectors—the model assumes that "similar" words have similar vectors, with the probability function being smooth regarding word vectors, so small changes in word vectors produce only small probability changes. Second, vector-based models inherently include smoothing functionality, as p(w|Context) is non-zero, eliminating the need for additional smoothing processing.

In the vector space model, word vectors from different citation metadata categories (author names, titles, journal names, dates, etc.) cluster in relatively stable regions, making automatic classification and indexing feasible.

(2) Classification Feature Training Since each metadata category clusters in the same spatial region, we performed clustering calculations on word vectors for each category in the training data to determine cluster centers—the most representative metadata for each category. New words are classified based on their distances to these category centers in the space model. The experiments organized training data by category and used word vectors with the K-means clustering algorithm to compute cluster centers for each category.

K-means is a commonly used clustering algorithm. For a given dataset containing n d-dimensional data points, each partition represents a class Ck with a class center ti. Using Euclidean distance as the similarity metric, the algorithm calculates the sum of squared distances from points within each class to its center ti. The clustering objective minimizes the total sum of squared distances across all clusters.

According to least squares and Lagrangian principles, the cluster center tk should be the mean of all data points in class Ck. The K-means algorithm begins with an initial K-class partition, then assigns data points to classes to reduce the total distance sum. Since the total sum decreases as K increases (J(C)=0 when K=n), the minimum can only be achieved at a specific K value. Clustering algorithms identify the positions (cluster centers) of the most characteristic data for each metadata type, guiding metadata classification and reducing feature dimensions to shorten training time.

Since foreign citations commonly use ",.:" as metadata separators, with each

separator containing the same data type, we used the Euclidean distance between the centroid (cluster center) of each segmented part and the centroids of various categories as classification features. This approach reduces feature quantity while strengthening feature description, as illustrated in [Figure 5: see original paper].

Because each metadata element's category also correlates with its position in the citation, we incorporated relative position as a classification feature, calculated by dividing each segment' s position by the total number of segments. Assuming segmented citation data is represented as (i/n), feature collection examples for classification training are shown in .

Integrating CBOW, K-means, and metadata position features consolidates vector space characteristics, enabling similar metadata categories to cluster in relatively concentrated regions for automatic identification and annotation.

(3) Support Vector Machine Classification SVM represents a significant achievement in machine learning research and serves as an important classification algorithm primarily addressing binary pattern recognition problems. Developed from Statistical Learning Theory (SLT), its core content was proposed by Stitson et al. [15] between 1992 and 1995. SVM' s main advantages include: First, it specifically addresses finite sample situations, aiming for optimal solutions based on available information rather than asymptotic optimality as sample sizes approach infinity. Second, theoretically, it achieves global optima, solving the local extremum problem inherent in neural network methods. Third, it transforms practical problems through nonlinear mapping into high-dimensional feature spaces, constructing linear discriminant functions to realize nonlinear discrimination in original space, with special properties ensuring good generalization while elegantly solving dimensionality issues—algorithm complexity becomes independent of sample dimensionality.

After comprehensive comparison of neural network and SVM models, we selected the SVM algorithm for citation metadata feature classification training. For preprocessed citation data, we segmented data using common metadata separators, computed distances from each segment' s cluster center to various category centers using clustering algorithms, and combined position feature values for automatic classification of segment categories.

Experimental Evaluation

Evaluation Metrics

This experiment adopted precision, recall, and their harmonic mean (F1-score) as evaluation criteria, with formulas as follows:

Precision = (Number of correctly extracted information items) / (Number of extracted information items)

Recall = (Number of correctly extracted information items) / (Number of information items in sample)
F1 = 2 × (Precision × Recall) / (Precision + Recall)

Experimental Results

Using 2,000 foreign language citation records collected from CSSCI as experimental data, manually annotated as the training set, partial results are shown in [Figure 6: see original paper].

[Figure 6: see original paper] demonstrates that the proposed method effectively identifies multi-unit publisher names and segmented units using different citation styles. Semantic analysis accurately annotates temporal abbreviations such as "Nov" and "Dec," highlighting the advantage of semantic-based annotation over single template-based approaches. This avoids the need for continuous template adjustment and program complexity when new template formats emerge. Furthermore, semantic-based annotation effectively circumvents real-world separator misuse, improving algorithm fault tolerance and universality. However, limitations exist: distinguishing publication years from page numbers proves challenging since both consist of numerals with minimal semantic differences; combining template methods could yield better results.

Comparative Experimental Analysis

Common natural language processing approaches like Hidden Markov Models and Conditional Random Fields (CRF) address sequence labeling problems, with CRF being currently predominant. To highlight our method's effectiveness, we conducted comparative experiments using CRF as a reference. Proposed by Lafferty et al. [16], CRF is an undirected graphical model combining maximum entropy and hidden Markov model characteristics, achieving excellent results in word segmentation, POS tagging, and named entity recognition tasks.

Since CRF also requires extensive manual annotation, and considering both methods achieve high accuracy for author names and numeric dates/page numbers (making differentiation difficult), we focused comparative experiments on publisher name extraction. Using Stanford Parser from Stanford' s NLP group for English citation POS tagging, we employed a five-tag marking pattern with specific annotation rules detailed in .

The comparative experimental results are shown in [Figure 7: see original paper]. The proposed algorithm outperformed standard CRF in both recall and precision, particularly in recognition accuracy. Since the CRF experiment only used POS features, this likely contributed to moderate performance. Classic pattern recognition algorithms like CRF typically require manual feature extraction and correlation analysis to identify the most representative features while removing irrelevant or autocorrelated ones. This process heavily depends on human experience and subjective judgment, where feature selection and even extraction order significantly impact classification performance. Our experimental

algorithm uses semantic features for classification with SVM, achieving notable results, especially for English name abbreviation issues. The approach demonstrates strong robustness against input data distortions using fuzzy semantic knowledge.

Results show the improved citation metadata annotation algorithm significantly enhances recognition accuracy through three main advantages: strong robustness against input distortions (e.g., English abbreviation recognition for institutional and publisher names), high fault tolerance (semantic recognition despite incorrect separators), and strong portability (good adaptability across languages). These advantages make our method superior to common machine learning algorithms like CRF. However, limitations exist: manual annotation for training sets is time-consuming, and insufficient training data may yield suboptimal word vector models, reducing accuracy and recall. For more precise recognition of numerically composed data like publication years and page numbers with minimal semantic differences, combining template methods could improve accuracy. Future work will explore hybrid intelligent algorithms integrating machine learning and rule-based models.

References

[1] Jiang Xin. Several Main Quotation Ways in British-American Academic Documents [J]. Library and Information, 2003(3): 26-30.

[2] Wei W, King I, Lee J H M. Bibliographic Attributes Extraction with Layer-upon-Layer Tagging[C]//Proceedings of the 9th International Conference on Document Analysis and Recognition. IEEE, 2007, 2: 804-808.

[3] Besagni D, Belaïd A, Benet N. A Segmentation Method for Bibliographic References by Contextual Tagging of Fields[C]//Proceedings of the 7th International Conference on Document Analysis and Recognition. IEEE, 2003: 384-388.

[4] Li Chaoguang, Zhang Ming, Deng Zhihong, et al. Automatic Metadata Extraction for Scientific Documents [J]. Computer Engineering and Applications, 2002, 38(21): 189-191, 235.

[5] Day M Y, Tsai R T H, Sung C L, et al. Reference Metadata Extraction Using a Hierarchical Knowledge Representation Framework [J]. Decision Support Systems, 2007, 43(1): 152-167.

[6] Cortez E, da Silva A S, Gonçalves M A, et al. FLUX-CIM: Flexible Unsupervised Extraction of Citation Metadata[C]//Proceedings of the 7th ACM/IEEE Joint Conference on Digital Libraries. ACM, 2007: 215-224.

[7] Huang I A, Ho J M, Kao H Y, et al. Extracting Citation Metadata from Online Publication Lists Using BLAST[C]//Proceedings of the 8th Pacific-Asia Conference, PAKDD 2004. Springer Berlin Heidelberg, 2004: 539-548.

[8] Chen C C, Yang K H, Kao H Y, et al. BibPro: A Citation Parser Based on Sequence Alignment Techniques[C]//Proceedings of the 22nd International Conference on Advanced Information Networking and Applications-Workshops (AINAW 2008). IEEE, 2008: 1175-1180.

[9] Han H, Giles C L, Manavoglu E, et al. Automatic Document Metadata Extraction Using Support Vector Machines[C]//Proceedings of the 2003 Joint Conference on Digital Libraries. IEEE, 2003: 37-48.

[10] Peng F, McCallum A. Accurate Information Extraction from Research Papers Using Conditional Random Fields[C]//Proceedings of the Human Language Technology Conference of the North American Chapter of the Association-for-Computational-Linguistics. 2004: 329-336.

[11] Yu J, Fan X. Metadata Extraction from Chinese Research Papers Based on Conditional Random Fields[C]//Proceedings of the 4th International Conference on Fuzzy Systems and Knowledge Discovery. IEEE, 2007, 1: 497-501.

[12] Mikolov T, Le Q V, Sutskever I. Exploiting Similarities Among Languages for Machine Translation [OL]. arXiv Preprint. arXiv:1309.4168, 2013.

[13] Mikolov T. Word2Vec Code [EB/OL]. [2015-09-18]. http://word2vec.googlecode.com/svn/trunk/.

[14] Zhou Lian. Exploration of the Working Principle and Application of Word2Vec [J]. Sci-Tech Information Development & Economy, 2015(2): 145-148.

[15] Stitson M O, Weston J A E, et al. Theory of Support Vector Machines [R]. Technical Report, CSD-TR-96-17, London: University of London, 1996.

[16] Lafferty J, McCallum A, Pereira F C N. Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data [EB/OL]. [2016-07-15]. http://repository.upenn.edu/cis_{papers}/159.

Author Contributions: Jiang Lin: Conceived research objectives and technical approach, implemented experimental programming, wrote the manuscript; Wang Dongbo: Collected and organized training data, refined research methodology, revised the manuscript.

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

Supporting Data: Supporting data is self-archived by authors, E-mail: 18205185622@163.com.
[1] Jiang Lin, Wang Dongbo. meteSplit_{SVM}.rar. Implementation of automatic citation metadata detection and annotation experimental program.
[2] Jiang Lin, Wang Dongbo. Train.rar. Training corpus.

Received Date: 2016-08-18
Revised Date: 2016-11-06

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