Automatic Discovery and Annotation of 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 review of existing citation metadata extraction methods, this study explores automatic extraction approaches for citation metadata by integrating semantic knowledge and machine learning techniques.

【Method】The experiment employs neural network models to train word vectors on a manually segmented corpus. Leveraging the phenomenon that metadata of the same type tend to cluster in specific regions of the vector space, a Support Vector Machine (SVM) classification algorithm is utilized to achieve automatic categorization and annotation of metadata.

【Results】In experiments using foreign-language citation data as the test set, the proposed method achieved high precision and recall rates, exhibiting particularly robust processing capability for citations containing multiple languages and abbreviations.

【Limitations】There exist certain limitations in the fine-grained extraction of temporal content within citation metadata.

【Conclusion】Experimental results demonstrate that this method achieves favorable performance in the automatic discovery and annotation of citation metadata, and can substantially enhance the method's applicability and fault tolerance.

Full Text

ChinaXiv Collaborative Journal, Issue 1, 2017

Research on Automatic Discovery and Tagging Methods for Citation Metadata: A Case Study of Foreign Language Citations

Jiang Lin1,2, Wang Dongbo3

1School of Information Management, Nanjing University, Nanjing 210023, China
2Jiangsu Key Laboratory of Data Engineering and Knowledge Service, Nanjing 210023, China
3College of Information Science and Technology, Nanjing Agricultural University, Nanjing 210095, China

Abstract: [**Objective**] This paper proposes a new method to automatically extract bibliographic metadata using semantic knowledge and machine learning technologies. [**Methods**] We employed a neural network model to create word vectors from manually segmented data, observing that metadata of the same type tends to cluster in specific regions of the vector space. Based on this finding, we developed a novel SVM classification algorithm for automatic classification and annotation of bibliographic metadata. [**Results**] The proposed method achieved high recall and precision rates, particularly for citations containing multiple languages and abbreviations. [**Limitations**] Fine-grained extraction of time-related content could be improved. [**Conclusions**] The method effectively detects and tags bibliographic metadata while enhancing system compatibility and fault tolerance.

Keywords: Bibliographic Metadata; Metadata Extraction; Machine Learning; Neural Network

Classification Number: G254

Scientific literature, particularly technical documents, contains extensive citation data. Such data not only reflects the continuity of scientific development but also demonstrates respect for and protection of intellectual property rights. Bibliographic references typically comprise numerous descriptive elements, including title, author, publisher, publication date, and others.

In the vast majority of document metadata standards, citation data is considered an important type of metadata with numerous applications in digital library and semantic web construction. In traditional libraries, bibliographic metadata often required manual extraction or entry after the fact. However, with the current exponential growth in literature volume, reliance on manual extraction has become impractical. Furthermore, the digitization of large volumes of legacy paper documents necessitates automatic metadata extraction from these sources. Citation metadata extraction serves as the foundation for research in domain retrieval, citation network analysis, article contribution evaluation, topic discovery, and other areas.

However, due to inconsistent standards adopted across different contexts, citation metadata often exhibits varying styles. For instance, citation styles differ across languages, subjects, and publication types such as books, journals, and conference proceedings. In terms of content, different citations contain varying numbers of metadata elements arranged in different sequences. In English scientific literature alone, six major styles are commonly encountered: APA, MLA, Chicago, AMA, IEEE, and ACM [1]. Precisely because of the importance of

citations and their stylistic diversity, analyzing and mining information contained within citation data has become a significant yet challenging task in information extraction. Consequently, this paper designs a machine learning-based approach for automatic citation metadata extraction and tagging. This method can circumvent inconsistencies arising from manual citation compilation using disparate templates and demonstrates excellent cross-linguistic platform applicability.

As a subtask of metadata extraction, citation metadata extraction holds significant research importance in fields such as computer science and library science, having evolved into multiple methodological approaches. Broadly speaking, citation metadata extraction methods can be categorized into three types: 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 example, Wei et al. [2] utilized a Layer-upon-Layer Tagging approach to extract metadata from citations, employing progressive annotation across format attribute layers and dictionary semantic layers to achieve automatic citation 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 establish 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 constructed a multi-layer knowledge representation framework called INFOMAP, upon which they developed 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 analysis tool used in DNA databases, to find similar DNA sequences for citations under analysis—that is, citation templates—and finally parsed citation data according to the matched templates.

These rule-based or template-based methods generally offer high analysis efficiency, particularly for citation styles covered by their rules or templates, achieving high recognition rates. However, researchers have recognized inherent limitations in these approaches: when new citation styles are introduced, additional rules or templates must be created. As citation styles proliferate, the burden of rule or template creation increases, leading to higher system redundancy and

reduced applicability.

In contrast to rule-based and template-based methods, many researchers have opted for machine learning approaches to automatically discover and index metadata. In natural language processing, numerous scholars have employed classification algorithms to solve text sequence labeling problems. For instance, Han et al. [9] treated metadata extraction as a classification problem and introduced Support Vector Machines (SVM) to metadata extraction tasks, improving upon the independence assumption limitations of HMM methods and achieving satisfactory results. However, this method simultaneously lost the close relationship between state transitions and observation sequences. Additionally, Conditional Random Fields (CRF) currently represent a widely used method. For example, Peng et al. [10] applied CRF to automatic citation metadata extraction, achieving excellent extraction results on the Cora dataset, a public test set for paper metadata extraction. Yu et al. [11] tested CRF methods for extracting paper header and citation metadata on Chinese scientific paper datasets, likewise achieving favorable results.

In summary, machine learning-based methods can achieve excellent results in metadata extraction but also introduce additional overhead such as manual annotation and lengthy training times. Moreover, due to the diversity of citation styles and languages in real-world applications, it is impossible to exhaustively cover all citation styles. Particularly when authors manually add citation data, they may inevitably misuse templates, substantially reducing automatic recognition accuracy. In this sense, neither manually created rules or templates nor those generated through machine learning training possess strong adaptability. Therefore, we aim to improve machine learning algorithms to enhance cross-linguistic adaptability and break free from template usage limitations, thereby increasing automatic annotation accuracy and universality.

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

To address existing problems in citation metadata extraction, this paper proposes an improved feature analysis-based method that eliminates traditional template dependencies and offers cross-linguistic platform advantages. The specific technical implementation roadmap is shown in Figure 1 [Figure 1: see original paper].

The citation data used in experiments primarily originates from the Chinese Social Sciences Citation Index (CSSCI) citation database. Since constructing word vector space models requires word segmentation for Chinese citations, and segmentation quality significantly impacts experimental results, this study mainly employs foreign language citation data for effect testing. Foreign language citation data was obtained primarily by constructing regular expressions to filter the citation data.

Through observation of extensive foreign language citation data, we found that

foreign language citations commonly use symbols such as ". , :" as separators between metadata elements. However, the "." symbol is also frequently used to indicate name abbreviations, tool version numbers, etc. To improve recognition accuracy for metadata separator symbols, the following data preprocessing rules were established in the experiments:

(1) Separator replacement rule: Since citation data often exhibits mixed usage of Chinese and English punctuation, increasing difficulty in recognizing data separators, all punctuation was replaced with English punctuation.

(2) Dot replacement rule: When a dot is preceded by an uppercase letter and followed by an English letter and punctuation, it typically indicates an English name. When a dot is surrounded by single digits, such as "Windows 3.0", it often represents software version numbers. Dots also combine with adjacent words to form abbreviations, such as "St.", "Vol.", "Aug." In these cases, the dot is replaced with a "*" symbol and no longer considered a separator between metadata elements.

3.2 Training Metadata Classification Feature Values

Current neural network-based word vector calculation has achieved excellent results. For example, Mikolov et al. [12] from Google developed an automatic generation technology for dictionaries and terminology tables that can transform one language into another, achieving remarkable success.

In the example considering English and Spanish, word vector spaces E (English) and S (Spanish) were obtained through training. Five words from English—one, two, three, four, five—were selected, with their corresponding word vectors in the left part of Figure 2 [Figure 2: see original paper] denoted as u1, u2, u3, u4, u5. For visualization convenience, Principal Component Analysis (PCA) was applied for dimensionality reduction to obtain corresponding two-dimensional vectors v1, v2, v3, v4, v5. From Spanish, the corresponding words uno, dos, tres, cuatro, cinco were selected and similarly processed with PCA, as shown in the right part (S) of Figure 2.

As shown in Figure 2, the five words occupy similar relative positions in both vector spaces, demonstrating structural similarity between the vector spaces of different languages. This further validates the reasonableness of using distance in vector space to characterize word similarity and indicates that words with similar functions tend to cluster in the same region. Based on this phenomenon, we constructed a word vector space model for words in metadata using neural network models. Similarly, words frequently appearing in the same type of metadata will cluster in the same region. This characteristic suggests that in the vector space model, different types of citation metadata—such as author names, titles, journal names, and dates—will respectively aggregate in different regions of the space, thereby enabling effective automatic indexing of citation metadata and reducing the impact of different language types on classification effectiveness. Since Chinese citations lack obvious word boundaries and must

rely on segmentation software, erroneous segmentation can introduce significant interference to experimental results. Therefore, this experiment primarily uses foreign language citations as examples.

In the experiments, preprocessed training data was first manually identified and annotated. The annotated data served two main purposes: providing training sets for word vector training and for SVM feature analysis classification. Specific examples of manually annotated data are shown in Figure 3 [Figure 3: see original paper].

In Figure 3, annotations were made according to metadata types, with each line representing one metadata type. The categories represented by the annotated metadata and their classification information are shown in Table 1 .

Table 1 Training set annotation description

Represented Classification	Annotation
Journal name or book title	<title>…</title>
Author name	<author>…</author>
Publisher or press	<publisher>…</publisher>
Publication date and page numbers	<date>…</date>

(1) Word Vector Training This paper primarily employs the CBOW model [13-14] to construct word vectors, using segmented metadata as training data for word vector construction. The main idea of this model is to predict the current word Wt given its context Wt-2, Wt-1, Wt+1, Wt+2. Figure 4 [Figure 4: see original paper] illustrates the network structure of the CBOW model, which resembles a neural network architecture and mainly includes three layers: input layer, projection layer, and output layer.

Input Layer: Contains the word vectors of 2c words in Context(w): v(Context(w)1), v(Context(w)2)…v(Context(w)2c) R^m, where m represents the word vector length and c indicates taking c words before and after word w.

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

$$\mathbf{v}_{\text{projection}} = \sum_{i=1}^{2c} \mathbf{v}(\text{Context}(w)_i)$$

Output Layer: Corresponds to a binary tree that uses words appearing in the corpus as leaf nodes, with word frequencies as weights to construct a Huffman tree. This tree contains N leaf nodes (N = |D|), each corresponding to a word in dictionary D.

The objective function typically uses the log-likelihood function shown in equation (2):

$$\mathcal{L} = \log P(w|\text{Context}(w))$$

Using neural network models to construct word vectors offers two main advantages: First, similarity between words can be represented through word vectors. Neural network probabilistic language models assume that "similar" words have similar word vectors, and the probability function is smooth with respect to word vectors—small changes in word vectors result in small changes in probability. Second, vector-based models inherently include smoothing functionality. Since P(w|Context) is non-zero, no additional processing is required.

In the vector space model, word vectors contained in different citation metadata categories—such as author names, titles, journal names, and dates—are relatively concentrated in stable regions, making it possible to use classification algorithms for automatic citation metadata classification and indexing.

(2) Classification Feature Training Since metadata of each category is relatively concentrated in the same spatial region, we performed clustering calculations on word vectors of each category in the training data to find cluster centers—the most representative metadata for each category. The classification of new words is then determined by their distance to each category center in the space model. In the experiments, we organized the training data by category and used word vectors with the K-means clustering algorithm to find the cluster center for each category.

K-means is a commonly used clustering algorithm. For a given dataset containing n d-dimensional data points X = {x1, x2, …, xn}, the algorithm partitions the data into K clusters C = {C1, C2, …, CK}, where each partition represents a class Ck with a class center tk. Euclidean distance is used as the similarity and distance metric to calculate the sum of squared distances from points within each class to the cluster center tk. The clustering objective is to minimize the total sum of squared distances across all clusters.

According to the least squares method and Lagrange principle, the cluster center tk should be the mean of all data points in class Ck. The K-means clustering algorithm starts with an initial K-class partition, then assigns data points to various classes to reduce the total sum of squared distances. Since the total sum of squared distances in K-means tends to decrease as the number of categories K increases (when K = n, J(C) = 0), the minimum value can only be obtained at a certain determined number of categories K. Using clustering algorithms, we can identify the positions (cluster centers) occupied by data with the most characteristic features of each metadata type, and use clustering algorithms to guide metadata classification, reducing data feature dimensions during classification to shorten training time.

Since foreign language citations commonly use ",.:" as separators for citation metadata blocks, and each separator contains data of the same type, the Euclidean distance from the centroid (cluster center) of each segmented part to the

centroids of various classes serves as the classification feature for citation metadata classification. As shown in Figure 5 [Figure 5: see original paper], using centroid-to-centroid distances as classification features can reduce the number of classification features while strengthening feature description.

Because the category of each element in metadata is also importantly related to its position information in the citation, during classification we also use the relative position of each segmentation block in the citation—calculated by dividing the block's position by the total number of segmentation blocks—as the position feature value in classification. Assuming the segmented citation data is represented as S1, S2, …, Sn, the relative position is (i/n). Sample feature collection examples for classification feature training are shown in Table 2 .

Table 2 Sample feature values for SVM

Segmentation	Distance to Cluster 1	Distance to Cluster 2	Distance to Cluster 3	Distance to Cluster 4	Distance to Cluster 5	Distance to Cluster 6	Position
Chatterjee	0.89	0.76	0.23	0.45	0.67	0.34	0.00
Regression and Analysis by Example	0.12	0.23	0.78	0.56	0.34	0.45	0.11
John Wiley & Sons Inc	0.45	0.34	0.23	0.12	0.67	0.89	0.78

By integrating CBOW algorithms, K-means algorithms, and metadata position features, the vector space features of metadata are consolidated, causing metadata of the same category to be distributed in relatively concentrated regions of the vector space, thereby enabling automatic identification and annotation of citation metadata using classification algorithms.

(3) Support Vector Machine Classification SVM represents a major achievement in machine learning research and serves as an important classification algorithm primarily used to solve binary classification pattern recognition problems. Developed based on Statistical Learning Theory (SLT), its core content was proposed by Stitson et al. [15] between 1992 and 1995. The

main advantages of using SVM include: First, SVM specifically addresses finite sample situations, aiming to obtain the optimal solution under existing information rather than merely the optimal value as sample size approaches infinity. Second, the algorithm ultimately transforms into a quadratic optimization problem, which theoretically yields a global optimum, solving the local extremum problem unavoidable in neural network methods. Third, it converts practical problems through nonlinear transformation into high-dimensional feature spaces, constructing linear discriminant functions in high-dimensional space to implement nonlinear discriminant functions in the original space. This special property ensures good generalization capability while cleverly solving dimensionality problems, as algorithm complexity is independent of sample dimensionality.

After comprehensively comparing the characteristics of neural network models and SVM models, this study selected the SVM algorithm for classification training of citation metadata features. For preprocessed citation data, we segmented the data using common metadata separators, calculated the distance from the cluster center of the current segment to the cluster centers of various classes through clustering algorithms, and combined this with the position feature value of the segment as classification features to automatically classify the category to which the segment belongs.

4 Experiments and Results

4.1 Experimental Evaluation Metrics

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

$$\text{Precision} = \frac{\text{Number of correctly extracted information items}}{\text{Number of extracted information items}}$$

$$\text{Recall} = \frac{\text{Number of correctly extracted information items}}{\text{Number of information items in the sample}}$$

$$F1 = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

4.2 Experimental Results

Using 2,000 foreign language citation data entries collected from CSSCI as experimental data, we manually annotated them to serve as the experimental training set. Partial experimental results are shown in Figure 6 [Figure 6: see original paper].

The experimental method achieved good recognition results for publisher names with multiple units and for segmentation units using different citation annota-

tion styles. Through semantic analysis, it accurately annotated time abbreviations such as "Nov" and "Dec," highlighting the advantages of combining semantic approaches for automatic citation metadata annotation over single template-based methods. This avoids the need for continuous template adjustment and increased program complexity when new template formats are added. Furthermore, combining semantic annotation with citation recognition can effectively circumvent real-world phenomena of incorrect separator usage, improving algorithm fault tolerance and universality. However, this method also has certain limitations. For instance, when identifying publication years and page numbers—both composed of numbers with minimal semantic differences—relying solely on semantic models makes metadata differentiation difficult. Combining template methods would yield better results.

4.3 Comparative Experimental Results Analysis

In natural language processing, Hidden Markov Models and Conditional Random Fields are commonly used to solve sequence labeling problems. The currently prevalent model is the Conditional Random Field (CRF). To better highlight the effectiveness of our approach, we conducted comparative experiments using CRF as a reference group. CRF, proposed by Lafferty et al. [16], is an undirected graphical model combining characteristics of maximum entropy models and Hidden Markov Models. In recent years, it has achieved excellent results in sequence labeling tasks such as word segmentation, part-of-speech tagging, and named entity recognition.

Since CRF experiments also require extensive manual annotation, and to reduce annotation workload while considering that both methods achieve high accuracy for author names and numeric dates/page numbers (making it difficult to demonstrate experimental effects), we conducted comparative experiments only on publisher name extraction. In the experiments, we used the Stanford Parser, a syntactic parsing tool from Stanford University's Natural Language Processing Group, as the part-of-speech tagging tool for English citations. When annotating the corpus, we used a five-tag marking pattern, with specific annotation rules shown in Table 3 .

Table 3 Annotation rule examples

Tag	Meaning
B	Beginning of publisher name
C	Continue, name not ended
E	End of publisher name
S	Single-word publisher name
N	Non-publisher name word

The specific annotation format for the CRF training set is shown in Table 4 .

Table 4 CRF training set annotation format

Word	Recognition Sequence Annotation
Ollman	B
Bertell	C
Academy	C
of	C
Marxist	C
Scholarship	C
on	C
American	C
Campuses	E
McGraw	B
Hill	E
Publishing	N
Company	N

The specific comparative experimental results are shown in Figure 7 [Figure 7: see original paper].

The experimental algorithm outperforms standard CRF algorithms in both recall and precision, particularly in recognition accuracy. Since the comparative experiment only used word part-of-speech features as extraction features, this may account for the mediocre results. Classic pattern recognition algorithms like CRF generally require feature extraction before model construction. After extracting numerous features, correlation analysis must be performed to identify the most representative features while removing irrelevant and self-correlated ones. Consequently, feature extraction becomes overly dependent on human experience and subjective judgment, with different extracted features significantly impacting classification performance—even the order of feature extraction affects final results. The experimental algorithm uses word semantic features as classification features and employs SVM for automatic metadata identification, achieving certain effectiveness, particularly for the common problem of name abbreviations in English. The experimental algorithm utilizes fuzzy semantic knowledge, demonstrating strong robustness to distortions in input data space.

Experimental results reveal that the improved citation data element annotation algorithm can significantly improve recognition accuracy. Its advantages manifest in three aspects: strong robustness to input data distortions (e.g., recognition of English abbreviations, including institutional names and publisher abbreviations); high fault tolerance, enabling semantic identification even when incorrect separators are used as metadata delimiters; and strong portability, offering good adaptability for different languages. These three advantages make the experimental method significantly superior to common machine learning algorithms such as CRF. However, the method also has some shortcomings. Since

manual annotation is required to obtain training sets, training data acquisition is relatively time-consuming compared to other algorithms. Moreover, if the training data volume is too small, the word vector model constructed using neural network algorithms may be unreasonable, preventing optimal classification results and reducing recognition precision and recall. For more precise citation data recognition—such as distinguishing publication years and page numbers, both composed of numbers with minimal semantic differences—combining template methods could more effectively improve recognition accuracy. In future metadata automatic recognition experiments, constructing hybrid intelligent recognition algorithms that combine machine learning and rule-based models will achieve better results.

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): 75-93.

[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 roadmap, completed experimental programming, and wrote the paper; Wang Dongbo: Collected and organized training data, refined research methodology, and revised the paper.

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

Supporting Data: Supporting data is self-archived by the authors, E-mail: 18205185622@163.com. [1] Jiang Lin, Wang Dongbo. meteSplit_{SVM}.rar. Implementation of automatic discovery and tagging experimental program for citation metadata. [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.