Identification High Influential Articles by Considering the Topic Characteristics of Articles

MINGYANG WANG¹, SHIJIA JIAO¹, JIAQI ZHANG¹, XIANGRONG ZHANG², AND NA ZHU³

¹College of Information and Computer Engineering, Northeast Forestry University, Harbin 150040, China
²College of Economics and Business Administration, Heilongjiang Institute of Technology, Harbin 150040, China
³Library, Harbin University, Harbin 150086, China

Corresponding author: Mingyang Wang (wmynefu@163.com)

This work was supported in part by the National Natural Science Foundation of China under Grant 71473034 and Grant 717D1063, in part by the Heilongjiang Provincial Natural Science Foundation of China under Grant LH2019G001, in part by the Financial Assistance from Postdoctoral Scientific Research Developmental Fund of Heilongjiang Province under Grant LBH-Q16003, and in part by the Heilongjiang Province Art Planning Project: Research on Discipline Theme Evolution Based on Multi-source Data Fusion under Grant 2019C027.

ABSTRACT The topic of one article reflects its main semantic content, which is also the main guidance for researchers to choose reference literature. In order to explore whether the topic of an article will affect its citation trend in future, this paper establishes a machine learning framework to study the role of topic characteristics in the prediction of future high influential articles. Articles from four different disciplines are collected as experimental samples to verify whether the framework proposed in this paper can be applied to the prediction task in different disciplines. The Latent Dirichlet Allocation (LDA) is used to determine the topic characteristics of sample articles. LDA can map sample articles to current hot topics and generate the mapping probability of sample articles under different hot topics. The maximum mapping probability of the sample article under the hot topics is extracted as the topic feature of the article. Then the feature space for the prediction task is constructed by combining the topic feature and some bibliometrics indices of articles. Three feature selection algorithms, Fisher Score, Relief-F and Spectral Feature Selection (SPEC), are taken to select the important features in the feature space. The prediction performance of these features is finally tested by three classifiers, SVM, KNN and Bagging. The experimental results show that the topic characteristics of article, the early citation characteristics of article, and the reputation of the author are the key factors that determine whether an article can grow into a highly influential one. The important value of topic characteristics in articles’ citation activities shows that the content of the article is an important factor in attracting more citations.

INDEX TERMS High influential articles, topic characteristics, LDA, feature extraction.

I. INTRODUCTION

Scientific research article is an important materialized form of scientific research achievements and serves as an important guidance in evaluating scientific research status of different academic entities, such as countries [1]–[4], institutions [5]–[7] and individuals [8]–[14], etc. Researchers tend to pay more attentions to articles with high academic influence, and try to explore whether there are characteristics that can help identify the influential articles in advance [15]–[27], which makes the research of predicting potential high influential articles become an important issue in the field of bibliometrics. Because it is difficult to define the influence of articles quantitatively, researchers usually use articles’ citation counts to represent the citation influence of them [21], [22], [28]–[30]. Therefore, the prediction of future high influential articles is naturally transformed into the prediction of highly-cited articles in advance.

Tahamta et al. [20] made a comprehensive overview on the work of researchers in predicting of articles’ future citation counts. They detected 198 relevant papers and summarized that the indices from authors [31]–[36], papers [29], [36]–[39], journals [40]–[45], and some alternative metrics [30], [46]–[51] are related to the number
Three classification algorithms of SVM, KNN and Bagging are taken to verify the performance of the selected key features in identifying high influential articles.

II. RELATED WORK

Since Garfield regarded the citation counts as the basic index of bibliometrics, the citation counts provide an important basis for researchers to quantitatively evaluate the impact of scientific research papers [36], [63]–[71], academic institutions [5] and academic individuals [8], [9], [12], [72]. Although many studies have shown that it is difficult to fully reflect the impact of research literature only with its total number of citations, considering the universality and availability of indicators, researchers still tend to use the number of citations to express the influence of an article in the academic community [21], [22], [47], [48], [73]. In recent years, more and more research began to focus on what kind of articles can grow into influential ones [74]–[77]. Researchers extracted features from the authors of the article [74], [76], [78]–[82], the bibliometrics indices of the article [37], [39], [83]–[85], the journal where the article was published [44], [86]–[88], the attention [89] and the download [51], [90], [91] of the article in website and other web usage indexes [24], [48], [92], [93], so as to identify papers which will obtain high academic influence in future.

Among the author related characteristics, researchers pointed out that the author’s reputation [21], [22], [24], [94]–[96], the number of authors [17], [84], [97]–[99] and the academic ranking of authors [100] have important impacts on the citation behavior of articles. In addition, the author’s gender [31] and whether there is international cooperation among authors [6], [17], [84], [88], [96], [98], [101], [102] have also been widely studied. Among the indexes related to journals, journal impact factor is the most widely used feature for prediction [17], [84], [88]. The open access [40], [103], the accessibility of periodicals [40], [96], the ranking of periodicals [98], [100], [104], the reputation of periodicals [87], [99] and other factors [76], [105] also affect the citation activity of articles to a certain extent. Researchers also discussed the relationship between the indicators of the article itself and the citation activity. These indicators include the length of the research article [87], [97], [106], the type of the article [107], the length of the title and abstract [108], the number and diversity of the references in...
article [17], [19], [84], the time characteristics of the article [75], [109], [110], and the web usage statistic indicators of the article [24], [30], [91], [111]. They pointed out that the scientific research work in developed countries [109], the relatively short title and abstract of article [112], the diversity of the references [108] play important roles in attracting more citations. In our previous studies, we discussed the influence of articles’ early citation behavior on their subsequent citations [21]–[23]. It is found that if the paper can be cited by more countries, institutions, journals and disciplines in a short period of time after publication, the paper will grow into a highly influential paper to a large extent. Although the above studies show widely discussion on the factors that affect the citation from different perspectives, most of these factors are related to the external characteristics of the article, and do not involve the content of it. In recent years, some researchers discussed the role of the content of the article in its citation behavior, but these studies mainly analyzed the popularity [100], variety [101], novelty and diversity [100], [113] of the article, and the exploration of article’s semantic content is still not enough. The semantic content of the article embodies the research theme of the article, which is of great significance for researchers to choose references. The content of the article determines whether the article can arouse the interest of researchers, and then stimulate the cited behavior.

In recent years, with the development of natural language processing technology and deep learning technology, researchers proposed many technologies and models that can effectively extract the characteristics of text content [100], [113], [114]. Among them, LDA model is one of the most widely used method that can effectively explore the theme of documents [55]. LDA can transform the text into a three-tier structure of “Document-Topic-Word”. By extracting the “Document-Topic matrix” in the three-tier structure, the topic distribution of the text could be obtained. In the field of citation analysis, LDA model has been widely used in topic detection,[56]–[58], hotspot discovery [59]–[60], automatic document construction [61] and article recommendation [62].

In this paper, the LDA model is introduced to extract and generate the topic features of the article. On the basis of the extracted topic features and some commonly used bibliometrics features, this paper establishes a machine learning framework to test the value of articles’ topic characteristics in predicting their citation trend in future.

### III. DATA SET

This paper selects articles published in 2010 as experimental samples, and uses the total citation counts of these articles obtained from the beginning of publication to 2018 as the representation of the future citation influence of these articles. In order to ensure the universality of the experimental results, this paper collects the article data from four different fields: Physics, Chemistry, Oncology and Neurology. One journal from each field is selected as the representative of the corresponding field. Table 1 shows the information of the selected articles. There are total 421 articles, including 1792 authors, are used in our experiments.

The reasons for choosing these four fields are the following. Physics and Chemistry have strong interdisciplinary characteristics with other disciplines, while Oncology and Neurology are relatively professional disciplines with relatively low interdisciplinary characteristics with other fields. Choosing articles from these four fields can better explore the common characteristics that affect the citation behavior of articles. In order to meet the needs of classification tasks, the selected articles from each field are divided into three kinds of highly-cited papers (HCPs), medium-cited papers (MCPs) and low-cited papers (LCPs) according to their total citations.

1) HCPs: Articles whose total citation counts is at least 7 times the average citations of articles in the same journal are selected as HCPs. The choice of multiple seven mainly ensures that an appropriate number of articles are selected as HCPs. According to this standard, in each journal, the selected HCPs happen to be the top 1% articles in the journal according to articles’ total number of citations, which conforms to the definition of highly-cited papers in the Web of Science (WoS). In Web of Science, papers received enough citations to place it in the top 1% in the same subject area and in the same publication year are classified as highly-cited papers.

2) LCPs: Articles whose total citation counts is less than the average citations of articles in the same journal are taken as LCPs.

| ID | Journal                        | Journal Abbreviation | Research field | Number of HCPs | Number of MCPs | Number of LCPs | Total  |
|----|--------------------------------|----------------------|----------------|----------------|----------------|----------------|--------|
| 1  | Applied Physics Letters       | APL                  | Physics        | 19             | 43             | 61             | 123    |
| 2  | Journal of the American       | JACS                 | Chemistry      | 20             | 40             | 60             | 120    |
| 3  | Lancet Oncology               | LO                   | Oncology       | 12             | 24             | 48             | 84     |
| 4  | Lancet Neurology              | LN                   | Neurology      | 5              | 15             | 74             | 94     |
| Total | -                          | -                    | -              | 56             | 122            | 273            | 421    |
3) MCPs: The articles whose citation counts located between the ones for HCPs and LCPs in the same journal are taken as MCPs.

Due to the great differences in the topics involved in each article, it is difficult to establish the relationship between a specific topic of the article and its future citation trend, and it is also difficult to draw general conclusions in different articles. Therefore, this paper does not examine the specific theme of the article itself, but on the basis of examining the relationship between the theme of the article and the current research hotspots, generates the topic characteristics of the sample article. And to do this, it is necessary to extract the research focus in the field of sample articles before determining the topic characteristics of sample articles. This paper selects highly-cited papers published in 2005-2009 from the representative journal in each field, and extracts the research topics reflected by these highly-cited papers through LDA model, which are regarded as the current research hotspots in the domain of this journal. These highly-cited papers are still defined as papers whose citation counts is at least 7 times of the average citation frequency of journal papers. After determining the current research hotspots in each field, the LDA model maps the work of the sample articles to the research hotspots and obtains the mapping probability of the topic of sample articles under different research hotspots. Then the topic characteristics of the sample articles are generated based on the results of the mapping probability.

IV. METHOD

A. CONSTRUCTING FEATURE SPACE FOR PREDICTING FUTURE HIGH INFLUENTIAL ARTICLES

Previous studies have shown that some commonly used bibliometric features have an important impact on the citation behavior of articles. This paper first extracted these commonly used bibliometric features from sample articles. At the same time, by using LDA model, this paper extracts the topic feature of sample articles. Combined with the topic feature and the extracted bibliometrics features, the feature space for predicting high influential articles is established. Each feature in the feature space is normalized to the range of [0.01, 0.99] to eliminate the influence of the value range of different features on classification.

1) BIBLIOMETRIC FEATURES

Twenty bibliometric indicators are collected for the sample articles. These bibliometric indicators do not involve the content of the article, but all belong to the external factor indicators collected from the statistical point of view. Table 2 shows the information on these 20 bibliometric indicators.

Indices \{x_1 - x_6\} describe the citation characteristics two years after publication. These characteristics give the early feedback of the scientific community to the paper. Indices \{x_7 - x_{13}\} discuss the features related to authors. Indices \{x_7 - x_8\} are the h-index of authors to represent the reputation of them. Indicator \{x_{10}\} is the maximum number of citations of all the authors in the last 5 years before the publication of this paper. These indicators are used to indicate the influence and reputation of the authors in the corresponding domain. Indices \{x_{11} - x_{13}\} discuss the authors’ affiliation characters. Indices \{x_{14} - x_{20}\} give the features of the article itself, such as the number of references, the length, the language and the type of the article.

By using the functions of “Create Citation Report” and “Analyze the Indexing Results” in Web of Science, it is easy to collect these bibliometrics indices.

2) TOPIC FEATURE

In order to extract the topic feature of articles, this paper first uses LDA to train the highly-cited articles published in 2005-2009 to get the distribution of current hot topics. During the training process, the parameters of LDA model are set to Gibbs = 200, Dirichlet hyper-parameter β = 0.1, α = 50/N (N is the dimension of topics). By inputting the sample articles into the trained LDA model, and calculating the mapping probability of each article under the current hot topics, the topic characteristics of the sample articles are obtained on this basis.

In order to determine the dimension of topics, the topic coherence score is calculated to examine the coherence of the model under different number of topics. The topic coherence captures the semantic interpretability of a topic based on its corresponding descriptor terms (for example, the first n topic words with the largest probability in the topic estimated by LDA) [115]. It was shown that there is a good correlation between topic coherence and human expert labeling [116]. The higher the topic coherence score is, the higher the semantic interpretability of the documents assigned to the same topic, the more reasonable the result of topic division is. The coherence score is calculated as follows:

\[
\text{coherence}(V) = \sum_{(v_i, v_j) \in V} \text{score}(v_i, v_j) \quad (1)
\]

\[
\text{score}(v_i, v_j) = \frac{1}{D(v_i) + 1} \quad (2)
\]

where \(V\) is a topic described by a set of words, \(\{v_i\}\) represents the topic words. \(D(v_i)\) counts the number of documents containing word \(v_i\), \(D(v_i, v_j)\) counts the number of documents containing words \(v_i\) and \(v_j\).

We calculate the coherence score of each topic and average them to get the average coherence score of all topics. By changing the number of topic dimensions, we can get the average score of topic coherence under different topic dimensions. According to the comparison of the average coherence scores of different topic dimensions, the topic dimension with the highest score is selected for the final dimension.

Figure 2 shows the results of the average coherence score of the model under different number of topics. It shows that when the number of topics is 10, the average coherence score of topics reaches a small peak, and then tends to be flat with
TABLE 2. The 20 bibliometric indicators.

| Indices | Detailed information |
|---------|----------------------|
| $X_1$   | Total citation counts obtained in the first 2-years after publication |
| $X_2$   | Number of citing subjects in the first 2-years after publication |
| $X_3$   | Number of citing journals in the first 2-years after publication |
| $X_4$   | Number of citing countries in the first 2-years after publication |
| $X_5$   | Number of citing institutions in the first 2-years after publication |
| $X_6$   | The first-cited-age of the article |
| $X_7$   | The maximum $h$-index of the authors |
| $X_8$   | The $h$-index of the first author |
| $X_9$   | Number of authors |
| $X_{10}$| The maximum total citation counts of the authors before publishing the paper |
| $X_{11}$| Whether there is international cooperation |
| $X_{12}$| Whether there is American |
| $X_{13}$| Whether there is team cooperation |
| $X_{14}$| Number of subject distribution of references |
| $X_{15}$| Number of references |
| $X_{16}$| Number of cooperating countries |
| $X_{17}$| Location of research |
| $X_{18}$| Length of the article |
| $X_{19}$| Language of the article |
| $X_{20}$| Type of the article |

Figure 3 shows the probability distribution of some sample articles under the 10 dimensional research hotspots. This paper mainly discusses whether the relationship between the topic of the article and the current research hotspots will affect the future citation behavior of the article. Therefore, we take the maximum membership degree of each sample article under the 10 dimensional research hotspots as the topic feature of it, which is labeled as $Topic$, and added it to the feature space.

B. FEATURE SELECTION

In order to overcome the “dimensionality curse” [117] which may probably existed in the feature space, a feature selection process is applied to find the subset of features which maintain the essential characteristics of the dataset. Considering that each feature selection technique may be biased to some features due to their initial mechanisms, three different selection techniques of Fisher Score, Relief-F and SPEC method are used to detect the significance of each feature and verify whether the core feature subset can achieve good results under different classification algorithms.

1) FISHER SCORE

Fisher Score calculates the variance of samples of different categories to determine the contribution of one feature to
the classification. If a feature has a strong ability to distinguish categories, the variance difference of samples in the same category should be as small as possible, and the variance difference between different categories of samples should be as large as possible.

\[ F(f_i) = \frac{\sum_{k=1}^{C} n_k (\mu_{f_i}^k - \mu_i)^2}{\sum_{k=1}^{C} \sum_{j=1}^{n_k} (j_{f_i,j} - \mu_{f_i}^k)^2} \]  

(3)

where \( \mu_{f_i}^k \) expresses the average of the \( i \)-th feature \( f_i \) in the sample, \( \mu_i \) represents the mean of the \( i \)-th feature \( f_i \) in the sample in the \( k \)-th category. \( n_k \) describes the number of samples in the \( k \)-th category, and \( j_{f_i,j} \) gives the value of the \( i \)-th feature in the \( j \)-th sample.

Based on Equation (3), we calculate the value of Fisher Score for each feature to represent the contribution of each feature to classification. All the features are ranked by their Fisher Score in descending order.

2) RELIEF-F

Relief-F [118] weights features by calculating their ability to distinguish different types of samples. Relief-F randomly selects one sample from the sample set, and then finds its \( k \) nearest neighbor samples in the same category and each different category as the random sample. It calculates the ability of each feature to distinguish random samples and their neighborhood samples, and gives more weight to the features with higher ability to distinguish.

To be specific, for the randomly selected sample \( x_i \), it first finds \( k \) nearest neighbor samples \( \{h_j\} \) from the homogeneous sample set \( C \), and then finds \( k \) neighboring samples \( \{m_j\} \) in the sample set \( S \) which belongs to different class from \( x_i \). Finally, Relief-F calculates the probability estimation of the above formula and assigns it as the weight of feature \( A \). Where \( p(C) \) is the probability of class \( C \), and \( p(S) \) is the probability of the other class \( S \) besides \( C \).

\[
W(A) = W(A) - \sum_{j=1}^{k} \frac{\text{diff}(A, x_i, h_j)}{(nk)} + \sum_{S \neq C} \left[ \frac{p(C)}{1 - p(S)} \sum_{j=1}^{k} \text{diff}(A, x_i, m_j) \right] / (nk)
\]

\[
diff(A, x_i, x_j) = \begin{cases} |x_i[A] - x_j[A]| / \max(A) - \min(A) & A \text{ is continuous} \\ 0 & A \text{ is discrete and } x_i[A] = x_j[A] \\ 1 & A \text{ is discrete and } x_i[A] \neq x_j[A] 
\end{cases}
\]

(4)

Taking the order of weights from large to small, we can get all the features ranked.

3) SPEC

SPEC selects features based on “spectrum theory”, and uses mutual information to calculate the correlation between features and categories [119]. The basic idea of SPEC is that a good feature should have a structure similar to that of a graph composed of initial data. SPEC calculates the correlation between the initial features by evaluating the consistency of the features of the spectral matrix derived from the similarity matrix between samples. For \( n \) samples \( X = (x_1, x_2, \ldots, x_n) \) in the data sets \( X \), \( \{f_1, f_2, \ldots, f_m\} \) are \( m \) features of the data sets, \( \{f_1, f_2, \ldots, f_m\} \) are \( m \) feature vectors. Following the computational procedure, the weights of all features could be computed:

Step 1: Calculate the similarity matrix \( S \) of \( X \):

\[
S_{ij} = e^{-\frac{|x_i[A] - x_j[A]|^2}{\sigma^2}}
\]

(5)

Step 2: Further acquire the representation \( G \) of the graph of \( S \), and calculate the adjacency matrix \( W \) of \( G \):

\[
w_{ij} = S_{ij}
\]

(6)

\[
W(i, j) = w_{ij}
\]

(7)

Step 3: Compute the degree matrix \( D \) of \( G \):

\[
d_i = \sum_{k=1}^{n} w_{ik}
\]

(8)

\[
D(i, j) = \begin{cases} d_i & \text{if } i = j \\ 0 & \text{otherwise} 
\end{cases}
\]

(9)

Step 4: Compute the feature weights for each feature vector \( f_i \):

\[
\hat{f}_i \sim \frac{D^2 f_i}{||D^2 f_i||}
\]

(10)
Step 5: Sort each feature using the ranking function $\tilde{\varphi}(x)$:

$$SF_{SPEC}(i) \leftarrow \tilde{\varphi}(F_i)$$  (11)

Through the above steps, we can get the feature set arranged in the order of feature weight from large to small. According to the ranking results in each of the feature selection algorithm, we take the intersection of the top 10 features in each algorithm as the key features. These features are of great value for identifying future high influential articles.

C. CLASSIFICATION MODELS

Three classification models of support vector machine (SVM), $k$-nearest neighbor classifier (KNN) and decision tree-based Bagging classifier are used to test the classification performance of the selected key features. SVM tends to find the best separating hyper-plane in the feature space, so as to maximize the interval between positive and negative samples in the training set [120]. The classical SVM algorithm is a supervised binary learning algorithm. The multiclassification problem can be solved by the combination of SVM classifiers [121]. In this paper, Gaussian function is used as the kernel function of SVM. The hyper-parameters of Gaussian function is set to $\gamma = 9$, $c = 5$.

KNN first determines the $k$ nearest neighbors of the samples to be classified in the feature space, and assigns the category of most of the $k$ nearest neighbors to the samples to be classified [122]. In this paper, the parameter $k$ of KNN classifier is set to 12.

Bagging is a classifier based on decision tree. It is a parallel integrated learning method, which uses multiple trees for training and prediction, and outputs prediction value combined with training results [123]. In this paper, ten decision trees are integrated in Bagging.

In “Supplementary Material”, we show the detailed discussion on determining the optimal parameters for the three classifiers.

TABLE 3. Feature selection results.

| Feature selection techniques | Feature subsets |
|------------------------------|-----------------|
| Fisher Score                 | $\{x_1, x_2, x_3, x_4, x_5, x_6, x_7, x_8, x_9\}$ |
| Relief-F                     | $\{Topic, x_2, x_3, x_4, x_5, x_6, x_7, x_8\}$ |
| SPEC                         | $\{Topic, x_2, x_3, x_4, x_5, x_6, x_7, x_8\}$ |

V. EXPERIMENTAL RESULTS AND DISCUSSION

Table 3 shows the top 10 features selected under three feature selection algorithms. It can be seen that topic feature Topic and seven bibliometrics features of $\{x_1, x_2, x_3, x_4, x_5, x_6, x_7, x_8\}$ appear in the feature subset obtained by three different feature selection algorithms, which shows that these 8 features are more important than other features in determining the future citation trend of the article. Among them, the topic feature Topic reflects the semantic content of the article. The features of $\{x_1, x_2, x_3, x_4, x_5\}$ give the early citation characteristics of the article, which are denoted as Early-Citation features in this paper. The features of $x_7, x_8$ give the $h$-index of the author to represent the academic reputation of the author, which are denoted as Author-Reputation features in this paper.

In addition, the sample articles are selected from different fields, which have different interdisciplinary characteristics. Some of the sample articles come from the fields of Physics and Chemistry. These two areas have a high degree of intersection with other areas. The rest of the articles are from Oncology and Neuroscience. These two fields belong to the fields with strong specialization, and they have a low interdisciplinary with other fields. In order to explore the applicability of the framework proposed in this paper to articles from different interdisciplinary areas, the above three feature selection algorithms are used to process articles from two kinds of fields respectively and identify the key features in different kinds of fields. Under each algorithm, according to the order of feature weight calculated by the algorithm, the top 10 features are selected.

TABLE 4. Feature selection results in fields with high interdisciplinary.

| Feature selection techniques | Feature subsets |
|------------------------------|-----------------|
| Fisher Score                 | $\{x_1, x_2, x_3, x_4, x_5, x_6, x_7, x_8\}$ |
| Relief-F                     | $\{x_1, x_2, x_3, x_4, x_5, x_6, x_7, x_8\}$ |
| SPEC                         | $\{x_1, x_2, x_3, x_4, x_5, x_6, x_7, x_8\}$ |

TABLE 5. Feature selection results in fields with low interdisciplinary.

| Feature selection techniques | Feature subsets |
|------------------------------|-----------------|
| Fisher Score                 | $\{x_1, x_2, x_3, x_4, x_5, x_6, x_7, x_8\}$ |
| Relief-F                     | $\{x_1, x_2, x_3, x_4, x_5, x_6, x_7, x_8\}$ |
| SPEC                         | $\{x_1, x_2, x_3, x_4, x_5, x_6, x_7, x_8\}$ |

TABLE 6. Confusion matrix.

| Output is positive | Positive samples | Negative samples |
|--------------------|------------------|------------------|
| TP                 |                  | FN               |
| FP                 |                  | TN               |

Table 4 and Table 5 show the results of feature selection in the two kinds of fields respectively. The results show that for
TABLE 7. Classification performance of the key features.

|       | APL | JACS | LO  | LN  | Ave. |
|-------|-----|------|-----|-----|------|
| SVM   | 0.94| 0.97 | 0.96| 0.97| 0.96 |
| KNN   | 0.93| 0.94 | 0.92| 0.96| 0.938|
| Bagging | 0.95| 0.96 | 0.95| 0.97| 0.959|
| Ave.  | 0.94| 0.956| 0.943| 0.967| -   |

Table 7 shows the classification performance of three classifiers in each journal. The last column in Table 7 shows the average classification performance of the three classifiers under each journal. The last row gives the average classification performance of each classifier for articles in different journals.

In order to verify the value of selected key features in identifying high influential articles, this paper uses three classifiers of SVM, KNN and Bagging to test the classification performance of these features. The classification performance is determined by 10 times cross validation. In this study, AUC under ROC curve is used to evaluate the classification model. ROC curve is drawn with FPR (False Positive Rate) as horizontal axis and TPR (True Positive Rate) as vertical axis [124]. AUC value gives the area under ROC curve [125]. The larger AUC value is, the better the classification performance of the model is. Table 6 shows the confusion matrix used to calculate FPR and TPR.

Figure 4 shows the results in Table 7 more intuitively in the form of histogram. It can be seen that the key features extracted in this paper have achieved good classification performance for sample articles in different fields. Under each journal and each classifier, the average classification performance represented by AUC is above 0.9, which proves that these features can effectively identify and distinguish articles with different citation influences.

Among the three classifiers used in this paper, SVM achieves better classification results. We believe that there are two reasons. Firstly, SVM can map the data to high-dimensional space, and construct the optimal separation hyperplane in high-dimensional feature space, so it can realize the separation of the indivisible data in the plane. Secondly, by adjusting the kernel functions and selecting the appropriate parameters, SVM can still have strong robustness even if there is a certain deviation in the training samples. These two points should be the reason why SVM is superior to other classifiers.

TABLE 8. Classification performance after removing features.

|       | APL | JACS | LO  | LN  | Initial |
|-------|-----|------|-----|-----|---------|
| SVM   | -Early-citation | -Topic | -Author-Reputation | 0.74 | 0.83 | 0.87 | 0.94 |
| KNN   | 0.87 | 0.9   | 0.89 | 0.97 |
| Bagging | 0.86 | 0.88 | 0.89 | 0.96 |
| Ave.  | 0.75 | 0.77 | 0.85 | 0.97 |

In order to further verify the role of the three kinds of key features in the recognition of high influential articles, this paper successively removes these three types of features, and uses the best classifier, SVM, to test the classification performance of the feature subset after feature removal. Table 8 shows the classification performance after excluding features. In Table 8, the minus sign ‘‘−’’ in front of the features indicates that deletion of the corresponding feature, and the values in the table give the classification performance after deleting the corresponding feature.

It can be seen that after deleting one of the three key features, the classification performance of the four journals is greatly reduced. Among them, the classification performance of the four journals decreases the most after excluding the Early-citation features and the AUC value of the four journals decreases by 10%-20%. The AUC value of two journals of APL and LN even decreases from 0.94 and 0.97 to 0.74 and 0.75, respectively. This shows that the Early-citation features of the article is a very important feature to determine whether it has a high influence in the future. The role of Topic feature is second only to that of Early-citation features. The AUC value decreases significantly after the
Topic feature is deleted. Especially in journal LN, its classification performance decreased from 0.97 to 0.77, which shows that the topic feature is also an important feature that affects the article’s citation trend in the future. After removing the Author-Reputation features from the key feature subset, the classification performance of the four journals also decreases significantly, which indicates that the Author-Reputation features also have a great impact on the citation behavior of articles.

Therefore, the topic feature, the early citation features and the author’s reputation are the key factors that affect the future citations of the article. Basing on these three indicators, we can effectively identify the articles with high influence in the future.

The topic characteristics of the article reflect the semantic content of the article. The important value of topic feature on the citation behavior shows that, in addition to some external bibliometrics characteristics from the statistical perspective, the content of the article also has a great impact on the future citation trajectory of the article. To some extent, this conclusion reflects the researchers’ reading and research habits. When researchers make research directions and choose reading literature, they usually first understand the current research hot spots, and then choose articles that meet their own research interests to read. The role of topic feature in citation behavior reflects researchers’ preference for current research hot spots.

The early citation features of the article represent the academic feedback on the article in a short time after publication. If an article can be cited by more countries, institutions, disciplines, journals, etc. in a short time after publication, it means that the knowledge carried by the article is widely recognized and accepted by the academic community. The wide spread of knowledge brought by the citation activity will bring more opportunities for the article to be cited, and make the article more likely to grow into a highly influential document in the future [69]. The conclusion of this paper is consistent with that of some previous researches [21], [23], [94]–[96], [126].

The author’s $h$-index was proposed in 2005 by Hirsch to assess the academic level of scientists [127]. The index is proved quite useful and widely used in quantifying an individual’s scientific output [128]. The high $h$-index indicates that researchers have high academic level and high academic reputation. The important influence of the author’s $h$-index on the articles’ citation activity shows that the articles published by the authors with high $h$-index will receive more attentions, which will bring many potential opportunities for citation. This conclusion is consistent with the research results in some previous work [21]–[23], [94]–[96].

VI. CONCLUSIONS

The purpose of this paper is to explore which characteristics of the article will produce important value for its future citation behavior. Compared with the previous works of determining the influencing factors of citation, this paper focuses on the influence of the topic characteristics of the article on its citation trend.

There are three main contributions in this paper. 1) The LDA model is introduced to generate the topic feature of the sample articles. By training the highly-cited papers at present and the sample articles, LDA model can map the content of the sample articles to the current research hot spots determined by the highly-cited articles. Based on the probability distribution of the content of sample articles in the research hot spots, the maximum membership degree of each sample article under the research hot spots are extracted as its topic feature, which lays the foundation for the subsequent recognition of high influential articles. 2) On the basis of the extracted topic feature and some commonly used bibliometrics features, the feature space is constructed to identify the future high influential articles. Three feature selection algorithms are used to extract the key features which are of great value to distinguish articles with different citation influences. 3) Three classifiers are utilized to test the classification performance of the key features to verify whether these features, especially the topic feature, are valuable to identify the high influential articles in the future.

In order to get more reliable experimental results, this paper selects articles in different fields as experimental samples. The experimental results show that the characteristics of early citations, topics and authors’ reputation are the key features that affect articles’ citations. The topic characteristics of the article is very helpful to attract more citations, and is also very important to the identification of high influential papers in the future. As the topic of the article reflects its content, it shows that in addition to some external bibliometrics characteristics of the article, the content of the article will also have a great impact on its citation behavior, and to a certain extent determines whether the article can grow into an influential one in the future.

The work of this paper is to study the influence of the content of the article on its future citation behavior. Although the conclusion of this paper shows that the content characteristics of the article do have an important impact on its future citation behavior, the conclusion is a little rough because the content of the article is only based on its topic. In the following work, we will continue to explore how to better quantify the content characteristics of the article to further investigate the value of the content in its citation behavior.

REFERENCES

[1] R. J. W. Tijssen, “Africa’s contribution to the worldwide research literature: New analytical perspectives, trends, and performance indicators,” *Scientometrics*, vol. 71, no. 2, pp. 303–327, May 2007.

[2] K. Jonkers and R. Tijssen, “Chinese researchers returning home: Impacts of international mobility on research collaboration and scientific productivity,” *Scientometrics*, vol. 77, no. 2, pp. 309–333, Nov. 2008, doi: 10.1007/s11192-007-0017-1-x.

[3] H. Li and Y. Ho, “Use of citation per publication as an indicator to evaluate pentachlorophenol research,” *Scientometrics*, vol. 75, no. 1, pp. 97–110, 2008.
[4] W. Glänzel, B. Thijs, and K. Debackere, “The application of citation-based performance classes to the disciplinary and multidisciplinary assessment in national comparison and institutional research assessment,” *Scientometrics*, vol. 101, no. 2, pp. 939–952, Nov. 2014.

[5] J. Bar-Ilan, “A microscopic link analysis of academic institutions within a country—the case of israel,” *Scientometrics*, vol. 59, no. 3, pp. 391–403, 2004.

[6] L. Bornmann, M. Stefaner, F. de Moya Anegón, and R. Mutz, “What is the effect of country-specific characteristics on the research performance of scientific institutions? Using multi-level statistical models to rank and map universities and research-focused institutions worldwide,” *J. Informetrics*, vol. 8, no. 3, pp. 581–593, Jul. 2014.

[7] L. Bornmann and W. Glänzel, “Applying the CSS method to bibliometric indicators used in (University) rankings,” *Scientometrics*, vol. 110, no. 2, pp. 1077–1079, Feb. 2017.

[8] L. L. Hargens and H. Schuman, “Citation counts and social comparisons: Scientists’ use and evaluation of citation index data,” *Social Sci. Res.*, vol. 19, no. 3, pp. 205–221, Sep. 1990.

[9] M. Moin, M. Mahmoudi, and N. Rezaei, “Scientific output of Iran at the threshold of the 21st century,” *Scientometrics*, vol. 62, no. 2, pp. 239–248, Jan. 2005.

[10] W. Glänzel, S. Heefer, and B. Thijs, “A triangular model for publication and citation statistics of individual authors,” *Scientometrics*, vol. 107, no. 2, pp. 857–872, May 2016.

[11] B. Thijs, K. Debackere, and W. Glänzel, “Improved author profiling through the use of citation classes,” *Scientometrics*, vol. 111, no. 2, pp. 829–839, May 2017.

[12] D. H. Lee, “Predicting the research performance of early career scientists,” *Scientometrics*, vol. 121, no. 3, pp. 1481–1504, Dec. 2019.

[13] M. Wang, S. Jiao, K.-H. Chai, and G. Chen, “Building journal’s long-term impact: Using indicators detected from the sustained active articles,” *Scientometrics*, vol. 121, no. 1, pp. 261–283, Oct. 2019.

[14] J. Zhang, B. Xu, J. Lii, A. Tolba, Z. Al-mahkadmeh, and F. Xia, “PePSI: Personalized prediction of Scholars’ impact in heterogeneous temporal academic networks,” *IEEE Access*, vol. 6, pp. 55661–55672, 2018.

[15] G. D. Walters, “Predicting subsequent citations to articles published in twelve crime-psychotherapy journals: Author impact versus journal impact,” *Scientometrics*, vol. 69, no. 3, pp. 499–510, 2006.

[16] L. Bornmann and H.-D. Daniel, “Citation speed—a measure to predict the attention an article receives: An investigation of the validity of editorial decisions at angewandte chemie international edition,” *J. Informetrics*, vol. 4, no. 1, pp. 83–88, Jan. 2010, doi: 10.1016/j.joi.2009.09.001.

[17] N. Haslam and P. Koval, “Predicting long-term citation impact of articles in social and personality psychology,” *Psychol. Rep.*, vol. 106, no. 3, pp. 891–900, Jun. 2010.

[18] K.-Y. Chuang, M.-H. Wang, and Y.-S. Ho, “High-impact papers predicted through the use of citation classes,” *Scientometrics*, vol. 87, no. 3, pp. 551–562, Jun. 2011.

[19] T. Chakraborty, S. Kumar, P. Goyal, N. Ganguly and A. Mukherjee, “Towards a stratified learning approach to predict future citation counts,” in *Proc. ACM/IEEE Joint Conf. Digit. Libraries (JCDL)*, Feb. 2014, pp. 351–360.

[20] I. Tahamtan, A. Safipour Afshar, and K. Ahmazadeh, “Factors affecting number of citations: A comprehensive review of the literature,” *Scientometrics*, vol. 107, no. 3, pp. 1195–1225, Jun. 2016, doi: 10.1007/s11192-016-1889-2.

[21] M. Wang, G. Yu, S. An, and D. Yu, “Discovery of factors influencing citation impact based on a soft fuzzy rough set model,” *Scientometrics*, vol. 93, no. 3, pp. 635–644, Dec. 2012.

[22] M. Wang, G. Yu, J. Xu, H. He, D. Yu, and S. An, “Development a case-based classifier for predicting highly cited papers,” *J. Informetrics*, vol. 6, no. 4, pp. 586–599, Oct. 2012.

[23] M. Wang, Z. Wang, and G. Chen, “Which can better predict the future success of articles? Bibliometric indices or alternative metrics,” *Scientometrics*, vol. 119, no. 3, pp. 1575–1595, Jun. 2019.

[24] L. Waltman, “A review of the literature on citation impact indicators,” *J. Informetrics*, vol. 10, no. 2, pp. 365–391, May 2016.

[25] T. Yu, G. Yu, P.-Y. Li, and L. Wang, “Citation impact prediction for scientific papers using stepwise regression analysis,” *Scientometrics*, vol. 101, no. 2, pp. 1233–1252, Nov. 2014, doi: 10.1007/s11192-014-1279-6.
M. Wang et al.: Identification High Influential Articles by Considering the Topic Characteristics of Articles
M. So, J. Kim, S. Choi, and H. W. Park, “Factors affecting citation
P. D. Turney and P. Pantel, “From frequency to meaning: Vector
B. McGillivray and M. Astell, “The relationship between usage and
M. D. Ginsberg, “Introspection: An analysis of the citation impact of
J. Xie, K. Gong, Y. Cheng, and Q. Ke, “The correlation between paper
M. E. Falagas, A. Zarkali, D. E. Karageorgopoulos, V. Bardakas, and
M. Wang, J. Zhang, G. Chen, and K.-H. Chai, “Examining the influ-
A. Annalingam, H. Damayanthi, R. Jayawardena, and P. Ranasinghe,
M. E. Falagas, A. Zarkali, D. E. Karageorgopoulos, V. Bardakas, and
C. J. C. Jabbour, A. B. L. D. S. Jabbour, and J. H. C. de Oliveira, “The
T. F. Frandsen and J. Nicolaisen, “The ripple effect: Citation chain
I. Ayres and E. Vars, “Determinants of citations to articles in elite
W. Glänzel, R. Rousseau, and L. Zhang, “A visual representation of
A. Perianes-Rodríguez and C. Olmeda-Gómez, “Effects of journal
T. Fawcett, “An introduction to ROC analysis,” Pattern Recognit. Lett,
C. X. Ling, H. Jin, and H. Zhang, “AUC: A better measure than accuracy
T. Hastie and R. Tibshirani, “Discriminant adaptive nearest neighbor
to SVM multi-class classifier,” J. Informetrics, vol. 5, no. 4, pp. 583–593,
M. Wang, B. Thijs, and W. Glänzel, “Characteristics of interna-
J. Adams, “Early citation counts correlate with accumulated impact,” Sci-
J. E. Hirsch, “An index to quantify an individual’s scientific research
I. Kononenko, “Estimating attributes: Analysis and extensions of
K. Stevens, P. Kegelmeyer, D. Andrzejewski and D. Buttler, “Exploring
eat Forestry University, China. Her main areas
She is currently an Associate Professor with the
SHIJIA JIAO received the B.S. degree from Liaoning Shihua University, China, in 2017. She is currently pursuing the M.S. degree with the College of Information and Computer Engineering, Northeast Forestry University. She has a number of publications in international journals and conferences. Her current research interests include text mining, knowledge management, and scientometrics. She is a reviewer of several international journals.

MINGYANG WANG received the Ph.D. degree from the School of Economics and Management, Harbin Institute of Technology, China, in 2011. She is currently an Associate Professor with the College of Information and Computer Engineering, Northeast Forestry University. She has a number of publications in international journals and conferences. Her current research interests include text mining, knowledge management, and scientometrics. She is a reviewer of several international journals.

[94] P. S. Pagel and J. A. Hudetz, “Scholarly productivity of united states aca-
demic cardiothoracic anesthesiologists: Influence of fellowship accred-
itation and transesophageal echocardiographic credentials on h-index and other citation bibliometrics,” J. Cardiovascular Anaesthesia, vol. 25, no. 5, pp. 761–765, Oct. 2011.
[95] F. Frandsen and J. Nicolaisen, “The ripple effect: Citation chain
[96] M. Wang, J. Zhang, G. Chen, and K.-H. Chai, “Examining the influ-
[97] A. Annalingam, H. Damayanthi, R. Jayawardena, and P. Ranasinghe,
[98] S. Youk and H. S. Park, “Where and what do they publish? Editors’ and
[99] M. E. Falagas, A. Zarkali, D. E. Karageorgopoulos, V. Bardakas, and
[100] C. J. C. Jabbour, A. B. L. D. S. Jabbour, and J. H. C. de Oliveira, “The
[101] T. F. Frandsen and J. Nicolaisen, “The ripple effect: Citation chain
[102] I. Ayres and E. Vars, “Determinants of citations to articles in elite
[103] W. Glänzel, R. Rousseau, and L. Zhang, “A visual representation of
[104] A. Perianes-Rodríguez and C. Olmeda-Gómez, “Effects of journal
[105] T. Fawcett, “An introduction to ROC analysis,” Pattern Recognit. Lett,
[106] C. X. Ling, H. Jin, and H. Zhang, “AUC: A better measure than accuracy
[107] M. Wang, B. Thijs, and W. Glänzel, “Characteristics of interna-
[108] J. Adams, “Early citation counts correlate with accumulated impact,” Sci-
[109] J. E. Hirsch, “An index to quantify an individual’s scientific research
[110] I. Kononenko, “Estimating attributes: Analysis and extensions of
[111] K. Stevens, P. Kegelmeyer, D. Andrzejewski and D. Buttler, “Exploring
even publication in the field of communication,” Scientometrics, vol. 105, no. 2, pp. 843–862, Nov. 2015.
[112] T. Asubiario, “How collaboration type, publication place, funding and
author’s role affect citations received by publications from africa: A bibliometric study of LIS research from 1996 to 2015,” Scientometrics, vol. 120, no. 3, pp. 1261–1287, Sep. 2019.
[113] M. Wang, J. Zhang, G. Chen, and K.-H. Chai, “Examining the influ-
ence of open access on journals’ citation obsolescence by modeling the
actual citation process,” Scientometrics, vol. 119, no. 3, pp. 1621–1641, Jun. 2019.
[114] A. Perianes-Rodríguez and C. Olmeda-Gómez, “Effects of journal
choice on the visibility of scientific publications: A comparison between subscription-based and full open access models,” Scientometrics, vol. 121, no. 3, pp. 1737–1752, Dec. 2019.
[115] I. Tahamtan and L. Bornmann, “What do citation counts measure? An
updated review of studies on citations in scientific documents published
between 2006 and 2018,” Scientometrics, vol. 121, no. 3, pp. 1635–1684, Dec. 2019.
[116] J. Xie, K. Gong, Y. Cheng, and Q. Ke, “The correlation between paper
length and citations: A meta-analysis,” Scientometrics, vol. 118, no. 3, pp. 763–786, 2019.
[117] M. D. Ginsberg, “Introspection: An analysis of the citation impact of
stroke,” Stroke, vol. 43, no. 6, pp. 1695–1699, Jun. 2012.
[118] I. Ayres and E.Vars, “Determinants of citations to articles in elite
law reviews author(s),” J. Legal Stud., vol. 29, no. 1, pp. 427–450, 2014.
[119] S.-C.-J. Sin, “International coauthorship and citation impact: A biblio-
metric study of six LIS journals, 1980-2008,” J. Amer. Soc. Inf. Sci. Tech-
ol., vol. 62, no. 9, pp. 1770–1783, Sep. 2011.
[120] W. Glänzel, R. Rousseau, and L. Zhang, “A visual representation of
relative first citation times,” J. Amer. Soc. Inf. Sci. Technol., vol. 63, no. 7, pp. 1420–1425, Jul. 2012.
[121] B. McGillivray and M. Astell, “The relationship between usage and
citations in an open access mega-journal,” Scientometrics, vol. 121, no. 2, pp. 817–838, Nov. 2019.
[122] I. Ayres and E. Vars, “Determinants of citations to articles in elite
law reviews author(s),” J. Legal Stud., vol. 29, no. 1, pp. 427–450, 2014.
[123] T. Hastie and R. Tibshirani, “Discriminant adaptive nearest neighbor
classification,” in Proc. 16th Conf. Can. Soc. Comput. Stud. Intell., Halifax, AI, Canada, 2003, pp. 329–341.
[124] J. Adams, “Early citation counts correlate with accumulated impact,” Sci-
entometrics, vol. 63, no. 3, pp. 567–581, Jun. 2005, doi: 10.1007/s11192-005-0228-9.
[125] J. E. Hirsch, “An index to quantify an individual’s scientific research
output,” Proc. Nat. Acad. Sci. USA, vol. 102, no. 46, pp. 16569–16572, Nov. 2005, doi: 10.1073/pnas.0507655102.
[126] L. Zhang, B. Thijs, and W. Glänzel, “The diffusion of H-related litera-
ture,” J. Informetrics, vol. 5, no. 4, pp. 583–593, Oct. 2011.
[127] J. E. Hirsch, “An index to quantify an individual’s scientific research
output,” Proc. Nat. Acad. Sci. USA, vol. 102, no. 46, pp. 16569–16572, Nov. 2005, doi: 10.1073/pnas.0507655102.
[128] L. Zhang, B. Thijs, and W. Glänzel, “The diffusion of H-related litera-
ture,” J. Informetrics, vol. 5, no. 4, pp. 583–593, Oct. 2011.
JIAQI ZHANG received the B.S. degree from Northeast Forestry University, China, in 2017, where she is currently pursuing the M.S. degree with the College of Information and Computer Engineering. Her main areas of interest are text mining, scientometrics, and machine learning.

XIANGRONG ZHANG received the Ph.D. degree from the Harbin Institute of Technology, China, in 2016. She is currently an Associate Professor with the School of Management, Heilongjiang Institute of Technology. Her current research interests include text mining, natural language processing, and deep learning.

NA ZHU received the master’s degree from the School of Information Management, Heilongjiang University, China, in 2010, where she is currently pursuing the Ph.D. degree in literature informatics. She is also an Associate Research Librarian with the Library of Harbin University. Her research interests include bibliometrics and data policy.