Journal of Informetrics 14 (2020) 101066

Contents lists available at ScienceDirect

Journal of Informetrics

journal homepage: www.elsevier.com/locate/joi

How do authors select keywords? A preliminary study of author keyword selection behavior

Wei Lu[a,b,1], Zhifeng Liu[a,b,1], Yong Huang[a,b], Yi Bu[c,d], Xin Li[a,b], Qikai Cheng[a,b,*]

[a] *School of Information Management, Wuhan University, Wuhan, Hubei, China*
[b] *Information Retrieval and Knowledge Mining Laboratory, Wuhan University, Wuhan, Hubei, China*
[c] *Department of Information Managemenet, Peking University, Beijing, China*
[d] *Center for Complex Networks and Systems Research, Luddy School of Informatics, Computing, and Engineering, Indiana University, Bloomington, IN, USA*

ARTICLE INFO

Article history:
Received 11 January 2020
Received in revised form 25 April 2020
Accepted 25 May 2020

Keywords:
Author keywords
author keywords selection behavior
automatic keyword extraction
citation behavior
citation
scientometrics

ABSTRACT

Author keywords for scientific literature are terms selected and created by authors. Although most studies have focused on how to apply author keywords to represent their research interests, little is known about the process of how authors select keywords. To fill this research gap, this study presents a pilot study on author keyword selection behavior. Our empirical results show that the average percentages of author keywords appearing in titles, abstracts, and both titles and abstracts are 31%, 52.1%, and 56.7%, respectively. Meanwhile, we find that keywords also appear in references and high-frequency keywords. The proportions of author-selected keywords appearing in the references and high-frequency keywords are 41.6% and 56.1%, respectively. In addition, keywords of papers written by core authors (productive authors) are found to appear less frequently in titles and abstracts in their papers than that of others, and appear more frequently in references and high-frequency keywords. The percentages of keywords appearing in titles and abstracts in scientific papers are negatively correlated with citation counts of papers. In contrast, the percentages of author keywords appearing in high-frequency keywords are positively associated with citation counts of papers.

1. Introduction

Author keywords (AKs) for scientific literature are terms selected and created by authors, and are, in general, considered as a core element that summarizes and represents the scientific publications' content (Kwon, 2018). AKs have been widely utilized in the fields of bibliometrics, information retrieval, and knowledge organization (e.g., Liu, Tian, Kong, Lee, & Xia, 2019; Raamkumar, Foo, & Pang, 2017; Shen, Nguyen, & Hsu, 2018). Bibliometricians, for instance, quantified the frequency of AK and conducted co-occurrence analyses to elucidate scientific intellectual structures, as well as research hotspots, of different disciplines (Farooq et al., 2018; Haunschild, Leydesdorff, Bornmann, Hellsten, & Marx, 2019; Hu & Zhang, 2015; Keramatfar & Amirkhani, 2019; Olmeda-Gómez, Ovalle-Perandones, & Perianes-Rodríguez, 2017; Pinto et al., 2019; Yoo,

[*] Corresponding author at: School of Information Management, Wuhan University, Wuhan, Hubei, China.
E-mail addresses: weilu@whu.edu.cn (W. Lu), zhifengliu@whu.edu.cn (Z. Liu), yonghuang1991@whu.edu.cn (Y. Huang), buyi@pku.edu.cn (Y. Bu), lucian@whu.edu.cn (X. Li), chengqikai0806@163.com (Q. Cheng).
[1] Wei Lu and Zhifeng Liu contributed equally to this work.

https://doi.org/10.1016/j.joi.2020.101066
1751-1577/© 2020 Elsevier Ltd. All rights reserved.

Jang, Byun, & Park, 2019). In information retrieval, keyword search has been a popular search metric in databases and search engines, in which AKs have often been employed as a tool for indexing (Ghanbarpour & Naderi, 2019). Additionally, AKs constitute valuable information for both human indexing and automatic indexing systems to better organize information and knowledge resources (Fadlalla & Amani, 2015).

The above-mentioned studies focus primarily on how to apply AKs as a representative of authors' research interests to resolve specific research questions (e.g., assessment, retrieval, and organization of information) in various domains. There are, however, only a limited number of studies to date that have addressed how authors select terms as keywords in their papers. From a qualitative perspective, Gbur and Trumbo (1995) provided guidelines for the selection of optimal AKs, such as not repeating keywords from the title and avoiding terms that are too common. Later, more quantitatively, Mao, Lu, Zhao, and Cao, (2018) investigated the number of keywords that authors assign to research papers at the discipline level, which provided practical implications for keyword selection in co-word analysis. However, these investigations do not offer an in-depth understanding of how authors select keywords.

Selection terms as keywords for scientific papers are also known as *keyword tagging* or *keyword indexing*. According to the mental model of article indexing (Chen & Ke, 2014), AK selection behavior is characterized as follows: Based on the content of the paper and the background of the topic, authors select appropriate terms as keywords of their papers according to their previous experience and knowledge. Indeed, AK selection behavior is multi-dimensional and depends upon numerous factors, such as: (1) the content of the paper; (2) the background of the topic; and (3) the previous experience and knowledge of the authors. These three factors correspond to three channels for AKs selection, which are the content channel, the background channel, and the prior knowledge channel, respectively. When authors choose keywords, these three factors may play dissimilar roles, which will lead to various distributions of the three channels. Therefore, one important yet basic research objective of this paper is to examine the share of AKs from these three channels.

We are also quite intersted in the potential differences of distributions of the three channels between core and non-core authors (i.e., productive and less productive authors). To this end, we use the Price Law (Egghe, 1987) as a theoretical support to select core authors. In addition, it is worthwhile to examine the relationships between distributions of the three channels and citation counts, because it assists us to select appropriate keywords to improve the visibility of papers to attract more citations. The specific research questions are as follows:

RQ1. What are the distributions of the three channels and how do the distributions change over time?

RQ2. Do the distributions of the three channels exhibit differences between core and non-core authors?

RQ3. Are the distributions of the three channels correlated with the citation counts of papers?

The remainder of this paper proceeds as follows. We present related research on patterns of AKs. Then, we define AK selection behavior, and present a detailed description of the data and methods utilized for our analysis. We next provide and discuss our findings, and their interpretations. In addition, we discuss the results of the paper and provide implications from this study. Finally, we point out the limitations of this study, and suggest directions for future research.

2. Literature review

2.1. Patterns of AKs

Related work on patterns of AKs can be assigned to one of the two categories. In the first category, the quantitative characteristics of AKs are investigated. Gil-Leiva and Alonso-Arroyo (2007), for instance, examined the presence of keywords given by authors of scientific papers in descriptors assigned by professional indexers. They noticed that approximately 46% of keywords appeared in the same or normalized form as descriptors, which indicates that keywords provided by authors constitute a valuable source of information for automatic indexing systems. Chen and Ke (2014) explored the mental models of article indexing of taggers and experts in keyword usage, which enable to inspire better selection of appropriate keywords for organizing information resources. Tripathi, Kumar, Sonker, and Babbar, (2018) assessed probabilities of occurrences of AKs in titles and abstracts of articles in social sciences and humanities disciplines in India, and reported that 61.1% of research papers have one or two AKs in their titles, and 42% of research papers have one or two AKs in their abstracts. Based on their findings, they argued that readers can rely on titles and abstracts to understand the content of papers if the AKs are not given. Mao et al. (2018) studied the distributions of the number of keywords in six different domains, and discussed the significance of the selection for keyword co-occurrence analysis. Peset et al. (2019) applied the survival analysis technique to detect new AKs that appear in the library and information science field for a period of one year to quantify their probabilities of survival over a period of 10 years. Their results demonstrated that measurements of appearance and disappearance of AKs can assist to understand the evolution of a discipline.

In addition to focusing on AK patterns, several studies have focused on the application perspective of AKs. For example, Lu, Huang, Bu, and Cheng, (2018) statistically analyzed the distribution of keywords in different structural functions of papers. Based on the distribution characteristics of keywords in various structural functions of papers, the weights of structural-functional features were quantified and integrated into a model of automatic keyword extraction, which significantly improved the accuracy of automatic keyword extraction. Meng et al. (2017) conducted a statistical analysis of the proportions of keywords in abstracts in four different datasets, and determined that a large proportion of keywords do not appear in the abstracts of the papers. For that reason, a new deep learning model was proposed to predict keywords that do not appear within the paper. Uddin and Khan (2016) researched the impact of keyword characteristics, including keyword growth, keyword diversity, number of keywords, percentage of new keywords, and network centrality of keyword co-occurrence networks on the citation counts of papers. They concluded that keyword growth, number of keywords, and network centrality all have positive relations with citation counts, while the percentage of new keywords has a negative relation. These findings assist authors to select keywords to improve the visibility of papers.

In summary, patterns of AKs can provide new insights into the selection and application of AKs. However, current studies on patterns of AKs primarily address analysis of AKs themselves, i.e., there is a paucity of analysis to elucidate what factors influencing AKs selection. Therefore, the current study will comprehensively analyze the distributions of AKs appearing in the three channels to measure the impact of the three channels on AKs selection.

2.2. Keyword tagging behavior

As an important way of organizing information and knowledge, people use keywords to represent the cognitive understanding of information resources and their content (Chen & Ke, 2014). Examples of information resources that authors or readers use keywords to tag are scientific papers, webpages, and multi-media. Tsai, Hwang, and Tang, (2011) analyzed differences of keyword-based tagging behaviors between experts and novices, and unsurprisingly found that tags chosen by experts can more accurately represent content. Ke and Chen (2012) employed social network analysis and the frequent-pattern tree method to reveal the structure and pattern of social tags for keyword selection behaviors. Moreover, the mental model of article indexing assumes that, in interacting with information resources, users and experts will assign keywords with an implicit knowledge structure to organize information according to previous experience and knowledge. Chen and Ke (2014) explored differences in mental models between taggers and experts in article indexing based on the analysis of keyword usage. Other studies also exist that have examined differences of tagging behaviors between users and experts. For instance, Kipp (2018) compared tagging of biomedical articles on CiteULike between users, authors, and professionals. Lu, Park, and Hu, (2010) investigated differences and connections between social tags and expert-assigned subject terms. Moreover, the relationship between social tagging and controlled vocabulary-based indexing was also examined in detail (Wu, He, Qiu, Lin, & Liu, 2013).

From the aforementioned studies, it can be seen that many investigations about keyword tagging behavior focused on social tagging. These studies mainly investigated differences and connections between social tags, expert-assigned subject terms, and controlled vocabulary. However, few researches have examined how authors choose terms as keywords for their scientific papers. Therefore, a deeper understanding of the influencing factors of AKs selection is still needed to fill in this gap.

3. Methodology

The framework of this study is shown in Fig. 1. First, we define AK selection behavior and three influencing channels. Then, the data are collected from the ACM digital library proceeding collections, the fields of which include title, author, author_id, abstract, AKs, citation counts, and references. Additionally, the dataset is stored in the database and preprocessed. Next, the percentages of AKs appearing in the three corresponding channels are calculated. Moreover, we investigate the differences of distributions of the three channels between core and non-core authors. Finally, the relationships between the distributions of the three channels and citation counts are examined.

3.1. AKs selection behavior and its operationalization

AKs are generally regarded as one of the most important elements of a scientific paper (Li, 2018; Raamkumar et al., 2017). When writing a paper, authors should select appropriate terms as keywords of the paper, which is known as *keyword selection*, *keyword annotation*, or *keyword index*. Chen and Ke (2014) illustrated that in interacting with the information sources, authors will assign keywords to represent the content according to their previous experience and knowledge. In addition, keywords are usually research field specific and reflect the understanding of their papers in the thematic context of their research fields (Uddin & Khan, 2016). The process of selecting keywords by authors is influenced by numerous factors, among which the main ones are the content of paper, the background of the topic, and the previous experience and knowledge of the authors. Therefore, in the current study, AK selection behavior is conceptualized as follows: Based on the content of the paper and the background of the topic, authors choose appropriate terms as keywords of their papers according to their previous experience and knowledge. Since the process of author keywords selection is missing, it is hard to measure keyword selection behaviors directly. However, we can calculate the proportion of keywords appearing in the three channels to measure the impact of these three channels on the author's keyword selection behavior.

To reveal potential quantitative patterns of the three channels, we use different metadata to represent the three channels, respectively. Firstly, the title of a scientific article plays a critical role in signalling the article's content (Guo, Ma, Shi, & Zong, 2018). In addition, the abstract of the paper reports the main topics and results of the authors' research (Ku, 2019). Hartley and Kostoff (2003) also illustrated that both title and abstract can describe an article's content with varying degrees of detail. Overall, the title and abstract of the paper constitute the central sections of a scientific article, which perform the significant function of conveying the content of the paper (Gil-Leiva & Alonso-Arroyo, 2007; Hudson, 2016; Rashidi & Meihami, 2018). Consequently, the title and abstract of the paper are used to represent the content of the paper.

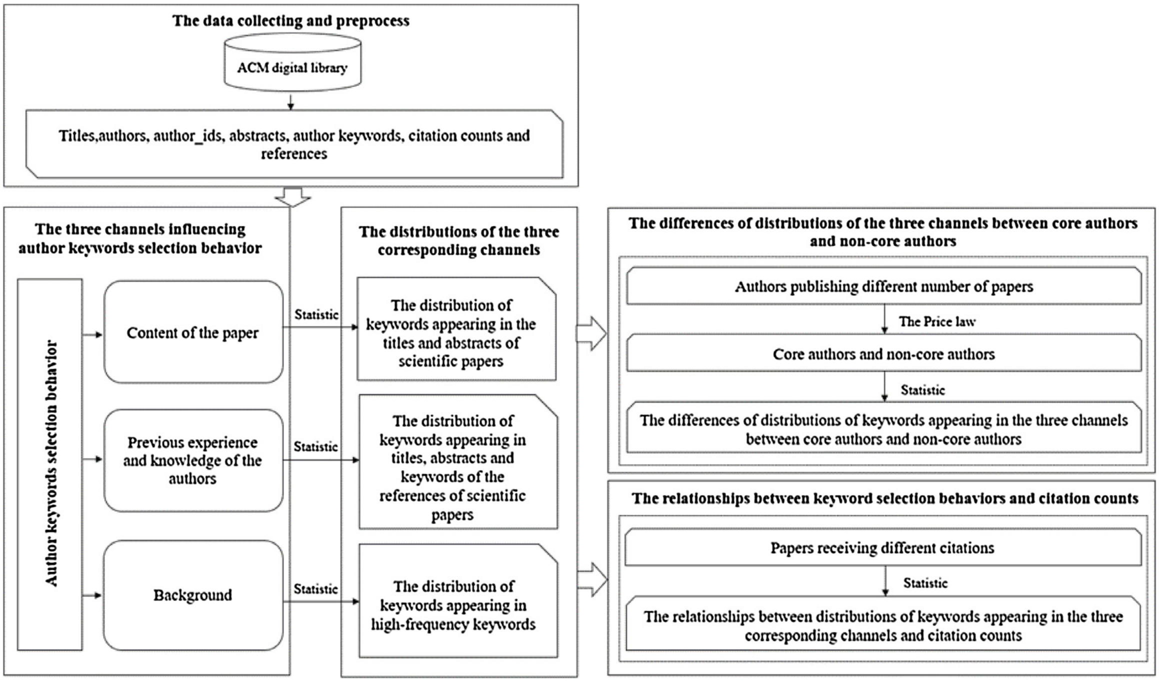
The data collecting and preprocess
ACM digital library
Titles,authors, author_ids, abstracts, author keywords, citation counts and references
The three channels influencing author keywords selection behavior
Author keywords selection behavior
Content of the paper
Previous experience and knowledge of the authors
Background
Statistic
Statistic
Statistic
The distributions of the three corresponding channels
The distribution of keywords appearing in the titles and abstracts of scientific papers
The distribution of keywords appearing in titles, abstracts and keywords of the references of scientific papers
The distribution of keywords appearing in high-frequency keywords
The differences of distributions of the three channels between core authors and non-core authors
Authors publishing different number of papers
The Price law
Core authors and non-core authors
Statistic
The differences of distributions of keywords appearing in the three channels between core authors and non-core authors
The relationships between keyword selection behaviors and citation counts
Papers receiving different citations
Statistic
The relationships between distributions of keywords appearing in the three corresponding channels and citation counts

Fig. 1. The framework of this study.

Secondly, references are the knowledge base of a scientific article, and are usually utilized to track the historical development of scientific ideas (Min et al., 2018). When authors write papers, they have to read related articles and cite them for improving the authority of their research, concept explanation, giving credit for related work and so on (Wang, Bu, & Xu, 2018). Bornmann and Daniel (2008) argued that a paper's references represent intellectual or cognitive influence on the scientific work, which can also indicate the previous experience and knowledge of authors concerning the scientific work. Therefore, references of papers can be used to represent the prior knowledge channel. It is important to note that the title, abstract, and keywords of references are all included to characterize the prior knowledge channel.

Thirdly, scientometricians frequently use high-frequency keywords to identify the state of a certain field or discipline (Song, Zhang, & Dong, 2016). High-frequency keywords also have been used to reveal the knowledge structures of research fields (Jeon & Kim, 2020). Lozano, Calzada-Infante, Adenso-Díaz, and García, (2019) reported that frequently used terms refer to the specific context of the research. In other words, high-frequency keywords can represent the background of the paper.

3.2. Data and processing

In this study, we first manually collected 295,561 academic papers from the ACM digital library proceeding collections during December 2018 and January 2019, and excluded papers without keywords and abstracts. Finally, there were 196,142 academic papers, including 854,206 keywords and 261,023 distinct authors. As shown in Fig. 2, it can be seen that most of the papers (98.49%) were published after 2000 and the range of AKs for most (95.93%) of the papers varied from 2 to 8 with the average number 4.36. In Fig. 3, 48.59% of papers have received at least one citation and most authors (96.32%) have published fewer than 10 papers. It is worth noting that only papers in the dataset that comprised references were included to compute the percentages of AKs appearing in the references, the size of which equals 124,100. In addition, the size of references of the papers is 521,892.

The bibliographic records of the dataset, including titles, authors, author_ids, abstracts, AKs, citation counts, and references were obtained from the ACM digital library proceeding collections. The author_id is unique, which can be utilized to distinguish different authors. Subsequently, an in-house database was established by using MySQL.

3.3. Selection of high-frequency keywords

To reveal the distribution of AKs appearing in high-frequency keywords, the frequency of each AK in the entire dataset was counted, and growth analysis (Uddin, Khan, & Baur, 2015) was employed to identify the high-frequency keywords set. Growth score is calculated using: $Growth = \sum_{i=1}^{n-1} \frac{(f_{i+1} - f_i)}{f_i}$, where f_i is frequency of a keyword in the i^{th} time segment. The time

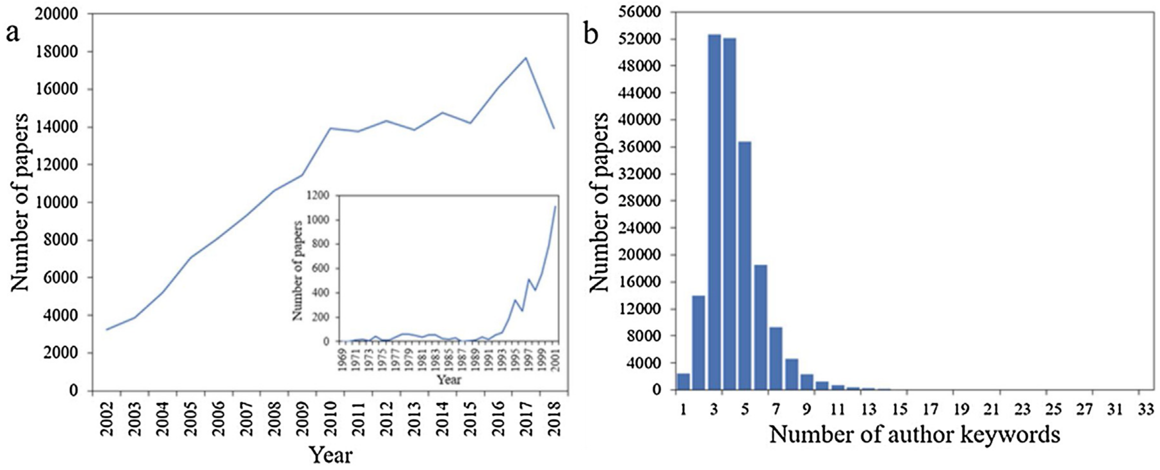

Fig. 2. (a) Time distribution of papers from the ACM digital library proceeding collections. (b) Distribution of the number of AKs in the dataset from the ACM digital library proceeding collections.

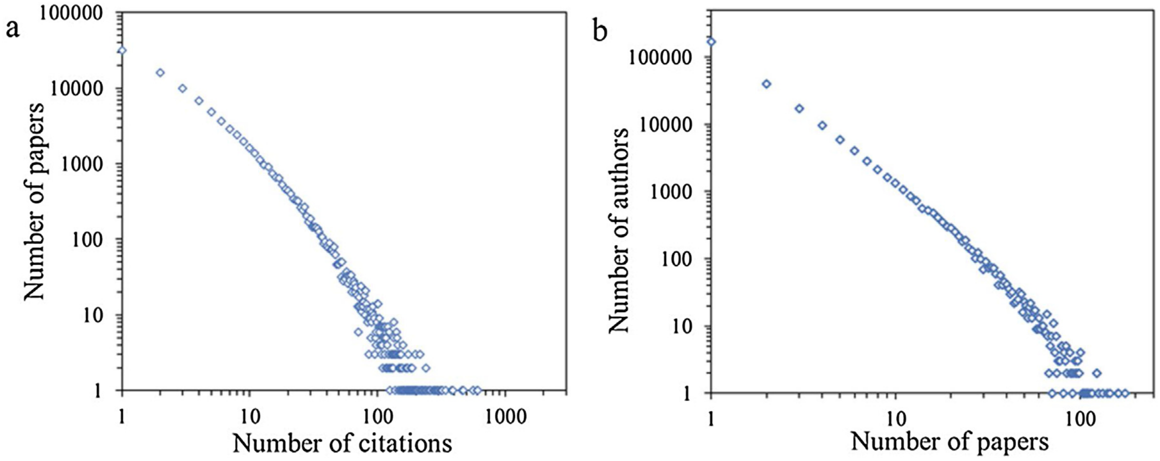

Fig. 3. (a) Distribution of the number of citations in the dataset from the ACM digital library proceeding collections. (b) Distribution of the number of papers published by different number of authors in the dataset from the ACM digital library proceeding collections.

span of our dataset range from 1969 to 2018, thus we divided the dataset into ten time slots based on the publication year. Keywords occured in less than two contigous time segements was omitted since these keywords cannot be included with the criterion of growth analysis. According to Pareto principle, we choose keywords scoring within the top 20% of valid keywords as high-frequency keywords, and the size of high-frequency keywords set is 7,917.

3.4. Classification of core and non-core authors

This research will investigate whether differences in the distributions of the three channels exist among papers written by different research productivity. In this study, the Price Law was employed to select core authors in the dataset. The Price Law (Egghe, 1987) holds that half of the papers were written by a group of highly productive authors, whose number is approximately equal to the square root of the total number of authors. The Price Law can be expressed as the following equation: $M = 0.749 * \sqrt{Nmax}$, where $Nmax$ is the largest number of papers published by the same author in the dataset. If the number of papers published by an author is greater than M, the author is a core author; whereas, if the number of papers published by an author is less than M, the author is a non-core author. In the present research, $Nmax$ equals 176. This yields 9,598 core and 251,425 non-core authors.

3.5. Measurement calculation and annotations

To investigate the impact of the three channels on AKs selection, the percentages of AKs appearing in the three channels were computed and visualized, respectively. At first, we used PorterStemmer to stem the words in AKs and the three

Table 1
The descriptions of indicators in this paper.

Indicator	Description	Indicator	Description
ktp	The percentage of AKs appearing in the title of a paper.	*krtakp*	The percentage of AKs appearing in titles, abstracts, and AKs of references of a paper.
kap	The percentage of AKs appearing in the abstract of a paper.	*khp*	The percentage of AKs of a paper appearing in high-frequency keywords.
ktap	The percentage of AKs appearing in the title and the abstract of a paper.	*avg_per_title*	The average of *ktp*.
krtp	The percentage of AKs appearing in titles of references of a paper.	*avg_per_abstract*	The average of *kap*.
krap	The percentage of AKs appearing in abstracts of references of a paper.	*avg_per_title_abstract*	The average of *ktap*.
krkp	The percentage of AKs appearing in AKs of references of a paper.	*avg_per_reference*	The average of *krtakp*.

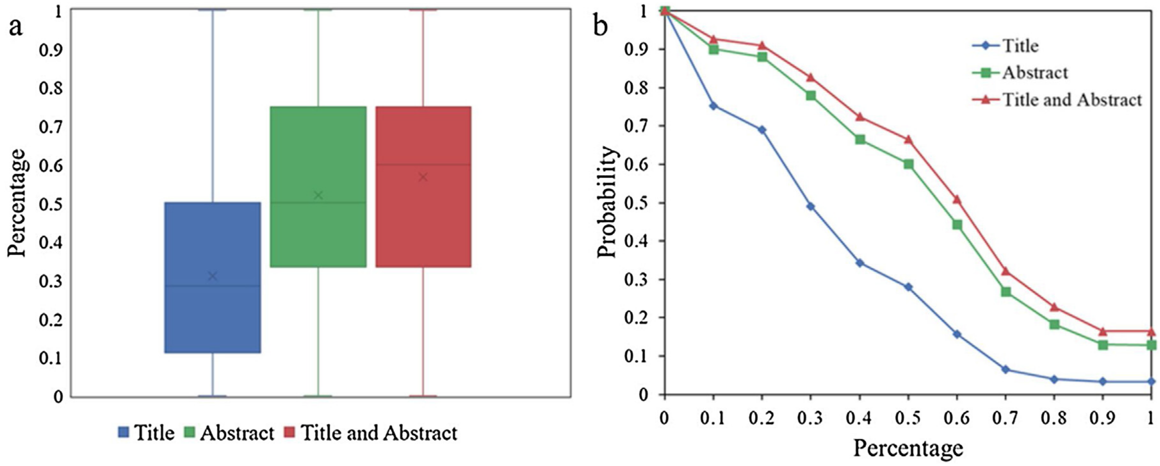

Fig. 4. (a) Distributions of *ktp*, *kap*, and *ktap*. (b) Complementary cumulative distribution functions(CCDFs) of *ktp*, *kap*, and *ktap*.

channels. We then matched the AKs with the three channels for each paper, and calculated the ratio of the number of matched keywords to the number of AKs of the paper. The indicators calculated in this paper are shown in Table 1. Finally, we computed average percentage of different channels: $Avg = (\sum_{i=1}^{n} p_i)/n$, where p_i is the percentage of keywords of paper i appearing in the three channels and n is the total number of papers.

4. Results

4.1. Overview of AK selection behavior

4.1.1. Distribution of keywords appearing in the content channel

As presented in Fig. 4(a), *avg_per_title*, *avg_per_abstract*, and *avg_per_title_abstract* were 31%, 52.1%, and 56.7%, respectively. Fig. 4(b) shows complementary cumulative distribution functions of *ktp*, *kap*, and *ktap*, from which it can be seen that *ktp*, *kap*, and *ktap* for most of papers (93.5%, 73.3%, and 67.8%, respectively) are less 0.7. The results shown in Fig. 4 indicate that most AKs (43.3%) did not appear in the title and abstract of the paper, which indicates that the absent AKs cannot be extracted from previous keyword automatic extraction approaches, which further prompts the development of a more powerful AKs prediction model.

Fig. 5 presents the mean, median, and standard deviation of *ktp* and *kap* from 2000-2018. It can be observed that the mean of *ktp* and *kap* tend to increase from 2000-2018, which means that increasing numbers of AKs appear in titles and abstracts. On the other hand, the standard deviations of *ktp* and *kap* are relatively stable from 2000-2018, which suggests that the changes of *ktp* and *kap* among different papers in each year are relatively stable.

4.1.2. Distribution of keywords appearing in the prior knowledge channel

Fig. 6(a) shows the distributions of *krtp*, *krap*, and *krkp* and *krtakp*, from which it can be seen that AKs may appear in titles, abstracts and author-selected keywords of the references. *Avg_per_reference* is 41.6%, and only keywords appearing in author-selected keywords of the references accounts for 31.8%. In Fig. 6(b), it can be observed that *krtp*, *krap*, *krkp*, and *krtakp*

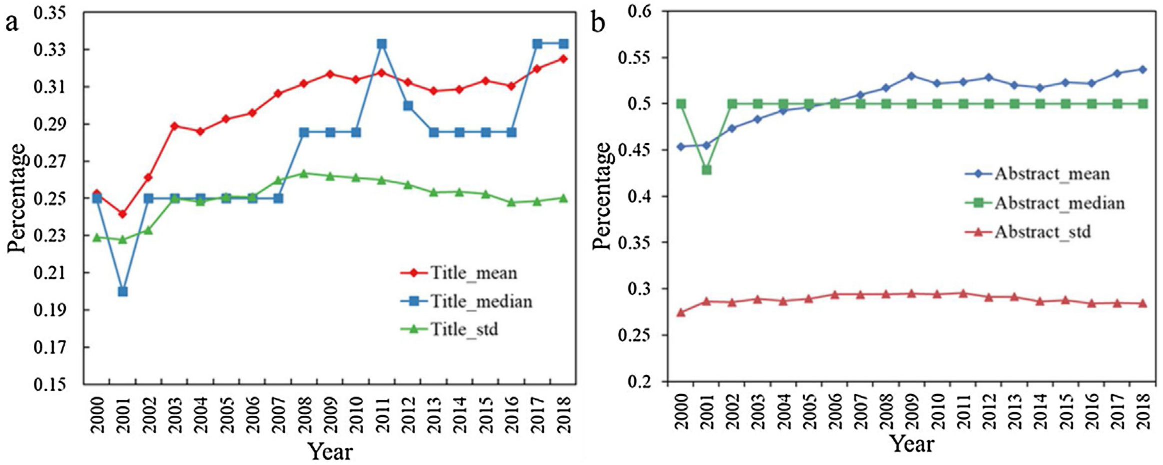

Fig. 5. (a) The mean, median, and standard deviation of *ktp* from 2000-2018. (b) The mean, median, and standard deviation of *kap* from 2000-2018.

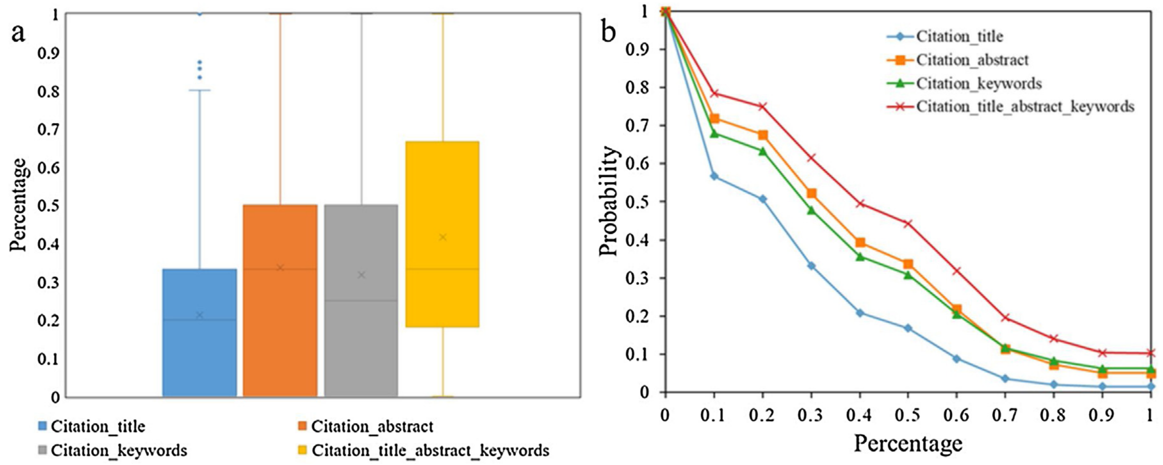

Fig. 6. (a) Distributions of *krtp*, *krap*, *krkp*, and *krtakp*. (b) Complementary cumulative distribution functions(CCDFs) of *krtp*, *krap*, *krkp*, and *krtakp*.

for a few papers (16.8%, 33.8%, 30.9%, and 44.3%, respectively) are more than 0.5. This means that the AKs may appear in AKs of the references. This interesting finding from Fig. 6 implies that we can use the bibliographic records of references as a supplement to the candidate keyword set in order to improve the accuracy of automatic keyword extraction approaches. In addition, Keywords Plus, provided by the Web of Science, are words and phrases that appear in the titles of references cited by authors (Tripathi et al., 2018). Consequently, Keywords Plus can also be utilized as a candidate keyword set for automatic keyword extraction, which may substantially enhance the accuracy of automatic keyword extraction methods.

4.1.3. Distribution of keywords appearing in the background channel

High-frequency author-selected keywords are commonly used by scientometricians to reveal the research hotspots and knowledge structure of discipline, which also enables researchers to document changes in a subject over time (Gbur & Trumbo, 1995; Xu & Liu, 2018). The percentages of AKs appearing in high-frequency keywords can reveal whether high-frequency keywords can be utilized as a candidate keyword set for automatic keyword extraction. Therefore, it is worthwhile to investigate the distribution of keywords appearing in high-frequency keywords. Fig. 7(a) presents the mean and median of percentages of AKs appearing in high-frequency keywords are 56.1% and 60%, respectively. From Fig. 7(b), one can see that around 16.6% of papers' AKs all appear in high-frequency keywords and it is also worth noting that the values of *khp* for 65.5% of papers are greater than 0.5, which indicates that a few AKs may appear in high-frequency keywords.

4.1.4. The Interactions among three channels

Usually, an overlap exists among keywords from the channels. To reveal the interactions among the three channels, we computed percentages of keywords appearing in two or three channels, respectively. As shown in Fig. 8, the mean and median of percentages of keywords in content channel and prior knowledge channel are 70.3% and 75%, which are 13.6% and 15% higher than those of percentages of keywords appearing in content channel, respectively. Besides, there are 78.1% of

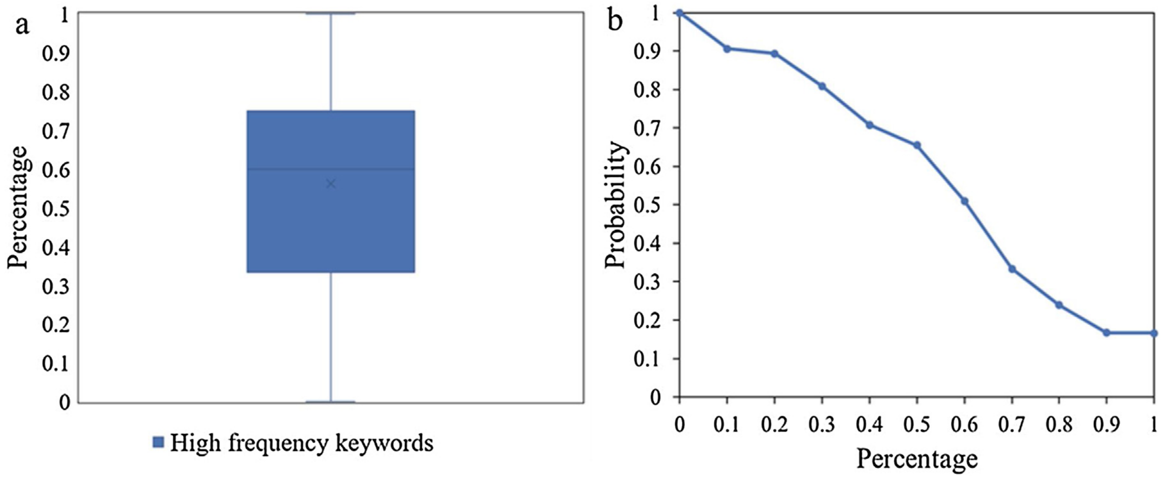

Fig. 7. (a) Distribution of *khp*. (b) Complementary cumulative distribution function(CCDF) of *khp*.

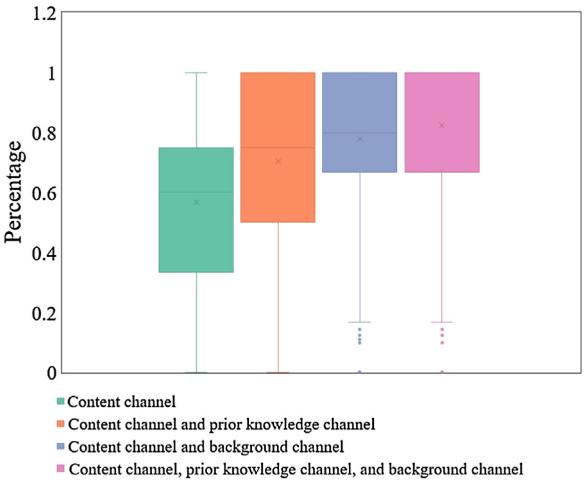

Fig. 8. The interactions among the three channels.

keywords appearing in content channel and background channel, with 21.4% higher than that of content channel. Moreover, keywords appearing in content channel, prior knowledge channel, and background channel account for 82.4%, which is 25.7% higher than that of content channel.

4.2. AK selection behavior of core vs. non-core authors

Different types of authors may possess different writing styles. For example, Lu et al. (2019) found that differences exist between authors from different ethnic backgrounds concerning the linguistic complexity of scientific writing. Since AKs constitute a core element of a paper, it seems important to determine if different types of authors exhibit differences in the distributions of the three channels.

4.2.1. Differences between core and non-core authors in the content channel

As shown in Fig. 9, one can see that the mean *ktp* of core authors is smaller than that of non-core authors. However, the standard deviation of *ktp* of core authors is higher than that of non-core authors. Similarly, the mean of *kap* of core authors is smaller than that of non-core authors, and the standard deviation of *kap* of core authors is higher than that of non-core authors. We employed a double-tailed t-test to verify the differences of the distributions of content channel between core and non-core authors. The tests identified a statistically significant difference between core and non-core authors at the 0.001 level. From Figs. 9 and 10, it can be seen that fewer AKs of paper written by core authors appear in titles and abstracts; however, authors among them may exhibit some differences. More AKs of papers written by non-core authors generally appear in titles and abstracts, and the distribution is relatively stable. This is probably because core authors possess more domain knowledge and have different strategies for keyword selection.

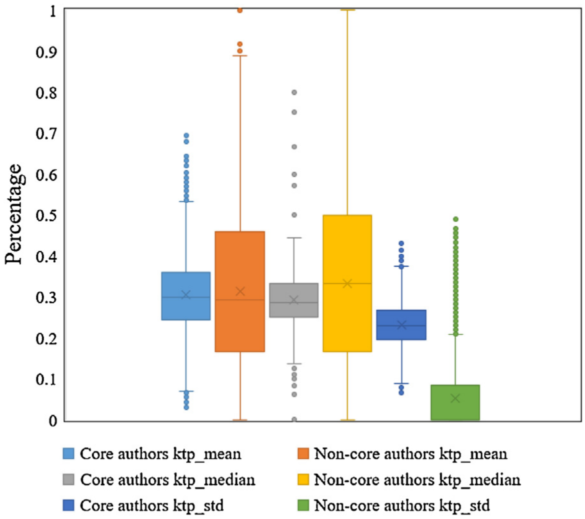

Fig. 9. Differences of mean, median, and standard deviation of *ktp* of core and non-core authors.

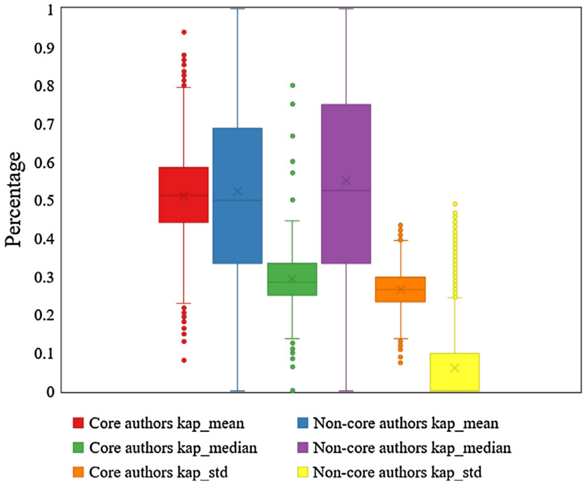

Fig. 10. Differences of mean, median, and standard deviation of *kap* of core and non-core authors.

4.2.2. Differences between core and non-core authors in the prior knowledge channel

The differences of the distributions of prior knowledge channel between core and non-core authors are presented in Fig. 11, from which it can be observed that the mean and median of *krtakp* of core authors are higher than those of non-core authors. Similarly, the standard deviation of *krtakp* of core authors is higher than that of non-core authors. We performed a double-tailed t-test to verify the differences of the distributions of prior knowledge channel between core and non-core authors. The test revealed a statistically significant difference between core and non-core authors at the 0.001 level, which indicates that, in general, more AKs of papers written by core authors appear in the references of papers. However, the distribution of prior knowledge channel of non-core authors is more stable than that of core authors.

4.2.3. Differences between core and non-core authors in the background channel

Fig. 12 shows that the mean and median of *khp* of core authors are higher than those of non-core authors. Similarly, the standard deviation of *khp* of core authors is higher than that of non-core authors. We conducted a double-tailed t-test to confirm the differences of distributions of the background channel between core and non-core authors. The test identified a statistically significant difference between core and non-core authors at the 0.001 level. Overall, slightly more AKs of papers written by core authors appear in high-frequency keywords than that of non-core authors; whereas, the distribution of background channel of core authors exhibit greater differences than that of non-core authors.

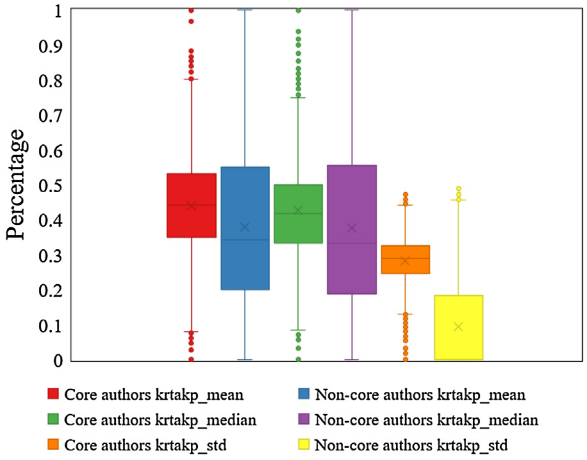

Fig. 11. Differences of mean, median, and standard deviation of *krtakp* of core and non-core authors.

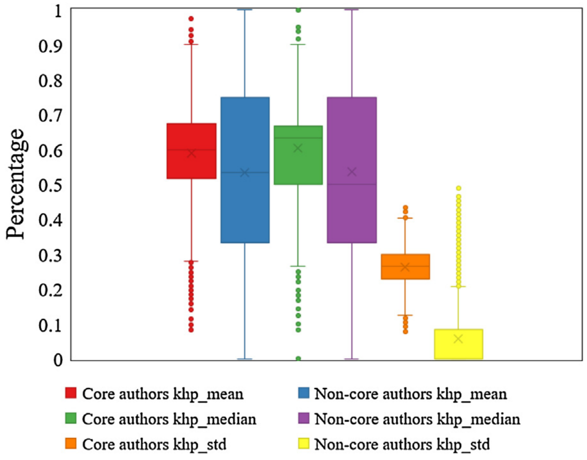

Fig. 12. Differences of mean, median, and standard deviation of *khp* of core and non-core authors.

4.3. AK selection behavior and citation counts

In recent decades, citation counts have become a major indicator of the impact of a paper, and thus has drawn intense attention from scholars. Researchers have investigated a variety of factors that influence the citation counts of paper, including scientific quality and abstract readability (Lei & Yan, 2016; Uddin & Khan, 2016). In addition, Uddin and Khan (2016) studied the impact of several statistical properties of author-selected keywords and the network attributes of their co-occurrence networks on citation counts. They reported that the choice of keywords has a significant relationship to citation counts. This paper will investigate the relationships between distributions of the three channels and citation counts.

To investigate the relationships between distributions of the three channels and citation counts, we drew the complementary cumulative distribution functions of citation counts and average percentages. Note that citation counts of most papers (99.8%) is no more than 100, and the number of papers with more than 100 citations is too limited, which may lead to the randomness of relationships between distributions of the three channels and citation counts, and thus we mainly analysed the papers with citation counts no more than 100. The relationships between distributions of the three channels and citation counts of all papers are shown in the Appendix A. Moreover, there are some overlapping keywords among content channel, prior knowledge channel, and background channel, and there might be partial correlation between a pair of content channels in regards to the number of keywords they contain. Therefore, we separated them and explore their impact on citation counts, respectively.

4.3.1. The relationship between keyword selection behavior and citation counts in the content channel

As shown in Fig. 13, based on the number of keywords containing in the content channel, the relationships between keyword selection behavior and citation counts in the content channel can be divided into four situations, which are keywords appearing in content channel, keywords appearing in content channel excluding the overlapping keywords between content

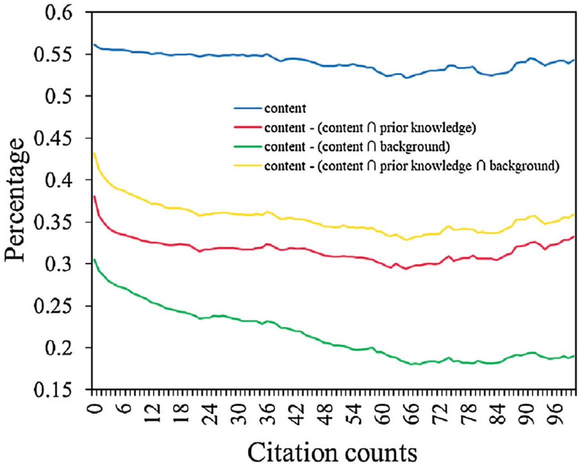

Fig. 13. Relationships between the mean of percentages of keywords appearing in content channel and citation counts (< = 100) in the four situations.

channel and prior knowledge channel, between content channel and background channel, and among content channel, prior knowledge channel, and background channel, respectively.

It can be seen that the mean of percentages of keywords appearing in content channel are all negatively correlated with citation counts in the four situations. When excluding the overlapping keywords between content channel and background channel, the falling range of mean of percentages of keywords appearing in content channel is largest. This means that, in highly cited papers, few AKs appear in the titles and abstracts. Gbur and Trumbo (1995) provided guidelines for the selection of optimal AKs, which include that authors should not repeat keywords from the titles of papers. Moreover, it is suggested authors should select keywords based on the semantic meaning of the paper and their prior knowledge, instead of following the written content in the paper, which may improve the paper's visibility. In this way, papers can be read and cited by more scholars. This discovery inspires us to minimize the direct selection of terms from titles and abstracts as keywords in papers, which may augment the visibility of papers and thus garner more citations.

In addition, when excluding the overlapping keywords between content and background channels, the green curve falls down more obviously, which may imply that if an author selects keywords from the content channel cooperating with the background channel, the paper may obtain more citations. Similarly, we observe that the red and yellow curves also decrease as the citation counts increase; this indicates that papers might receive more citations when their authors select more keywords from content, prior knowledge, and background channels.

4.3.2. *The relationship between keyword selection behavior and citation counts in the prior knowledge channel*

According to the number of keywords containing in the prior knowledge channel, the relationships between keyword selection behavior and citation counts in the prior knowledge channel can be divided into four situations, which are keywords appearing in prior knowledge channel, keywords appearing in prior knowledge channel excluding the overlapping keywords between prior knowledge channel and content channel, between prior knowledge channel and background channel, and among prior knowledge channel, content channel, and background channel, respectively.

Fig. 14 presents the relationships between the mean of percentages of keywords appearing in prior knowledge channel and citation counts in the four situations, from which it can be observed that the mean of percentages of keywords appearing in prior knowledge channel exhibit a negative relation with citation counts when excluding the overlapping keywords between prior knowledge channel and background channel. The red, yellow, and blue curves all increase when the citation counts are fewer than 5, and then tend to decrease. In Fig.15, it can be seen that the mean of percentages of keywords appearing in the background channel are all positively correlated with citation counts, which may trigger slight increases in the other three situations in the prior knowledge channel. Overall, the mean of percentages of keywords appearing in the prior knowledge channel has a negative relation with citation counts.

4.3.3. *The relationship between keyword selection behavior and citation counts in the background channel*

Based on the number of keywords containing in the background channel, the relationships between keyword selection behavior and citation counts in the background channel can also be divided into four situations, which are keywords appearing in background channel, keywords appearing in background channel excluding the overlapping keywords between background channel and content channel, between background channel and prior knowledge channel, and among background channel, content channel, and prior knowledge channel, respectively.

Fig. 15 shows the relationships between the mean of percentages of keywords appearing in background channel and citation counts. It is found that the mean of percentages of keywords appearing in background channel are all positively correlated with citation counts. This indicates that more AKs of highly cited papers appear in high-frequency keywords.

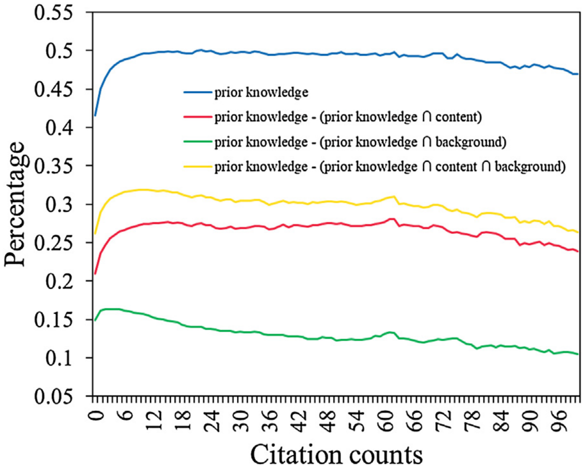

Fig. 14. Relationships between the mean of percentages of keywords appearing in prior knowledge channel and citation counts (< = 100) in the four situations.

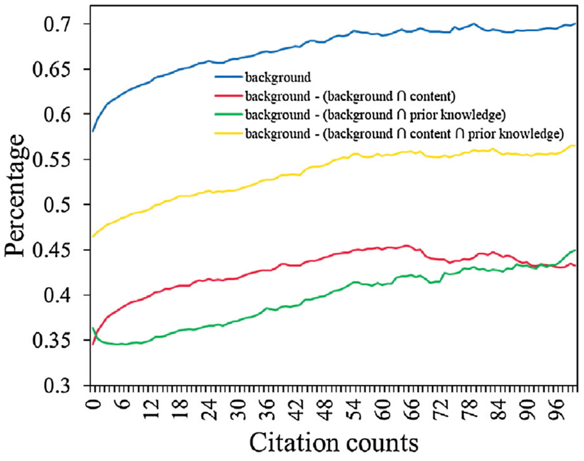

Fig. 15. Relationships between the mean of percentages of keywords appearing in background channel and citation counts (< = 100) in the four situations.

Overall, the larger is the percentage of high-frequency keywords in the AKs, the more are the citations that the paper will receive. One of the interpretations of this finding may be that high-frequency keywords increase the visibility of papers in search engines and databases, and thus the papers are read and cited by more scholars.

Moreover, we see from Fig. 15 that when not excluding the overlapping keywords between background and content channels, between background and prior knowledge channels, or among background, content, and prior knowledge channels, the growth rate and range of the blue curve is greater than those of the other three curves, Which illustrates that higher cited papers tend to have more keywords in two or three channels.

4.3.4. The relationship between keyword selection behavior and citation counts in the overlaps among the three channels

The overlaps among the three channels can be divided into four situations, which are the overlaps between content channel and prior knowledge channel, between content channel and background channel, and among content channel, prior knowledge channel, and background channel. As shown in Fig. 16, it can be observed that the average percentages of keywords appearing in the overlapping keywords among the three channels increases when the citation counts are less than 10, and tend to be stable when it is greater than 10. Besides, the rising range of mean of percentages of keywords appearing in the overlapping keywords between prior knowledge channel and background channel is largest. In addition, the percentages of overlapping keywords between background channel and the other two channels are both about 0.4, with about 0.1 higher than that between content channel and prior knowledge channel, which indicates that AKs usually appear in high-frequency keywords, while fewer keywords appear in the titles, abstracts, and keywords of references.

The overlapping keywords among the three channels essentially indicate the keywords appearing in at least one channel. The three channels represent different sources and different channels are complementary, which may improve the visibility of papers, and thus the papers can gain more citations.

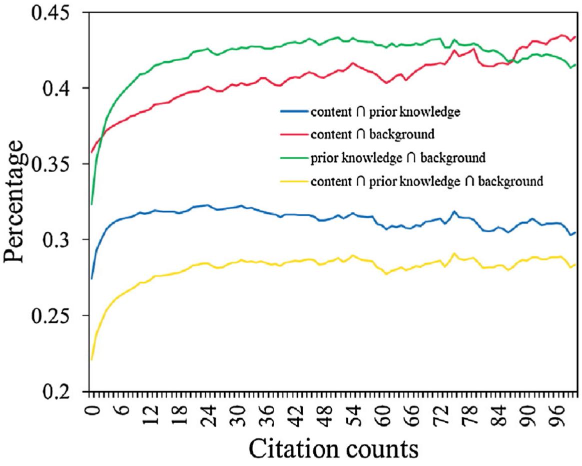

Fig. 16. Relationships between the mean of percentages of keywords appearing in the overlapping keywords among the three channels and citation counts (< = 100) in the four situations.

5. Discussion and Conclusion

AKs are considered as an important transmitter of academic concepts, ideas, and knowledge. In this article, we investigated the distributions of the three channels. On the one hand, the quantitative distributions, as well as the source analysis, of AKs on a relatively large scholarly data showed a wide perspective of AK appearing in academic manuscripts. This might be utilized as an effective means for improving the accuracy of automatic keywords extraction. On the other hand, the relationships between distributions of the three channels and citation counts provide invaluable guidance for authors to select keywords to enhance the visibility of papers. Below we discuss two implications from the perspectives of automatic keyword extraction and guiding authors to select keywords in their papers.

5.1. Implications for automatic keyword extraction

Keyword candidate selection and ranking are two major steps of many existing keyword extraction algorithms (He, Fang, Cui, Wu, & Lu, 2018). A major drawback of these approaches is that these methods only extract keywords that appear in the title or abstract of papers (Shah, Perez-Iratxeta, Bork, & Andrade, 2003). However, authors may select quite different keywords based on the semantic meaning of paper, their prior knowledge, and the background of the topic, instead of strictly adhering to the content in the paper, which leads to many keywords not appearing in the paper. In this study, we found that *avg_per_title*, *avg_per_abstract*, and *avg_per_title_abstract* were 31%, 52.1%, and 56.7%, respectively. In a previous investigation, Meng et al. (2017) computed the proportions of the keywords in abstracts in the four datasets of Inspec, Krapivin, NUS, and SemEval. They found that the proportions of the keywords present in the four datasets are 55.69%, 44.74%, 67.75%, and 42.01%, respectively. Overall, the proportion of author-selected keywords appearing in the titles and abstracts is low, which negatively affects the accuracy of automatic keyword extraction approaches. However, we found that AKs may appear in references and high-frequency keywords, and the proportions of author-selected keywords appearing in the references and high-frequency keywords are 41.6% and 56.1%, respectively. The net increase through references and high-frequency keywords are 13.6% and 21.4%, respectively. Therefore, we can utilize bibliographic records of references and high-frequency keywords as a supplement to the candidate keyword set in order to improve the accuracy of automatic keyword extraction methods.

In addition, we determined that differences exist in AKs selection between core and non-core authors, which leads to different distributions of author-selected keywords in the three corresponding channels. The findings show that more keywords of papers written by core authors appear in references and high-frequency keywords than that of papers written by non-core authors; whereas, fewer keywords of papers written by core authors appear in the titles and abstracts. According to the above results, we can adopt different approaches to extract keywords from papers of core and non-core authors.

5.2. Implications for guiding authors to select keywords

One of the most popular methods for evaluating the impact of academic papers is citation count, which refers to the number of citations received. It is assumed that papers which receive more citations have higher impact (Fu & Aliferis, 2010). Therefore, authors are eager to attract more citations for their papers, and this can be of great significance for promotions in

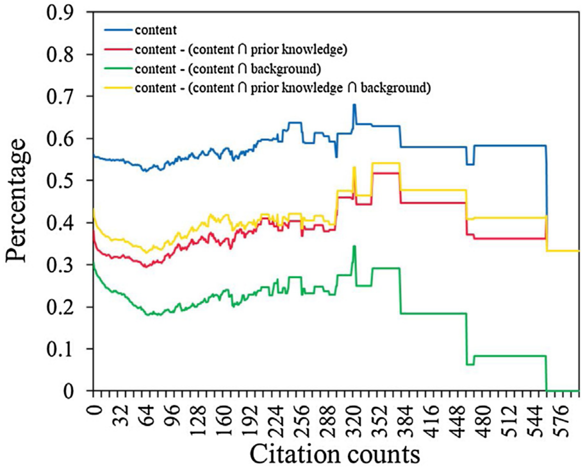

Fig. A1. Relationships between the mean of percentages of keywords appearing in content channel and citation counts in the four situations.

their careers. Numerous previous studies have examined factors that influence the citation counts of paper, which include scientific quality, abstract readability, etc. (Lei & Yan, 2016; Uddin & Khan, 2016). Unlike the extant literature, however, we investigated the relationships between distributions of the three corresponding channels and citation counts. The results demonstrate that the mean of percentages of keywords appearing in content channel is negatively correlated with the citation counts of papers, which implies that the fewer are the keywords appearing in the title and abstract of the paper, the more citation counts the paper will obtain. On the other hand, the mean of percentages of keywords appearing in background channel exhibit a positive association with the citation counts of papers, which indicates that the more are the keywords appearing in high-frequency keywords, the more citations the paper will attract. The abovementioned findings can assist authors to intelligently select keywords to improve the visibility of their papers and attract more citations.

There are, however, some limitations in this study. One major limitation of the current study is that the dataset only includes the field of computer science. Differences of patterns between various disciplines are not taken into account. To increase the robustness and generalizability of our findings, additional research can be performed that comprises datasets covering various fields. In addition, core author identification is another potential limitation in this paper. There are a lot of methods such as *h*-index and expert review, however, all these can not be implemented in our research context due to the lack of data. We have to utilize Price Law, an traditional and empirical approach, to identify core and non-core authors, which may not deal well with outliers. Moreover, since a purely quantitative investigation of the data might be insufficient, in the future, we will employ qualitative strategies, such as questionnaires and interviews, to more precisely illustrate the process of how authors select keywords (Lu et al., 2019). Finally, we plan to apply the findings to applications such as automatic keyword extraction, article indexing and document summarization.

Author contributions

Wei Lu: Conceived and designed the analysis, contributed data or analysis tools, and performed the analysis.

Zhifeng Liu: Conceived and designed the analysis, contributed data or analysis tools, performed the analysis, and wrote the paper.

Yong Huang: Conceived and designed the analysis, performed the analysis, and wrote the paper.

Yi Bu: Performed the analysis and wrote the paper.

Xin Li: Conceived and designed the analysis, and performed the analysis.

Qikai Cheng: Conceived and designed the analysis, contributed data or analysis tools, and performed the analysis.

Acknowledgements

This work was supported by the Major Project of the National Social Science Foundation of China (grant no. 17ZDA292). The authors are grateful to the anonymous referees and editors for their invaluable and insightful comments.

Appendix A. The relationships between keyword selection behaviors and citation counts of all papers

Figs. A1–A4

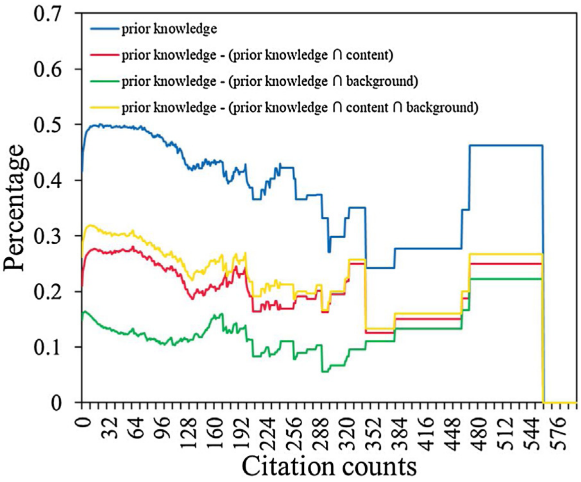

Fig. A2. Relationships between the mean of percentages of keywords appearing in prior knowledge channel and citation counts in the four situations.

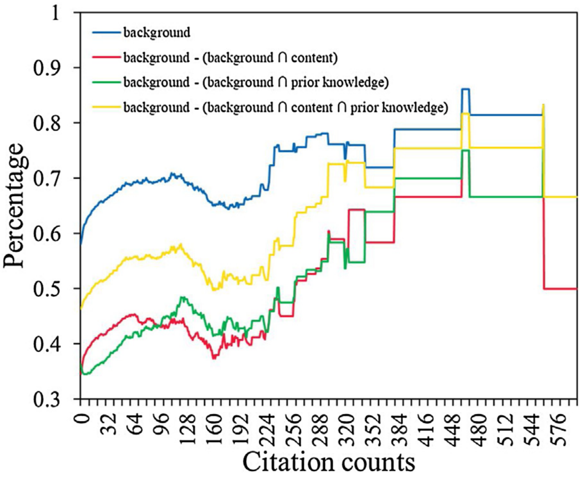

Fig. A3. Relationships between the mean of percentages of keywords appearing in background channel and citation counts in the four situations.

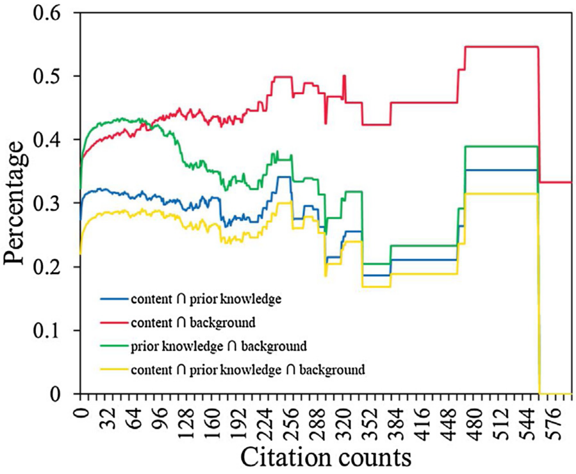

Fig. A4. Relationships between the mean of percentages of keywords appearing in the overlapping keywords among the three channels and citation counts in the four situations.

References

Bornmann, L., & Daniel, H. D. (2008). What do citation counts measure? A review of studies on citing behavior. *Journal of documentation*, *64*(1), 45–80.

Chen, Y. N., & Ke, H. R. (2014). A Study on Mental Models of Taggers and Experts for Article Indexing Based on Analysis of Keyword Usage. *Journal of the Association for Information Science and Technology*, *65*(8), 1675–1694.

Egghe, L. (1987). An exact calculation of Price's law for the law of Lotka. *Scientometrics*, *11*(1–2), 81–97.

Fadlalla, A., & Amani, F. (2015). A keyword-based organizing framework for ERP intellectual contributions. *Journal of Enterprise Information Management*, *28*(5), 637–657.

Farooq, M., Asim, M., Imran, M., Imran, S., Ahmad, J., & Younis, M. R. (2018). Mapping past, current and future energy research trend in Pakistan: a scientometric assessment. *Scientometrics*, *117*(3), 1733–1753.

Fu, L. D., & Aliferis, C. F. (2010). Using content-based and bibliometric features for machine learning models to predict citation counts in the biomedical literature. *Scientometrics*, *85*(1), 257–270.

Gbur, E. E., & Trumbo, B. (1995). Key words and phrases – the key to scholarly visibility and efficiency in an information explosion. *The American Statistician*, *49*(1), 29–33.

Ghanbarpour, A., & Naderi, H. (2019). A model-based method to improve the quality of ranking in keyword search systems using pseudo-relevance feedback. *Journal of Information Science*, *45*(4), 473–487.

Gil-Leiva, I., & Alonso-Arroyo, A. (2007). Keywords Given by Authors of Scientific Articles in Database Descriptors. *Journal of the China Society for Scientific and Technical Information*, *58*(8), 1175–1187.

Guo, F., Ma, C., Shi, Q., & Zong, Q. (2018). Succinct effect or informative effect: the relationship between title length and the number of citations. *Scientometrics*, *116*(3), 1531–1539.

Haunschild, R., Leydesdorff, L., Bornmann, L., Hellsten, I., & Marx, W. (2019). Does the public discuss other topics on climate change than researchers? A comparison of explorative networks based on author keywords and hashtags. *Journal of Informetrics*, *13*(2), 695–707.

Hartley, J., & Kostoff, R. N. (2003). How useful are key words' in scientific journals? *Journal of Information Science*, *29*(5), 433–438.

He, G., Fang, J., Cui, H., Wu, C., & Lu, W. (2018). Keyphrase Extraction Based on Prior Knowledge. In *Proceedings of the 18th ACM/IEEE on Joint Conference on Digital Libraries* (pp. 341–342).

Hu, J., & Zhang, Y. (2015). Research patterns and trends of Recommendation System in China using co-word analysis. *Information processing & management*, *51*(4), 329–339.

Hudson, J. (2016). An analysis of the titles of papers submitted to the UK REF in 2014: authors, disciplines, and stylistic details. *Scientometrics*, *109*(2), 871–889.

Jeon, S. W., & Kim, J. Y. (2020). An exploration of the knowledge structure in studies on old people physical activities in Journal of Exercise Rehabilitation: by semantic network analysis. *Journal of Exercise Rehabilitation*, *16*(1), 69–77.

Ke, H. R., & Chen, Y. N. (2012). Structure and pattern of social tags for keyword selection behaviors. *Scientometrics*, *92*(1), 43–62.

Keramatfar, A., & Amirkhani, H. (2019). Bibliometrics of sentiment analysis literature. *Journal of Information Science*, *45*(1), 3–15.

Kipp, M. E. I. (2018). Tagging of Biomedical Articles on CiteULike: A Comparison of User, Author and Professional Indexing. *Knowledge Organization*, *38*(3), 245–261.

Ku, M. C. (2019). A Comparative Analysis of English Abstracts and Summaries of Chinese Research Articles in Three Library and Information Science Journals Indexed by the Taiwan Social Science Citation Index. *Journal of Library and Information Studies*, *17*(1), 37–81.

Kwon, S. (2018). Characteristics of interdisciplinary research in author keywords appearing in Korean journals. *Malaysian Journal of Library & Information Science*, *23*(2), 77–93.

Lei, L., & Yan, S. (2016). Readability and citations in information science: evidence from abstracts and articles of four journals (2003–2012). *Scientometrics*, *108*(3), 1155–1169.

Li, M. (2018). Classifying and ranking topic terms based on a novel approach: role differentiation of author keywords. *Scientometrics*, *116*(1), 77–100.

Liu, J., Tian, J., Kong, X., Lee, I., & Xia, F. (2019). Two decades of information systems: a bibliometric review. *Scientometrics*, *118*(2), 617–643.

Lozano, S., Calzada-Infante, L., Adenso-Díaz, B., & García, S. (2019). Complex network analysis of keywords co-occurrence in the recent efficiency analysis literature. *Scientometrics*, *120*(2), 609–629.

Lu, C., Bu, Y., Wang, J., Ding, Y., Torvik, V., Schnaars, M., et al. (2019). Examining scientific writing styles from the perspective of linguistic complexity. *Journal of the Association for Information Science and Technology*, *70*(5), 462–475.

Lu, C., Park, J. R., & Hu, X. (2010). User tags versus expert-assigned subject terms: A comparison of LibraryThing tags and Library of Congress Subject Headings. *Journal of information science*, *36*(6), 763–779.

Lu, W., Huang, Y., Bu, Y., & Cheng, Q. (2018). Functional structure identification of scientific documents in computer science. *Scientometrics*, *115*(1), 463–486.

Mao, J., Lu, K., Zhao, W., & Cao, Y. (2018). How many keywords do authors assign to research articles–a multi-disciplinary analysis? *iConference 2018 Proceedings*.

Meng, R., Zhao, S., Han, S., He, D., Brusilovsky, P., & Chi, Y. (2017). Deep keyphrase generation. In *Annual Meeting of the Association for Computational Linguistics* (pp. 582–592).

Min, C., Ding, Y., Li, J., Bu, Y., Pei, L., & Sun, J. (2018). Innovation or imitation: The diffusion of citations. *Journal of the Association for Information Science and Technology*, *69*(10), 1271–1282.

Olmeda-Gómez, C., Ovalle-Perandones, M. A., & Perianes-Rodríguez, A. (2017). Co-word analysis and thematic landscapes in Spanish information science literature, 1985–2014. *Scientometrics*, *113*(1), 195–217.

Peset, F., Garzón-Farinós, F., González, L. M., García-Massó, X., Ferrer-Sapena, A., Toca-Herrera, J. L., et al. (2019). Survival analysis of AKs: An application to the library and information sciences area. *Journal of the Association for Information Science and Technology*.

Pinto, M., Fernández-Pascual, R., Caballero-Mariscal, D., Sales, D., Guerrero, D., & Uribe, A. (2019). Scientific production on mobile information literacy in higher education: a bibliometric analysis (2006–2017). *Scientometrics*, *120*(1), 57–85.

Raamkumar, A. S., Foo, S., & Pang, N. (2017). Using author-specified keywords in building an initial reading list of research papers in scientific paper retrieval and recommender systems. *Information processing & management*, *53*(3), 577–594.

Rashidi, N., & Meihami, H. (2018). Informetrics of Scientometrics abstracts: a rhetorical move analysis of the research abstracts published in Scientometrics journal. *Scientometrics*, *116*(3), 1975–1994.

Shah, P. K., Perez-Iratxeta, C., Bork, P., & Andrade, M. A. (2003). Information extraction from full text scientific articles: where are the keywords? *BMC bioinformatics*, *4*, 20–28.

Shen, C. W., Nguyen, D. T., & Hsu, P. Y. (2018). Bibliometric networks and analytics on gerontology research. *Library Hi Tech*, *37*(1), 88–100.

Song, J., Zhang, H., & Dong, W. (2016). A review of emerging trends in global PPP research: analysis and visualization. *Scientometrics*, *107*(3), 1111–1147.

Tripathi, M., Kumar, S., Sonker, S. K., & Babbar, P. (2018). Occurrence of author keywords and keywords plus in social sciences and humanities research : A preliminary study. *COLLNET Journal of Scientometrics and Information Management*, *12*(2), 215–232.

Tsai, L. C., Hwang, S. L., & Tang, K. H. (2011). Analysis of keyword-based tagging behaviors of experts and novices. *Online Information Review*, *35*(2), 272–290.

Uddin, S., & Khan, A. (2016). The impact of author-selected keywords on citation counts. *Journal of Informetrics*, *10*(4), 1166–1177.

Uddin, S., Khan, A., & Baur, L. A. (2015). A framework to explore the knowledge structure of multidisciplinary research fields. *PloS one*, *10*(4), Article e0123537.

Wang, B., Bu, Y., & Xu, Y. (2018). A quantitative exploration on reasons for citing articles from the perspective of cited authors. *Scientometrics*, *116*(2), 675–687.

Wu, D., He, D., Qiu, J., Lin, R., & Liu, Y. (2013). Comparing social tags with subject headings on annotating books: A study comparing the information science domain in English and Chinese. *Journal of Information Science*, *39*(2), 169–187.

Xu, J., & Liu, Y. B. (2018). A bibliometric analysis for global research trends on ectomycorrhizae over the past thirty years. *Electronic Library*, *36*(4), 733–749.

Yoo, S., Jang, S., Byun, S. W., & Park, S. (2019). Exploring human resource development research themes: A keyword network analysis. *Human Resource Development Quarterly*, *30*(2), 155–174.