View metadata, citation and similar papers at core.ac.uk
brought to you by CORE
provided by Illinois Digital Environment for Access to Learning and Scholarship Repository

How many keywords do authors assign to research articles – a multi-disciplinary analysis?

Jin Mao[1,*], Kun Lu[2], Wanying Zhao[1], and Yujie Cao[1]

[1] Center for Studies of Information Resources, Wuhan University, Wuhan, China
[2] School of Library and Information Studies, University of Oklahoma, Norman, OK, USA
*maojin@whu.edu.cn

Abstract. Author keywords are one important data source for co-word analysis. The distribution of author keywords in papers has not been investigated at the discipline level. We analyzed six research fields from soft science to hard science to reveal the underlying quantitative patterns of author keywords. Normal distribution, Poisson distribution, and Weibull distribution were fitted by applying Maximum Likelihood Estimation. Chi-Square tests and Kolmogorov-Smirnov tests were used to evaluate the goodness of fit. The results show that a large portion of papers have no keyword or only one keyword in all these fields. The author keyword distributions of the six fields are represented. It's shown that Weibull distribution is the best fitted. This study provides practical implications for keyword selection in co-word analysis.

Keywords: Author Keywords, Keyword Frequency Distribution, Co-word Analysis, Distribution Fitting.

1 Introduction

Co-word analysis has been widely used to reveal knowledge structures of a research field [1]. Semantic items used in the co-word analysis are assumed to represent the content of articles, such as terms from titles [2] and abstracts, or subject terms assigned by professional indexers [3]. Author keywords have also been selected as one crucial data source for establishing co-occurrence relationships between keywords in many co-word analysis studies [4]. The number of author keywords influences the results of co-word analysis. In the co-word analysis, not all author keywords are used to form co-occurrence relations. Only the keywords from papers that have two or more keywords are used. Therefore, investigating on the number of author keywords in research articles would help understand the coverage of papers by the co-word network using author keywords.

Author keywords are often requested or encouraged by many journals to highlight the content of articles in their publication policies. Some journals even specify the number of keywords authors should assign. However, in a research field, what is the distribution of author keywords in papers? Further, to our best knowledge, there are no studies that compare author keywords distributions in multiple disciplines. In this study, we

will investigate on how many author keywords are assigned to articles in multiple disciplines with the aim of revealing the underlying quantitative patterns.

2 Data and methods

Six research fields from soft science to hard science were analysed, including Ethnology, Sociology, Library and Information Science (LIS), Economics, Physics, Fluids & Plasma (Physics), and Acoustics. Articles from top journals in these six fields were retrieved from the Web of Science database. Bibliographic records of articles with the types of article were downloaded. Author keywords in the papers were parsed and stored into MySQL databases. The number of keywords for each article in the six fields was counted. Normal distribution, Poisson distribution, and Weibull distribution were fitted to the number of keywords per paper by applying Maximum Likelihood Estimation. Chi-Square(χ^2) tests and Kolmogorov-Smirnov(KS) tests were applied to evaluating the goodness of fit.

3 Results

The six disciplines have different magnitudes of selected research articles (Table 1). Table 1 shows that a significant proportion of papers across all these fields do not have any author keywords. Sociology has the lowest rate of keyword absence with 23.53% and Acoustics has the highest rate with 72.78%. In addition, a very small proportion of papers in these fields have only one keyword. Generally speaking, research fields in hard science (e.g., Physics and Acoustics) tend to have more papers without author keywords than those of soft science (e.g., Ethnology and Sociology).

Table 1. The number of papers without/with one author keyword in the six disciplines.

Discipline	# of papers	# of papers without keywords(percentage)	# of papers with one keyword(percentage)
Ethnology	6,655	2,211(33.22%)	3(0.05%)
Sociology	7,307	1,719(23.53%)	2(0.03%)
LIS	12,010	4,200(34.97%)	53(0.44%)
Economics	71,455	47,492(66.46%)	104(0.15%)
Physics	63,155	43,168(68.35%)	184(0.29%)
Acoustics	93,451	68,011(72.78%)	61(0.07%)

Fig. 1 presents the author keyword distributions in the six fields. Papers without any keywords are excluded, i.e., the dots with the number of keywords equaling to zero are not included. The results show that the six fields have similar author keyword distributions. Most papers have 3 to 7 keywords, and fewer papers have more than 9 author keywords. Interestingly, almost in every field, a few papers have a large quantity of keywords. For example, one paper in Ethnology has 33 keywords.

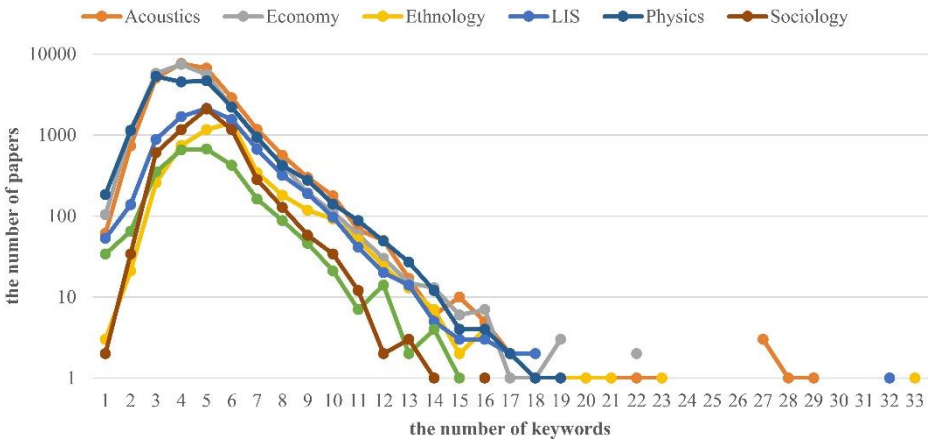

Fig. 1. Empirical author keyword distributions in the six fields. The number of papers is logarithmized.

Both Chi-Square(χ^2) tests and Kolmogorov-Smirnov(KS) tests show that for the author keyword distributions in the six fields, neither Normal distribution, Poisson distribution, nor Weibull distribution can be fitted with all p-values less than 0.05. Further, the distance values between observed data and expected data from the fitted distributions by KS tests are presented in Table 2. It suggests that Weibull distribution is the closest distribution among the three distributions, which further verifies the six fields have distributions.

Table 2. The distance values by KS tests.

Discipline	Poisson	Normal	Weibull
Ethnology	0.2677	0.2441	**0.2326**
Sociology	0.3208	0.2069	**0.2005**
Economics	0.3136	0.1966	**0.1777**
LIS	0.2662	0.1747	**0.1601**
Physics	0.2832	0.1695	**0.1626**
Acoustics	0.2929	0.1974	**0.1891**

*Bold values indicate the smallest distances.

4 Discussion

The significant portion of papers without any keywords or with one keyword in the six fields indicates that purely using author keywords in co-word analysis is insufficient due to the shortage of paper coverage. Other data sources, such as keywords from titles

and abstracts, could be adopted in co-word analysis or used to complement the paper coverage. This will guide the selection of keywords when applying co-word analysis.

One major reason for the absence of keywords could be the publication rules of the journals. Many journals do not allow authors to assign any keywords to articles, such as *Acta Acustica united with* Acustica in Acoustics and The *World Economy* in Economics. Some journals did not allow keywords in earlier years but started to require author keywords more recently. For example, *Applied Acoustics* did not ask for author keywords before 1995, but did later. Another factor comes from the authors. Even some journals require author keywords, some papers still have no keywords. For example, 14 of 40 papers published in Youth & Society in 2014 have no keywords. Therefore, it should be noted that the empiricial observation is a result of multiple factors, not the author indexing behavior alone. Nevertheless, this study provides a picture of the author keyword distribution in multiple disciplines for those who intend to analyze author keywords.

From the perspective of the entire field, author keywords in papers demonstrate similar quantitative patterns revealed by the fitted distribution (Table 2). The emergence of the patterns could be attributed to the dual effects of journal rules and individual decision of authors. Statistically, the mean of the number of author keywords mostly depend on journal rules and the deviation comes from individual decision. The selected distribution functions are not well fitted. One possible reason could be the sample size is too large that makes the observed data significantly depart from theoretical distributions [5]. The number of papers with some given number of keywords, e.g., from 3 to 7, is much higher than the expectation (the number of papers is logarithmized in Fig. 1). Thus, log distributions are worthy of attempt in future.

5 Conclusions

Author keywords in six fields are analysed quantitatively in this study. A large portion of papers have no keyword or only one keyword, informing the careful use of author keywords when applying co-word analysis. The six fields have similar distributions of author keywords with Weibull distributions best fitted although not strictly. The quantitative patterns of author keywords provide practical implications for keyword selection in co-word analysis. The general quantitative law of author keywords in research fields not only describes the status of current author keywords in a field but also can be used to predict the scale of future author keywords. This quantitative law together with other quantitative patterns about author keywords, such as keywords growth and decaying, could be used to disclose the dynamics of knowledge in a field. We will put the investigation on more quantitative patterns about keywords in the list of our research agenda. Journal rules on author keywords will be also further analysed to interpret the patterns. In addition, we will explore whether and how co-word analysis is influenced by the papers without/with one author keyword.

Acknowledgments

This study is supported by the National Natural Science Foundation of China funded projects under Grant Nos. 71603189, 71420107026, and 71403190.

References

1. Callon, M., Courtial, J., Turner, W., Bauin, S.: From translations to problematic networks: An introduction to co-word analysis. Information (International Social Science Council) 22(2), 191-235(1983).
2. Bhattacharya, S., Basu, P.: Mapping a research area at the micro level using co-word analysis. Scientometrics 43(3), 359-372(1998).
3. Ocholla, D. N., Onyancha, O. B., Britz, J.: Can information ethics be conceptualized by using the core/periphery model?. Journal of Informetrics 4(4), 492-502(2010).
4. Cho, J.: Intellectual structure of the institutional repository field: A co-word analysis. Journal of Information Science 40(3), 386-397(2014).
5. Ajiferuke, I.: A probabilistic model for the distribution of authorships. Journal of the American Society for Information Science 42(4), 279(1991).