Using WoS Keywords to Analyze Topic of Articles Collection with Different Interdisciplinary Degree

Tong Tiantian
Nanjing University of Science and Technology, China.
tongtiantiannjau@163.com

Zhang Chengzhi*
Nanjing University of Science and Technology, China.
zhangcz@njust.edu.cn

Zhang Lin
Wuhan University, China.
zhanglin_1117@126.com

ABSTRACT

Interdisciplinary scientific research (IDR) tends to be a seedbed for breakthroughs in science and technology. Topic analysis in IDR can provide a better understanding of knowledge flow among disciplines. Keywords Plus terms annotated by *Web of Science* (WoS) can capture articles' content and summarize their topics efficiently. In this study, we investigate the difference of IDR topics with interdisciplinary degree by WoS keywords. Firstly, we use feature selection method to contrast the distinctive keywords of articles with different interdisciplinarity. Then, time series analysis by a cluster analysis is conducted to evaluate differences between frequency patterns of individual keywords. Experimental results show that articles with high interdisciplinarity tend to include more general keywords than articles with lower interdisciplinarity. Furthermore, our study can be applied to understand research topics of IDR and provide suggestions to interdisciplinary researchers.

KEYWORDS

WoS Keywords, interdisciplinary research, interdisciplinarity, research topic.

INTRODUCTION

Interdisciplinary scientific research (IDR) has shown impressive growth for decades. Understanding what knowledge has exchanged between disciplines and how research topics change over time can help scholars better understand knowledge structure of scientific fields and development of new disciplines. Existing research on IDR knowledge structure only analyzed research topics according to titles and abstracts (Xu, Liu, Lei, Li, & Fang (2014). Since results of topic extraction are different by various extraction method, there are no uniform topic terms for summarizing interdisciplinary topics. It is not convenient to compare topic mining results across disciplines. Instead, Keywords Plus terms provided by WoS, which capture articles' content with greater depth and variety (Zhang et al. (2016), make it possible to provide systematic comparison across various disciplines.

In this paper, we focus on articles covering various disciplines published in PLOS One, and investigate research hotpot based on Keywords Plus terms. We examine the distribution of keywords in research articles with different interdisciplinary degree. Our goal is to identify changes of IDR topics with time and interdisciplinary degree. Our findings can expand the understanding of the development of interdisciplinary research across a multitude of fields.

METHODOLOGY

We propose a two-step model to make a comparative analysis of research topics in different interdisciplinary articles: CHI feature selection method is used to contrast the distinctive keywords of articles with different interdisciplinarity, and time series analysis by a cluster analysis is conducted to evaluate differences between frequency patterns of individual keywords.

Data

We download metadata (references and keyword etc.) of articles from WoS. These articles published from 2007 to 2015 in PLOS One (http://journals.plos.org/plosone/). Interdisciplinarity of each article is computed according to the method proposed by Zhang, Rousseau, & Glänzel (2016).The total number of articles is 143, 910 and we divide these articles into three groups according to interdisciplinarity: H-IDR group (the top 10%), M-IDR group (80%) and L-IDR group (the last 10%). We further check that keyword distributions in H-IDR and M-IDR Group are highly correlative (Pearson Coefficient r =.898, p=. 00), followed by M-IDR/L-IDR Group (r =.816, p=. 00), H-IDR Group/ L-IDR Group (r = .688, p=. 00), which indicates that there is a great difference between topics of articles with different interdisciplinarity.

Key technologies

Comparative analysis of distinctive topics: Distinctive topics which can differentiate research topics of each group, refer to keywords with CHI (Yang & Pederson (1997) value greater than a certain threshold value, expressed as $Dis_{topic} = \{\chi^2 > \Theta\}$. Θ is our threshold value. The CHI value of a keyword is defined as formula (1).

$$\chi^2(g,k) = \frac{N(AD-BC)^2}{(A+C)(B+D)(A+B)(C+D)} \quad (1)$$

Where, *A* is the number of articles in the group *g* that contain keyword *k*, *B* is the number of articles outside group g that contain k, C is the number of articles in group *g* that don't contain *k*. *D* is the number of articles outside group *g* that don't contain *k*. *N* is the total number of articles in the dataset.

* Corresponding author: Chengzhi Zhang, Email: zhangcz@njust.edu.cn.

Time series analysis of distinctive topics: K-Means (Macqueen (1967) is chosen for time series clustering. Let $X(i) = [x_1, x_2, \dots, x_N]$ and $Y(i) = [y_1, y_2, \dots, y_N]$ be two time-series of length N respectively. If the distance between two time-series is defined across all N attributes, then $D(X,Y)$ between two time-series defined as formula (2).

$$D(X,Y) = \sqrt[c_N^2]{\prod_{j>i} max(\frac{x_j/x_i}{y_j/y_i}, \frac{y_j/y_j}{x_j/x_i}) - 1} \qquad (2)$$

EXPRIMENTAL RESULTS ANALYSIS

In comparative analysis of distinctive topics, we regard keywords with CHI score exceeding 2 ($\Theta = 2$) as distinctive topics in each group. In Table 1 (only Top 10 is shown), keywords in H-IDR Group are concentrated in application research, including more general words, such as *Regulatory networks* and *Set enrichment analysis*. Keywords in L-group are concentrated in theoretical study. It indicates that the theory-related research topics are quite similar between groups, and the margins between them become larger since utilizing theory and technology in other subjects, especially *Computer Science* and *Physics*.

Group	Top 10 Keywords in Chi-square Score
H-IDR	*Acute lung injury; Drug-delivery;* **Oxidative stress;** Rna-seq; *Heart-rate-variability;* Metabolomics; **Gene-expression data; *Nanoparticles;Set enrichment analysis*; *Regulatory networks***
M-IDR	*optical coherence tomography;* ***Myoblast fusion;*** *Cystic-fibrosis;* ***ca2+-activated cl-channels;*** *Filtered back-projection; Phase-iii trial; Dendritic cells; Acute lung injury;* ***Messenger-Rna decay; Rna-seq;***
L-IDR	*Ligand cd55;optical coherence tomography;* ***Ca2+-activated cl-channels;*** *Phase-iii trial;* ***receptor cd97;*** *Open-angle glaucoma; Filtered back-projection; Computed-tomography; Long-term potentiation; Radiotherapy;*

Table 1. Top 10 keywords obtained by Chi-square across three groups.

In time-series analysis, keywords obtained by removing duplicated keywords obtained by Chi-square, are divided into four kinds of categories (denoted as A, B, C, D respectively) automatically .In Table 2(only Top 10 is shown), keywords in A and B, are more mentioned in high interdisciplinary articles than lower interdisciplinary articles. In A, the fluency of keyword increases with time in each IDR group; while in B, the fluency of keyword in H-IDR increases slowly until a point and decreases sharply afterwards. Keywords in category C and D, are more mentioned in low interdisciplinary articles. In C, keywords show continued growth in all three groups; while in D, keywords attract increased attention from L-IDR group until a certain year and decrease sharply.

Category	Top 10 keywords for each category
A	*Cells; Gene-expression; Acute lung injury; Air-pollution exposure; Amorphous calcium-carbonate; Atomic-force microscopy ； Drug-delivery; Electronic medical-records; Expression data; Female physical attractiveness;*
B	*Cystic-fibrosis; Dynamics; Heart-rate-variability; Inflammation; Ischemia-reperfusion injury; Mass-spectrometry; Maximum-likelihood; Respiratory-distress-syndrome; Complex networks; Human-immunodeficiency-virus;*
C	*Arabidopsis-thaliana; Care; Cell lung-cancer; Computed-tomography; Developing-countries; Mellitus; Open-angle glaucoma; Optical coherence tomography; Randomized controlled-trials; Risk-factors;*
D	*Abscisic-acid; Basal ganglia; Beta-cell function; Event-related fmri; Long-term depression; Long-term potentiation; Messenger-rna decay;Posterior parietal cortex; Transcranial magnetic stimulation; Visual-cortex*

Table 2. Top 10 keywords for each category.

CONCLUSION AND FUTURE WORKS

We investigate scientific articles' topics based on Keywords Plus terms across a multitude of fields to expand the understanding of interdisciplinary research. We find that the group of H-IDR articles tends to cover more fields and include more general keywords than the lower-IDR articles. In future, additional metrics might be considered to measure the interdisciplinarity, such as indicators based on citation. By contrast, the keywords chosen by authors could be applied to capture the content of articles. Also, although time series analysis provides some insight into a series of groups identified by a cluster analysis, some illustration of aggregate patterns within each cluster could be further studied for interdisciplinary research in the future.

ACKNOWLEDGEMENT

This work is supported by Major Projects of National Social Science Fund (No.17ZDA291).

REFERENCES

Macqueen, J. (1967).*Some methods of classification and analysis of multivariate observations*. In:Proceedings of the 5th Berkeley Symposium on Mathematical Statistics and Probability, Berkeley, USA, 1967: 281-297.

Xu, H et al. (2014). Measurement,Visualization and Application of Interdisciplinary Research. *Library & Information Service*, 8 (3), 21-27

Yang, Y et al. (1997). Feature selection in statistical learning of text categorization. *Planta*, 230(230), 671-685

Zhang, J et al. (2016). Comparing keywords plus of WOS and author keywords: A case study of patient adherence research. *Journal of the Association for Information Science & Technology*, 67(4), 967-972

Zhang, L et al. (2016). Diversity of references as an indicator of the interdisciplinarity of journals: Taking similarity between subject fields into account. *Journal of the Association for Information Science & Technology*, 67(5), 111-112