Interdisciplinary Topics Extraction and Evolution Analysis

Zhongyi Wang[1], Jing Chen[1], Jiangping Chen[2] and Haihua Chen[2,*]

[1]School of Information Management, Central China Normal University, Wuhan, China 430079
[2]Department of Information Science, University of North Texas, Denton, Texas, USA 76203

Abstract
In order to clarify the current interdisciplinary development process, this paper takes Library and Information Science and Management as an example, and firstly constructs an interdisciplinary literature set based on K-Means clustering algorithm, and then extracts interdisciplinary topics using LDA model; finally, it combines the first/last discrete time method to portray the interdisciplinary topic intensity and topic content evolution trend. The results shows that topics such as "government data openness, blockchain and public opinion governance" and "value co-creation and supply chain" are common research hotspots for scholars in both fields in recent years.

Keywords
Interdisciplinary research, Topic extraction, Topic evolution, LDA, Clustering analysis

1. Introduction

Extracting potential interdisciplinary research topics in different disciplines can promote disciplinary breakthroughs and innovations. Topic evolution research reveals the laws and processes of topic evolution through the comparison of knowledge structures or contents of different time windows and can understand the current development trends of disciplinary fields [1].

Existing studies mainly focus on macroscopic research on interdisciplinary intersection, so this study takes Library and Information Science and Management as an example, extracts interdisciplinary research topics in the literature of two disciplines, and constructs a dynamic LDA model of interdisciplinary research topic evolution research, so as to uncover the development trend of interdisciplinary research topics.

2. Methodology

The framework diagram of the research methodology in this paper is shown in Figure 1.

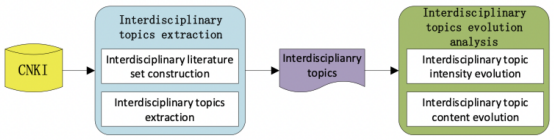

Figure 1: Research framework

2.1. Interdisciplinary topic extraction

2.1.1. Interdisciplinary literature set construction

The interdisciplinary literature set construction includes two steps: (1) text representation of each document and (2) clustering the documents. In the first step, we used Term frequency-inverse document frequency (TF-IDF) weighting algorithm for the text representation. In the second step, we use the K-means clustering algorithm based on SSE (sum of squares) [2] to determine the optimal number of clusters. Using the K-means clustering method, the ratio of the number of literature from two disciplines in each type of clusters is calculated, and the class cluster whose ratio is closest to the literature data from two disciplines in the original dataset is selected as an interdisciplinary literature set.

2.1.2. Interdisciplinary topic extraction based on LDA

We use the pyLDAvis topic model from the scikit-learn library to subjectively determine the best topic number on the overlap of the number of topics. After that, we apply LDA's variational inference EM algorithm for topic extraction.

3rd Workshop on Extraction and Evaluation of Knowledge Entities from Scientific Documents (EEKE2022), June 20-24, 2022, Cologne, Germany and Online
*Corresponding author.
wzywzy13579@163.com (Z. Wang); 2363238156@qq.com (J. Chen); jiangping.chen@unt.edu (J. Chen); haihua.chen@unt.edu (H. Chen)

Table 1
Distribution of core journals and number of papers

Discipline	Journal Source	Journal count	Papers count	Percentage
Library and Information Science	CSSCI	20	23,633	58.6%
Management	CSSCI	36	33,459	41.4%

Table 2
Top two topics from the interdisciplinary topic extraction

Topic id	Feature words (partial)
Topic #0	Change Indicator Status Logic Farmers Sector Revenue Banks Strategic emerging industries Change Systemic risk
Topic #1	User Information Data Factors Community WeChat Platform Model Mobile Library Satisfaction

2.2. Interdisciplinary topic evolution analysis

2.2.1. Interdisciplinary topic intensity evolution analysis

Let D_t be the set of texts on time window t, and θ_z^d be the proportion of topic z in document d. Formula (1) gives the method for calculating the intensity of topic z on time window t.

$$\theta_z^t = \frac{\sum_{d=1}^{D^t} \theta_z^d}{D_t} \tag{1}$$

2.2.2. Interdisciplinary topic content evolution analysis

We choose the cosine similarity metric to calculate the similarity.

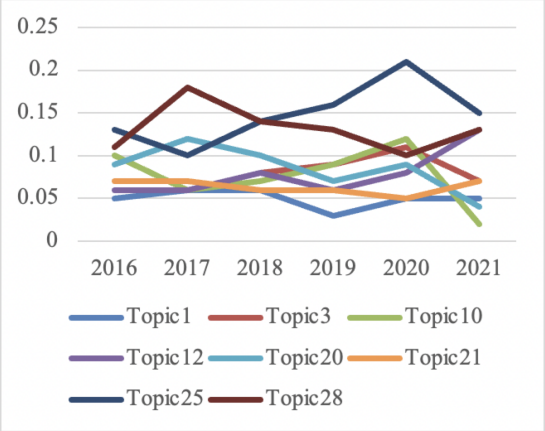

Figure 2: Evolution of the intensity of the interdisciplinary topics of Library and Information Science and Management

3. Experiment and results

3.1. Interdisciplinary topic extraction

The dataset of this paper was obtained from the CNKI database with the years 2016-2021, the specific data are as follows.

First, we use the method of Section 2.2.1 to construct interdisciplinary literature set. In our experiments, we define a K-Means algorithm with a K value of 15 to perform K-means clustering on the dataset, and finally select class cluster 8 as the basic dataset for interdisciplinary topic extraction and evolutionary research, which contains 3,006 documents.

According to the interdisciplinary topic extraction method introduced in Section 2.2.2, the results are partially shown below.

The interdisciplinary topics between the two fields were obtained through this process: “elements of information technology”, “user information and digital libraries”, and “information literacy education and industrial structure upgrading”.

3.2. Interdisciplinary topic evolution analysis

In this paper, we adopt the method of Section 2.2.1 for interdisciplinary topic intensity evolution. First, we set the threshold as 0.3, and obtains 8 valid interdisciplinary topics. According to the publication time of the texts, we divide the literature data from 2016 to 2021 into six corresponding time windows in units of years. The intensity of each topic is obtained according to Formula (1). The interdisciplinary topic intensity evolution trend is plotted as shown in Figure 2.

By analyzing the intensity of the interdisciplinary topics, we found that the attention to the topic of “Internet public opinion and emergencies” has increased since the beginning of 2020, and the attention to the topic of “user

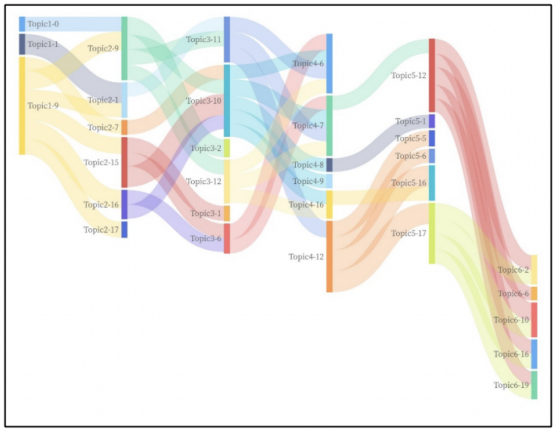

Figure 3: Evolutionary pathway of interdisciplinary topics in the field of Library and Information Science and Management

information and mobile libraries" is stable, but the overall intensity of the topic is not high. The attention to the topic of "Innovative Social System Establishment" starts to decline from 2021.

Then, we apply the method of Section 2.2.2 for interdisciplinary topic content evolution. We calculate the cosine similarity between topics in each adjacent time window of the literature in the two disciplines. According to the results, only topic 0, topic 1 and topic 9 of 2016 were more than 0.2 similar to the topic of 2017. Therefore, we select these three topics for topic evolution analysis. The results are shown in Figure 3.

By analyzing the evolution of the content of interdisciplinary topics, it was found that during the six-year period from 2016 to 2021, topics such as "corporate policies and consumer strategies" and "value co-creation and supply chains" have been hot issues of general interest to scholars from both disciplines in recent years.

Acknowledgments

This study was supported by Humanities and Social Science Research Foundation of Ministry of Education of China (Grant number 21YJA870003) and National Social Science Foundation of China (Grant number 19ZDA345).

References

[1] S. Chen, Research on Knowledge Evolution Analysis Based on Overlapping Structure, 2010.

[2] G. Du, Z. Xu, Mapping the knowledge map of innovation theory research:keyword co-occurrence analysis, Science & Technology Progress and Policy 26 (2009) 135–139.