Using Full-text Content of Academic Articles to Classify Research Methods in Library and Information Science

Ruping Wang, Liang Tian, Chengzhi Zhang

Department of Information Management, Nanjing University of Science and Technology, Nanjing, China

Introduction

The methodology is a key factor to connect research problems and results. Most of the existing studies have constructed many schemas of research methods in Library and Information Science (LIS). For example, Luo & McKinney (2015) divided the research methodology into the research design, research model, and research theory. Chu & Ke (2017) summarized the research methods of Library Information Science into 16 categories including 'Experiment', 'Bibliometrics', 'Questionnaire', etc.

Identifying the research methods used in academic papers is the basis for analysing usage behaviour and scenarios of different research methods. The research methods of papers usually need to be manually annotated, which is time-consuming and costing a lot. This paper uses a supervised multi-label classification method to identify the research methods in LIS based on the full-text content of academic articles.

Methodology

Dataset

Table 1. The number of articles using different research methods in training set (2001-2010)

No.	Method	Papers' Number
1	Experiment	553
2	Theoretical Research Method	377
3	Questionnaire	366
4	Content Analysis	339
5	Metrics	377
6	Interview	256
7	Transaction Log Analysis	120
8	Other	270

This paper collects research articles from three kinds of journals in LIS field, namely Journal of the Association for Information Science and Technology (JASIST), Library & Information Science Research (LISR), and Journal of Documentation (JOD). After filtering, we get 5,273 research papers from 1990 to 2019 for the experimental dataset. Among them, 1,977 papers from 2001 to 2010 were annotated with the research methods by Chu et al. (Chu & Ke, 2017). Table 1 shows the distribution of different research methods in the training corpus. The full-text content of the article is divided into sentences and remove the stop-words. Then, the full-text content is represented as a vector after feature selecting by Chi-Square and feature weight computing by TF-IDF.

Classification model of Research methods

Problem transformation and algorithm adaptation (Tsoumakas & Katakis, 2007) are two common kinds of multi-label classification strategies. The strategy of problem transformation includes Binary Relevance, Classifier Chain, Label Powerset, and Random k-label sets. Algorithm adaptation is a way to perform multi-label classification by directly improving or modifying the existing classification algorithm. We use these two multi-label classification strategies to classify the research methods of academic articles in LIS. For problem transformation, we use Naïve Bayes, SVM, and XGBoost as the basic classification algorithms. For algorithm adaptation, we used KNN as the basic classifier. We generate a total of 13 combined classification algorithms, namely BR-NB, BR-SVM, BR-XGB, CC-NB, CC-SVM, CC-XGB, LP-NB, LP-SVM, LP-XGB, RAKEL-NB, RAKEL-SVM, RAKEL-XGB, and MLKNN.

To evaluate and select the optimal classification model, we divide the annotated data into training set, validation set, and test set. The training set accounts for 80% of the total annotated corpus, the validation set and the test set account for 10% respectively. This paper trains each algorithm on the training set and optimize each algorithm on validation set. According to the results of the 13 combined classification models in Table 2, the Classifier Chain based upon XGBoost (CC-XGB) performs the best with the highest F1 value at 80.77%.

Table 2. Comparison of multi label classification algorithms

Model	P(%)	R(%)	F1(%)
BR-XGB	77.65	78.31	76.56
CC-XGB	80.18	84.58	80.77
LP-XGB	76.77	71.69	72.99
RAKEL-XGB	74.45	77.79	74.33
BR-NB	50.93	78.67	57.76
CC-NB	51.66	76.99	57.80
LP-NB	65.15	59.87	60.99
RAKEL-NB	59.57	71.43	62.30
BR-SVM	59.05	56.17	56.05
CC-SVM	40.08	68.28	47.96
LP-SVM	34.34	30.68	31.68
RAKEL-SVM	48.91	44.81	45.59
MLkNN	53.70	53.11	52.16

Experiments and Results Analysis

We use CC-XGB to train the final model upon all the annotated articles to predict the remaining 3,296 unlabelled articles. Using 1,977 papers that have been manually annotated with research methods and 3,296 papers that have been automatically classified by CC-XGB, we analyse the frequency of different research methods in 1990-1999, 2000-2009, and 2010-2019. The number of papers in each period affects the use of research methods. Hence, we analyse the relative frequency of different research methods in the three periods. Figure 1 shows the result.

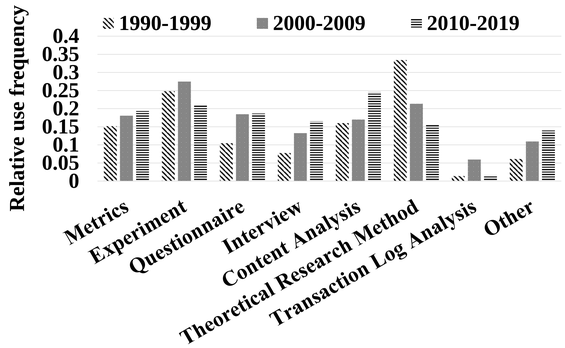

Figure 1. The relative frequency of different research methods in each period.

We can see that the relative usage frequency of 'Metrics', 'Questionnaire', 'Interview', 'Content analysis' and 'Other' is increasing year by year. The usage of 'Experiment' and 'Transaction Log Analysis' is increasing and then decreasing; the usage of 'Theoretical Research Method' is decreasing gradually. These changes reflect that researchers pay more attention to experiments and data in the process of research.

References

Chu, H., & Ke, Q. (2017). Research methods: What's in the name? Library & Information Science Research, 39(4), 284–294.

Luo, L., & McKinney, M. (2015). JAL in the Past Decade: A Comprehensive Analysis of Academic Library Research. The Journal of Academic Librarianship, 41(2), 123–129.

Madjarov, G., Kocev, D., Gjorgjevikj, D., & Džeroski, S. (2012). An extensive experimental comparison of methods for multi-label learning. Pattern Recognition, 45(9), 3084–3104.

Tsoumakas, G., & Katakis, I. (2007). Multi-Label Classification: An Overview. International Journal of Data Warehousing and Mining, 3(3), 1–13.

Contact

Ruping Wnag(wangrp@njust.edu.cn)

Chengzhi Zhang (Corresponding author, zhangcz@njust.edu.cn)