SCIREA Journal of Electrical Engineering

http://www.scirea.org/journal/DEE

April 6, 2022

Volume 7, Issue 2, April 2022

https://doi.org/10.54647/dee47263

Evaluation of Manuscript Preference of Sci-Tech Periodicals: Structural Topic Models of Computer Engineering Related Publications in China

Li Xiang

McGill University, Canada

Email: xiang.li8@mail.mcgill.ca

Abstract

To help scholars in computer engineering get to know the state-of-the-art topics and choose suitable journals to publish their work, the method that uses structural topic model (STM) to extract the topics of papers and uses the generalized linear model to analyze the influence of time on topic prevalence. Based on these results, the manuscript preferences of a journal are extracted. As an empirical experiment, a total number of 10320 papers are extracted from 3 different Chinese core computer engineering journals to analyze. As a result, 6 hotspots in the past 6 years in the Chinese computer engineering field are extracted, and the changes in topic prevalence in the Chinese computer engineering field over time are also studied. We believe our method could help researchers learn the most cutting-edge topics and choose suitable journals for them.

Keywords Information processing, computer engineering, academic papers publishing, manuscript preference, structural topic model

1. Introduction

As the number of researches in computer engineering grows fast, the number of journals in China is also booming year by year. Many old problems get new progressions and many new problems arise. The computer engineering topic is growing as a tree, and it is hard to tell which subtopic is the cornucopia. Also, for authors and researchers, it is difficult to choose where to submit a paper from a huge number of journals. However, it is worth mentioning that there exists a kind of preferences when the peer reviewers estimate papers. The bias might come from the preference of peer reviewers, the subject of journals and the public opinion on a certain field. This research will bring out a view that what topics are beloved by the journals and what journals are suitable for a certain topic.

Manuscript preference evaluation, a way to evaluate the bias in journals choosing the papers to publish, can assist researchers choose the best journal to submit, help lessen the work load of peer reviewers, and aid readers in choosing the fittest journal for them to refer.

In 2011, Lee J. et al. study the preference in journals selecting promising scientific technologies (Lee J. M. et al., 2011). In 2014, Dai J. analysis the logic of accept a paper of journals using SPSS19.0 (Dai J., 2014). In 2020, Lei Y. use Publons platform to investigate the influence on the journals from peer reviewers by producing the profile of them (Lei Y., 2021). In 2020, X. Zhou et al. discuss about the bias presented in the Chinese Economics Journals (Zhou X. et al., 2020). In 2021, Li Y. et al. analyze the characteristics of papers published on 6 journals in the past 20 years using CiteSpace(Li Y. et al., 2021).

The current researches mostly use statistical methods to analyze the characteristics. These methods require huge manual work. And they are not adequate to mine the latent information within the context of the journals. To solve the above problems, we use the topic models.

Topic model is a type of statistical model for discovering the abstract “topics” that occur in a collection of documents. Topic models, such as latent Dirichlet allocation (LDA), can be useful tools for the statistical analysis of document collections and other discrete data. Topic models can extract surprisingly interpretable and useful structure without any explicit "understanding" of the language by computer (Blei D. M. et al., 2007). Some commonly seen topic models including LDA, CTM, LSA, and so on.

Topic models can also be used for analyzing preferences. In 2008, XU G. et al. use lda model to investigate the user behavior for web recommendation (Xu G. et al., 2008). In 2010, Séaghdha D O uses latent variable models to analyze selectional preference (Séaghdha D. O., 2010). In 2015, Liu X. models the user's dynamic preference for personalized recommendation with LDS model (Xu G. et al., 2008). In 2018, L. Qiu et al. use CLDA to mine the user interest preference under Big Data Background (Qiu L. et al., 2018). In 2020, L. Shi et al. use topic model to extract user's preference and intention in social network (Shi L. et al., 2020).

However, (1) LDA may not be suited for tasks that require consistent evaluation of similarity (Alexander E. et al., 2015). And LDA cannot work well in modeling texts that are short and noisy (Zhao F. et al., 2016). (2) CTM is modelled using logistic Normal distribution. And this distribution is not conjugate to the multinomial, which complicates the corresponding approximate posterior inference procedure (Blei D. M. et al., 2007). (3) LSA and pLSA have several problems such as overfitting and inappropriate generative semantics (Blei D. M. et al., 2003; Sriurai W. et al., 2009).

The structural topic model (STM) was introduced by M. Roberts et al. in 2013 (Roberts M. E. et al., 2013). Comparing to other topic models, the STM's key innovation is that it permits users to incorporate arbitrary metadata, defined as information about each document, into the topic model (Roberts M. E. et al., 2019). People can name many scenarios of the application of the STM. In 2014, M. Roberts et al. proposed an open-ended survey and analysed the responses using STM (Roberts M. E. et al., 2014). In 2018, M. Chandelier applied STM to analyze news on wolf recolonization in France (Chandelier M. et al., 2018). In 2020, X. Chen et al. detected the latent topics and trends within Computers & Education using STM (Chen X. et al., 2020). STM has become a new trend in text analysis now. In this research, we will use STM to analyze the preference of journals accepting papers.

The rest of paper will be organized in the following structure: the second paragraph introduces the theory of STM; the third paragraph is the design of our research; the fourth paragraph describes the experiment of our research; the fifth paragraph shows a conclusion of our research.

2. Structural Topic Model

2.1 The principle

STM is a kind of Bayesian generative topic model, which assume that each topic is a distribution over words and each document is a mixture of corpus-wide topics (Hu N. et al., 2019) (Blei D. et al., 2010). Compared to LDA, the document-level structure information is introduced, thereby emphasizing the suitability of investigating how covariates affect text content (Hu N. et al., 2019).

The structure of STM is shown in Figure 1. Where θ denotes the topic proportions. The prevalence of those topics can be influenced by some set of covariates X through a standard regression model with covariates $\theta \sim LogisticNormal(X\gamma, \varepsilon)$. w is a word, and for each word in the response, a topic z is drawn from the response-specific distribution, and conditional on that topic, a word is chosen from a multinomial distribution over words parameterized by φ, which is formed by deviations from the baseline word frequencies m in log space, i.e., $\varphi_k \propto \exp{(m + \tau_k)}$. μ is another set of covariates, which provide a way of "structuring" the prior distributions in the topic model, injecting valuable information into the inference procedure (Roberts M. E. et al., 2014). The shaded nodes $X, w,$ and μ are observed variables. The rectangles denote replication: $n \in \{1, 2, ..., N\}$ indexes words within a document; $k \in \{1, 2, ..., K\}$ indexes each topic assuming the user-specified number of topics is K; and $d \in \{1, 2, ..., D\}$ represents the document indexes.

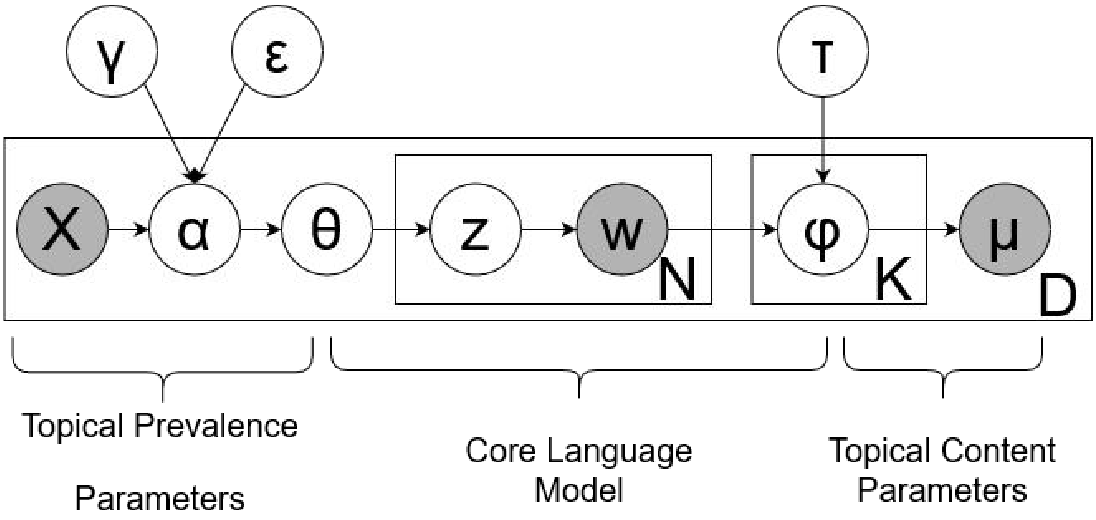


Fig.1. The Structure of STM

2.2 Metrics for Evaluating STM-Based Topic Quality

2.2.1 Semantic Coherence

Semantic Coherence is closely related to pointwise mutual information: It is maximized when the most probable words in a given topic frequently co-occur together (Roberts M. E. et al., 2019).

Let $D(w, w^{'})$ be the number of times that words w and w' appear together in a document. For a list of the W most probable words in topic k, the semantic coherence for topic k is given as (Roberts M. E. et al., 2019):

$$C_k = \sum_{i=2}^{W} \sum_{j=1}^{i-1} \log\left(\frac{D(w_i, w_j) + 1}{D(w_j)}\right) \tag{1}$$

2.2.2 Held-Out Llikelihood

Held-out likelihood indicates the estimation of the probability of words appearing within a document when those words have been removed from the document in the estimation step (Roberts M. E. et al., 2019).

Similar to cross—validation, when some of the data is removed from estimation and then later used for validation, the held-out likelihood helps the user assess the model's prediction performance (Roberts M. E. et al., 2019).

The formula for log-likelihood is given as (Asuncion A. et al., 2012):

$$l = \sum_i \log \sum_{z_i} P(\varphi_i|z_i, \tau) P(z_i|d_i, \theta) \quad (2)$$

Where (Asuncion A. et al., 2012):

$$\varphi_i \sim \tau_{w,z_i} \qquad z_i \sim \theta_{k,d_i} \quad (3)$$

2.2.3 Exclusivity

A word's exclusivity to a topic is its usage rate relative to a set of comparison topics (Bischof J. et al., 2012). Exclusivity can be used to calculate topic quality.

Exclusivity is calculated using the FREX metric, thus word frequency is balanced (Roberts M. E. et al., 2019):

$$FREX_{k,v} = \left(\frac{\omega}{ECDF\left(\frac{\beta_{k,w}}{\sum_{j=1}^{K} \beta_{j,w}}\right)} + \frac{1-\omega}{ECDF(\beta_{k,w})}\right)^{-1} \quad (4)$$

Where ECDF is the empirical CDF and ω is the weight. This parameter is set to 0.7 to favor exclusivity (Roberts M. E. et al., 2019).

2.2.4 Residual

Residuals are calculated by (Taddy M., 2012):

$$\frac{x_{ij} - \hat{x}_{ij}}{s_{ij}} \quad (5)$$

$$s_{ij}^2 = m_i \hat{q}_{ij}(1 - \hat{q}_{ij}) \quad (6)$$

Where x_i is a vector of counts for terms (words and phrases) in a document with total term-count m_i. And $\hat{q}_i = \hat{k}\hat{\omega}_i$, ω_i is probability vectors (Taddy M., 2012).

3. RESEARCH DESIGN

To reveal the manuscript preferences of different journals, the ideas of using STM for research topics extraction and further analysis to the relationships between research topics and journals, as well as the effect of publication year to topics are proposed, as shown in Figure 2.

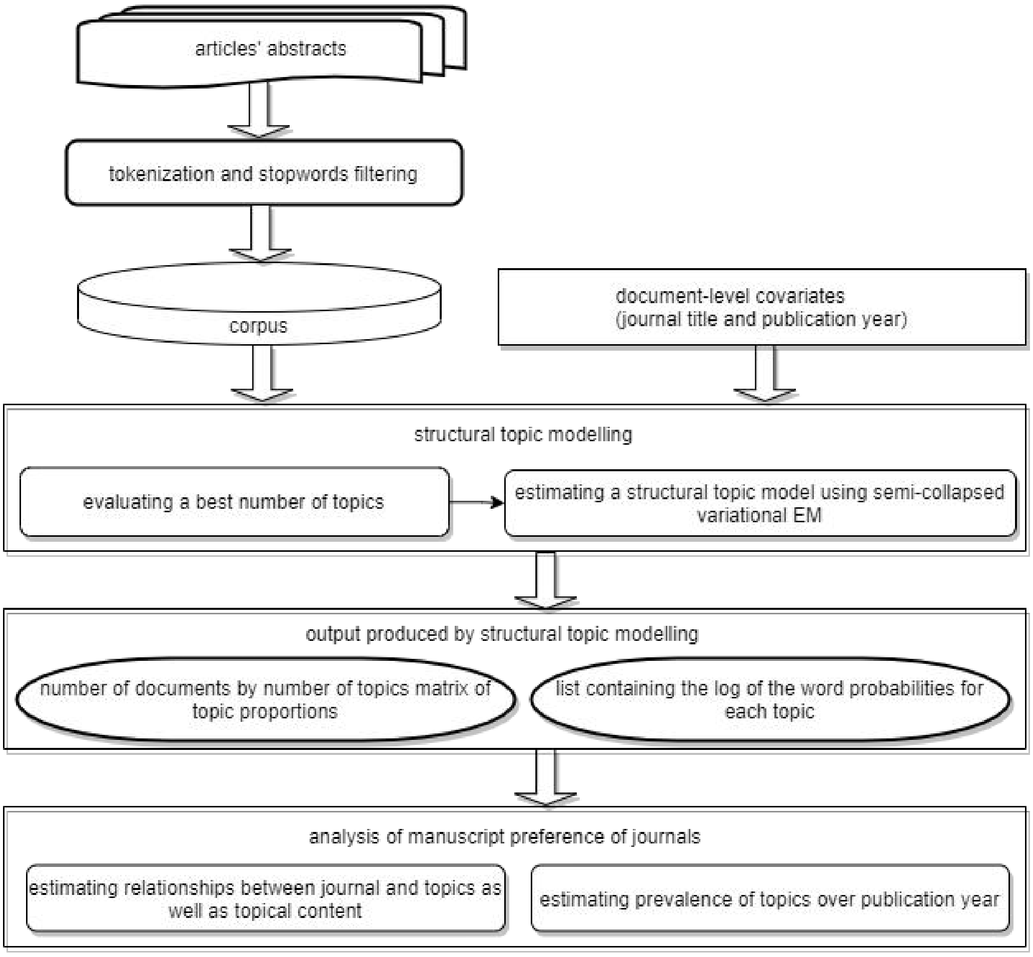

articles' abstracts
tokenization and stopwords filtering
corpus
document-level covariates
(journal title and publication year)
structural topic modelling
evaluating a best number of topics
estimating a structural topic model using semi-collapsed variational EM
output produced by structural topic modelling
number of documents by number of topics matrix of topic proportions
list containing the log of the word probabilities for each topic
analysis of manuscript preference of journals
estimating relationships between journal and topics as well as topical content
estimating prevalence of topics over publication year

Fig.2. Research Methodology

In this research, the abstracts of the articles are used as the text corpus for topic modelling. The research framework includes text data preprocessing, the selection of document-level covariates mixed into the prior distribution of the topic model, fitting a topic model, and estimating the effect of journal and publication year to topics, etc.

3.1 The Role of Abstracts on Extracting the Hidden Research Topics

According to Carole Slade, an abstract is "a concise summary of the entire paper." In most cases, the abstracts contain the most important key words referring to method and content. Abstracts make it easier for readers to know exactly the problems, the aims, the methods, the major findings, and the conclusions reached of the research. So, it is reasonable and feasible for us to extract the hidden research topics from large volumes of abstracts by using topic modelling techniques.

3.2 Text Preprocessing

For topic modelling, tokenization, stopwords filtering and lemmatization are routine text preprocessing tasks.

Tokenization is the process of splitting a text object into smaller units known as tokens. Examples of tokens can be words, characters, numbers, symbols, or n-grams. The most common tokenization process is whitespace/ unigram tokenization. In this process entire text is split into words by splitting them from whitespaces.

Lemmatization is the process of converting a word to its base form. The process that makes this possible is having a vocabulary and performing morphological an alysis to remove inflectional endings.

Stopwords filtering means removing some common words that generally do not contribute to the meaning of a sentence, at least for the purposes of topic modelling or other natural language processing.

To maintain semantic integrity, some special terms should not be split. Examples can be “machine learning”, “big data”, “artificial intelligence” and so on. During text preprocessing, we keep such compound words untouched.

After data wrangling, a STM-compatible document-term matrix is created.

3.3 Selection of Document-Level Covariates

The goal of the Structural Topic Model is to allow researchers to discover topics and estimate their relationships to document metadata where the outputs of the models can be used for hypothesis testing of these relationships.

According to STM, due to the effects of covariates, each paper has its own prior distributing over research topics rather than sharing a global mean and word use within a topic can also vary by covariates. Covariates are user-defined variables based on their hypotheses.

We hypothesize that:

(1) Although some journals belong to the same field, they have their own manuscript bias, which means that one journal may be more interested in one research topic than another.

(2) The prevalence of topics can vary over time.

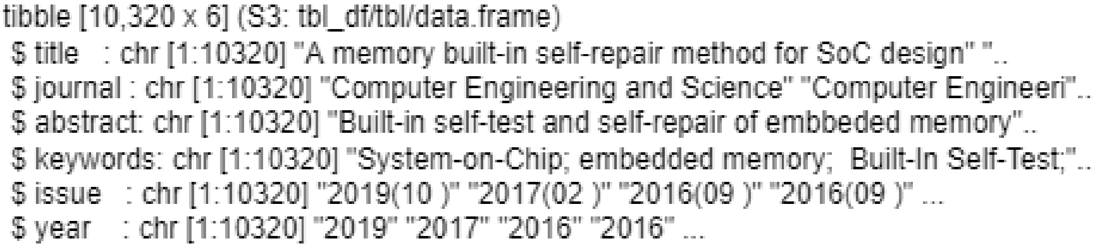
Thus, we choose journal title and publication year from the papers being studied as document-level covariates, after extracting the STM-based topic models, these two hypotheses will be tested by GLM regression analysis.

4. EXPERIMENT

4.1 Research Data and Word Frequency Analysis

We crawled 10320 papers published in three monthly journals of Computer Engineering, Computer Engineering and Design, and Computer Engineering and Science from 2015 to 2021 as research data. These monthly journals are journals with large influence factors in the field of computer engineering in China, and the papers published in them can explain the research hotspots in the field of computer engineering in China to some extent.

Under R language environment, we organize the collected data of all papers into a data frame whose internal structure is shown as Figure 3.



```
tibble [10,320 x 6] (S3: tbl_df/tbl/data.frame)
 $ title   : chr [1:10320] "A memory built-in self-repair method for SoC design" "..
 $ journal : chr [1:10320] "Computer Engineering and Science" "Computer Engineeri"..
 $ abstract: chr [1:10320] "Built-in self-test and self-repair of embbeded memory"..
 $ keywords: chr [1:10320] "System-on-Chip; embedded memory;  Built-In Self-Test;"..
 $ issue   : chr [1:10320] "2019(10 )" "2017(02 )" "2016(09 )" "2016(09 )" ...
 $ year    : chr [1:10320] "2019" "2017" "2016" "2016" ...
```

Fig.3. The Structure of the Research Data

Firstly, we group the papers by the name of journal and count the frequencies of the words used in abstracts of different journals. A word cloud that visualizes the top 20 used words in abstracts of different journals is shown as Figure 4.

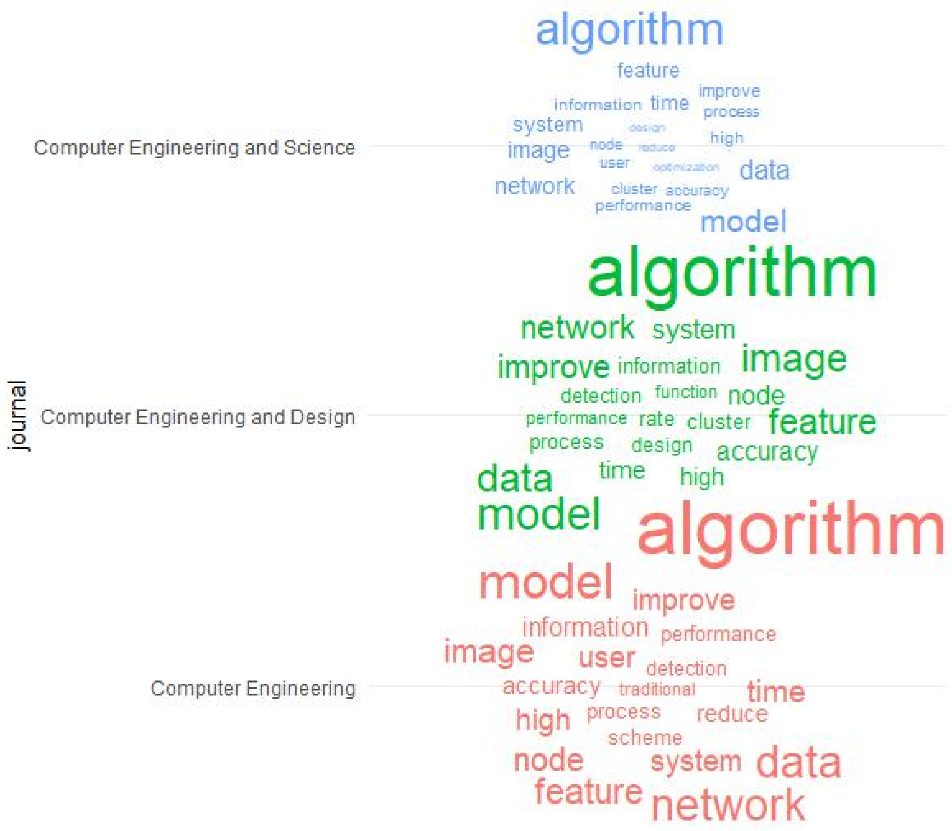


Fig.4. The Top 20 Used Words in Abstracts of Different Journals

We also make relative word frequency analysis using Chi square test to compare frequencies of words among journals.

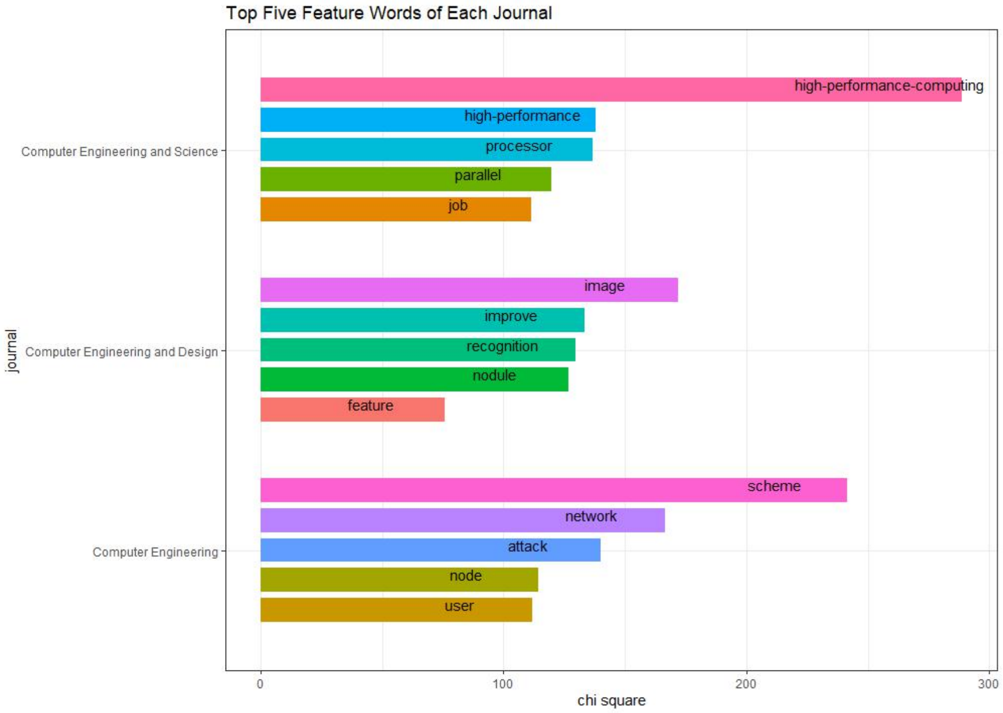


Fig.5. Top 5 Feature Words of Each Journal

Figure 4 and Figure 5 give us a glance into the most important feature words in papers of different journals and how words occur differentially among three journals. For example， the top five words which have significant relationship with the Journal of Computer Engineering and Science are : high-performance-computing, high-performance, processor, parallel, and job, which means these words appear more frequently in the Journal of Computer Engineering and Science than in the other two journals. And the words that appear most frequently in these three journals are “algorithm” and “model”, which fully embodies the disciplinary characteristics of the research.

Further, we cluster the words based on STM to make the research topics hidden in the papers clear.

4.2 Research Topic Extraction from Abstracts of Papers

4.2.1 Evaluating an Appropriate Number of Topics

Topic modeling is an unsupervised machine learning technique, we don’t know in advance what the number of topics should be. Researchers says there is no right answer for the number of topics that is appropriate for any given corpus.

To find the best number of topics for our corpus, we try a number of different values from 4 to 20, stepped by one. This will take a long time to generate numerous topic models with different numbers of topics. To save time, we took full advantage of the parallel processing functions of the furrr package and the purrr package in R.

After fitting all topic models with different given numbers of topics, we explore how many topics are best for our corpus by evaluating the semantic coherence of the topics, the likelihood for held-out datasets, the exclusivity, and the residuals. Our evaluation results are shown as Table 1.

Table 1. Diagnostic Values for Models with Different Values of K

K (the number of topics)	Semantic coherence	Held-out likelihood	Exclusivity	residual
4	-85.32773756	-6.935969114	8.761631631	7.945081582
5	-86.19127173	-6.893167211	8.998142553	7.465400322
6	-84.72031049	-6.832231781	9.04589802	7.004572367
7	-87.13827353	-6.813642038	9.155297098	6.710833409
8	-88.63614322	-6.781543407	9.314256851	6.505600294
…	…	…	…	
19	-99.01027709	-6.608272869	9.720050524	5.026329499
20	-104.2755236	-6.603884743	9.721338178	4.895237816

We visualize the evaluation results to compare the performance of the models at different numbers of topics, as shown in Figure 6.

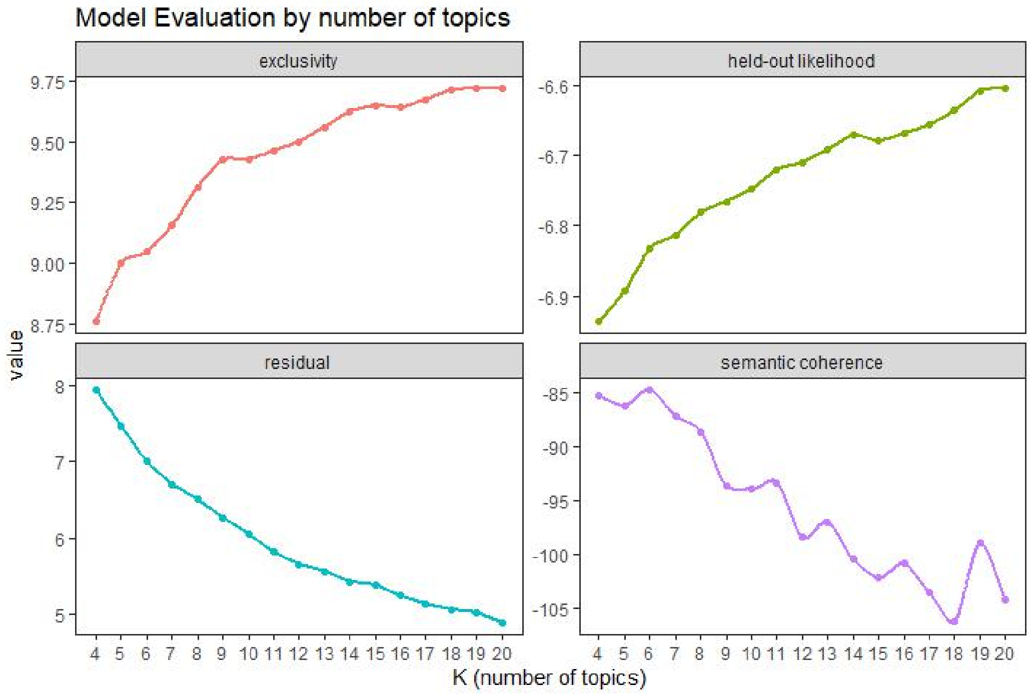


Fig.6. Model Evaluation by Number of Topics

There is a trade-off. Many researchers have verified that Topic Coherence measure is a good way to compare difference topic models based on their human-interpretability. So, we give priority to this metric. From Figure, we can see that the semantic coherence is highest at 6, so we think 6 would be an appropriate number of topics for our corpus.

4.2.2 The Extracted Research Topics

We pick the model with 6 topics to explore. The most important results of STM-based topic modelling are theta Matrix and beta matrix. For our corpus the theta matrix is a 10320 appropriate (number of documents) by 6 (user specified number of topics) matrix representing distribution of topics in documents. Each row of theta matrix is document and values are proportions of corresponding topics. And Beta is a 6(user specified number of topics) by 6430 (number of words in the corpus) matrix containing the natural log of the probability of seeing each word conditional on the topic.

After tidying theta matrix and beta matrix, we get the expected topic proportion and Top N label words of each topic. To make topics more human-interpretable, we use three labeling algorithms: “highest prob”, “frex” and “score” to generate a set of words describing each topic. “Highest prob" words are those which have the highest probabilities of association with the topic. “Frex” stands for “frequent” and “exclusive”, and “frex” words are those which are the most topic distinguishing as they appear frequently only in this topic and not in others. “Score” words are the those the ratio of whose log frequency in a particular topic to their marginal probability in the whole corpus is the highest.

The proportion, label words and interpretation of each topic are shown in Table 2. In Table 2, label words are grouped into three categories: highest prob, frex and score, and are listed in descending order of importance to the topic

Table 2. The Proportion, Label Words and Interpretation of Each Topic

Topic	Expected Topic proportion	Top 5 Label Words	Interpretation
Topic 1	0.1971	Highest Prob: image, feature, algorithm, detection, model Frex: image, texture, histogram, contour, illumination Score: image, feature, recognition, segmentation, texture	Image feature detection
Topic 2	0.1836	Highest Prob: algorithm, model, cluster, data, feature Frex: recommendation, sentence, sentiment, micro-blog, word Score: recommendation, word, text, classification, feature	Text analysis (sentiment, classfication,etc.)
Topic 3	0.1708	Highest Prob: algorithm, optimization, improve, signal, search Frex: particle-swarm, population, particle, mutation, optimum Score: algorithm, signal, optimization, particle-swarm, convergence	Improved optimization algorithm
Topic 4	0.1499	Highest Prob: model, system, data, design, analysis Frex: fault, forecast, passenger, enterprise, visualization Score: software, service, system, fault, model	Software system design
Topic 5	0.1482	Highest Prob: network, node, algorithm, energy, resource Frex: route, relay, allocation, slot, lifetime Score: node, route, network, schedule, protocol	Network routing protocols and algorithms
Topic 6	0.1504	Highest Prob: data, scheme, security, system, attack Frex: encryption, authentication, secret, IO, ciphertext Score: security, encryption, protocol, scheme, attack	Data security

Figure 7 more intuitively shows the prevalence of each topic in the corpus.

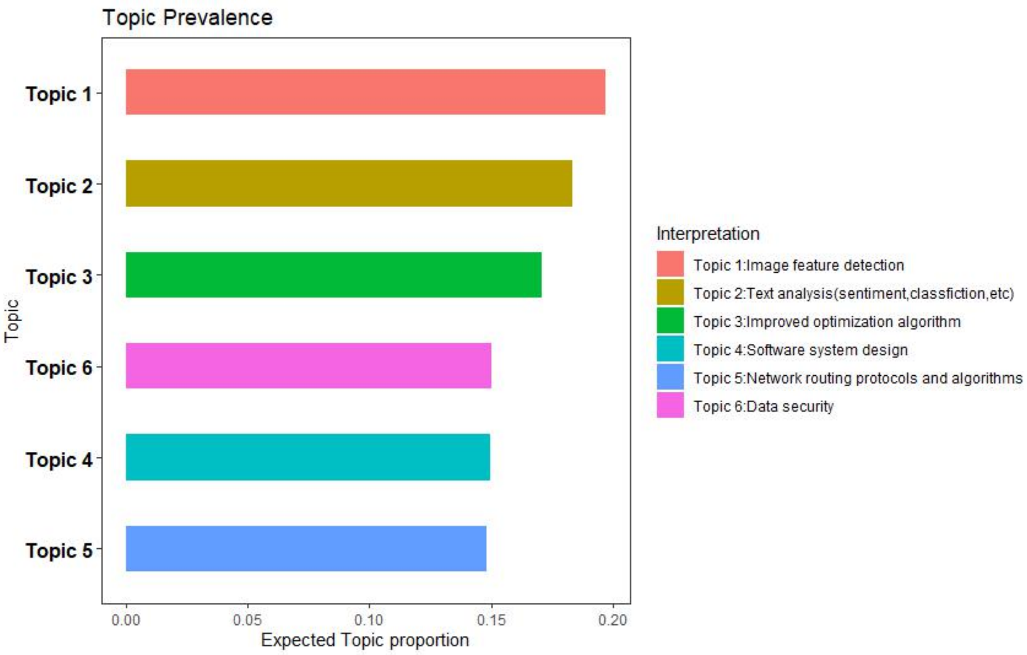


Fig.7. The Prevalence of Each Topic

From the above results of STM-based topic modelling, we conclude that from 2015 to 2021, the research in the field of computer engineering in China mainly focused on the following six topics:

(1) Image feature detection (or extraction).

(2) Text analysis, such as text sentiment analysis, text classification, recommendation, and so on.

(3) Improved optimization algorithms, especially the improvement of particle swarm optimization algorithm.

(4) Software system design, especially the design of software for commercial purposes or enterprise management.

(5) Network routing protocols and algorithms; here, network refers not only to computer networks, but also to sensor networks, wireless sensor networks, etc.

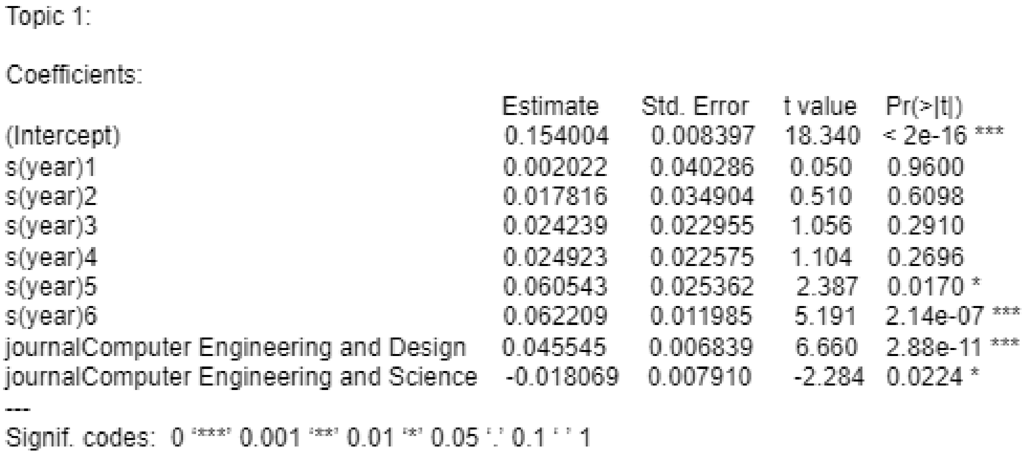
(6) Data security, such as encryption, authentication, cyber-attack, and so on.

4.3 Analysis of Journals' Preference to Research Topics

We continue to explore the relationship between the research topics and journals, as well as the relationship between the research topics and publication years. Once we answer these questions, we can find the journals' preference to research topics and the dynamic changes of their preferences. The most immediate way is to estimate a regression where the outcome is the proportion of each document about a topic and the covariates are journal titles and publication year, running the built-in estimateEffect function in STM R package.

```
effect ← estimateEffect(1:6~s(year)+journal, model.stm, meta = dfm2stm$meta)
```

Taking topic 1 as an example, the summary of the results of estimateEffect is shown as Figure 8.



```
Topic 1:

Coefficients:
                                     Estimate   Std. Error  t value  Pr(>|t|)
(Intercept)                          0.154004   0.008397   18.340   < 2e-16 ***
s(year)1                             0.002022   0.040286    0.050   0.9600
s(year)2                             0.017816   0.034904    0.510   0.6098
s(year)3                             0.024239   0.022955    1.056   0.2910
s(year)4                             0.024923   0.022575    1.104   0.2696
s(year)5                             0.060543   0.025362    2.387   0.0170 *
s(year)6                             0.062209   0.011985    5.191   2.14e-07 ***
journalComputer Engineering and Design   0.045545   0.006839    6.660   2.88e-11 ***
journalComputer Engineering and Science  -0.018069   0.007910   -2.284   0.0224 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Fig.8. The summary of the Results of Estimate Effect

4.3.1 The Effect of Journal to Topics

To visually understand the effect of Journal to the expected topic proportion, we plot the point estimation results of journal covariate under each topic, as shown in Figure 9.

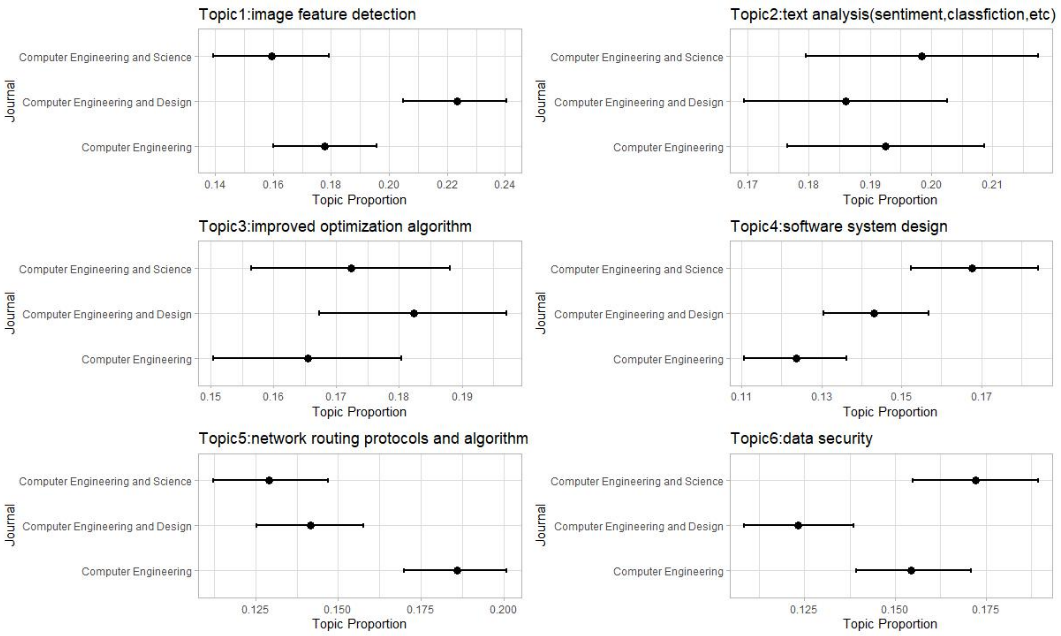


Fig.9. The Effect of Journal on Topics

It can be seen from Figure 8. that for the publication of papers on different research topics in the field of engineering, these three journals have significant difference. For example,

(1) Compared with the other two journals, the journal of Computer Engineering and Design paid more attention to the research of image feature extraction and improved optimization algorithm.

(2) Compared with the other two journals, the journal of Computer Engineering and Science paid more attention to the research of software system design, data security and text analysis.

(3) Compared with the other two journals, the journal of Computer Engineering paid more attention to the research of network routing protocols and algorithm.

4.3.2 The effect of Publication Year to Topics

To visually understand how the prevalence of topics varies over time, we plot the continuous estimation results of publication year covariate under each topic, as shown in Figure 10.

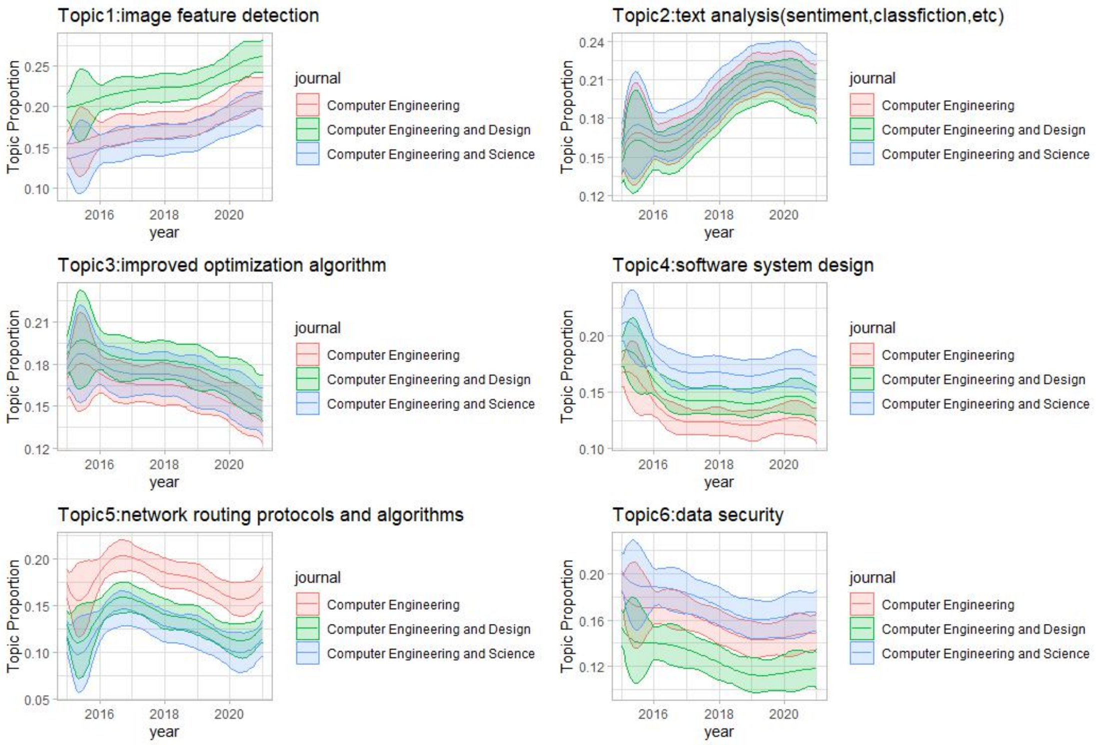


Fig.10. The Effect of Publication Year to Topics

It can be seen from Figure 10. that from 2015 to 2021, the fluctuation trend of the prevalence of each topic over time is roughly the same in these journals. For example,

(1) From 2015 to now, in the three journals, the research popularity of image feature detection has increased year by year.

(2) From 2015 to now, in the three journals, the research popularity of improved optimization algorithm, software system design and data security has shown an overall downward trend.

(3) The research popularity of network routing protocols and algorithms began to decline year by year after reaching the highest point around 2017 and began to rise again after 2020.

(4) The research popularity of text analysis has been increasing year by year since 2015, but it began to decline after 2020.

5. CONCLUSION

At present, academic journals in various disciplines emerge one after another. However, for researchers, this does not mean that it is getting easier and easier to publish their own research results. On the one hand, researchers, especially academic newcomers, do not know how to find suitable publication channels for their papers. On the other hand, journals are also undertaking too much processing work because they receive a large number of inappropriate manuscripts. Scholars have their own research interests, and journals also have their own manuscript preference. Only by combining the two well, can scholars' papers and journals become a perfect match.

This paper proposes a method of using Structure Topic Model to extract hot research topics from abstracts of papers in sci-tech journals and using Generalized Linear Model to analyze the effect of journals and publication year to topics.

Using the proposed method, an empirical analysis was performed on 10320 papers published between 2015 and 2021 in three Chinese journals: “Computer Engineering”, “Computer Engineering and Science” and “Computer Engineering and Design”. The results show that:

(1) Although all three journals belong to the field of computer engineering, there are obvious differences in their manuscript preferences. For example, the journal that most welcomed research in image processing is “Computer Engineering and Design”, the journal that most welcomed research in network routing is “Computer Engineering”, and he journal that most welcomed research in data security is “Computer Engineering and Science”.

(2) With the passage of time, the prevalence of each topic is also constantly changing. According to the predicted results of regression analysis, in the near future, in the field of computer engineering in China, the research topics whose research heat will be still on the rise maybe include “image processing” and “network routing” (here, network refers more to wireless sensor networks).

REFERENCE

[1] Alexander, E., & Gleicher, M. (2015). Task-driven comparison of topic models. *IEEE transactions on visualization and computer graphics*, 22(1), 320-329.

[2] Asuncion, A., Welling, M., Smyth, P., & Teh, Y. W. (2012). On smoothing and inference for topic models. *arXiv preprint arXiv*:1205.2662.

[3] Bischof, J., & Airoldi, E. M. (2012). Summarizing topical content with word frequency and exclusivity. In *Proceedings of the 29th International Conference on Machine Learning (ICML-12)* (pp. 201-208).

[4] Blei, D., Carin, L., & Dunson, D. (2010). Probabilistic topic models. *IEEE signal processing magazine*, 27(6), 55-65.

[5] Blei, D. M., & Lafferty, J. D. (2007). A correlated topic model of science. *The annals of applied statistics*, 1(1), 17-35.

[6] Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. *the Journal of machine Learning research*, 3, 993-1022.

[7] Chandelier, M., Steuckardt, A., Mathevet, R., Diwersy, S., & Gimenez, O. (2018). Content analysis of newspaper coverage of wolf recolonization in France using structural topic modeling. *Biological conservation*, 220, 254-261.

[8] Chen, X., Zou, D., Cheng, G., & Xie, H. (2020). Detecting latent topics and trends in educational technologies over four decades using structural topic modeling: A retrospective of all volumes of Computers & Education. *Computers & Education*, 151, 103855.

[9] Dai J. (2014, 10). The inner logic of core academic journals accepting manuscript within university field——Based on the statistic of the publication of 14 higher education type Chinese core journals (2006~2012). *Heilongjiang Researches on Higher Education* (pp. 29-33). doi: CNKI: SUN: HLJG.0.2014-10-010.

[10] Hu, N., Zhang, T., Gao, B., & Bose, I. (2019). What do hotel customers complain about? Text analysis using structural topic model. *Tourism Management*, 72, 417-426.

[11] Lee, J. M., Kwon, O. J., Lee, H. S., Coh, B. Y., & Park, Y. W. (2011, November). A research on the method to select promising scientific technologies in the condensed matter physics by using journal's editing preference. In *Proceedings of the 2011 ACM Symposium on Research in Applied Computation* (pp. 216-219).

[12] Lei Y. (2021, 32). Association between impacts of English scientific journals and reviewers' academic and reviewing performance: A Publons-based empirical study of medical journals. *Chinese Journal of Scientific and Technical Periodicals* (pp. 206-213). doi: 10.11946/cjstp.202006170598

[13] Li Y., Wen L., Song X., & Guan X.. (2021, 04). Research Status, Theoretical Hotspots and Forward Trends of Tea Science——Based on CiteSpace Visual Analysis of Articles Published in Six Journals Over the Past 20 Years. *Journal of Tea Communication* (736-743).

[14] Qiu, L., & Yu, J. (2018). CLDA: An effective topic model for mining user interest preference under big data background. *Complexity*, 2018.

[15] Roberts, M. E., Stewart, B. M., & Tingley, D. (2019). Stm: An R package for structural topic models. *Journal of Statistical Software*, 91(1), 1-40.

[16] Roberts, M. E., Stewart, B. M., Tingley, D., & Airoldi, E. M. (2013, December). The structural topic model and applied social science. In *Advances in neural information processing systems workshop on topic models: computation, application, and evaluation* (Vol. 4, pp. 1-20).

[17] Roberts, M. E., Stewart, B. M., Tingley, D., Lucas, C., Leder-Luis, J., Gadarian, S. K., ... & Rand, D. G. (2014). Structural topic models for open-ended survey responses. *American Journal of Political Science*, 58(4), 1064-1082.

[18] Séaghdha, D. O. (2010, July). Latent variable models of selectional preference. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics* (pp. 435-444).

[19] Shi, L., Song, G., Cheng, G., & Liu, X. (2020). A user-based aggregation topic model for understanding user's preference and intention in social network. *Neurocomputing*, 413, 1-13.

[20] Sriurai, W., Meesad, P., & Haruechaiyasak, C. (2009). Recommending related articles in wikipedia via a topic-based model. In *9th International Conference On Innovative Internet Community Systems* I2CS 2024. Gesellschaft für Informatik eV.

[21] Taddy, M. (2012, March). On estimation and selection for topic models. In *Artificial Intelligence and Statistics* (pp. 1184-1193). PMLR.

[22] Xu, G., Zhang, Y., & Yi, X. (2008, December). Modelling user behaviour for web recommendation using lda model. In *2008 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology* (Vol. 3, pp. 529-532). IEEE.

[23] Zhao, F., Zhu, Y., Jin, H., & Yang, L. T. (2016). A personalized hashtag recommendation approach using LDA-based topic model in microblog environment. *Future Generation Computer Systems*, 65, 196-206.

[24] Zhou X., Chen D., & Liang W.. (2020, 39). Does the Editorial Bias Exist in Elite Chinese Economics Journals? *South China Journal of Economics* (pp. 105-124).