Postprint: A Study on LDA Noise Topic Filtering Method Based on Keyword Relevance Index (KRI)

Authors: Jiang Tian, Liu Xiaoping, Liu Huizhou

Date: 2023-04-01T16:15:49+00:00

Abstract

[Purpose/Significance] To address the problem that LDA model topic identification results typically contain noise topics, a scientifically effective topic filtering method is established to exclude noise topics and ensure the accuracy of topic identification and subsequent evolution analysis. [Method/Process] Based on the co-occurrence relationships among keywords, a Keyword Relevance Index (KRI) is constructed to facilitate topic screening and filtering through quantitative means. Taking the single-cell research field as a case study, the KRI values for each topic-keyword distribution are calculated and compared with manual interpretation results. [Results/Conclusion] Experimental results indicate that this method can effectively exclude noise topics from LDA model identification results, improve the accuracy of topic identification, and reduce the dependency on manual interpretation in the topic identification process to a certain extent.

Full Text

Preamble

Volume 64, Issue 3, February 2020

Research on Filtering LDA Noise Topics Based on the Keyword Relevance Index (KRI)

Jiang Tian, Liu Xiaoping, Liu Huizhou

National Science Library, Chinese Academy of Sciences, Beijing 100190

Abstract: [Purpose/Significance] To address the problem that LDA model topic identification results often contain noise topics, this study establishes a scientific and effective topic filtering method to eliminate noise topics and ensure the accuracy of topic identification and subsequent evolution analysis.

[Method/Process] Based on the co-occurrence relationships between keywords, a Keyword Relevance Index (KRI) was constructed to enable quantitative topic screening and filtering. Taking the single-cell research field as an example, KRI values for each topic-keyword distribution were calculated and compared with manual interpretation results. [Result/Conclusion] Experimental results demonstrate that this method can effectively eliminate noise topics without clear meaning from LDA model recognition results, improve the accuracy of topic identification, and reduce dependence on manual interpretation in the topic identification process.

Keywords: topic filtering; LDA model; Keyword Relevance Index (KRI)

Classification Number: TP393

DOI: 10.13266/j.issn.0252-3116.2020.03.010

Scientific literature serves as an important knowledge carrier in the development of science and technology, containing rich thematically valuable content. In recent years, many researchers have attempted to employ various methods to analyze massive literature knowledge bases and identify textual topics, thereby assisting researchers in quickly grasping document themes, tracking evolution patterns of scientific domains, and improving research efficiency. Topic modeling methods can deeply mine the implicit relationships among "document-topic-word" at the semantic level and constitute an important approach for disciplinary topic evolution research. The LDA model, or Latent Dirichlet Allocation Model, is a classic and effective probabilistic generative model containing a three-layer Bayesian structure of document-topic-word that can mine latent topic information from large-scale document collections [1]. The LDA model is widely applied in machine learning, information retrieval, biometric identification, and other fields, particularly playing a crucial role in scientific literature topic identification and evolution research. However, LDA model topic identification results often contain a few invalid topics that are difficult to interpret through manual reading, causing serious interference to evolution analysis and requiring topic filtering. This study constructs a Keyword Relevance Index (KRI) based on statistical analysis of multi-word co-occurrence relationships in topic-keyword distributions from LDA model identification results, using it as a basis for screening and filtering topics to remove meaningless noise topics and avoid interference with topic evolution research.

2 Related Research

Since its proposal, the LDA model has received extensive attention and continuous improvement, leading to classic variants such as the Dynamic Topic Model (DTM) [2], Topic Over Time (TOT) model [3], and in recent years, BTM (Biterm Topic Modeling) for short text analysis on social media [4], Hashtag-LDA model [5], and DOLDA (Diagonal Orthant Latent Dirichlet Allocation) for supervised multi-class classification [6].

Beyond improvements to traditional models, optimizing model quality has also been a research focus. As an unsupervised machine learning method, LDA model-generated topics are not always satisfactory—some topics cannot be parsed to extract specific meanings, referred to as noise topics. The existence of noise topics directly affects how LDA models interpret textual data, making it necessary to filter noise topics from LDA model identification results. Current main research methods include:

(1) **Topic Word Determination Method.** This approach assumes that words frequently appearing in the current corpus but not commonly occurring in general English are topic words, while non-topic words are considered noise words to be excluded. Xie Yan et al. [7] utilized an external corpus (Wikipedia 2014) to generate word vectors, calculated semantic similarity between two words based on these vectors, and combined this with the co-document word frequency matrix in topic coherence to achieve external corpus guidance for topic coherence evaluation, enabling more precise topic quality assessment and noise topic filtering through threshold setting.

(2) **Topic Probability Distribution Method.** Qu Jiabin et al. [8] proposed filtering out topics with low occurrence probability across all documents by calculating the probability of topics appearing in all literature. This method assumes that only topics with high probability across all documents are core topics reflecting main content within a timeframe, important for topic evolution analysis. Conversely, topics with low probability are likely marginal or meaningless, interfering with analysis. However, emerging or declining topics within a certain period do not have high occurrence probabilities and would be easily filtered out by this method, negatively impacting the accuracy and scientific validity of topic evolution analysis.

(3) **Information Entropy-Based Filtering Method.** This commonly used method calculates topic information entropy based on the "topic-word" distribution output by LDA models. The more uniform the probability distribution of words under a topic, the higher the topic's information entropy. By setting an entropy threshold, semantically broad topics can be filtered. The calculation is shown in Formula (1) [9]:

$$Entropy(T) = -K \sum_{j=1}^{m} P_j \ln(P_j) \quad \text{Formula (1)}$$

where K is a constant, P_j represents the occurrence probability of the j-th word in topic T, and the topic contains m words total. While this method can exclude invalid topics to some extent, it has significant limitations: first, threshold determination is highly subjective; second, it cannot effectively filter topics with skewed topic-keyword distributions that still cannot be manually interpreted.

(4) **"Garbage Topic"-Based Filtering Method.** Li Baoli et al. [10] proposed filtering topics by calculating similarity between LDA-generated topics and predefined "garbage topics" that cannot highlight document content. Smaller similarity indicates better content representation. Setting an appropriate threshold filters out topics with high similarity. "Garbage topics" can be defined from either "topic-word" or "document-topic" perspectives.

(5) **Heuristic Methods.** Y. L. Chang et al. proposed using Spike-and-Slab prior distributions for feature extraction based on documents [11-12], where words belonging to the slab distribution are retained as features for topic estimation, while words in the spike distribution are filtered out. This improves model interpretability and sparsity but lacks guidance for topic semantic extraction.

In summary, current topic filtering methods have respective limitations and unsatisfactory filtering effects, particularly for emerging or declining topics with fewer documents, which are easily misidentified as noise topics. Therefore, exploring new topic filtering methods to improve accuracy is necessary. This study constructs a Keyword Relevance Index (KRI) for topic filtering through statistical analysis of multi-keyword co-occurrence frequencies in documents, assigning different weights to different co-occurrence word counts to strengthen the "contribution rate" of multi-keyword co-occurrence in revealing topic semantics.

3 LDA Noise Topic Filtering Based on KRI

3.1 Topic Identification Based on LDA

In the LDA model topic identification process, the number of topics directly affects identification effectiveness [13-14]. Too many topics lead to overly sparse distributions and high similarity; too few result in overly broad topics that cannot accurately reveal core content. This study employs a combination of average topic similarity and perplexity to determine the optimal topic number.

Perplexity evaluates language model quality by assigning higher probabilities to test sets [15]. The LDA perplexity formula is:

$$Perplexity(D) = \exp\left\{-\frac{\sum_{d=1}^{M} \log p(w_d)}{\sum_{d=1}^{M} N_d}\right\} \quad \text{Formula (2)}$$

where D represents the test corpus, M is the number of documents in the test set, N_d is the word count of document d, and p(w_d) is the word probability distribution in document d.

Average topic similarity measures the mean difference degree among all topics, typically using Jensen-Shannon divergence [16], calculated as:

$$avg\_sim(T_i, T_j) = \frac{\sum_{i=1}^{K} \sum_{j=i+1}^{K} JS(T_i||T_j)}{K \times (K-1)/2} \quad \text{Formula (3)}$$

where T_i and T_j represent two topics, and JS(T_i||T_j) is the JS divergence between them.

From a generalization perspective, lower perplexity indicates stronger LDA model generalization ability [14]; from a topic extraction perspective, smaller average topic similarity indicates greater differences between topics and fewer duplicate topics, corresponding to better LDA identification effects [17]. Typically, as topic numbers increase, average similarity increases while perplexity generally decreases, with obvious inflection points reflecting significantly enhanced model generalization at those topic counts [8]. This study comprehensively considers both metrics, selecting inflection points where the perplexity curve's decline slows, comparing average similarity values, and combining actual identification effects to determine the optimal topic number.

3.2 KRI Construction

To achieve effective topic filtering, an objective evaluation metric is needed. In information science, co-occurring words are considered to reveal the same topic meaning, enabling text mining through co-word analysis [18-19]. This study screens effective topics and filters noise topics by statistically analyzing keyword co-occurrence relationships in topic-keyword distributions.

Traditional co-word analysis only considers pairwise co-occurrence. The pairwise co-occurrence frequency formula for keywords in a topic is:

$$coof(W_a, W_b) = \frac{n}{N} \quad \text{Formula (4)}$$

where W_a represents keyword a, W_b represents keyword b, n is the number of documents containing both keywords a and b, and N is the total number of documents containing either keyword.

Beyond pairwise co-occurrence, keyword distributions also exhibit higher-order co-occurrence (three-word, four-word, etc.). Similar formulas yield:

$$coof(W_a, W_b, W_c) = \frac{n}{N} \quad \text{Formula (5)}$$

$$coof(W_a, W_b, W_c, W_d) = \frac{n}{N} \quad \text{Formula (6)}$$

This pattern extends to all higher-order co-occurrence frequencies.

In a topic-keyword distribution, more keywords co-occurring in the same document and more documents containing these keywords indicate higher topic

"concentration," greater topic revelation accuracy, and clearer meaning. Higher-order co-occurrence reveals greater topic concentration than lower-order co-occurrence. To emphasize the "contribution rate" of multi-word co-occurrence in revealing topic semantics, the square of the co-occurring keyword count is used as weight. Based on this discussion, the Keyword Relevance Index (KRI) is constructed as:

$$KRI = 2^2 \sum coof(W_a, W_b) + 3^2 \sum coof(W_a, W_b, W_c) + \cdots + n^2 \sum coof(W_a, W_b, W_c, \ldots W_n) \quad \text{Formula (7)}$$

KRI reflects the strength of keyword co-occurrence within a topic, revealing the distribution concentration of the topic's keywords across different documents, and providing a quantitative means for noise topic identification.

4 Empirical Study

4.1 Dataset Construction

Single-cell research is a hotspot in life sciences and an interdisciplinary field integrating life science, materials science, and chemistry. Single-cell technology is widely applied in preimplantation genetic diagnosis [20], stem cell and regenerative medicine [21-22], cancer diagnosis and treatment [23], environmental monitoring [24], and other areas, involving numerous subfields that place high demands on topic identification methods. To verify the effectiveness of the KRI-based topic filtering method, the single-cell field was used for LDA topic identification and filtering. From the Web of Science Core Collection, 54,848 single-cell related documents from 1990-2018 were retrieved (document types: Review, Article, Proceedings Paper, and Letter). Titles, abstracts, and author keywords were extracted as the corpus for topic identification and analysis. Python programs were written to call the NLTK library for text preprocessing including tokenization, part-of-speech tagging, stemming, lemmatization, and stopword removal.

4.2 LDA-Based Topic Identification and Manual Interpretation

The LDA model was applied to the constructed dataset. Perplexity and average topic similarity were calculated for topic numbers K ranging from 5 to 100 in increments of 5, with the resulting curve shown in Figure 1 [FIGURE:1]. As topic numbers increase, perplexity shows a downward trend, slowing after K=30, indicating enhanced model generalization at K=30 [8], and plateauing after K=45.

Comprehensively considering perplexity and average similarity values, the LDA model output with 30 topics was selected. Manual interpretation was performed by analyzing high-probability keywords in the topic-keyword distributions and their semantic relationships.

Table 1 lists some of the 30 topics identified by the LDA model, showing only the top 10 high-probability words per topic. Some topics' keywords effectively reveal content: Topic 13 contains mostly "gene expression regulation"-related terms, while Topic 23 contains "microbial fuel cell"-related vocabulary. However, not all topics can be interpreted. For example, Topic 7 contains high-probability words like "comparison," "datum," "reaction," "extent," "degree"—all very broad terms that cannot represent specific meanings, necessitating topic filtering to exclude such noise topics.

4.3 Topic Filtering Using KRI

Based on the KRI constructed in Section 3.2, the LDA model results for K=30 were filtered by calculating and ranking KRI values for each topic-keyword distribution (Table 2). Topics with KRI values >100 are marked (H), while those <20 are marked (L).

Table 2 shows that topics uninterpretable through manual reading all have low KRI values, in this case below 20, demonstrating that the KRI index effectively filters topics.

4.4 Comparison with Word Co-occurrence Clustering Methods

The KRI topic filtering method draws from co-word analysis but differs from traditional approaches. Co-word analysis assumes that when two keywords appear together in a document, they are related, with more co-occurrences indicating stronger relationships [25]. This method only considers pairwise co-occurrence, ignoring actual multi-keyword co-occurrence situations, which better reflect topic concentration and validity. The KRI method employs multi-word co-occurrence analysis, calculating KRI values through statistical analysis of multi-word co-occurrence frequencies.

For comparison, co-word clustering analysis was also performed using VOSviewer software, yielding the keyword co-occurrence network shown in Figure 2

. This analysis identified six major topics: gene expression regulation (red), neural regulation and calcium control (green), microbial fuel cells (blue), single-cell culture and analysis (yellow), single-cell dynamic modeling (purple), and single-cell gel electrophoresis (cyan). Further analysis of keyword clusters within each major topic revealed subtopics. For instance, gene expression regulation could be divided into stem cell gene expression regulation, tumor cell heterogeneity, cancer diagnosis and treatment, single-cell in situ hybridization, and flow cytometry—five subtopics. Single-cell culture and analysis contained four subtopics: culture, live imaging, microfluidic chips, and cell migration.

While co-word analysis accurately identifies major research directions and their hierarchical relationships, it has limitations: (1) many isolated words lack connections with others, affecting topic parsing (e.g., core keywords like "single-

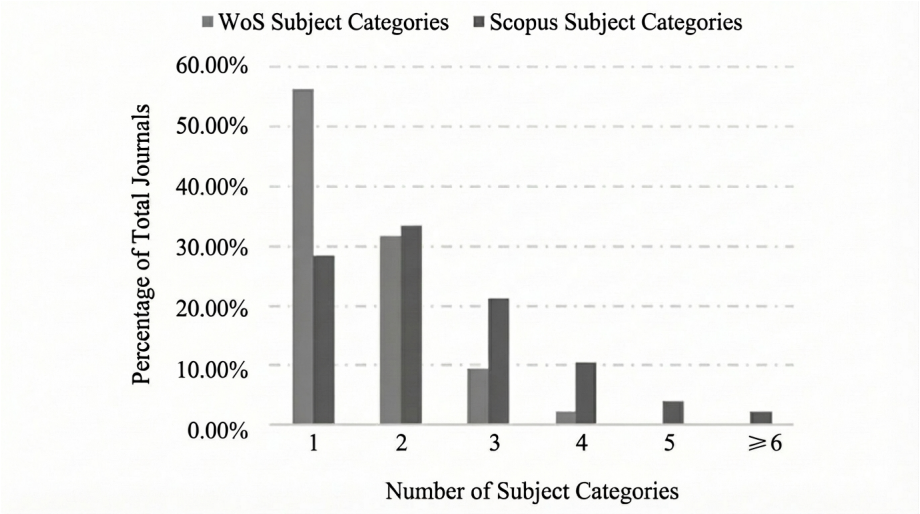

Figure 1: Figure 2

cell oil," "preimplantation genetic diagnosis," "immune response," "cell cycle," "single-cell whole genome sequencing," and "embryogenesis" appear as isolated nodes); (2) it is constrained by word frequency, easily identifying conventional and hot topics but struggling with smaller, marginal topics, compromising comprehensiveness; (3) controlling topic granularity is difficult—LDA models can adjust granularity through topic number selection, while co-word analysis results are sensitive to co-occurrence frequency threshold settings; (4) for scientific literature mining, the focus extends beyond topic identification to topic evolution, where LDA models offer three main evolutionary analysis approaches (discrete-then-discrete, time-integrated) enabling quantitative evolution study through topic intensity and similarity calculations, whereas co-word analysis applies temporal elements simply, poorly reflecting finer developmental changes.

4.5 Comparison with "Edge Topic Identification and Filtering Based on Topic Distribution"

Statistical analysis of the document-topic distribution for K=30 shows each topic's document count ratio to the total dataset. As shown in Figure 3

, the KRI index curve and document-topic probability curve show similar trends, particularly for high-probability topics. However, they diverge for low-probability topics—for example, Topic 19 has low probability but a valid KRI value within the effective range and was manually interpreted as meaningful.

For comparative visualization, log KRI values and topic probabilities were sorted in descending order (Table 3). Some high-probability topics lack clear meaning,

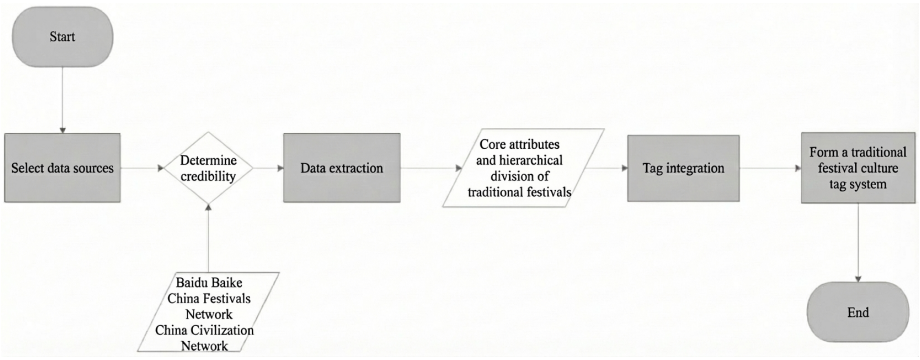

Figure 2: Figure 3

while some low-probability topics are valid. Comparing manual interpretation, KRI index, and topic probability demonstrates that KRI filtering outperforms the "edge topic identification based on topic distribution" method.

4.6 Significance of KRI for Determining Optimal Topic Number

In LDA models, the topic number K directly affects model quality and topic generation—too many or too few topics both impact results. In Figure 1, inflection points appear at both K=30 and K=45, indicating significantly enhanced model generalization. After K=45, the perplexity curve flattens, suggesting better identification effects at K=45. However, from an average similarity perspective, K=30 shows greater average JS distance and lower average similarity (Table 4), indicating better identification.

Manual interpretation and KRI calculation for both K=30 (Table 2) and K=45 (Table 5) reveal that K=45 produces 19 invalid topics (42.2%), far exceeding the 20% invalid topic rate at K=30 (Table 6). While both identify some common valid topics (high-KRI core topics like "microbial fuel cell," "single-cell gel electrophoresis," "single-cell oil," "preimplantation genetic diagnosis"), K=30 yields more valid non-common topics. At K=45, topics like "air-cathode microbial fuel cell" and "microbial fuel cell performance" are subtopics of "microbial fuel cell," indicating overly fine granularity. Comprehensive analysis shows K=30 produces better LDA identification results. Thus, when selecting optimal topic numbers via perplexity-similarity curves, calculating KRI indices at different inflection points can assist in determining the optimal topic number.

Conclusion

The presence of noise topics in LDA model results affects topic identification accuracy and subsequent evolution analysis. This paper proposes a KRI-based topic filtering method that effectively eliminates meaningless noise topics, improving identification precision and ensuring scientific validity of subsequent

evolution path construction. Comparative analysis shows KRI outperforms the "edge topic identification based on topic distribution" method. The KRI index reduces over-reliance on manual interpretation and provides reference value for optimal topic number selection.

However, KRI lacks a clear boundary between valid and invalid topics, with values declining gradually rather than precipitously, serving only as a reference requiring final manual judgment. This study only verified KRI's effectiveness for LDA models; its applicability to other topic models requires further research.

References

[1] Blei DM, Ng AY, Jordan MI. Latent Dirichlet allocation [J]. Journal of machine learning research, 2003(3): 993-1022.

[2] Blei DM, Lafferty JD. Dynamic topic model [C]//Proceedings of the 23rd international conference on machine learning. New York: ACM, 2006: 113-120.

[3] Wang XR, McCallum A. Topic over time: A non-markov continuous-time model of topical trends [C]//Proceedings of the 12th international conference on machine learning. New York: ACM, 2006: 424-433.

[4] Yan XH, Guo JF, Lan YY, et al. A biterm topic model for short texts [C]//Proceedings of the 22nd international conference on World Wide Web. New York: ACM. 2013: 1445-1455.

[5] Zhao F, Zhu YJ, Jin H, et al. A personalized hashtag recommendation approach using LDA-based topic modeling in microblog environment [J]. Future generation computer systems, 2016, 65: 196-206.

[6] Magnusson M, Jonsson L, Villani M. DOLDA: a regularized supervised topic model for high-dimensional multi-class regression [EB/OL]. [2019-09-08]. https://doi.org/10.1007/s00180-019-00891-1.

[7] Xie Yan. Topic optimization filtering method and research application [D]. Dalian: Dalian Maritime University, 2015: 26-27.

[8] Qu Jiabin, Ou Shiyan. Disciplinary topic evolution analysis based on topic filtering and association [J]. Data analysis and knowledge discovery, 2018, 2(1): 64-75.

[9] Mackay DJC. Information theory, inference, and learning algorithms [M]. Cambridge: Cambridge University Press, 2003.

[10] Li Baoli, Yang Xing. Research topic evolution analysis based on LDA model and topic filtering [J]. Small microcomputer systems, 2012, 3(12): 2738-2743.

[11] Ishwaran H, Rao JS. Spike and slab gene selection for multigroup microarray data [J]. Journal of the American Statistical Association, 2005, 100(471): 764-780.

[12] Chang YL, Lee KF, Chien JT. Bayesian feature selection for sparse topic model [C]//IEEE international workshop on machine learning for signal processing (MLSP). Santander: IEEE, 2011: 1-6.

[13] Pönweiser M, Grün B. Finding scientific topics revisited [C]//Carpita M, Brentari E, Qannari EM. Advances in latent variables. Berlin: Springer, 2014: 93-100.

[14] Guan Peng, Wang Yuefen. Research on determining optimal topic number for LDA topic model in scientific intelligence analysis [J]. Modern library and information technology, 2016(9): 42-50.

[15] Grossman DA, Frieder O. Information retrieval: algorithms and heuristics [M]. Berlin: Springer, 2004.

[16] Lee L. On the effectiveness of the skew divergence for statistical language analysis [C]//Richardson TS, Jaakkola TS. Proceedings of the Eighth International Workshop on Artificial Intelligence and Statistics. Key West: Society for Artificial Intelligence and Statistics, 2001: 65-72.

[17] Cao J, Xia T, Li J, et al. A density-based method for adaptive LDA model selection [J]. Neurocomputing, 2009, 72(7/9): 1775-1781.

[18] Callon M, Courtial JP, Laville F. Co-word analysis as a tool for describing the network of interactions between basic and technological research: the case of polymer chemistry [J]. Scientometrics, 1991, 22(1): 155-205.

[19] Wang ZY, Li G, Li CY, et al. Research on the semantic-based co-word analysis [J]. Scientometrics, 2012, 90(3): 855-875.

[20] Turner K, Lynch C, Rouse H, et al. Direct single-cell analysis of human polar bodies and cleavage-stage embryos reveals no evidence of the telomere theory of reproductive ageing in relation to aneuploidy generation [J]. Cells, 2019, 8(2): 1-17.

[21] Fletcher RB, Das D, Gadye L, et al. Deconstructing olfactory system cell trajectories at single-cell resolution [J]. Cell system cell, 2017, 20(6): 817-830.

[22] Jacobsen SEW, Nerlov C. Haematopoiesis in the era of advanced single-cell technologies [J]. Nature cell biology, 2019, 21(1): 2-8.

[23] Gerdes MJ, Gökmen-Polar Y, Sui Y, et al. Single cell heterogeneity in ductal carcinoma in situ of breast [J]. Modern pathology, 2018, 31(3): 406-417.

[24] Davis KM, Isberg RR. Defining heterogeneity within bacterial populations via single-cell approaches [J]. Bioessays, 2016, 38(8): 782-790.

[25] Kostoff RN. Co-word analysis [C]//Bozeman B, Melkers J. Evaluating R&D impacts: methods and practice. New York: Springer, 1993: 63-78.

Author Contributions

Jiang Tian: Proposed research ideas and technical approach, conducted experiments, analyzed data, wrote the paper.

Liu Xiaoping: Revised the paper.

Liu Huizhou: Proposed research direction, revised the paper and final version.

Figures

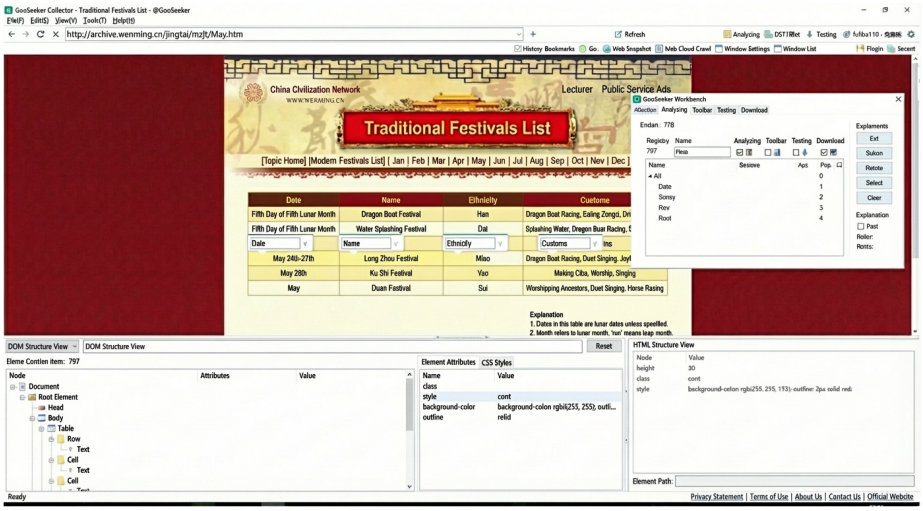

Figure 3: Figure 4

Source: ChinaXiv — Machine translation. Verify with original.