# In-Depth Text Topic Mining Based on Association Rules: An Applied Research Postprint

**Authors:** Ruan Guangce, Xia Lei

**Date:** 2017-11-08T00:00:00+00:00

## Abstract

[Objective] To accurately comprehend latent knowledge associations within textual information and enrich methodologies for text knowledge mining. [Method] This study integrates topic modeling and association rule mining, employing the LDA topic model to extract topic sets from texts, thereby achieving dimensionality reduction while enabling semantic space representation; association rules are then utilized to further excavate semantic associations among topics. [Results] With appropriately set support and confidence thresholds, latent knowledge associations in texts can be effectively mined, facilitating a profound "understanding" of the texts. [Limitations] During data preprocessing, the design of user-defined dictionaries may influence experimental outcomes. [Conclusion] This paper proposes a novel approach for mining latent semantic associations in unstructured text information, enhancing the effectiveness of knowledge discovery from textual data.

## Full Text

### Mining Document Topics Based on Association Rules

Guangce Ruan and Lei Xia
(Department of Information Management, East China Normal University, Shanghai 200241, China)
(Shanghai Library, Shanghai 200031, China)

**Abstract:**
[Objective] This study aims to accurately identify potential knowledge correlations within textual information and enrich the methodology of text mining. [Methods] We integrate topic modeling with association rules, employing the LDA topic model to extract topic sets from texts, thereby achieving dimensionality reduction while simultaneously representing documents in semantic space. Association rules are then applied to further mine the semantic relationships

among these topics. [Results] By setting reasonable support and confidence thresholds, we can effectively uncover latent knowledge associations in texts, enabling deeper "understanding" of the content. [Limitations] During data preprocessing, the design of the user-defined dictionary may influence experimental outcomes. [Conclusions] This paper proposes a novel approach for mining latent semantic associations in unstructured text information, improving the effectiveness of knowledge discovery from textual sources.

**Keywords:** Association rules; Topic model; Text topics
**Classification Number:** G350

With the development and popularization of information technology and internet communications, massive amounts of textual information have been generated. This rapid growth presents unprecedented challenges for information processing and retrieval. Understanding text not only facilitates information retrieval and content discovery but also provides valuable insights for effective classification and organization. However, the loose organizational structure of textual information reduces usability for general users, who often become lost in complex information spaces. The volume of text has far exceeded human capacity for comprehension and summarization, making manual information extraction and knowledge condensation impossible. Consequently, effectively organizing and managing these resources with computational assistance and leveraging information technology to mine implicit knowledge from large text collections has become a major challenge.

As understanding of text has evolved, researchers have pursued deeper comprehension to enable both computers and humans to better "understand" textual content. Deep text understanding can facilitate text mining and natural language processing tasks, such as automated question-answering systems, while also uncovering latent semantics to provide technical support for information professionals. Before the advent of topic models, text representation primarily relied on the vector space model and statistical language models. Despite methodological differences, both approaches map documents through a "text→word" transformation. Traditional methods represent texts in dictionary space, ignoring much important information and failing to achieve semantic understanding. Topic models introduce a semantic dimension, condensing textual information at the semantic level and enabling a "text→semantic→word" mapping. This paper combines association rules with topic models, constructing topic sets from large text collections and using association rule algorithms to build relationships among topics, thereby achieving deep mining of textual themes. We demonstrate this approach through experiments on news reports about the "Belt and Road Initiative."

**Corresponding Author:** Guangce Ruan, ORCID: 0000-0001-8685-5234, E-mail: rgc1976@126.com.
*This work is supported by the Shanghai Philosophy and Social Science General Project "Research on Interdisciplinary Knowledge Discovery Based on Topic Models" (Project No.: 2016BTQ002).*

### 3.3 Deep Text Topic Mining Based on Association Rules

In this framework, the text collection forms a transaction group where each document represents a transaction. By applying association rule algorithms with appropriate support and confidence thresholds, we can identify topic associations within the collection and discover latent knowledge. Texts typically revolve around specific themes, and information within a domain often contains direct or indirect semantic connections. Identifying semantically related entities helps users better comprehend the collection and understand its implicit knowledge.

Topic models represent texts in semantic space, while our approach aims to discover association rules among these topics, calculating the strength of semantic connections between entities and describing strongly associated themes. Consider a text space D with topic set T and vocabulary set W, where D = {$d_1$, $d_2$, ···, d } represents documents, T = {$t_1$, $t_2$, ···, t } represents topics, and w W denotes topic terms in document i. In association rule processing, D represents transactions, d represents individual transactions with a unique transaction identifier (TID) and item list, and W represents the itemset composed of topic terms. The transaction set containing w is denoted as {d | W d , d D}.

Combining association rules with topic models offers three main advantages: (1) It resolves semantic relationships between keywords. Traditional keyword extraction relies on statistical methods that may identify high-frequency terms lacking semantic connections to other words. (2) It achieves dimensionality reduction in semantic space. LDA serves as a dimensionality reduction tool that learns document representations in topic space through machine learning, converting documents from term space to topic space. (3) It discovers knowledge associations among terms. Association rule algorithms can mine multi-word associations through support and confidence settings, revealing direct and indirect connections.

## 4 Experiments and Discussion

**4.1 Experimental Data and Procedure** The fundamental approach is illustrated in [Figure 1: see original paper]. The process begins by acquiring the target text dataset and constructing a domain-specific lexicon for preprocessing tasks such as segmentation and stop-word removal. The preprocessed collection undergoes topic extraction via the LDA model. Based on the generated "document-topic" distribution, high-probability topic feature words are selected to represent documents, achieving dimensionality reduction while preserving semantic features. This reduces feature vector dimensions while improving information extraction efficiency and accuracy. The resulting feature term collection can be treated as an itemset for association rule mining.

We retrieved 13,392 news articles about the "Belt and Road Initiative" from the National Library' s WiseSearch newspaper database, totaling 73.7MB. During preprocessing, we constructed a custom dictionary and stop-word list, using Python and the Jieba segmentation component. The custom dictionary included domain-specific terms such as "Belt and Road Initiative," "Maritime Silk Road," "Silk Road Economic Belt," and "Chinese Dream" to prevent inappropriate segmentation. To reduce dimensionality, we defined a stop-word list removing high-frequency journalistic terms like "reporter," "daily," "evening news," and "correspondent."

**4.2 Text Topic Mining** Effective topic identification forms the foundation of our experiments. We used one-third of the documents for model training and the remaining for topic identification and dimensionality reduction. In LDA, the number of topics T must be predefined, typically increasing with corpus size. We determined the optimal topic count using perplexity, a common metric in statistical language models that measures the inverse geometric mean of sentence similarity, decreasing as similarity increases. Lower perplexity indicates better performance. [Figure 2: see original paper] shows the perplexity calculation results with 1,000 iterations and 10 topic words per topic.

The perplexity curve exhibits an inflection point at 230 topics before stabilizing, so we selected 230 topics with parameters $\alpha = 50/230$, $\beta = 0.01$, k = 230, 10 topic words per topic, and 1,000 iterations. After modeling, we selected the top 3 highest-probability topics for each document, representing each text with 30 topic terms in semantic space. Partial results are shown in [Figure 3: see original paper].

**4.3 Deep Topic Mining Based on Association Rules** We employed the Apriori algorithm for mining topic associations. Based on the dimensionality reduction results, each document t $= \{w_1, w_2, \cdots, w\}$ represents a transaction where w denotes topic terms. We used R for association rule analysis on the reduced data, with basic information shown in .

During mining, setting only minimum support and confidence may generate uninteresting rules. To address this, we incorporated the lift metric, which measures rule value through correlation analysis and describes the influence of itemset X on Y. Lift is calculated as:

lift(X Y) = support(X Y) / (support(X) × support(Y))

A lift value of 1 indicates independence between X and Y, while values less than 1 suggest mutual exclusivity. Rules with lift $> 3$ are generally considered valuable. We therefore integrated lift when determining support thresholds. Experiments with different support and confidence settings are visualized in [Figure 4: see original paper].

Figures 4(a) and 4(b) use support values of 0.1 and 0.2 with 80% confidence, while Figures 4(c) and 4(d) use the same support values with 95% confidence.

With support = 0.1, both experiments generated over 100,000 rules, while support = 0.2 yielded over 10,000 rules. Figure 4(b) shows most high-strength rules with lift $> 3$ and higher confidence, leading us to adopt these parameters.

The experiment produced 10,228 rules with average confidence of 0.9982 and average lift of 3.862. After sorting and manual analysis of high-association rules, partial results are shown in .

To analyze implicit knowledge, we classified rules by lift value, using the right-hand side (Rhs) as topic knowledge. Different lift ranges revealed distinct topic patterns: high lift values (lift $> 8$) described "ports,""economic belts,""railways," and "logistics" ; medium lift ($5 <$ lift $< 8$) covered "infrastructure," "funds," "industries,"and "innovation"; lower lift ($3 <$ lift $< 5$) described "concepts,""trade," and "Silk Road." This demonstrates varying association strengths among Belt and Road news themes.

For deeper analysis, we treated Rhs as topic knowledge and Lhs as descriptive information, computing descriptions for each topic. As shown in , the results reveal semantic descriptions of key topics. For instance, fund-related topics involve infrastructure construction and Silk Road investment projects, while innovation topics relate to industrial and ecological projects. These semantic connections enhance understanding of topic knowledge descriptions.

The results enable discovery of latent knowledge in domain-specific collections, representing content at the semantic dimension and helping information professionals identify valuable implicit knowledge. Our approach combines topic modeling with association rules to mine implicit theme connections, using LDA for semantic representation and dimensionality reduction, then applying association rules to uncover semantic relationships. Experiments validate the method, enriching knowledge mining approaches and assisting analysts in understanding large text collections.

---

## References

[1] Lazer D, Pentland A, Adamie L, et al. Computational Social Science [J]. Science, 2009, 323(5915): 721-723.
[2] Salton G, Wong A, Yang C. A Vector Space Model for Automatic Indexing [J]. Communications of the ACM, 1975, 18(11): 613-620.
[3] Ponte J M, Croft W B. A Language Modeling Approach to Information Retrieval [C]. In: Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. 1998: 275-281.
[4] Agrawal R, Imieliński T, Swami A. Mining Association Rules Between Sets of Items in Large Databases [C]. In: Proceedings of the 1993 ACM SIGMOD International Conference on Management of Data. 1993: 207-216.
[5] Wang Jianquan, Ji Shaobo. Research and Application on Auto-word

Building [J]. Computer Science, 2014, 41(11): 256-259.
[6] He Yu, Feng Jianlin, Wang Yuanzhen. Text Classification Based on Maximal Association Rule [J]. Computer Science, 2006, 33(11): 143-145.
[7] Cherfi H, Napoli A, Toussaint Y. Towards a Text Mining Methodology Using Association Rule Extraction [J]. Soft Computing, 2006, 10: 431-441.
[8] Sekhavat Y A, Hoeber O. Visualizing Association Rules Using Linked Matrix, Graph, and Detail Views [J]. International Journal of Intelligence Science, 2013, 3(1): 34-49.
[9] Liu Fei, Huang Xuanjing, Wu Lide. Approach for Extracting Thematic Terms Based on Association Rules [J]. Computer Engineering, 2008, 34(7): 81-83.
[10] Maedche A, Staab S. Discovering Conceptual Relations from Text [C]. In: Proceedings of the 14th European Conference on Artificial Intelligence (ECAI), Berlin, Germany. 2000: 321-325.
[11] Schutz A, Buitelaar P. RelExt: A Tool for Relation Extraction from Text in Ontology Extension [C]. In: Proceedings of the 4th International Semantic Web Conference. 2005: 593-606.
[12] Blei D M, Ng A Y, Jordan M I. Latent Dirichlet Allocation [J]. The Journal of Machine Learning Research, 2003, 3(3): 993-1022.
[13] Zaki M J. Scalable Algorithm for Association Mining [J]. IEEE Transactions on Knowledge and Data Engineering, 2000, 12(3): 372-390.
[14] Wu Yongliang, Chen Lian. Valid Association Rules Based on Lift-calculation [J]. Computer Engineering, 2003, 29(8): 98-100.

---

**Author Contributions:**
Guangce Ruan: Conceptualization, methodology, original draft preparation.
Lei Xia: Data collection, experimental analysis, revision of final manuscript.

**Conflict of Interest Statement:**
All authors declare no conflict of interest.

**Supporting Data:**
The supporting data is self-archived by the authors, E-mail: cgruan@infor.ecnu.edu.cn.
[1] Guangce Ruan, Lei Xia. userdict.txt. User-defined dictionary.
[2] Guangce Ruan, Lei Xia. model-final.twords. Topic results.

**Received:** 2016-09-07
**Revised:** 2016-10-18

*Note: Figure translations are in progress. See original paper for figures.*