Online-Journals.org

iJET International Journal of Emerging Technologies in Learning

iJET | eISSN: 1863-0383 | Vol. 18 No. 17 (2023) | OPEN ACCESS

https://doi.org/10.3991/ijet.v18i17.43489

PAPER

Multi-Level Text Clustering in Subject Knowledge Library and its Visualization

Ying Liu[1](✉), Yaofei Hao[2]

[1]Department of Tourism & Commerce, Handan Polytechnic College, Handan, China

[2]Department of Management, Handan Polytechnic College, Handan, China

liuying@hdvtc.edu.cn

ABSTRACT

The large-scale and complex data generated in the teaching field of business administration poses challenges for decision-makers and managers of companies, and how to effectively extract and manage the useful information contained in these data has become a problem to be solved. Currently available methods of subject knowledge library clustering and visualization struggle to handle the complexity and multi-hierarchies of such subject data effectively or meet users' requirements for advanced semantic understanding and retrieval. In view of these matters, this study aims to probe deeper into the problem of multi-level text clustering in the subject knowledge library and its visualization. Firstly, an innovative strategy-based subject semantic representation method for knowledge libraries was proposed to better interpret and represent the semantic information of subject data. Secondly, a subject clustering model of the knowledge library was constructed based on an improved hierarchical Dirichlet polynomial distribution, enabling efficient and accurate clustering of subject data. Lastly, visualization technology was employed to display the cluster results, allowing users to gain a clear understanding of the internal relationships and structure of the subject data. The research findings of this study could provide valuable new tools and methods for solving the problem of subject knowledge library management and utilization, analyzing the subject data, and supporting decision-making. As a result, they hold both theoretical and practical significance.

KEYWORDS

subject knowledge library, multi-level text clustering, semantic representation, hierarchical Dirichlet polynomial distribution, visualization

1 INTRODUCTION

A huge amount of data has been accumulated in the teaching field of business administration major with the progress of informationization and networking, and it covers many subjects that have complex relationships and hierarchical structures [1–3]. How to effectively manage, structure, and use these subject data so as to make efficient and accurate decisions has become an important challenge for modern

Liu, Y., Hao, Y. (2023). Multi-Level Text Clustering in Subject Knowledge Library and its Visualization. *International Journal of Emerging Technologies in Learning (iJET)*, 18(17), pp. 251–265. https://doi.org/10.3991/ijet.v18i17.43489

Article submitted 2023-06-16. Revision uploaded 2023-07-23. Final acceptance 2023-07-23.

society [4–7]. For instance, companies need to quickly and accurately acquire, analyze, and process information on various business subjects to improve the accuracy and efficiency of their decisions, and in the meantime, as the size of knowledge libraries gets bigger, how to enable the users to quickly locate and acquire the knowledge they need is also an urgent issue to be solved [8].

For this reason, the creation and management of subject knowledge libraries have become particularly important [9–14]. The main task of a subject knowledge library is to collect large amounts of disorderly scattered subject information so that users can search for and utilize this subject information easily through reasonable organization and classification. In the meantime, to meet users' personalized needs, the subject knowledge library needs to provide subject representation and analysis from multiple levels and angles to support users' diverse decisions, and in this process, text clustering and visualization technologies are playing a key role [15–17].

Nevertheless, the current available methods of subject knowledge library clustering and visualization exhibit evident deficiencies. Firstly, many existing methods struggle to effectively handle the complexity and multi-hierarchies present in subject data, resulting in unsatisfactory accuracy and stability of clustering results [18] [19]. Secondly, the existing visualization technology fail to clearly display the intrinsic relationship and structure of subject data, limiting users from gaining a comprehensive or thorough understanding of subject data [20–22]. Lastly, existing methods have limited capabilities in interpreting and representing the semantics of subjects, thus falling short of meeting users' demands for advanced semantic understanding and retrieval.

In view of these matters, this study probed deep into the problem of multi-level text clustering of the subject knowledge library and its visualization. First, a strategy-based subject semantic representation method for knowledge libraries was innovatively proposed, aiming at better interpreting and representing the semantic information of subject data. Then, a subject clustering model of the knowledge library was constructed based on an improved hierarchical Dirichlet polynomial distribution to realize efficient and accurate clustering of subject data. Third, with the help of visualization technology, the clustering results were visualized so that users could clearly understand the internal relationships and structure of the subject data. The research findings of this study could provide new tools and methods for solving the problem of subject knowledge library management and utilization, analyzing subject data, and supporting decision-making thus, they have valuable theoretical and practical significance.

2 STRATEGY-BASED SUBJECT SEMANTIC REPRESENTATION OF KNOWLEDGE LIBRARY

The creation and management of a subject knowledge library is especially important in this era of big data since the data in the library are mostly texts, and these textual data usually contain rich semantic information. How to extract and interpret this semantic information has a large impact on the application value, management, and retrieval efficiency of the knowledge library, but due to the diversity and ambiguity of language, direct processing and analysis of text data can hardly get accurate, comprehensive semantic information, so it's of both theoretical and practical significance to research an efficient text semantic representation method to better understand and represent the semantic information of subject data.

The strategy-based subject semantic representation method of the knowledge library is an important topic in this paper; the method takes strategy as an intermediary and gives efficient and accurate semantic representation of subject text data through the learning and application of the strategy. Compared with conventional semantic representation methods, strategy-based methods can better capture the deeper-level semantic information of text data and improve the accuracy and completeness of semantic representation. Besides, the strategy-based subject semantic representation method of the knowledge library can effectively solve the complexity and multi-hierarchies of subject data. Through the adjustment and optimization of the strategy, it can provide flexible presentation and processing of subject data at different levels and from different angles, thereby meeting users' personalized needs. Figure 1 gives a diagram showing the flow of the subject semantic representation model of the knowledge library.

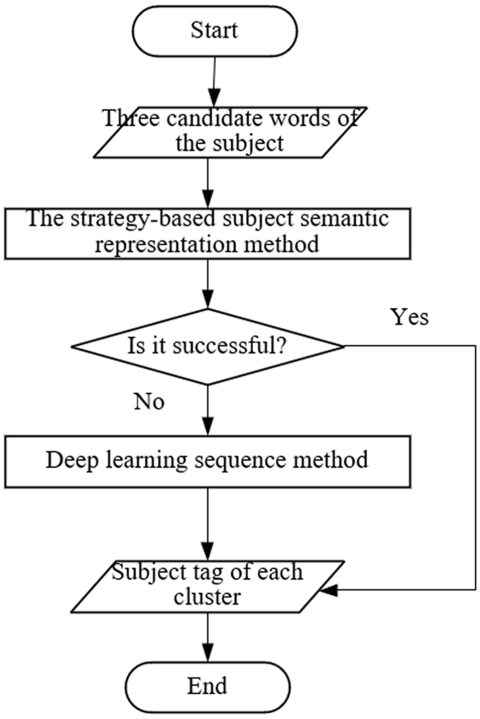

Fig. 1. Flow of the subject semantic representation model of knowledge library

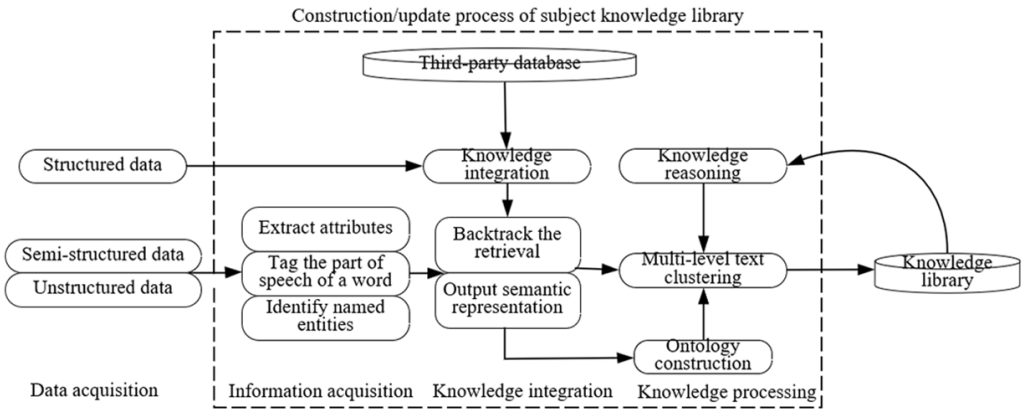

Fig. 2. Technical architecture of subject knowledge library

In this study, the data sources of the subject knowledge library were divided into three types: structured data, semi-structured data, and unstructured data. Regardless of the type, they all reflect the content and features of the subject to a certain extent, so a subject tag that can represent the subject meaning of a knowledge library must be contained in the text. In other words, subject tags are an important representation form of subject data, and they describe the core content and main features of subject data. Through the analysis and processing of subject tags, we can better understand and represent the subject data so as to effectively improve the management and utilization efficiency of the subject knowledge library. Figure 2 shows the technical architecture of the subject knowledge library.

Based on the above assumptions, this study proposed a strategy-based subject semantic representation method for knowledge libraries, and here is a detailed explanation of the specific flow of the method.

(1) Attain the model input. In this study, the model input was three candidate words for the subject in the knowledge library. The candidate words can be extracted based on the analysis of subject texts; for instance, natural language processing technologies such as the tagging of parts of speech of words and the identification of named entities could be adopted to analyze the subject texts and extract the potential candidate words. The input of the subject semantic representation model of the knowledge library was three candidate words, q_1, q_2, and q_3, which represent the subject of a certain knowledge library subject cluster, x_u. Assuming: F_u represents the text set consisting of all texts in x_u, and F_u can be split into B_u text sub-sets, $f_1^u, f_2^u, ..., f_{Bu}^u$ then it can be written as:

$$F_u = f_1^u, f_2^u, ..., f_{Bu}^u \tag{1}$$

(2) Backtrack the retrieval. This step is a core in the proposed method, aiming at finding out the most representative subject semantic words through the searching strategy. The $f_1^u, f_2^u, ..., f_{Bu}^u$ were divided into sentences according to punctuation marks, that is, the u-th text f_k^u in the text set corresponding to x_u can be represented by L_{uk} sentences, that is:

$$f_k^u = A_1^{uk}, A_2^{uk}, ..., A_{L_{uk}}^{uk} \tag{2}$$

Each sentence in A_j^{uk} can be represented by BL_{ukj} words:

$$A_j^{uk} = q_1^{ukj}, q_2^{ukj}, ..., q_{BQ_{ukj}}^{ukj} \tag{3}$$

Next, walk through each A_j^{uk} of each f_k^u in knowledge library subject cluster x_u, and judge whether input q_1, q_2, q_3 appear simultaneously in the BL_{ukj} word sets characterizing A_j^{uk}.

If q_1, q_2, q_3 appear simultaneously in the BL_{ukj} word sets characterizing A_j^{uk}, then the result state *R-S* is set as successful, namely *SU*, and the corresponding sentence A_{RE} is recorded and saved; otherwise, *R-S* is set as failed, namely *FA*.

(3) Output results. After attaining the final subject words, they were taken as the semantic representations of the subject. These subject words can effectively express the main content and features of the subject and provide users with an intuitive understanding of the subject. In addition, these subject words can also

be used as indexes in the knowledge library to improve the retrieval efficiency and accuracy of subject data. Also, these subject words can be taken as the basis for subsequent analysis and application, such as clustering, classification, and recommendation of subject data. If *R-S* is successful, then the corresponding sentences will be taken as the output. If *R-S* fails, then q_1, q_2, and q_3 will be taken as the input of the deep learning sequence model, and the output result of the model will be taken as the output of the final result of the subject semantic representation of the knowledge library.

3 MULTI-LEVEL TEXT CLUSTERING IN THE SUBJECT KNOWLEDGE LIBRARY

Currently, available text clustering algorithms are mostly designed based on the distribution characteristics of long texts. These algorithms generally assume that each document is a polynomial distribution of the knowledge library subject, and each knowledge library subject can be represented by a polynomial distribution of words. These algorithms make use of the consensus phenomenon of words in documents and adopt mathematical statistics to infer the distribution of documents to knowledge library subjects or the distribution of knowledge library subjects to words. However, when dealing with short texts, due to their characteristics of being short in length and sparse in language, the consensus phenomenon of words gets fewer, so the distribution counted based on it becomes even sparser. Therefore, the distribution attained is based on less evidence, which can result in poor performance of the algorithm in clustering short texts, and this is why the currently available algorithms are not suitable for the scenario studied in this paper.

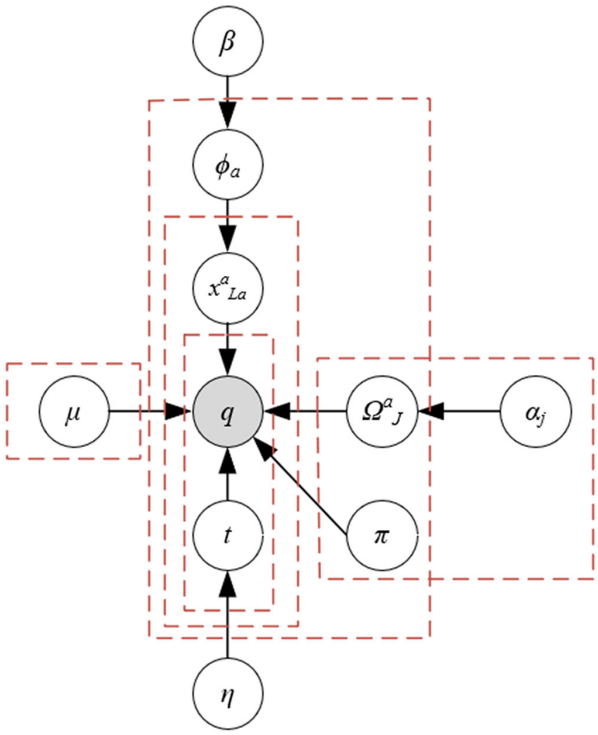

Fig. 3. A schematic diagram of the clustering model

In this study, a subject clustering model of the knowledge library was constructed based on improved hierarchical Dirichlet polynomial distribution and was taken as the multi-level text clustering method of the subject knowledge library. The proposed method was designed and improved specifically for the characteristics of short texts. By introducing the hierarchical Dirichlet distribution, it can solve the sparse feature problem of short texts that cannot be solved by conventional models and effectively improve the clustering effect of short texts. Besides, this method also considered the multi-hierarchies of data in the subject knowledge library, and it is able to find and describe the multi-layer structure and relationship of subject data. In this way, the semantic information of the subject knowledge library has been further enriched.

Based on the constructed model mentioned above, this study designed a two-layer generation structure and added two potential feature weight matrices. This structure can better capture and describe the hierarchy and heterogeneity of subject data. The model can find and represent not only the common laws of subject data but also their special laws, thereby giving a deeper understanding and more accurate representation of subject data. Figure 3 gives a diagram showing the clustering model.

Assuming: π and μ represent the two potential feature weight matrices added to the model, π is related to knowledge library subject j and μ is related to word q; $CATR(\ )$ represents a categorical distribution with a logarithmic spatial parameter; π_j and μ_q represent vectors of potential feature weights, then the formula below calculates the probability of generating word q under the condition of a given knowledge library subject:

$$CATR\left(q \mid \pi_j \mu^Y\right) = \frac{\exp\left(\pi_j \cdot \mu_q\right)}{\sum_{q' \in C} \exp\left(\pi_j \cdot \mu_{q'}\right)} \tag{4}$$

The proposed model defaults to the assumption that words in a document in the knowledge library are jointly determined by word distribution and the potential feature matrix, so when constructing the model, a Bernoulli distribution parameter η was set in the model to judge the main source of the word.

Detailed steps for the proposed model to generate documents are introduced below:

(1) Generate $\phi^a \sim DIR(\beta/J, \beta/J, ..., \beta/J)$ under each data source a;
(2) Generate $\alpha_{j,q} \sim B(\omega, \delta^2 U)$, $q = 1,2, ..., C$ under each knowledge library subject j;
(3) Generate $\Omega_j^a \sim DIR(\alpha_{j,1}, \alpha_{j,2}, ..., \alpha_{j,c})$ under each knowledge library subject j in a;
(4) For each document f in a, generate $x_l^a \sim DIS\left(\phi_1^s, \phi_2^s, ..., \phi_k^s\right)$;
(5) For each word u in each document f, generate $t_{fu} \sim$ Bernoulli(η), and $u \sim (1-t_{fu}) CATR\left(\Omega_{xcl}^a\right) + t_{fu} CATR\left(\pi_{xal}^a \mu^Y\right)$.

Wherein, t_{fu} represents the indicator of the word u generated by the Bernoulli distribution parameter η; when t_{fu} is equal to 0, word u is generated by the knowledge library subject-word distribution; when t_{fu} equals 1, word u is generated by the potential feature matrix.

Since the constructed model defaults that words in the text can be either generated by the polynomial distribution or the potential feature matrix, the conditional

probability density function of document f_l^a can be calculated based on the following formula:

$$d\left(f_l^a \mid \theta_j^a, t\right) \approx \frac{\left(\sum_{q=1}^{C}(1-t_{fq})z_{f,q}^a\right)!}{\prod_{q=1}^{C}\left(1-t_{fq}\right)z_{f,q}^a!}\prod_{q=1}^{C}\theta_{j,q}^{a\ z_{f,q}^a\left(1-t_{fq}\right)}\frac{\exp\left(\pi_j^a\cdot\mu_q\right)}{\sum_{q'\in C}\exp\left(\pi_j^a\cdot\mu_{q'}\right)}^{x_{d,w}^s y_{dw}} \tag{5}$$

Further, through the integral operation on $\Omega_1^a, \Omega_2^a, ..., \Omega_j^a$ shown by the following formula, the value of the conditional probability density function of data set F^a under the given condition $\left\{x_1^a, x_2^a, ..., x_{La}^a\right\}$ can be estimated:

$$d\left(F^s \mid X, t\right) \approx \prod_{f=1}^{L^a}\int d\left(f_l^a \mid \theta^a, t\right)d\left(\theta^a \mid \vec{\alpha}\right)f\theta^a \approx \prod_{f=1}^{L^a}\frac{\left(\sum_{q=1}^{C}\left(1-t_{fq}\right)z_{f,q}^a\right)!}{\prod_{q=1}^{C}\left(1-t_{fq}\right)z_{f,q}^a!}$$
$$\prod_{j=1}^{J}\frac{\Pi\left(\sum_{w=1}^{C}\alpha_j^q\right)}{\Pi\left(\sum_{q=1}^{C}\left(B_j^q+\alpha_j^q\right)\right)}\prod_{q=1}^{C}\frac{\Pi\left(B_j^q+\alpha_j^q\right)}{\Pi\left(\alpha_j^q\right)}\frac{\exp\left(\pi_j^a\cdot\mu_q\right)}{\sum_{q'\in C}\exp\left(\pi_j^a\cdot\mu_{q'}\right)}^{z_{f,q}^a t_{fq}} \tag{6}$$

Assuming: B_j^q represents the number of times a word q is generated by distributed control under knowledge library subject j, that is, $B_j^q = \sum_{l:x\ ^a{}_{l=j}}\left(1-t_{l,q}\right)z_{l,q}^a$ and $B_j = \sum_{q=1}^{C} B_j^q$.

A subject knowledge library usually contains large amounts of text data coming from many different data sources. To deal with such high-dimensional complex problems, an effective sampling method is needed to ensure that the data sets of all sources can be regarded as a whole during the clustering process and that cluster tag numbers in each data set can be matched one to one. Gibbs sampling can meet such requirements while avoiding consuming a lot of computation time. When processing large-scale datasets, direct computation can take a significant amount of time, while by carrying out Gibbs sampling after multiple samplings, the computation time can be greatly reduced without sacrificing clustering quality.

The sampling and reasoning process is detailed as follows:

(1) By changing the value of t_{fq}, generate a new candidate value t_q^{NE}; let $d(t_{fq}|f,x) \propto d(f|t_{fq}, x)o(t_{fq})$, the formula below calculates the probability for a new candidate value being accepted:

$$MIN\left\{1, \frac{d\left(t_{fq}^{NE} \mid f, x\right)}{d\left(t_{fq}^{OL} \mid f, x\right)}\right\} \tag{7}$$

(2) When the number of iterations meets a certain condition, update parameter α_j;

(3) In data source a, for each knowledge library subject j, if j is not in X^a, then Ω_j^a is attained through Dirichlet distribution with α_j as the parameter; if j is in X^a, then

Ω_j^a is attained via the sampling through the Dirichlet distribution based on the following formula:

$$\left\{\alpha_j^1+\sum_{l:x_l^a=j}\left(1-t_{l,1}\right)z_{l,1}^a,\alpha_j^2+\sum_{l:x_l^a=j}\left(1-t_{l,2}\right)z_{l,2,\ldots,}^a\alpha_j^C+\sum_{l:x_l^a=j}\left(1-t_{l,C}\right)z_{l,C}^a\right\} \tag{8}$$

(4) Find the optimal knowledge library subject vector π_j in data source a;

(5) Update ϕ^a through the Dirichlet distribution in data source a, when X_l^a is equal to j, $U\left(X_l^a=j\right)$ is equal to 1; when X_l^a is not equal to j, $U\left(X_l^a=j\right)$ is equal to 0;

$$\left\{\beta\big/_J+\sum_{l=1}^{L_a}U\left(x_l^a=1\right),\beta\big/_J+\sum_{l=1}^{L_a}U\left(x_l^a=2\right),\ldots,\beta\big/_J+\sum_{l=1}^{L_a}U\left(x_l^a=J\right)\right\} \tag{9}$$

(6) For document f, by carrying out sampling on the discrete distribution of parameter $\{O_{f,1}, O_{f,2}, \ldots, O_{fJ}\}$, the update of X_l^a can be realized, wherein $O_{f,j}\propto\phi_j^a d\left(f_l^a\,|\,\Omega_j^a,t\right)$ and $\Sigma_{j=1}^J O_{f,j}=1$.

Radial visualization (*Radviz*) is an efficient data visualization method suitable for the visualization processing of high-dimensional data. It maps each data point to a point in a two-dimensional space to achieve an intuitive representation of high-dimensional data. Based on *Radviz*, this study performed visualization analysis on the multi-level text clustering data of the subject knowledge library, and the specific flow is explained below:

(1) Data preprocessing: Pre-process the multi-level text data of the subject knowledge library, including text cleaning, word segmentation, removal of stop words, and other steps. These steps help extract effective information about the data and prepare it for subsequent data analysis and visualization.

(2) Feature selection: since the multi-level text data of the subject knowledge library contains lots of features, direct visualization may give too complicated and difficult results, so feature selection is needed to pick the most representative features for display, and one common method is to use feature selection algorithms, such as information gain or statistics-based approaches, to choose the most representative data features.

(3) Data mapping: after selecting representative features, use the *Radviz* method to map each data point to a point in a 2D space. Specifically, the *Radviz* method will select a circle in the 2D space and correspond each feature to a point on the circle, namely the anchor point. The position of each data point is determined by the vector superposition of its values on each feature.

(4) Data visualization: visualize the mapped data points in 2D space and use visual features such as color, shape, and size of the points to describe other attributes of the data; in the meantime, interactive methods (such as scaling and dragging) could be used to explore the data and discover their hidden structure and laws.

(5) Result analysis: a deeper understanding of the data could be attained through the observation and analysis of visualization results; for instance, points that are close to each other may have similar attributes, while points far from each other may be quite different. This can help us understand the relationship between subjects in the subject knowledge library and the features of the multi-level text data.

4 EXPERIMENTAL RESULTS AND ANALYSIS

Table 1 shows the number of subject clusters estimated by different algorithms based on different-types of data sources. The estimation results of *K*-means clustering, Density-Based Spatial Clustering of Applications with Noise (*DBSCAN*) clustering, spectral clustering, and the proposed method were close, wherein the number estimated by *DBSCAN* clustering was obviously higher than that of other algorithms. In terms of semi-structured and unstructured data sources, the number estimated by the subject model was higher, while in terms of structured and integrated data sources, the subject model didn't give a result. In terms of integrated data sources, all algorithms (except for the subject model) gave the same estimated cluster number, and this result indicates that different algorithms performed differently based on different types of data sources, which has emphasized that choosing an appropriate clustering algorithm is an important thing, and this is determined by the specific application scenario and data type. The proposed method performed well in terms of all data source types, indicating that it has good adaptability and stability in dealing with different-type text data of the subject knowledge library and is expected to achieve more accurate and stable text clustering in practical applications.

Table 1. Number of subject clusters estimated by different algorithms based on a real text dataset of subject knowledge library

	K-Means Clustering	*DBSCAN* Clustering	Subject Model	Spectral Clustering	The Proposed Method
Structured data source	6	31	\	4	3
Semi-structured data source	6	31	32	2	2
Unstructured data source	5	31	32	4	3
Integrated data source	28	31	\	21	21

Figure 4 illustrates a comparison of the clustering performance among various clustering algorithms, namely *K*-means clustering, *DBSCAN* clustering, the proposed method, subject model, and spectral clustering. The evaluation is conducted under different subject numbers. Here, the metric of performance is accuracy; the closer its value to 1, the better the clustering effect. From the figure, it can be seen that the performance of the proposed method and the subject model was better in most cases, and the accuracy was always above 0.8. Especially in the case of higher subject numbers, the two methods performed better than other methods, showing great advantages in dealing with large-scale subject data. The performance of *DBSCAN* clustering was relatively stable, with an accuracy of around 0.85, indicating that it is robust to changes in subject numbers. The performance of *K-means* clustering and spectral clustering fluctuated greatly when dealing with different subject numbers. The accuracy of *K*-means clustering was low when dealing with lower subject numbers, but as the subject numbers grew, its accuracy increased, suggesting that it's better at dealing with large-scale data. The accuracy of spectral clustering fluctuated during the entire process as well, and its accuracy increased significantly when dealing with higher subject numbers.

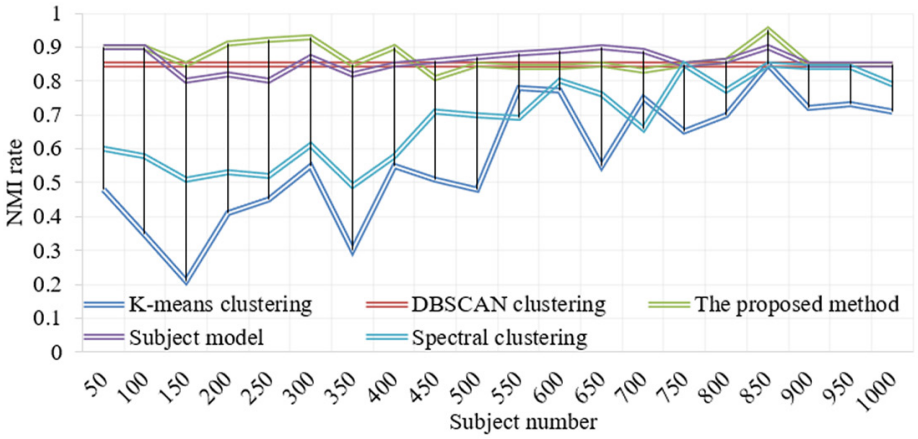

Fig. 4. Clustering results of various algorithms in case of different subject numbers

Table 2. Clustering accuracy of different clustering algorithms

	P Value	*R* Value	*F* Value	*NMI* Value
K-means clustering	0.8468	0.8979	0.8453	0.8768
DBSCAN clustering	0.6450	0.8672	0.7650	0.8972
Subject model	0.8560	0.8772	0.8972	0.8543
Spectral clustering	0.9673	0.9562	0.9134	0.8744
The proposed method	0.9888	0.9864	0.9671	0.9858

Table 2 gives the performance of the above-mentioned five clustering algorithms in terms of clustering accuracy. The metrics used here include the *P* value (precision rate), *R* value (recall rate), *F* value (comprehensive evaluation value), and Normalized Mutual Information (*NMI*) value, which measures the consistency of the clustering result with the real situation. According to the data given in the table, the proposed method outperformed other algorithms in terms of all metrics; its *P*, *R*, *F*, and *NMI* values all exceeded 0.98, indicating its superiority in dealing with text clustering tasks in subject knowledge libraries. The performance of spectral clustering was not bad, especially in terms of *P* value (0.9673) and *R* value (0.9562), and both were high, indicating its advantages in clustering precision and recall. The performance of the subject model and *K-means* clustering was close; their *P*, *R*, *F*, and *NMI* values were all above 0.84 but slightly lower than those of spectral clustering and the proposed method. This is because the performance of the subject model and *K-means* clustering declined a bit when dealing with large-scale and high-dimensional data. The performance of *DBSCAN* clustering was the worst in terms of all four metrics, especially its *P* value of only 0.6450, indicating low precision, and this is because *DBSCAN* clustering relies on density connectivity and has difficulty in processing complex and high-dimensional data. To sum up, the proposed method performed the best in terms of clustering accuracy; spectral clustering, subject model clustering, and *K*-means clustering performed stably, while *DBSCAN* clustering needs to be selected carefully in specific application scenarios. These results provide important references for the selection of text clustering algorithms.

Table 3. Clustering accuracy of different clustering algorithms in case of multiple data sources

Data Source	Subject Model		Spectral Clustering		The Proposed Method	
	F Value	*NMI* Value	*F* Value	*NMI* Value	*F* Value	*NMI* Value
Structured data source	0.876	0.857	0.877	0.836	0.965	0.874
Semi-structured data source	0.845	0.897	0.834	0.847	0.862	0.898
Unstructured data source	0.854	0.834	0.897	0.867	0.862	0.872
Integrated data source	0.851	0.862	0.812	0.907	0.862	0.912

Table 3 shows the clustering accuracy of the above-mentioned five clustering algorithms in the case of different-type data sources. Metrics used here included *P* value, *R* value, *F* value, and *NMI* value. According to table data, the proposed method also outperformed other algorithms in terms of all metrics; its *P*, *R*, *F*, and *NMI* values all exceeded 0.98, indicating its advantages in dealing with text clustering tasks in the subject knowledge library. The performance of spectral clustering was good, especially in terms of *P* value (0.9673) and *R* value (0.9562), and both were high, indicating its advantages in clustering precision and recall. The performance of the subject model and *K-means* clustering was similar; their *P*, *R*, *F*, and *NMI* values were all above 0.84 but slightly lower than those of spectral clustering and the proposed method. This is because the performance of the subject model and *K-means* clustering declined a bit when dealing with large-scale and high-dimensional data. The performance of *DBSCAN* clustering was the worst in terms of all four metrics, especially its *P* value of only 0.6450, indicating low precision. This is because *DBSCAN* clustering relies on density connectivity and has difficulty in processing complex, high-dimensional data. Comprehensively speaking, the proposed method was the best in terms of clustering accuracy; the performance of spectral model clustering, subject model, and *K-means* clustering was relatively stable; and the *DBSCAN* clustering needs to be selected carefully in specific application scenarios. These results provide important references for the selection of text clustering algorithms.

Next, this study adopted two evaluation indicators the Rand Index and Mutual Information (*MI*), to compare the clustering performance of different clustering algorithms in the case of different-type data sources. Figure 5 gives the Rand Index of the five algorithms in the case of different-type data sources (structured, semi-structured, unstructured, and integrated data sources). The Rand Index is an evaluation metric of the clustering effect; the closer its value is to 1, the better the clustering effect. As can be seen from the figure, the proposed method performed the best in terms of all types of data sources; its Rand Index had always exceeded 0.8, showing its superior clustering performance under various conditions of data sources, and this is thanks to its ability to effectively capture the heterogeneity and differences of multi-source texts in the subject knowledge library. Spectral clustering also performed well in the case of each data source type, its Rand Index was above 0.7, indicating that its clustering effect was stable and excellent, especially on unstructured and integrated data sources. Subject model and *K*-means clustering had close values of the Rand Index in the case of all types of data sources; the value was generally between 0.5 and 0.6, and the performance was average, indicating

that the two methods only had average adaptability to different-type data sources. The Rand Index of *DBSCAN* clustering was the lowest among all algorithms, especially on structured data sources; its value was only 0.4, indicating a poor clustering effect on different-type data sources. Overall speaking, the clustering effect of the proposed method was the best for most data source types; spectral clustering performed well, while *K*-means clustering, subject model clustering, and *DBSCAN* clustering should be carefully chosen according to the characteristics of specific data sources. These results also provide important references for the selection of text clustering algorithms.

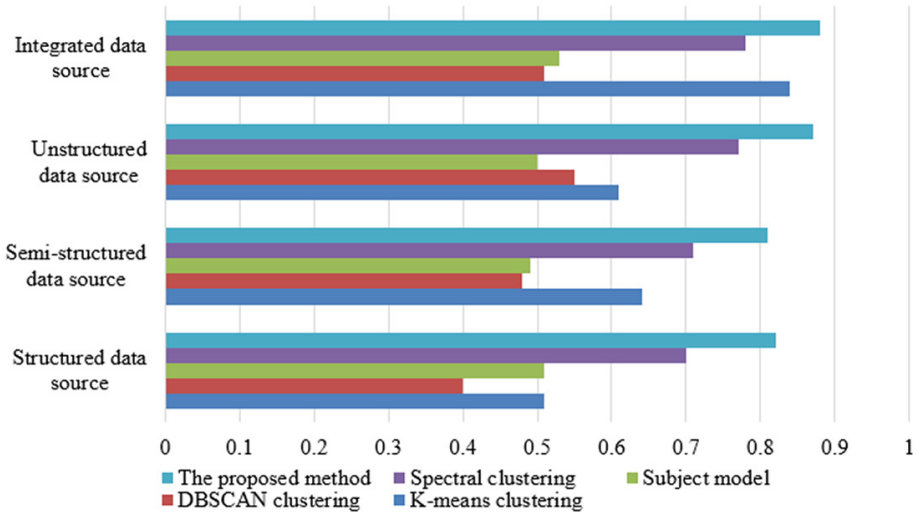

Fig. 5. Rand Index of different clustering algorithms in case of different-type data sources

Figure 6 shows the *MI* of the five algorithms in the case of different-type data sources (structured, semi-structured, unstructured, and integrated data sources). *MI* measures the correlation between the clustering result and the real category. The higher the *MI* value, the stronger the correlation between the two and the better the clustering effect. The proposed method got the highest *MI* value in terms of all data source types, exceeding 0.88, and even reached as high as 0.94 in the case of an integrated data source, indicating that the proposed method has obvious advantages in capturing the real category structure in data sources. The performance of spectral clustering was also outstanding; its *MI* value was 0.78 for all data source types, especially unstructured and integrated data sources, where its *MI* value was as high as 0.89, indicating good stability and accuracy in the clustering effect. The *MI* of *K-means* clustering and subject model was close in terms of all data source types; the value was generally between 0.5 and 0.7, indicating that these two methods have certain limitations in reflecting the real category structure of data sources. The *MI* of *DBSCAN* clustering was the lowest among all algorithms; its value was below 0.5 under various conditions, indicating a poor clustering effect in all cases. In summary, the clustering effect of the proposed method was good under conditions of all types of data sources, and it has very good practicality; the *MI* of *K-means* clustering, subject model clustering, and *DBSCAN* clustering was low under conditions of all types of data sources, and they need to be chosen carefully according to the features of specific data sources.

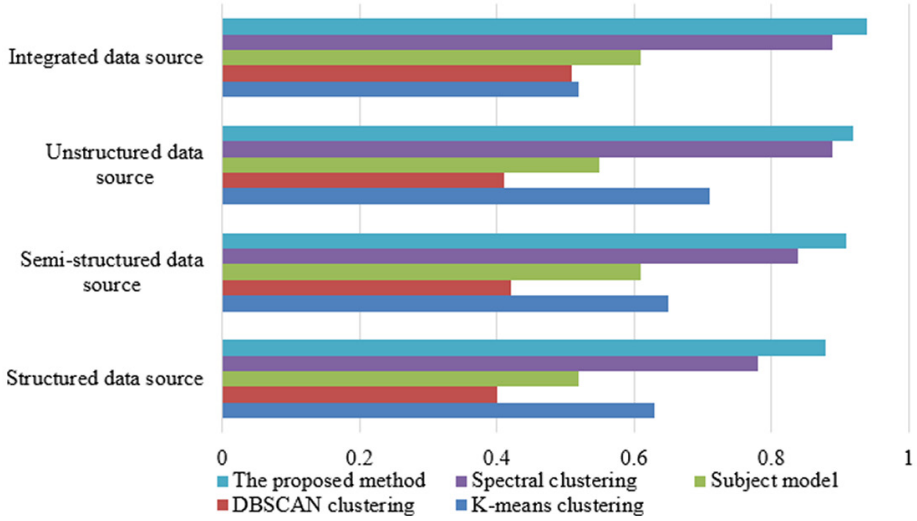

Fig. 6. Mutual information of different clustering algorithms in case of different-type data sources

5 CONCLUSION

In view of the difficulty of conventional clustering methods in dealing with the clustering analysis of short texts caused by the sparse feature problem, this study proposed a subject text clustering model for subject knowledge libraries based on the improved hierarchical Dirichlet polynomial distribution and used it to solve the said difficulty effectively. The proposed model adopted a two-layer generation structure to capture the heterogeneity and differences of texts and ensured the cluster tag numbers could be matched one to one in each data source with the help of the Gibbs sampling inference algorithm.

Experimental results demonstrated that the proposed method outperformed all other reference algorithms in terms of subject cluster number estimation, clustering results, clustering accuracy, and Rand Index and MI in the case of multi-type data sources. Regardless of subject numbers or data source types, the proposed method exhibited excellent stability and performance. Moreover, the *Radviz* visualization technology was adopted to visualize the multi-level text clustering data of the subject knowledge libraries, and the results further proved the advantages of the proposed method in processing multi-source short text data clustering.

Through the comparative analysis of experimental results, it's known that choosing a suitable clustering algorithm is very important for a text clustering task in a subject knowledge library. Although conventional methods such as *K*-means clustering, *DBSCAN* clustering, subject model clustering, and spectral clustering have all achieved certain effects, they have their respective limitations, while the proposed method exhibited strong advantages in short text processing and clustering effects.

In conclusion, this study innovatively proposed a new method and verified its effect experimentally, offering a new and effective means for the clustering analysis of multi-source short text datasets in subject knowledge libraries. The research findings of this study provide useful evidence for subsequent research in related fields.

6 REFERENCES

[1] Y. Liang and B. Lou, "Design of an information management system for the case library of business management courses," *International Journal of Emerging Technologies in Learning (iJET)*, vol. 16, no. 7, pp. 203–217, 2021. https://doi.org/10.3991/ijet.v16i07.22119

[2] F. Su, "Application research of college English teaching resources based on knowledge base management platform," *Journal of Physics: Conference Series*, vol. 1345, no. 5, p. 052072, 2019. https://doi.org/10.1088/1742-6596/1345/5/052072

[3] P.D. Larasati and A. Irawan, "Design of knowledge-base teaching activities for earlier diagnosis of health problem," In *7th International Conference on Cyber and IT Service Management (CITSM)*, Jakarta, Indonesia, pp. 1–4, 2019. https://doi.org/10.1109/CITSM47753.2019.8965377

[4] Y. Xi and L. Zhu, "Construction of college students' management informatization ecosystem based on data analysis technology," *International Journal of Emerging Technologies in Learning (iJET)*, vol. 18, no. 10, pp. 166–183, 2019. https://doi.org/10.3991/ijet.v18i10.40239

[5] L. Deng, "A study on distance personalized English teaching based on deep directed graph knowledge tracking model," *International Journal of Advanced Computer Science and Applications*, vol. 14, no. 2, pp. 761–770, 2023. https://doi.org/10.14569/IJACSA.2023.0140288

[6] M.A. Gómez, R.F. Herrera, E. Atencio, and F.C. Munoz-La Rivera, "Key management skills for integral civil engineering education," *International Journal of Engineering Pedagogy (iJEP)*, vol. 11, no. 1, pp. 64–77, 2021. https://doi.org/10.3991/ijep.v11i1.15259

[7] Y. Chen, X. Liu, X. Gao, J. Zhang, B. Liu, and H. Yang, "Personalized teaching knowledge recommendation system based on artificial intelligence algorithm," In *International Conference on Big Data Analytics for Cyber-Physical System in Smart City*, Bangkok, Thailand, pp. 185–192, 2022. https://doi.org/10.1007/978-981-99-1157-8_23

[8] H. Shi, X. Meng, J. Du, and L. Wang, "Design and realization of experiential teaching based on knowledge feature transformation of course teaching," *International Journal of Emerging Technologies in Learning (iJET)*, vol. 17, no. 7, pp. 226–239, 2022. https://doi.org/10.3991/ijet.v17i07.30403

[9] P. Nuankaew, P. Nasa-Ngium, K. Phanniphong, O. Chaopanich, S. Bussaman, W.S. Nuankaew, "Learning management impacted with COVID-19 at higher education in Thailand: Learning strategies for lifelong learning," *International Journal of Engineering Pedagogy (iJEP)*, vol. 11, no. 4, pp. 58–80, 2021. https://doi.org/10.3991/ijep.v11i4.20337

[10] T. Yuensuk, P. Limpinan, W. Nuankaew, and P. Nuankaew, "Information systems for cultural tourism management using text analytics and data mining techniques," *International Journal of Interactive Mobile Technologies (iJIM)*, vol. 16, no. 9, pp. 146–163, 2022. https://doi.org/10.3991/ijim.v16i09.30439

[11] Q. Xiao, "Resource classification and knowledge aggregation of library and information based on data mining," *Ingénierie des Systèmes d'Information*, vol. 25, no. 5, pp. 645–653, 2020. https://doi.org/10.18280/isi.250512

[12] I. Noorhasyimah, A.R. Azlan, M.A.R. Aina, M.J. Siti, L.Z. Nur, and A.R. Hanisah, "A qualitative study of data management development for hybrid organisation application," *International Journal of Interactive Mobile Technologies (iJIM)*, vol. 16, no. 13, pp. 184–191, 2022. https://doi.org/10.3991/ijim.v16i13.30617

[13] I. Wen and C.W. Liang, "Integration of safety knowledge into three-dimensional model design and construction plan from the perspective of project executors in petrochemical industry," *International Journal of Safety and Security Engineering*, vol. 11, no. 2, pp. 185–192, 2021. https://doi.org/10.18280/ijsse.110207

[14] G. Romagnoli, M. Galli, D. Mezzogori, and D. Reverberi, "A cooperative and competitive serious game for operations and supply chain management: Didactical concept and final evaluation," *International Journal of Online and Biomedical Engineering (iJOE)*, vol. 18, no. 15, pp. 17–30, 2022. https://doi.org/10.3991/ijoe.v18i15.35089

[15] H. Yang and H.M. Alabool, "Application of exercise-correlated knowledge proficiency tracing model based on multiple learning objectives in English teaching curriculum resources," *International Journal of Emerging Technologies in Learning (iJET)*, vol. 17, no. 24, pp. 157–173, 2022. https://doi.org/10.3991/ijet.v17i24.35417

[16] C. Wang and S. Xu, "Construction of the evaluation index system of physical education teaching in colleges and universities based on scientific knowledge graph," *Mobile Information Systems*, vol. 2022, no. 4225081, pp. 1–11, 2022. https://doi.org/10.1155/2022/4225081

[17] A. Petcharit, P. Sornsaruht, and P. Pimdee, "An analysis of total quality management (TQM) within the Thai auto parts sector," *International Journal of Online and Biomedical Engineering (iJOE)*, vol. 16, no. 2, pp. 131–145, 2020. https://doi.org/10.3991/ijoe.v16i02.11917

[18] K. Srinivasa and P.S. Thilagam, "Clustering and bootstrapping based framework for news knowledge base completion," *Computing & Informatics*, vol. 40, no. 2, pp. 318–340, 2021. https://doi.org/10.31577/cai_2021_2_318

[19] F. Song and E. de la Clergerie, "Clustering-based automatic construction of legal entity knowledge base from contracts," In *IEEE International Conference on Big Data (Big Data)*, Atlanta, GA, USA, pp. 2149–2152, 2020. https://doi.org/10.1109/BigData50022.2020.9378166

[20] S. Yogev, G. Shani, and N. Tractinsky, "HiveRel: Hexagons visualization for relationship-based knowledge acquisition," *CCF Transactions on Pervasive Computing and Interaction*, vol. 4, no. 4, pp. 408–436, 2022. https://doi.org/10.1007/s42486-022-00097-3

[21] H. Sun, "Interactive knowledge visualization based on IoT and augmented reality," *Journal of Sensors*, vol. 2022, no. 7921550, pp. 1–8, 2022. https://doi.org/10.1155/2022/7921550

[22] W. Lu, W. He, and Y. Zhang, "Visualization analysis of additive manufacturing in medical rehabilitation field based on knowledge graph," *Materialstoday: Proceedings*, vol. 70, pp. 66–71, 2022. https://doi.org/10.1016/j.matpr.2022.08.530

7 AUTHORS

Ying Liu, a graduate of Tiangong University, holds a Master's degree. She is currently employed at Handan Polytechnic College, where her primary research interests lie in the fields of commerce and management (E-mail: liuying@hdvtc.edu.cn; ORCHID: https://orcid.org/0009-0001-0638-0716).

Yaofei Hao, a graduate of Hebei University of Technology, holds a Master's degree. He is currently employed at Handan Polytechnic College, where his primary research interests lie in the fields of commerce and management (E-mail: haoyaofei@hdvtc.edu.cn; ORCHID: https://orcid.org/0009-0001-6973-979X).