Journal of Network Intelligence
Taiwan Ubiquitous Information
 ISSN 2414-8105
Volume **10**, Number **2**, May **2025**

Knowledge Graph-K-modes Clustering Algorithm for Archival Text Categorization

Xue-Jun Wu*

Archives
China Jiliang University, Hangzhou 310018, P. R. China
xuejw02@163.com

Huan-Bin Zhang

Faculty of Engineering and Information Technology
INTI International University, 71950 Jurai Newtown, Negeri Sembilan, Malaysia
gf5439@163.com

*Corresponding author: Xue-Jun Wu
Received 20240311, revised 20240704, accepted 20241126

Abstract. Automated classification of electronic archive text's can improve document management efficiency, support rapid information retrieval, and enhance data analysis capabilities, thus helping to achieve efficient use of information resources and knowledge management. However, medium-scale archival text classification needs to deal with complex natural languages as well as deal with noise, ambiguity and sparsity in unstructured text data. In order to solve the above problems, this work proposes a knowledge graph-K-modes clustering algorithm-CTransR-Full-K-Modes for archival text classification. Firstly, existing knowledge representation learning modes and dissimilarity measures in clustering algorithms are analysed. Next, we provide a representation learning model called CTransR. This model is based on hyperparameter tuning and relation matrix projection. Its major purpose is to address the issue of poor recognition capacity in the current TransR model while dealing with distinct entities that have great resemblance. Allocate items and relations into distinct vector spaces. The hyperparameter tuning principle and a priori probability based on the use of relations are employed to distinguish relations that encode similarity, thus facilitating the process of knowledge fusion in archival knowledge mapping. Finally, a full attribute-valued K-Modes algorithm-Full-K-Modes for archival knowledge graphs is proposed to achieve clustering processing of archival texts using preclustering-based multi-attribute-valued clustering centres and their corresponding dissimilarity measures to mitigate the effects of local optimal cases. The experimental results show that the proposed CTransR-Full-K-Modes algorithm can effectively improve the accuracy of the clustering results of ordered-type categorical data, and achieve the purpose of automatic categorisation mining for medium-sized digital archives.
Keywords: Archive classification; Ordered categorical data; Clustering algorithm; Knowledge graph; K-Modes; TransR

1. **Introduction.** The study of automatic classification of electronic archive text is of great significance and value because it can process and organise large amounts of electronic text information [1, 2], improve retrieval efficiency and accuracy, save human resources, and enhance information management and knowledge discovery. Through automatic classification, it can ensure sustainable management of archival information, improve the scientific and systematic nature of archival work, and provide convenient and accurate

information support for governmental decision-making, academic research, and business analyses [3, 4], as well as help to protect cultural heritage and social memory, with far-reaching impacts on the development of society.

Automated processing of large-scale electronic archive text data reduces the labour intensity of manual classification and sorting, improves work efficiency, and makes information retrieval and use more rapid [5, 6]. Computer algorithms can reduce human error and bias brought about by subjective judgement, and maintain the consistency and repeatability of classification results [7]. Through automatic classification, fragmented information can be organised into a structured knowledge system, which helps to transform data into valuable knowledge for further analysis and application. By automatically extracting themes and patterns in archival texts, potential knowledge points can be revealed, which is important for thematic research and academic exploration. Effective archival categorisation leads to a more rational allocation of storage and management resources and reduces the cost of information management [8]. The development of automated archival text classification tools and methods is a major contribution to the field of archival management in the information age. The research aim of this work is mainly to improve the classification of archival texts containing categorical data by combining the rich semantic relations of knowledge graphs and the ability of K-modes clustering algorithms to handle categorical variables. The Knowledge Graph is able to enhance the semantic understanding of the algorithm by mapping out the relationships and attribute links between archival texts, while the K-modes algorithm is suitable for dealing with these categorical features as it does not rely on the computation of means for numerical data. This combination of research is dedicated to achieving more accurate and efficient automatic classification of archival texts, especially when dealing with medium-sized archival data, which can significantly improve the quality of classification.

1.1. **Related work.** The current state of research on automatic text categorisation for electronic archives shows that it is an active and growing field. Current research focuses on the development of more accurate and efficient algorithms [9, 10] aimed at extracting and organising information from huge volumes of data. Using techniques such as machine learning, deep learning, and natural language processing, scholars are trying to overcome the challenges of linguistic diversity, semantic complexity, and data sparsity, and are realising multi-dimensional feature extraction and automatic annotation of text [11]. At the same time, many researches are also focusing on how to integrate domain knowledge and improve the adaptability and interpretability of algorithms to meet the growing demand in practical applications.

Vellino and Alberts [12] discusses the use of machine classification to assist in the appraisal of corporate e-mail archives, a task that has grown in difficulty due to increasing volumes of digital information. The study underscores the potential for automating the archival process to support records managers. However, this research may not fully address the complexity and variety of non-email archival texts, and the adoption of such machine classification systems may require significant technical infrastructure and expertise. Zhao [13] explores the use of text classification algorithms to manage and organize the increasing amount of electronic archives in academic institutions, highlighting computer technology's role in simplifying archival tasks. However, the study's algorithms may not be universal and might need adaptation for different types of archival content, and the research focus is somewhat narrow, being limited to university archives. Hadi et al. [14] presents enhancements to the typical Naive Bayes classification approach to improve

archival text classification. A new model is proposed and tested against traditional classification algorithms. However, the use of Naive Bayes may not be as effective with highly dimensional or unstructured data, which typifies many archival texts.

Franks [15] addresses the growing crisis in recordkeeping due to the surge in data volumes, proposing automatic text classification as a solution to managing the complexities of digital recordkeeping. While offering a solution to data volume challenges, the classification process may not fully take into account the historical or preservation contexts significant for cultural heritage archives. Brokensha et al. [16] serves as an investigation into the application of machine learning techniques for the classification of documents within a specific archive, showcasing the potential of text processing in heritage preservation. However, the focus on a specific archive might limit the generalizability of the findings, and machine learning techniques may require large, well-annotated corpora, which are not always available.

Mustafi et al. [17] proposed a new method for text clustering based on genetic algorithm with nearest neighbour heuristic. The method uses genetic algorithm to optimize the clustering results of text data and nearest neighbour heuristic for text classification. Experimental results show that the method achieves good performance in text clustering and classification tasks. However, the method may face efficiency problems when dealing with large-scale text data, and the performance of the genetic algorithm is affected by the selection and setting of parameters. Chang et al. [18] applied text mining, cluster analysis, and latent Dirichlet allocation techniques to classify the topics of environmental education journals in a classification study. The results of the study showed that the subject classification using the hierarchical K-means method was better than the LDA method. However, the method may face efficiency problems when dealing with large-scale text data and may have some limitations for domain-specific topic classification.

1.2. **Motivation and contribution.** Archival text classification deals with textual data rather than numerical data. k-medoids does not rely on the computation of the mean, so it is better suited to deal with the classification of data with non-numerical attributes [19, 20]. K-medoids algorithm has a higher computational cost as compared to k-means, as it requires distance calculation for all pairs of points [21]. Therefore, the computational effort is within acceptable limits for medium-sized datasets. Knowledge graphs can provide richer feature representations by mining entities and relationships between entities in text [22]. K-medoids algorithms perform clustering by calculating distances between texts, whereas knowledge graphs can provide richer and more complex relational information for texts, thus improving the effectiveness of clustering. Therefore, in this paper, knowledge graph and K-medoids algorithm are combined to achieve predictive classification of archival texts. The main innovations and contributions of this work include:

(1) Aiming at the problem that traditional knowledge representation models cannot accurately deal with the semantic relationships between high similarity entities in archival knowledge graphs, this paper proposes a new knowledge representation learning model CTransR based on the TransR model [23, 24] by combining the principle of hyperparameter tuning. Compared with the traditional knowledge representation model, the relationship matrix projection improves the recognition ability between similar entities and improves the accuracy of knowledge representation.

(2) Aiming at the problem of classification accuracy, this paper proposes a Full-K-Modes clustering algorithm based on full attribute values, which effectively improves the accuracy of the clustering results of ordered-type categorical data. The effect of local optimal solutions is mitigated using a multi-attribute value clustering centre initialisation method based on preclustering. Using the dissimilarity metric improves the shortcomings

of the simple 0-1 matching metric method of the traditional K-Modes algorithm, effectively prevents the loss of important attribute values in the clustering process, and strengthens the similarity between attribute values under the same dimensional attributes within the class. Using the information entropy theory to calculate the weights of different dimensional attributes reinforces the differences between different dimensional attributes.

2. Relevant technical principles.

2.1. Knowledge representation learning model.

The construction of the knowledge graph starts with the extraction of knowledge from the dataset, and then the extracted text information is processed using the relevant algorithms of text processing, and finally the text information is stored in the knowledge base.

In the previous knowledge representation process, the learning is usually based on the ternary (h, r, t) representation [25]. The structure of the knowledge representation is shown in Figure 1. However, there are still many shortcomings in the ternary-based representation only, such as low computational efficiency and incomplete data information. With the advancement of knowledge representation technology, models based on deep learning techniques have made great progress in application. The semantic content of an entity is usually represented as a dense low-dimensional vector, and then a series of computations are carried out in this space, which can improve the efficiency of computation to a certain extent. We classify the commonly used knowledge representations into traditional representation models, complex relationship models, and other knowledge representation models. Since archive classification is a more complex task, this work focuses on the complex relational model.

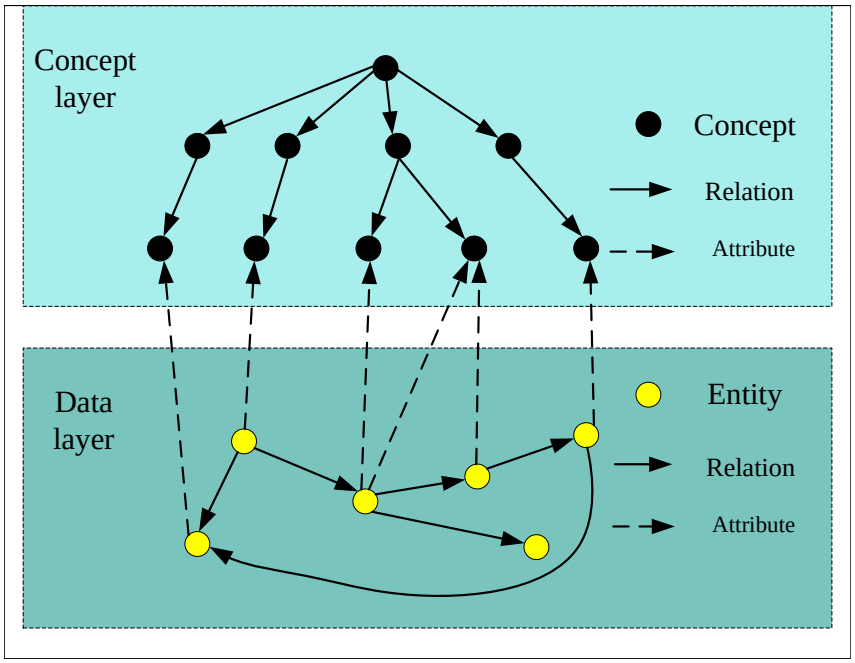

Figure 1. Structure of knowledge representation

We generally classify the entity relationships in a knowledge base into four types: 1-to-1, 1-to-N, N-to-1, and N-to-N, among which the latter three relationship types are mainly applied in complex relationship models. The TransH model becomes more effective in complex relationship mapping compared to the TransE model. Usually in the TransH model, we find that if the form of entities is different, then the entity structure will also change, even the same entity will change due to different relationships.

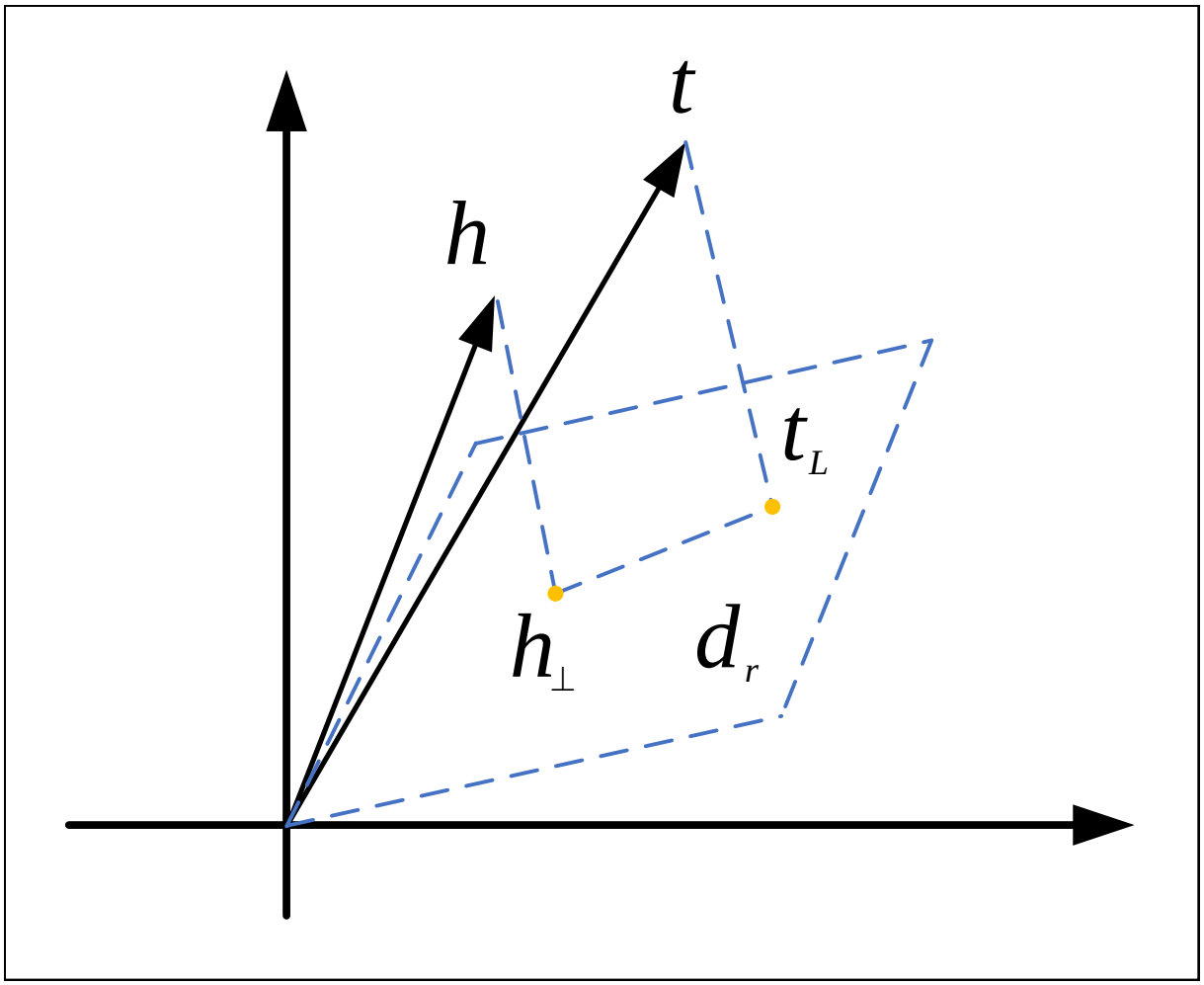

Figure 2. TransH vector space hypothesis

As shown in Figure 2, TransH employs the practice of transferring the vectors of the head and tail entities onto a chosen hyperplane by using the direction of the standard vector as a guide before making a scoring judgement. The representation of the entities under this model will be different as the relationship changes, but the location where the entity vectors are projected does not change, so the spatial latitude remains the same.

The components of the two entities after projection are:

$$h_{\perp} = h - W_r^{\top} h\, W_r \tag{1}$$

$$t_{\perp} = t - W_r^{\top} t\, W_r \tag{2}$$

The loss function for this model is:

$$f_r(h,t) = \left\| (h - W_r^{\top} h\, W_r) + d_r \; - \; (t - W_r^{\top} t\, W_r) \right\| \tag{3}$$

where d_r denotes the vector relationship of r.

Since entities and relations do not belong to the same vector, and the properties of relations are different if they belong to the same semantic space, it will reduce the expressive power of the model. Therefore, a new model—TransR—emerges on this basis. It maps the head and tail entities of the ternary (h, r, t) in the graph to the relation space, so that they conform to $I_{h_r} + I_r \approx I_{t_r}$, and then finally makes a judgement. The model considers

that entities should not be in the same space as relations, and then the relation space is independent. The loss function of the TransR model is as follows:

$$f_r(h,t) = \left\| h_r + r - t_r \right\| \tag{4}$$

The differences in meaning and attributes of entities in the knowledge base affect the representation of the model in the same matrix space. In addition, the TransR model does not take into account the links between entities and relationships in the knowledge representation. Instead, the TransD model relates the entities to the relationships by re-generating the projection matrices of the entities.

The mapping function for the head and tail entities of the model is as follows:

$$M_{rh} = r_p\, h_p^\top + I^{m\times n} \tag{5}$$

$$M_{rt} = r_t\, p_p^\top + I^{m\times n} \tag{6}$$

where $I^{m\times n}$ denotes the unit matrix; M_{rh} denotes the projection of the head entity under the relation r; M_{rt} denotes the projection of the tail entity; r_p and $h_p^\top/t_p^\top$ denote the mapping matrices of the head or tail entity.

The entity vectors and the loss function of the model are as follows:

$$h_\perp = M_{rh}\, h \tag{7}$$

$$t_\perp = M_{rt}\, t \tag{8}$$

$$f_r(h,t) = \left\| h_\perp + r - t_\perp \right\| \tag{9}$$

2.2. **Division based clustering algorithms.** Division-based clustering algorithms are a technique for dividing a dataset into groups or clusters, with the aim of making the data points within clusters as similar as possible and the data points in different clusters as different as possible. Common division-based clustering algorithms include K-means, K-medoids, DBSCAN and Mean Shift. Division-based clustering algorithms are usually suitable for medium-sized datasets and are easy to implement because they are relatively intuitive. The K-medoids clustering algorithm is similar to K-means, but instead of calculating the mean, it selects the most representative object among the data points as the centre of the cluster. This makes the algorithm more robust to noise and outliers. Archival text may contain irregular or rare text, and using K-medoids reduces the impact of these outliers on classification results.

In addition, when dealing with medium-sized archival text data, various distance measures need to be used to calculate the similarity between the texts. The K-medoids algorithm is relatively flexible, and different distance measures can be used to meet the specific needs of archival text classification. The dissimilarity measures used in clustering algorithms mainly include spatial distance measures for numerical data and attribute matching measures for ordered categorical data.

(1) Spatial distance metrics. The spatial distance metric formula is mainly used to calculate the degree of dissimilarity between two numerical object points. When the calculated spatial distance value is smaller, it means that the two object points are located in a closer spatial location. Minkowski Distance is calculated as follows:

$$d\big(x_i, x_j\big) = \big(|x_{i1} - x_{j1}|^r + |x_{i2} - x_{j2}|^r + \cdots + |x_{im} - x_{jm}|^r\big)^{1/r} \tag{10}$$

where $x_i = (x_{i1}, x_{i2}, \ldots, x_{im})$ is the vector representation of the object point x_i, and $x_j = (x_{j1}, x_{j2}, \ldots, x_{jm})$ is the vector representation of the object point x_j.

Manhattan Distance is often used to deal with the problem of calculating the distance of a path in a grid, and its calculation method is shown as follows.

$$d(x_i, x_j) = |x_{i1} - x_{j1}| + |x_{i2} - x_{j2}| + \cdots + |x_{im} - x_{jm}| \tag{11}$$

Euclidean Distance is the most commonly used dissimilarity measure formula in segmentation-based clustering algorithms, enabling good detection of spherical or convex class clusters, and is calculated as follows.

$$d(x_i, x_j) = \sqrt{|x_{i1} - x_{j1}|^2 + |x_{i2} - x_{j2}|^2 + \cdots + |x_{im} - x_{jm}|^2} \tag{12}$$

All the distance metric formulas mentioned above satisfy the following property: (1) non-negativity, (2) symmetry, and (3) triangular inequality.

(2) Attribute matching metrics. The attribute matching measure formula is often used to calculate the degree of difference between two ordered categorical data object points. Since this type of data has a limited number of attribute values, no sequential relationships, and no natural geometric properties, there are only two values of 0 and 1 for the degree of difference between two attribute values. A simple 0–1 match dissimilarity measure is shown as follows:

$$d(x_i, x_j) = \sum_{l=1}^{m} \delta\big(f(x_i, a_l), f(x_j, a_l)\big) \tag{13}$$

$$\delta\big(f(x_i, a_l), f(x_j, a_l)\big) = \begin{cases} 1, & f(x_i, a_l) \neq f(x_j, a_l), \\ 0, & f(x_i, a_l) = f(x_j, a_l). \end{cases} \tag{14}$$

where $f(x_i, a_l)$ denotes the attribute value taken by the object point x_i in the l^{th} dimension attribute. The simple 0–1 matching dissimilarity measure also satisfies non-negativity, symmetry and triangular inequality.

3. **CTransR modelling of archival texts.**

3.1. **Knowledge representation design.** The representation of knowledge graph is a Baidu encyclopaedia-like semantic network, which contains many ternary entities. The traditional representation based on entities and relations cannot satisfy efficient data computation, so we need to try other methods to improve the application depth of knowledge graph. Knowledge representation learning models require a large number of negative samples in the training process. Aiming at the problem that the samples obtained from traditional negative sampling are of low quality and not very helpful for training, this paper proposes a new knowledge representation learning model based on TransR–CTransR, which adopts the principle of hyperparameter tuning, generates negative samples based on the current embedding of entities and relations, and gives different weights to negative samples with different scores.

Knowledge representation is a key step in the process of archival knowledge graph construction, and the technical manual document data acquired through OCR and other technical means need to be represented in vectorial words through knowledge representation. As mentioned above, the TransR model projects the ternary (h, r, t) entity pairs in the knowledge base into the relational space through the mapping matrix, thus realising $h_r + r \approx t_r$. TransR model projects entity vectors into a relationship-specific space by

introducing a relationship-specific projection matrix, thus solving the problem of relationship vector sharing. In this way, the semantic information between different relationships can be better distinguished, and the vector representation of different entities under different relationships can be better preserved. Therefore, compared with the TransE and TransH models, the TransR model can represent the semantic association between entities and relationships more accurately when dealing with knowledge maps with multiple relationships.

3.2. **Principle of hyper-parameter adjustment.** This study introduces the CTransR model, which is designed to address the issue of the TransR model's limited capacity to differentiate between highly similar entities. The CTransR model is built on the notion of hyper-parameter tuning and utilizes a relation matrix projection approach. As shown in Figure 3, we propose the principle of hyper-parameter tuning for the first time. For any triple (h, r, t), given that r and t are known, we let the head entities have identical orientations but varying sizes, and ϕ is adjusted as a hyperparameter. Similarly, we make the assumption that when given h and t, r is allowed to have the same orientation but might vary in size, and β is adjusted as a hyperparameter. Once the values of h and r are established, we let t maintain the same orientation but vary in size, while λ is adjusted as a hyperparameter.

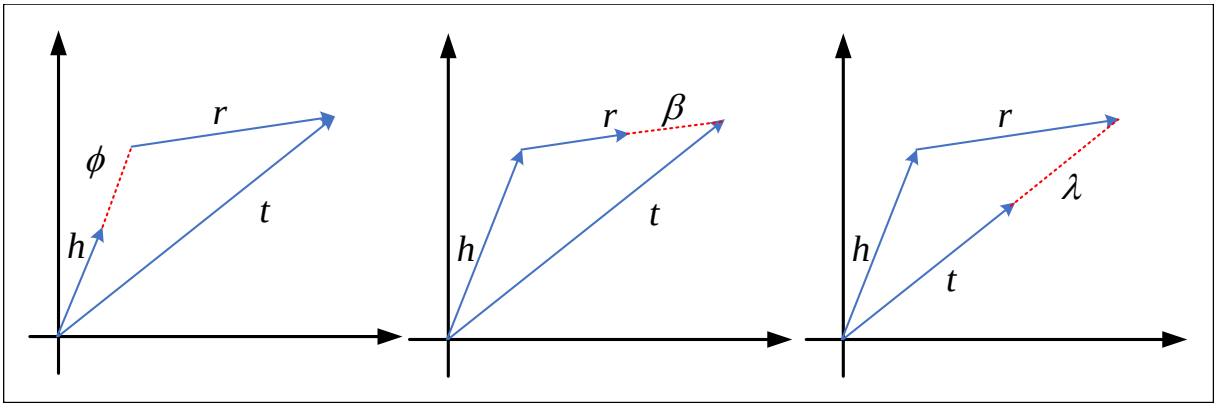

Figure 3. Principles of hyperparameter adjustment

In translation, entities and relations are embedded inside the same area. The adaptation rule is defined as follows:

$$(h + \phi + j_1) + (r + \beta + j_2) \approx (t + \lambda + j_3) \tag{15}$$

where $\phi, \beta, \lambda, j_1, j_2, j_3 \in \mathbb{R}^{m \times n}$. Accordingly, the evaluation function is shown as follows:

$$f_r(h, t) = \|h + r - t + \mu\| \tag{16}$$

$$\mu = \phi + j_1 + \beta + j_2 - (\lambda + j_3) \tag{17}$$

If the triad (h, r, t) exists, the function is small; otherwise, it is large. The score function uses the Euclidean distance measure between $h + r$ and t. Each dimension of the vector is considered equally important whether it is an entity vector or a relationship vector. To the TransR model's projection of the relation matrix, we add the principle of flexible transformation. In the relational representation space of r, for a given r and t_r, the head entity is denoted by $h_r + \phi + j_1$. For a given h_r and t_r, the relationship is denoted by $r + \beta + j_2$. For a given h_r and r, the tail entity is denoted by $t_r + \lambda + j_3$. Therefore, the evaluation function of CTransR is given by:

$$f_r(h,t) = \|h_r + r - t_r + \mu\| \tag{18}$$

3.3. **Structure of the model.** Building the negative triad is a crucial step in training a model. We use a probabilistic technique whereby we substitute the head entity with the tail entity. The preferred technique for constructing negative triples involves replacing entities that have various connection types. In N-to-1 connections, the entity with a greater probability is selected to represent the relationship. In 1-to-N connections, the head entity is represented by selecting a higher probability. When entities have numerous characteristics, replacing the tail entity in N-to-1 connections enables effective training of several properties of the entity. When handling 1-to-N connections, replacing the main entity allows for complete training of numerous characteristics of the main object.

We use a priori probability to address the issue of encoding analogous connections. The probability of (h,t) satisfying a relation is proportional to the frequency of the relationship occurs. The a priori probability of a relation is computed as follows:

$$p_r(r) = \frac{N_r}{N_r + N_r'} \tag{19}$$

where N_r denotes the number of occurrences and N_r' denotes the most analogous connection r. In order to distinguish the triples, we use the loss function of the edges to represent the optimisation objective function of the model: N_r denotes the number of occurrences.

$$L = \sum_{(h,r,t)\in S} \sum_{(h',r',t')\in S'} \max\bigl(f_r(h,t) + \gamma - f_r(h',t'),\, 0\bigr) \tag{20}$$

where S denotes the positive triad, the relation matrix projection S' denotes the negative triad, $\max(x,y)$ is the maximum of x and y, and γ denotes the distance between the loss function scores in the triad. The optimisation objective is to completely distinguish between positive and negative triples.

4. **Archival text classification models.**

4.1. **Archival data preprocessing.** Data preprocessing is very important in the data mining process, and many mature algorithms have stringent requirements on the experimental data set.

4.1.1. *Archive data preparation.* Image recognition processing of digital archives. The image recognition of this work mainly relies on the more mature OCR software, using ABBYY FineReader11 as the recognition tool. The resolution of the scanned documents must be above 300 dpi. Randomly selected part of the recognition of the text file as a sample, found that the recognition rate of a single document can reach about 80 per cent, the recognition rate is basically guaranteed to restore the main information contained in the archives themselves, and exclude a variety of noise word factors, I believe that the degree of restoration is relatively reliable.

In order to avoid directly selecting archive texts with large order of magnitude as experimental samples, which will increase uncertainty and experimental difficulty, and to ensure the effectiveness and feasibility of Chinese text preprocessing and classifier. In this paper, we use part of the files selected from the manually classified archive text as the training set for the experiment, and then randomly select part of the files from the training set data as the test set. The sample data are presented through the archive catalogue, and the actual samples are the archive text files corresponding to the catalogue entries, stored

in TXT format, and the actual classification of each archive text file has been marked in the corresponding entries.

4.1.2. *Construction of the deactivation lexicon.* After processing the archive corpus using the jieba participle algorithm, a deactivation lexicon is prepared to be introduced in order to prepare for the dimensionality reduction of the archive text data features and to improve the accuracy of the decision tree classification algorithm. Due to the professional and special nature of the archive itself, it is necessary to select a suitable deactivation dictionary for the archive industry to reduce the noise caused by various invalid words, further optimise the classification algorithm and reduce the algorithm execution time. In this paper, the joint entropy algorithm is chosen to select deactivated words.

Joint entropy requires the calculation of the frequency of occurrence of a word within a text and the frequency of occurrence of a sentence containing that word [26, 27]. By the probability of occurrence of a word within each text in the sample set and the probability of occurrence of a sentence containing the word in each text P_j, the corresponding entropy can be calculated.

Finally, the deactivated words are selected based on the ranking of the joint entropy results. The joint entropy is calculated as follows.

$$W(w_i) = H(w_i) + H(s \mid w_i) \tag{21}$$

$$H(w_i) = -\sum_{j=1}^{n} P_j(w_i) \lg P_j(w_i) \tag{22}$$

$$H(s \mid w_i) = -\sum_{l=1}^{n} P_l(s \mid w_i) \lg P_l(s \mid w_i) \tag{23}$$

$$P_j(w_i) = \frac{f_j(w_i)}{\sum_{j=1}^{n} f_j(w_i)} \tag{24}$$

$$P_l(s \mid w_i) = \frac{f_l(s \mid w_i)}{\sum_{l=1}^{m} f_l(s \mid w_i)} \tag{25}$$

where $f_j(w_i)$ is the frequency of occurrence of the word w_i in the sentence j; n is the number of sentences; $f_l(s \mid w_i)$ is the frequency of occurrence of the sentence containing w_i in the text l; m is the number of texts.

The rationale is that only if the average information content of a word occurring in a sentence and the average information content of the sentence containing the word are both large, does it mean that the word is more common.

4.2. **Full-K-Modes algorithm based on full attribute values.** Aiming at the shortcomings exposed by the traditional K-Modes algorithm, this paper proposes the Full-K-Modes clustering algorithm based on full attribute values. Instead of retaining only the attribute values with the highest frequency of occurrence in each dimension, all the attribute values and their frequencies occurring in the class are retained. The categorical attribute distance calculation method extends the probability-based calculation method to reinforce the similarity between attribute values under the same dimensional attribute within the class. Information entropy theory is used to calculate the weights of attributes of different dimensions to reinforce the differences between attributes of different dimensions.

Let $S = (U, A, V, f)$ be a question about classifying archive texts. Here, $U = \{x_1, x_2, \ldots, x_n\}$ represents a non-empty finite collection of data points, and n represents the number of

samples in the data set. A is a non-empty finite set of attribute dimensions, denoted as $\{a_1, a_2, \ldots, a_m\}$. Here, m represents the number of attributes. V is the full set of all attribute domains, $V_{a_j} = \{a_j^1, a_j^2, \ldots, a_j^{n_j}\}$ is the finite unordered non-repeating set of attribute values, where $1 \leq j \leq m$, n_j denotes the total number of attribute values that can be taken for the attribute dimension a_j. f is the mapping of the information function, including the data point mapping function f_1 and the clustering centre Modes mapping function f_2.

The degree of difference between attribute values under the same dimension attribute in Full-K-Modes clustering algorithm is calculated as follows:

$$\delta_1\big(f_1(x_i, a_j),\ f_2(z_l, a_j)\big) = 1 - p\big(a_j^{q_{jl}}\big) \tag{26}$$

where $f_1(x_i, a_j) = a_{jl}^q \in V_{a_j}$. When $p(a_{jl}^{n_j})$ takes the value 0, the attribute value $a_{jl}^{n_j}$ has the largest degree of dissimilarity with it, and the intra-class distance can be taken up to the maximum value 1. When $p(a_{jl}^{n_j})$ is 1, the attribute value $a_{jl}^{n_j}$ has the smallest degree of dissimilarity, and the distance within the class is the smallest value 0.

Information entropy is used to quantify the degree of difference in the composition of attribute values in order to measure the weight of an attribute [28]. As the calculated information entropy gets smaller, the weight of the attribute for that dimension gets larger. The weights of each dimension attribute are recalculated during each iteration as the clustering centre Modes are changed to ensure that the weights can be adjusted in time. The calculation method of attribute weights is shown below:

$$\varpi_{jl} = 1 \ - \ \frac{\sum_{i=1}^{n_j} p(a_{jl}^i) \log p(a_{jl}^i)}{n_j} \tag{27}$$

where ϖ_{jl} denotes the weight value of the attribute dimension a_j of the l-th class. When ϖ_{jl} is larger, it means that the difference of attribute values in the attribute dimension a_j of the l class is smaller. It means that the attribute can better identify the class.

The dissimilarity measure based on full attribute value clustering centres is calculated as shown below:

$$d_1(x_i, z_l) = \sum_{j=1}^{m} \varpi_{jl}\, \delta_1\big(f_1(x_i, a_j),\ f_2(z_l, a_j)\big) \tag{28}$$

5. **Experimental results and analyses.**

5.1. **Experimental environment and data.** The experiments in this work were done on a PC (Intel® CoreTM i5-4590 Processor 3.70 GHz, Windows 7 64-bit) with PyCharm 2018.3.3 (Professional Edition) development platform. Using the Python 3.6.1 environment.

The experimental data in this paper comes from some of the paper files received by a city archive management department. In practice, because the file formation units did not categorise these files according to the rules, it is assumed that the archivists need to reorganise this part of the files and manage to preserve them according to different categories, in order to improve the accuracy of the files to find and retrieve them in the service utilisation. There are 728 documents in the text base of the archive, which are divided into 6 categories. The text categories are "Administrative Documents", "Laws and Regulations", "Finance and Accounting", "Personnel Training", "Engineering and Technology", "Academic and Scientific Research", "Medical and Health", "Business Management" and "Other Types". The ratio of training and test sets is still 8:2 as discussed earlier in this paper. The detailed information of the experimental data is shown in Figure 4.

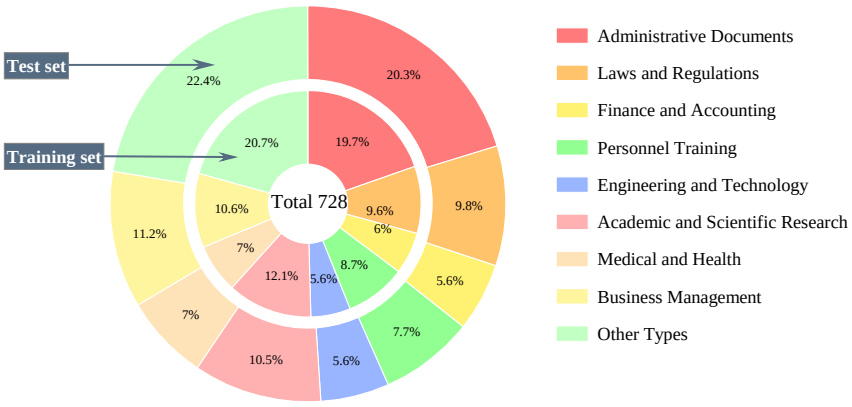

Figure 4. Experimental data details

5.2. **Comparative Analysis of Classification Algorithms.** Experiments were conducted to compare the traditional K-Modes algorithm, DP-K-Modes algorithm [28], Rough-Set-K-Modes algorithm [29] and CTransR–Full-K-Modes proposed in this paper. In order to validate the effectiveness of the algorithms, 100 trials were taken for each algorithm and the average Accuracy was calculated, Precision and Recall. The archival text classification results of various improved K-Modes algorithms are shown in Figure 5.

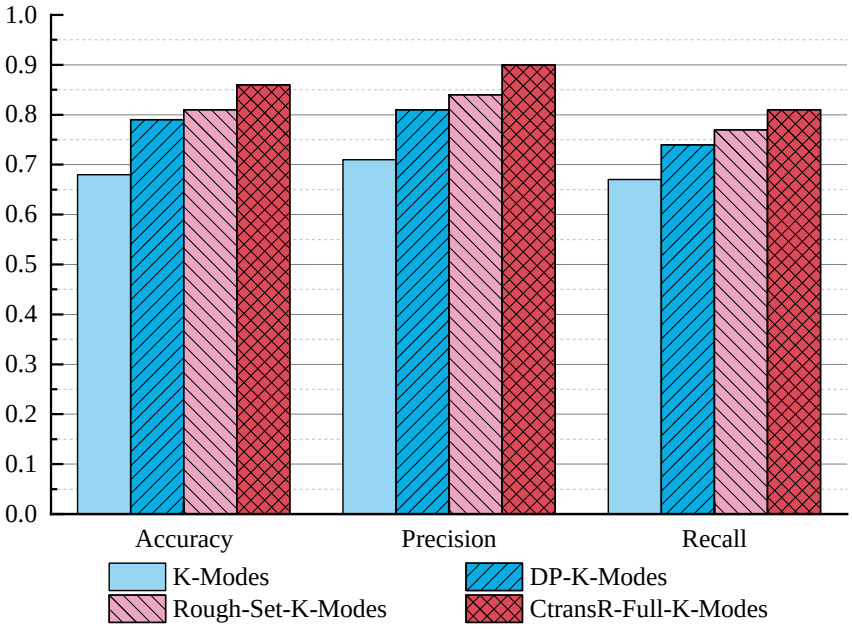

Figure 5. Archival Text Classification Effectiveness of Various Improved K-Modes Algorithms

It can be found that in the experiments on the archive classification dataset the CTransR–Full-K-Modes algorithm improves Accuracy by 16.7%, Precision by 19.1% and Recall by 14.3% over the traditional K-Modes algorithm. It also improves by 3.8%, 5.8% and 3.2% respectively compared to other improved algorithms with better results. Therefore, the clustering effect of the proposed algorithm is better than the traditional K-Modes algorithm and the improved K-Modes algorithm of other scholars. The total number of iterations (preprocessing iterations and algorithmic iterations) is slightly higher than other

algorithms, but within acceptable limits, as the CTransR–Full-K-Modes algorithm requires some additional computational overheads in the data preclustering stage due to the use of preclustering-based centroid selection method. Therefore, it is proved that the CTransR–Full-K-Modes algorithm proposed in this paper can effectively improve the accuracy of the clustering results of ordered type classification data.

6. **Conclusion.** In this work, the CTransR–Full-K-Modes algorithm is proposed for predictive classification of archival texts by combining knowledge graph and K-medoids algorithms. A new knowledge representation learning model CTransR is proposed based on the TransR model by incorporating the principle of hyperparameter tuning. Compared with the traditional knowledge representation model, the use of relationship matrix projection improves the recognition between similar entities and increases the accuracy of knowledge representation. A Full-K-Modes clustering algorithm based on full attribute values is proposed, which effectively improves the accuracy of clustering results for ordered categorical data. The effect of local optimal solutions is mitigated using a multi-attribute value clustering centre initialisation method based on preclustering. The shortcomings of the simple 0–1 matching metric method of the traditional K-Modes algorithm are improved using the dissimilarity metric. The use of information entropy theory to calculate the weights of different dimensional attributes reinforces the dissimilarity between different dimensional attributes. In the experiments on the archive classification dataset CTransR–Full-K-Modes algorithm improves Accuracy by 16.7%, Precision by 19.1%, and Recall by 14.3% over the traditional K-Modes algorithm. It also improves 3.8%, 5.8% and 3.2% respectively compared to other improved algorithms with better results.

REFERENCES

[1] B. E. L'Eplattenier, "An argument for archival research methods: thinking beyond methodology," *College English*, vol. 72, no. 1, pp. 67–79, 2009.

[2] M. C. Zhang, D. N. Stone, and H. Xie, "Text data sources in archival accounting research: Insights and strategies for accounting systems' scholars," *Journal of Information Systems*, vol. 33, no. 1, pp. 145–180, 2019.

[3] C. Hyde and J. Rezek, "Introduction: The Aesthetics of Archival Evidence," *The Journal of Nineteenth-Century Americanists*, vol. 2, no. 1, pp. 155–162, 2014.

[4] J. Clary-Lemon, "Archival research processes: A case for material methods," *Rhetoric Review*, vol. 33, no. 4, pp. 381–402, 2014.

[5] J. Xia and L. Sun, "Assessment of self-archiving in institutional repositories: Depositorship and full-text availability," *Serials Review*, vol. 33, no. 1, pp. 14–21, 2007.

[6] P. Eggert, "The archival impulse and the editorial impulse," *The Journal of the European Society for Textual Scholarship*, no. 14, pp. 3–22, 2019.

[7] J. Munday, "The role of archival and manuscript research in the investigation of translator decision-making," *International Journal of Translation Studies*, vol. 25, no. 1, pp. 125–139, 2013.

[8] K. D. Good, "From scrapbook to Facebook: A history of personal media assemblage and archives," *New Media & Society*, vol. 15, no. 4, pp. 557–573, 2013.

[9] J.-Y. Lee, J.-Y. Moon, and H.-J. Kim, "Examining the intellectual structure of records management & archival science in Korea with text mining," *Journal of the Korean Society for Library and Information Science*, vol. 41, no. 1, pp. 345–372, 2007.

[10] D. Beard, "From work to text to document," *Archival Science*, vol. 8, pp. 217–226, 2008.

[11] W. M. Duff, E. Monks-Leeson, A. Galey, and I. N. K. E. Team, "Contexts built and found: a pilot study on the process of archival meaning-making," *Archival Science*, vol. 12, pp. 69–92, 2012.

[12] A. Vellino and I. Alberts, "Assisting the appraisal of e-mail records with automatic classification," *Records Management Journal*, vol. 26, no. 3, pp. 293–313, 2016.

[13] Z. Zhao, "Classification tree algorithm and its application in general archives management system," *Procedia Computer Science*, vol. 228, pp. 946–951, 2023.

[14] W. e. Hadi, Q. A. Al-Radaideh, and S. Alhawari, "Integrating associative rule-based classification with Naïve Bayes for text classification," *Applied Soft Computing*, vol. 69, pp. 344–356, 2018.

[15] J. Franks, "Text Classification for Records Management," *Journal on Computing and Cultural Heritage (JOCCH)*, vol. 15, no. 3, pp. 1–19, 2022.
[16] S. Brokensha, E. Kotzé, and B. Senekal, "Machine learning for document classification in an archive of the National Afrikaans Literary Museum and Research Centre," *Journal of the South African Society of Archivists*, vol. 56, pp. 134–147, 2023.
[17] D. Mustafi, A. Mustafi, and G. Sahoo, "A novel approach to text clustering using genetic algorithm based on the nearest neighbour heuristic," *International Journal of Computers and Applications*, vol. 44, no. 3, pp. 291–303, 2022.
[18] I.-C. Chang, T.-K. Yu, Y.-J. Chang, and T.-Y. Yu, "Applying text mining, clustering analysis, and latent dirichlet allocation techniques for topic classification of environmental education journals," *Sustainability*, vol. 13, no. 19, 10856, 2021.
[19] T.-Y. Wu, H. Li, S. Kumari, and C.-M. Chen, "A Spectral Convolutional Neural Network Model Based on Adaptive Fick's Law for Hyperspectral Image Classification," *Computers, Materials & Continua*, vol. 79, no. 1, pp. 19–46, 2024.
[20] T.-Y. Wu, A. Shao, and J.-S. Pan, "CTOA: Toward a Chaotic-Based Tumbleweed Optimization Algorithm," *Mathematics*, vol. 11, no. 10, 2339, 2023.
[21] T.-Y. Wu, H. Li, and S.-C. Chu, "CPPE: An Improved Phasmatodea Population Evolution Algorithm with Chaotic Maps," *Mathematics*, vol. 11, no. 9, 1977, 2023.
[22] C. Zins, "Knowledge map of information science," *Journal of the American Society for Information Science and Technology*, vol. 58, no. 4, pp. 526–535, 2007.
[23] J. Watthananon and A. Mingkhwan, "Optimizing knowledge management using knowledge map," *Procedia Engineering*, vol. 32, pp. 1169–1177, 2012.
[24] D. Zhong, J. Fan, G. Yang, B. Tian, and Y. Zhang, "Knowledge management of product design: A requirements-oriented knowledge management framework based on Kansei engineering and knowledge map," *Advanced Engineering Informatics*, vol. 52, 101541, 2022.
[25] F. Liu, Y. Zhang, and L. Li, "Review of systematic financial risk research based on knowledge map," *Procedia Computer Science*, vol. 199, pp. 315–322, 2022.
[26] C.-F. Xie, L.-X. Xu, and F. Zhang, "Concentration control of SMB system in parameter space of region velocity based on adjusted fuzzy control," *Journal of Intelligent &Fuzzy Systems*, vol. 38, no. 5, pp. 5717–5729, 2020.
[27] Z. Cheng and F. Zhang, "Flower End-to-End Detection Based on YOLOv4 Using a Mobile Device," *Wireless Communications and Mobile Computing*, vol. 2020, 8870649, 2020.
[28] P. Mei, G. Ding, Q. Jin, F. Zhang, and Y. Jiao, "Quantum-Based Creative Generation Method for a Dancing Robot," *Frontiers in Neurorobotics*, vol. 14, 559366, 2020.
[29] J. Xie, M. Wang, X. Lu, X. Liu, and P. W. Grant, "DP-k-modes: A self-tuning k-modes clustering algorithm," *Pattern Recognition Letters*, vol. 158, pp. 117–124, 2022.
[30] S. Wang, Z. Yuan, C. Luo, H. Chen, and D. Peng, "Exploiting fuzzy rough entropy to detect anomalies," *International Journal of Approximate Reasoning*, vol. 165, 109087, 2024.