Application of Text Dimensionality Reduction Method in Information Filtering on Railway Transit Cloud Platform

Ye Qiu\textsuperscript{1,2}, Gang Li\textsuperscript{1,*} and Junfeng An\textsuperscript{1}

\textsuperscript{1}Jinan Rail Transit Group Co. LTD., Jinan, China
\textsuperscript{2}State Key Laboratory of High-end Server & Storage Technology, Jinan, China

*Corresponding author email: 172714484@qq.com

Abstract. This paper focuses on the useful information data extraction of urban rail transportation's urban rail cloud platform. Through the correlation and redundancy between feature items in the text vector space of a large amount of information data, a method of clustering text vector space dimensionality reduction is proposed and applied to the information data extraction of urban rail cloud platform. The vector aggregation theory and feature selection are applied to reduce the dimensionality of the feature space, so that the reduced dimensional feature items are more representative of the category. Through test experiments, it is proved that this text dimensionality reduction method algorithm is applied to the information data filtering of urban rail cloud platform, which effectively reduces the dimensionality of vector space and improves the accuracy and speed of text classification in urban rail cloud platform information data.

Keywords: Text dimensionality reduction; Feature selection; Vector aggregation; Intelligence operations; Urban Rail Transit.

1. Introduction

Urban rail transit adopts an integrated model of construction, deployment, development and management etc., which is responsible for planning and design, financing, construction, management, operation and property development of rail transit. In order to realize the unified deployment of information application systems, Intelligence operations, system security operation and maintenance, and independent control of business data, it has become an urgent need to build a data center room cloud platform, deploy IT hardware resources required for application systems, integrate network architecture, and unify operation and maintenance of computing, storage and network environment.

The urban rail cloud platform creates a unified portal for rail transit, unifies the bearing of rail transit safety production business, internal enterprise management business and passenger service business, relies on cloud computing and big data technology, builds a data sharing platform for information infrastructure construction and big data analysis, realizes the integration of line network resources, unified management, elastic expansion and on-demand deployment, ensures the safety, reliability and continuous operation of core business application systems, enhances the response to various It will ensure the safety, reliability and continuous operation of core business application systems, enhance the resilience to various emergencies, strengthen the ability to resist disasters and accidents, improve the response speed and operation management level, lay a good foundation for the construction of digital rail transit, provide more value-added data services and intelligent decision support, enhance resource management and improve economic efficiency\cite{1}. The construction of urban rail cloud platform is serving as an increasingly significant role in today's fast-pace development of Internet. The
information data in the cloud platform are far too complex and of huge scale, how to utilize these data effectively to make an thorough analysis is the focus of our research.

2. Related Work
The construction of urban rail cloud platform provides practical and user-friendly network management functions to make topology, fault, performance, configuration, security and other management functions possible on the basis of centralized management of network resources, which not only provides functions but also shows how to use the functions to meet business needs by way of process wizard. However, for the large amount of equipment information data, a variety of information data types, useful information data is widely distributed and relatively scattered, and the information text feature matrix always shows thousands or even larger dimensions, which makes the work of extracting useful data very complicated. Therefore, solution to this problem is reducing the dimensionality of the information text feature matrix. Dimensionality Reduction (DR) is a typical step in many text mining problems which involves transforming sparse data into a shorter and more compact one. DR can be done in 2 ways: feature reduction and feature selection. In reference [2], K-means algorithm is deployed to achieve feature selection, where InfoGain DR technique is implemented. Experimental analysis validated the efficiency of this approach with improved accuracy, precision and recall.

In literature [3], authors developed an enterprise-category matrix which manifests the features of enterprise users, and was lately validated by improved performance to be applicable to Railway B2B E-commerce platform. Reference [5] designed a text mining-based identification model for urban rail transit system infrastructure fault analysis, which uses the Naive Bayes algorithm to segment the long text data which has been preprocessed, and later identifies the cause of infrastructure fault. In reference [5], a naive Bayes classifier is constructed via space vector method to classify railway complaint texts automatically. And compared to other competitors, it achieved nearly 50% increase in accuracy of text categorization. In study [4], the author explored fuzzy clustering as a novel NFT-based approach to create a lower-dimensional representation of texts grabbed from documents. Furthermore, this study also took one step further into dimensionality reduction with the ability to work among both discrete and continuous data. To achieve dimensionality reduction without alteration of the inner semantic relation between data, reference [4] introduced a manifold learning-based method to explore the low-dimensional, neighborhood text.

In the IaaS service of urban rail cloud platform, information text classification usually adopts vector space model (VSM) to express text feature terms, i.e., the text-feature word matrix is constructed by calculating the weights of feature words in the text, and the number of feature words appearing in the text is large, and the high-dimensional feature terms may not be all crucial and beneficial for clustering, and may even greatly increase the clustering time and only produce clustering results associated with a much smaller subset of features. Moreover, the complexity of the algorithm and implementation of clustering increases rapidly as the dimensionality of the pattern space becomes larger. Therefore, choosing as less features to express more information can improve accuracy.

3. Urban Rail Cloud Platform Architecture
The network architecture of the cloud platform is composed of hardware and software, in which the hardware resource pool of the cloud platform provides independent hardware physical devices, and the physical devices are divided into multiple virtualized logical resources for use by various business systems. Both software resource pooling scheme and hardware resource pooling scheme can realize network virtualization and can divide logical resources flexibly according to tenants, and software resource pooling and hardware resource pooling methods can be selected independently for different security protection technologies. The current cloud computing platform includes IaaS layer, PaaS layer and SaaS layer, as can be seen Figure 1. IaaS layer consists of logical/pooled computing, storage, network and other hardware and software resource pools and encapsulated IaaS services, which can be used directly by cloud service users or combined to support more complex business scenarios.
Figure 1. Cloud Platform Architecture.

The PaaS layer provides customers with the environment and services to deploy, manage and run applications, and should provide application frameworks, middleware and corresponding deployment and control capabilities, while PaaS should also provide code management, compilation and packaging, release and deployment, continuous integration and continuous delivery, and other development and operation integration services. The SaaS layer covers safety production, internal management and external services, and it is appropriate for cloud service providers to provide application-based services in combination with the business scenario of smart city rail transportation cloud.

The urban rail cloud platform carries out safety production, internal management and external service business, as shown in Figure 2, which reserves the capacity expansion of future line business application system, video storage, enterprise portal, enterprise management platform and other business. Adopting technologies such as server virtualization, desktop virtualization, network virtualization, storage virtualization, cloud security and cloud computing management to build a data center in cloud computing mode that is easy to manage, dynamic and efficient, flexible and scalable, stable and reliable, and used on demand. Server, network, storage and security virtualization mainly provide computing/network/storage/security resource services for each city rail business system. Desktop virtualization provides scheduling/monitoring desktop workstation services for business system users.

The cloud platform deploys computing resource pools in the control center to unify the CPU, memory and other computing resources of physical servers in the control center to provide corresponding computing services for each business system. The computing resource pool can be divided into virtualized resource pool and physical machine resource pool, in which the virtualized resource pool can provide services with corresponding performance levels according to different needs of business systems and features flexible resource allocation, dynamic adjustment of specifications, and high resource utilization; the physical machine resource pool mainly provides physical machine services and features high performance of single device and strong security.
4. Traditional Text Classification Algorithm

The text classification algorithm, originally proposed by Cover and Hart in 1968, is a theoretically mature approach. The basic idea of the algorithm is that according to the traditional vector space model, the text content is formalized as a weighted feature vector in the feature space, i.e., $D = D(T_1, W_1; T_2, W_2; \cdots; T_n, W_n)$. For a text test, the similarity between it and each text in the training sample set is calculated to find the K most similar texts, and the category to which the test text belongs is judged according to the weighted distance and the specific algorithm steps are as follows:

- For a test text, form a text test vector based on the feature words.
- Calculate the text similarity between this test text and each text in the training set, and the calculation formula is:

$$\text{Sim}(d_i, d_j) = \frac{\sum_{k=1}^{M} W_{i,k} \times W_{j,k}}{\sqrt{\sum_{k=1}^{M} W_{i,k}^2} \times \sqrt{\sum_{k=1}^{M} W_{j,k}^2}}$$

(1)

As can be seen in equation (1), $d_i$ is the feature vector of the test text, $d_j$ is the central vector of the jth class; M is the dimension of the feature vector; $W_k$ is the kth dimension of the vector. k value is generally determined by using an initial value, and then adjusting the k value according to the results of experimental testing, generally the initial value is set at a few hundred to a few thousand.

- According to the text similarity, the k texts in the training text set are selected as the most similar to the test text.
- Among the k neighbors of the text, the weights of each class are calculated in turn, and the formula is as follows:

$$P(X_{ij}) = \begin{cases} 1 & \text{If } \sum_{d_i \in \text{knns}} \text{Sim}(X_{di})y(d_i, C_j) - b \geq 0 \\ 0 & \text{Other Conditions} \end{cases}$$

(2)

here: X is the feature vector of the text; $\text{Sim}(X_{di})$ is the similarity calculation formula; b is the threshold value, to be selected optimally; and $y(d_i, C_j)$ takes the value of 1 or 0. If $d_i$ belongs to $C_j$, the function value is 1, otherwise it equals to 0.

- Compare the weights of classes and assign the text to the one with the greatest weight.

This method is based on analogy learning, a non-parametric classification technique, which is very effective in statistical-based pattern recognition, and can achieve high classification accuracy for unknown and non-normal distributions, with the advantages of robustness and conceptual clarity. However, in text classification, there are shortcomings in the method, such as The algorithm is a lazy classification algorithm, and its time and space overhead is large; when calculating the similarity, the feature vector dimension is high, and the association relationship between feature words is not considered; when calculating the sample distance, the weights of each dimension are the same, which
makes the calculation of the distance between feature vectors not accurate enough and affects the classification accuracy.

Traditional text dimensionality reduction methods have a commonality, that is, the ability of individual features to distinguish between categories is calculated separately by some evaluation function, and the feature sets selected by these methods often have redundancy because the correlation between features is not considered. To address this problem, a clustering-based text dimensionality reduction method is proposed, in which the redundancy among features is first cut down using the clustering method, and then the features with strong category differentiation ability are selected using the TF-IDF method. This clustering-based feature selection method can effectively reduce the redundancy of feature sets and achieve the effect of text dimensionality reduction.

5. Application of Clustering Algorithm in Text Dimensionality Reduction Methods

To improve the speed of text classification, the number of feature terms cannot be changed with a certain number of texts, and it is also possible to start from feature term calculation. Therefore, a clustering-based feature term reduction algorithm is proposed to address the redundancy of features in category differentiation. The main idea is to incorporate the traditional comparison method of feature terms into the algorithm; the algorithm is based on the spatial vector model VSM, which represents each feature term as a vector, and first finds two words with the same or associated original feature terms and their weights, and then uses the corresponding weight vectors of the feature terms to calculate the similarity between these two feature vectors. Based on the similarity between the features, the features are clustered, and one feature in each cluster is selected to represent the whole cluster, and the other features in the cluster are eliminated from the candidate feature set, so that the redundancy in the feature set is greatly reduced. The specific algorithm is as follows.

Firstly, the feature items in the training text are represented as vectors, and let W be the \( m \times n \) text feature item matrix (where \( m \) is the number of text in the training set and \( n \) is the number of different feature items in the training set), \( i \) in \( W[i, j] \) is the vector space representation of the \( i \)th document, the \( j \)th column of \( W \) is the \( j \)th feature item and its frequency of occurrence in each document, and then let \( r \) be the number of dimensions to be reduced to.

Then the traditional clustering algorithm is used for \( r \)-way clustering of the training set, as follows. To improve the speed of text classification, the number of feature terms cannot be changed with a certain number of texts, and it is also possible to start from feature term calculation. Therefore, a clustering-based feature term reduction algorithm is proposed to address the redundancy of features in category differentiation. The main idea is to incorporate the traditional comparison method of feature terms into the algorithm; the algorithm is based on the spatial vector model VSM, which represents each feature term as a vector, and first finds two words with the same or associated original feature terms and their weights, and then uses the corresponding weight vectors of the feature terms to calculate the similarity between these two feature vectors. Based on the similarity between the features, the features are clustered, and one feature in each cluster is selected to represent the whole cluster, and the other features in the cluster are eliminated from the candidate feature set, so that the redundancy in the feature set is greatly reduced. The specific algorithm is as follows.

- Firstly, the feature items in the training text are represented as vectors, and let \( W \) be the \( m \times n \) text feature item matrix (where \( m \) is the number of text in the training set and \( n \) is the number of different feature items in the training set), \( i \) in \( W[i, j] \) is the vector space representation of the \( i \)th document, the \( j \)th column of \( W \) is the \( j \)th feature item and its frequency of occurrence in each document, and then let \( r \) be the number of dimensions to be reduced to.
- Then the traditional clustering algorithm is used for \( r \)-way clustering of the training set, as follows.

\[
\text{sim}(A,B) = \frac{A \cdot B}{|A||B|} = \frac{x_1 y_1 + y_2 y_2}{\sqrt{x_1^2 + y_1^2} \sqrt{x_2^2 + y_2^2}}
\]
• If the similarity of the feature and all cluster centers is less than the threshold, a new cluster is created with the feature as the center; otherwise the feature is added to the cluster with the greatest similarity.
• Repeat 2 and 3 until all the features are processed.
• Finally, the training set is divided into $r$ disjoint subsets $S_1, S_2, \cdots, S_r$. The center-of-mass point vector $C_i$ of each set $S_i$ is calculated separately, and $C_i$ is used as one coordinate axis of the reduced dimensional space to form an $r$-dimensional subspace, and the $r$-dimensional vector representation of each text can be obtained by mapping it to this $r$-dimensional subspace; therefore, the reduced dimensional text vector space $W'$.

$$W'_m \times r' = Wm \times n \cdot Cn \times r$$
Where $Cn \times r = (C1, C2, \cdots, Cr)$ \hspace{1cm} (4)

For a set of $r$ prime point vectors and a text, the $i$th dimensional coordinate of $d$ in the $r$-dimensional space is its cosine similarity to the $i$th prime point vector. The more rationally the $r$ initial clusters are divided, the more accurately the original matrix is represented after dimensionality reduction. In dimensionality reduction, the prime point vector specifies a mechanism for generalizing its contents, in particular, the major dimension of the vector corresponds to the most important feature term in the set. What should be noted is how much it should cost to find the $r$ clusters and their center-of-mass points. For each prime point, there are always a small number of related feature terms whose weights account for most of their total value. The same feature term often appears in multiple prime points, which can easily happen when two prime points are different parts of the same topic, but may also easily happen when many feature terms have multiple meanings. Therefore, each prime point can be described by a little number of key feature terms, and each prime point corresponds to a similar text cluster rather than just a random subset of the training set.

6. Experimental Results
The experimental process is elaborated as follows: By cutting the text after each word is considered as a neuron, numerous neurons are input to the neural network for optimization, and after processing in the middle layer of the network, the output gets the optimal feature subset, so that the purpose of dimensionality reduction is achieved. When all neurons have no more messages to be processed and at the same time each fact is used up, the computation of the neural network is considered to stop.

The specific steps are as follows:

![Figure 3. Neural network algorithm flow chart.](Image 202x117 to 392x329)
Step1: each word of the text content is a neuron, initialize the neuron fact base, communicate with the text content, and convert its information into facts to be stored in the fact base.

Step2: check whether there is a new message to be processed in the fact base, and if there is a new message in the message queue then call the message processing this function, check whether the existing facts match the rules of the above neural network computation algorithm, if they match then go to step 3; otherwise go to step 5.

Step3: activate this rule, each document as a layer, and execute a different neural network computation method computation for the i th neuron of each document according to the rule corresponding to the different syntactic analysis of the type of neural network computation.

Step4: deposit the computation results in the fact base in order of activation variable value X from largest to smallest, turn to the second step.

\[
E = \left(1 - \frac{N_r}{N}\right) \times 100\% \quad \text{(5)}
\]

\[
F = \frac{\text{Number of correctly attributed texts}}{\text{Total number of texts in the test set}} \times 100\% \quad \text{(6)}
\]

(where N is defined as the original feature dimension, N_r is the dimension of the feature vector after simplification); And also:

Step5: End of the algorithm.

The text used in the experiment is derived from the experimental test data. 1000 texts of 8 categories are selected, of which 600 are used as training samples and 400 are used as test texts, and the experiment is divided into three stages: feature vector acquisition, downsampling and correctness verification with an improved clustering algorithm. Meanwhile, in order to better reflect the effectiveness of the algorithm, we introduced two indexes, vector parsimony rate E and accuracy rate F, for evaluation.

In the text filtering stage, we performed the necessary pre-processing on the text, such as removing low-frequency words and deactivating words, and then conducted several experiments with the plain Bayesian classifier, of which eight experimental results are selected below, as shown in Table 1.

| No. | Total number of texts | Correct categorization | Original feature dimension | Number of feature dimension after simplification | Simplification rate | Accuracy rate |
|-----|-----------------------|------------------------|----------------------------|-----------------------------------------------|--------------------|---------------|
| 1   | 400                   | 380                    | 12000                      | 2780                                          | 66.64%             | 95.00%        |
| 2   | 400                   | 354                    | 12000                      | 2096                                          | 74.85%             | 88.5%         |
| 3   | 400                   | 347                    | 12000                      | 2969                                          | 75.26%             | 86.75%        |
| 4   | 400                   | 369                    | 12000                      | 2658                                          | 77.85%             | 92.25%        |
| 5   | 400                   | 375                    | 12000                      | 2926                                          | 75.62%             | 93.75%        |
| 6   | 400                   | 329                    | 12000                      | 2758                                          | 77.02%             | 82.85%        |
| 7   | 400                   | 381                    | 12000                      | 2893                                          | 75.91%             | 95.25%        |
| 8   | 400                   | 373                    | 12000                      | 2619                                          | 77.53%             | 93.25%        |

From the results, it can be seen that the clustering algorithm is applied to reduce the dimensionality of the text vector space of the cloud platform, which can effectively ensure the accuracy of the text classification of the urban rail cloud platform information on the basis of obtaining a high parsimony rate.

7. Future Work and Conclusion
This paper designs a clustering information text DR method based on the urban railway cloud platform, which can effectively filter redundant features and improve the quality of the final feature subset by clustering the features, so as to effectively extract useful data. Feature selection in information text
classification on cloud platform has been the key and bottleneck technology for text classification. At the same time, features with distinguishing ability can improve the efficiency and accuracy of the system, so the feature selection part of the cloud platform system is crucial. The dimensionality of the dimensionality reduction space has a great influence on the results of text processing, and how to determine a reasonable dimensionality reduction is also a problem worth studying in the future urban rail cloud platform.

Acknowledgment
This work was supported by Natural Science Foundation of Shandong Province (No.ZR2019LZH013). This work was supported by Industry and Information Technology Project of Shandong Province (No.202060102367). (Research on key technologies of smart city rail operation and maintenance System based on Internet of Things)
This work was supported by Natural Science Foundation of Shandong Province (No.ZR2020QE268). (Abp-iot technology in the basic application of subway tunnel fan)

References
[1] Wilk, A., Koc, W., Specht, C., Judek, S., Karwowski, K., Chrostowski, P., Czaplewski, K., Dabrowski, P.S., Grulkowski, S., and Licow, R.: ‘Digital Filtering of Railway Track Coordinates in Mobile Multi–Receiver GNSS Measurements’, Sensors, 2020, 20, (18), pp. 5018
[2] Ramkumar, A.S., and Poorna, B.: ‘Text document clustering using dimension reduction technique’, International Journal of Applied Engineering Research, 2016, 11, (7), pp. 4770-4774
[3] Yanhua, W., ZhiQiang, M., Zhengyang, Z., and Zhibo, C.: ‘Dual Preference Matrix Collaborative Filtering Algorithm and Its Application in Railway B2BE-commerce Platform’, in Editor (Ed.)’(Eds.): ‘Book Dual Preference Matrix Collaborative Filtering Algorithm and Its Application in Railway B2BE-commerce Platform’ (IEEE, 2020, edn.), pp. 562-569
[4] Zhen Yang, Fei Yao, Kefeng Fan, Jian Huang."Text Dimensionality Reduction with Mutual Information Preserving Mapping", Chinese15 Journal of Electronics, 2017
[5] Bo Zhang, Qing Li. "Text Mining-based Identification Model for Urban Rail Transit System Infrastructure Fault Analysis", 2019 CAA Symposium on Fault Detection, Supervision and Safety for Technical Processes (SAFEPROCESS), 2019