Application of Improved DBSCAN Algorithm in the Plan Compilation Management

Huaiguang Wu[1], Qing Lin[1], Baohua Jin[1] and Qi Liu[2]

*[1]School of Computer and Communication engineering,
Zhengzhou University of Light Industry, Zhengzhou 450000, China*

*[2]College of Computer and Software,
Nanjing University of Information Science and Technology, Nanjing 210044, China*

hgawu@126.com

Abstract

For further solving the problem of normative emergency plan, text mining should be combined with emergency plan compilation. Fixed ¨ threshold strategy is used by traditional DBSCAN, which would lead to the problem of cluster boundary wrong recognition. Improved DBSCAN algorithm is introduced in this paper. Least Square Fit is taken to fit plans similarity curve to find the best of initial ¨ threshold. According to the initial ¨, a new strategy is used to get dynamic ¨ threshold to improve the precision and recall. The simulations results show that the presented method is efficient for providing intelligent reference groups for government staff.

Keywords: *DBSCAN algorithm, Text similarity, Text management, Plan compilation*

1. Introduction

Clustering means the muster of similarity elements to mass data. Clustering analysis is a kind of statistical analysis method which is used to do research about classification problem of samples or index. Clustering analysis methods mainly include the division method, the hierarchical method, the method based on density, the method based on the grid and the method based on the model [1].

Clustering would not be stopped until the density of the number of objects in neighboring areas is more than a certain threshold. By this way, the method could be used to filter isolated data points and find out any shapes of category [2]. And these points are called noise points. The representative clustering algorithms based on density are DBSCAN algorithm, OPTICS algorithm, DENCLUE algorithm and so on. Among these algorithms, the DBSCAN algorithm is a typical case. The algorithm controls the increasing of category according to density threshold. The problem of generating reference groups during the process of plan compilation is described in this paper. After that, DBSCAN algorithm is improved to solve problem further and analyze experimental results.

Along with initiated emergency plans by governments of each level step by step, these plans are lack of standard exposed problems when faced with every kind of emergency. The process of plan compilation should be further supervised in order to consummate emergency plan mechanism. By making use of modern information technology, text mining technology and text clustering, plan reference groups would be generated and intelligent strategy would be provided to official plan makers. In this way, the reusable of plans would be raised.

At the same time, emergency plan should be emerged showed out as a kind of regulations and normative documents. Responsibility subject, range of performance and operating mechanism should be clearly defined in plans [3]. Although the actual condition is different in every area, scientific method and experience should be used for reference from plan at the same level during plan compilation. Based on this, this paper introduced the core problem: how to generate valid reference groups by clustering.

The first step of standardized plans is the introduction of the "frame theory" of artificial intelligence [4]. "Frame" is created to be encoded for each plan text. In general, plans include three partition methods. In China, the plans are divided into the type of disaster, natural calamities, public health and social security according to accident type. According to administrative division, the plans are also divided into the type of enterprise level, prefectural level and municipal level and so on. According to function, the plans are divided into the type of comprehensive plan, special plan and spot disposed plan [5]. Identified bits and note information of plan text are showed as the following Table 1.

Table 1. Identified Bits Information Table

Plan identified bits	Plan identified bits notes
1-2 bits	identified bits of accident classification, 00 is set as natural calamities, 01 is set as accident disasters, 10 is set as public health, 11 is set as social security
3-5 bits	identified bits of administrative division classification, 000 is set as national level, 001 is set as provincial level, 010 is set as municipal level, 011 is set as prefectural level, 100 is set as enterprise level
6-7 bits	identified bits of functional classification, 00 is set as special plan, 01 is set as comprehensive plan, 10 is set as disposal method
8-13 bits	flowing bits

2. Algorithm

Based on the above ideas, basic solution of the problem is set as followed. Corresponding frame information is read from identified bits. After that, N keywords are extracted. Therefore, N dimensionality space text set is formed by plans. Similarity of plans is taken as measure and plans are clustering analyzed based on density [6]. The flow figure of Specific intelligent plan compilation is showing in Figure 1.

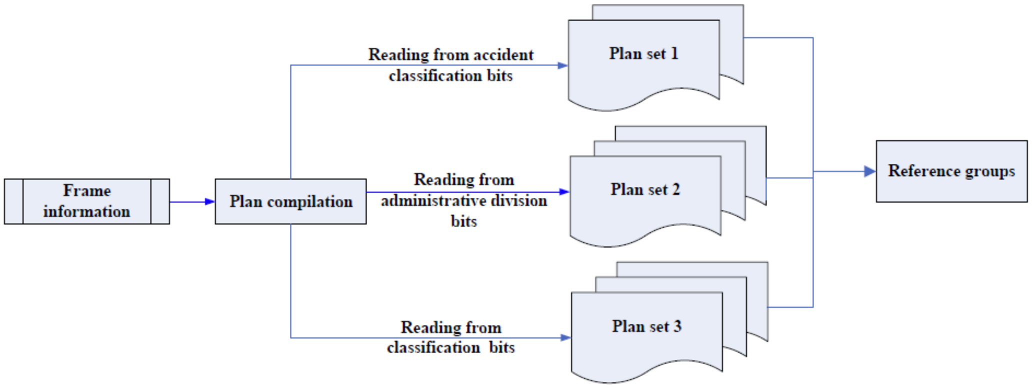
Frame information
Plan compilation
Reading from accident classification bits
Reading from administrative division bits
Reading from classification bits
Plan set 1
Plan set 2
Plan set 3
Reference groups

Figure 1. Flowing Chart of Plan Compilation

2.1. Similarity of Plans

Vector Space Model, which is usually used and turns out to be effective, is one of the IR (Information Retrieval) models. In the model of [7], plans are normalized showed as feature vector to form the whole plan space. The feature vector is showed as the equation (1).

$$V(d) = \sum_{i=0}^{n} (t_i, w_i(d)) \qquad (1)$$

In the equation, keywords are set as t_i. The weight of keyword is set as $w_i(d)$ to show how important the word could reflect theme. Every plan is composed of a group of independent words, which is showed as $d = \{t_1, t_2, \cdots, t_n\}$. w is regarded as the spot in N dimension coordinate system. The vector space of plan d is composed of orthogonal vectors which are got from $\{t_1, t_2, \cdots, t_n\}$. By this way of structured transformation, the handling of plans turns into the handling of keywords vectors. And the vectorial angle cosine is used to show the similarity of plans. $(w_{x1}, w_{x2}, w_{x3}, w_{x4} \cdots w_{xn})$ is used to show $plan_x$ and $(w_{y1}, w_{y2}, w_{y3}, w_{y4} \cdots w_{yn})$ is used to show $plan_y$. Therefore,$sim(plan_x, plan_y)$ is calculated by equation (2).

$$sim(plan_x, plan_y) = cos(plan_x, plan_y) = \frac{plan_x \cdot plan_y}{\| plan_x \| \cdot \| plan_y \|} = \frac{2\sum_{i=1}^{n} (w_{xi} \cdot w_{yi})}{\sqrt{\sum_{i=1}^{n} w_{xi}^2} \cdot \sqrt{\sum_{i=1}^{n} w_{yi}^2}} \qquad (2)$$

Based on this, if $(plan_1, plan_2, =cdots, plan_i)$is input, similarity inverted table is got and showed as $SimList(plan_i, plan_j) = \sum_{i=1, j=2, i<j}^{n} [Sim(Plan_i, Plan_j), Plan_i - Plan_j]$ .And the similarity inverted table between $Plan_i$ and other plans are showed as $SimList_i(plan_i, plan_j) = \sum_{j=1, i \neq j}^{n} [Sim(plan_i, plan_j), plan_i - plan_j]$,The similarity inverted table is listed from high to low by similarity. The output form is showed as $Sim(plan_i, plan_j), plan_i - plan_j$, the $Sim(plan_i, plan_j)$ is the similarity of $plan_i$ and $plan_j$, and the $plan_i - plan_j$ is the plan identified code.

2.2. Initialization of Neighborhood Threshold ε

The main title (on the first page) should begin 1 3/16 inches (7 picas) from the top edge of the page, centered, and in Times New Roman 14-point, boldface type. Capitalize the first letter of nouns, pronouns, verbs, adjectives, and adverbs; do not capitalize articles, coordinate conjunctions, or prepositions (unless the title begins with such a word). Please initially capitalize only the first word in other titles, including section titles and first, second, and third-order headings (for example, "Titles and headings" — as in these guidelines). Leave two blank lines after the title.

The advantage of DBSCAN algorithm is that the number of category is not needed. And the disadvantage of DBSCAN algorithm is that the global variable ε has great influence on clustering results. To initialize the value of ε, the first step is to build k-dist figure. When the k is given, the essence of k-dist is mapping $plan_x$ to the nearest distance from k_{th} point in N dimension space. Because the similarity represented the distance between plans, $SimList_i(plan_i, plan_j)$ could be used to get the solution of problems instance of the above step [8].

First of all, randomly a plan is picked as $plan_i$, its identified bits are read. $SimList_i(plan_i; plan_j)$ is gotten by identified bits. According to reference [9], the first concave point is the threshold point in k-dist figure in traditional DBSCAN algorithm. Here the k-dist figure is turned into similarity curve to find threshold ε.

Least Square Fit method is used to fit similarity curve, $(x_i; y_i) x \in [0; m]$ is given as data group. In $Span(\varphi_0(x); \varphi_1(x); \cdots; \varphi_n(x))$, there is a function which satisfies the condition $Sim(x) \in \Phi$ and $Sim(x) = \sum_{j=0}^{n} a_j \varphi_j(x); n < m$, and it makes the least value of the equation $\delta^2 = \sum_{i=0}^{m} Sim(x_i - y_i)^2$. Due to the similarity of the similarity distribution curve and hyperbolic curve, hyperbolic function $\varphi(x) = a_0 - \frac{a}{x}$ is selected. And linearly independent function $\varphi_0 = 1$, $\varphi_1 = -\frac{1}{x}$ is selected as primary function to solve a_0 and a_1.If the matrix notation is used and y, and G have the equations as followed:

$$y = [y_0; y_1; \cdots; y_m] \in R^{m+1} \tag{3}$$

$$a = [a_0; a_1]^T \in R^2 \tag{4}$$

$$G = \begin{bmatrix} \varphi_0(x_0) & \varphi_1(x_0) \\ \varphi_0(x_1) & \varphi_1(x_1) \\ \vdots & \vdots \\ \varphi_0(x_m) & \varphi_1(x_m) \end{bmatrix} \in R^2: \tag{5}$$

Therefore, δ^2 is turned into $(Ga - y)^T (Ga - y)$.

It could be verified that the necessary and sufficient conditions of a is the solution of the least squares is a satisfied a normal equations system $G^T Ga = G^T y$. And the normal equations system is showed as followed:

$$\begin{bmatrix} \varphi_0(x_0) & \cdots & \varphi_0(x_m) \\ \varphi_1(x_0) & \cdots & \varphi_1(x_m) \end{bmatrix} \begin{bmatrix} \varphi_0(x_0) & \varphi_1(x_0) \\ \varphi_0(x_1) & \varphi_1(x_1) \\ \vdots & \vdots \\ \varphi_0(x_m) & \varphi_1(x_m) \end{bmatrix} \begin{bmatrix} a_0 \\ a_1 \end{bmatrix} = \begin{bmatrix} \varphi_0(x_0) & \cdots & \varphi_0(x_m) \\ \varphi_1(x_0) & \cdots & \varphi_1(x_m) \end{bmatrix} \begin{bmatrix} y_0 \\ y_1 \\ \vdots \\ y_m \end{bmatrix} \tag{6}$$

Obviously, (6) has unique solution. And $Sim(x)$ curve could be fitted. To simplify the process of solution, although actual data is discrete, the similarity distribution curve is represented as continuous in Figure 2. The schema of the $Sim(x)$ and the similarity distribution curve is showed as Figure 2. When the σ is eligible, the method to find concave convex is used to find the first concave point in the similarity distribution curve. The inequation $(y_i - y_i')(y_{i+1} - y_{i+1}') \leq 0$, $y_i - y_i' \neq 0$ and $y_i - y_i' > 0$ are marked as formula (7). When formula (7) is as satisfied, y_{i+1} is the functional value as ε.

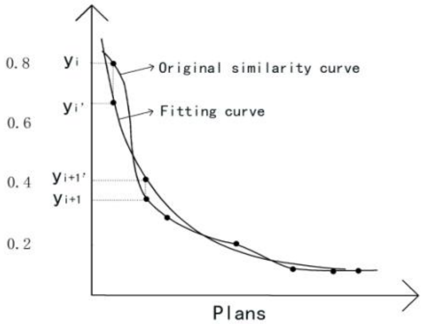

Figure 2. The Relationship Figure of Fitting Curve $Sim(x)$ and Original Similarity Curve

When the first point is concave point, the situation is not conforming to formula (7), which is showed as in Figure 3. The inequation $(y_1 ¡ y_1^0)(y_2 ¡ y_2^0) 6 0$, $y_1 ¡ y_1^0 6= 0$ and $y_1 ¡ y_1^0 > 0$ are marked as formula (8). In this case, formula (8) could be used to judge the situation of first point before formula (7).

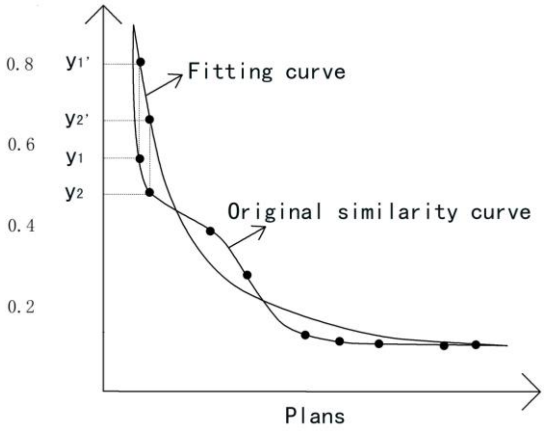

Figure 3. The Relationship Figure of Fitting Curve $Sim(x)$ and Original Similarity Curve (i=1)

2.3. Improved DBSCAN Algorithm

The overall variable ¨, Minpts should be initialized before using DBSCAN. The value of ´ has great influence on the clustering result. In general, the initialization of Minpts is relatively easy. Noise points would be disposed of boundary points by huger ¨. When Minpts is given, the way as Section 2.2 is used to initialize ¨.

In order to achieve this situation, DBSCAN algorithm is improved. Global ¨ is replaced by the adjustable ¨. The improved algorithm is shown as followed.

Step 1 ¨, Minpts and the interval of ¨ are given. And the category C is created and C is set to be a null set.

Step 2 the records of SimList are read. The first one of which number more than Minpts is selected to be marked visited. And its text identification code $Plan_i ¡ Plan_j$ is checked. If the $Plan_i$ is not included in C, the record would be thought to be seminal. $Plan_i$ is taken in the array of SeedList, and this record is marked visited.

Step 3 the front of the array SeedList is set to be $Plan_i$, and then array $SimList_i$ is had traversal. If the record of $SimList_i$ satisfies the condition that the similarity of $Plan_i$ is greater than Minpts and the similarity of other plans in C is greater than ´, $Plan_j$ would be taken in the array of SeedList. At the same time, the record of $Plan_i ¡ Plan_j$ of SimList would be marked visited.

Step 4 All the Plans in SeedList are put into C and deleted from SeedList.

Step 5 If the SeedList is empty, step 7 would be turned to. Among the equation, $¨_{i ¡ 1}$is the current radius threshold. Sim_{min} is the minimum value of similarity accord with the Minpts in current turn. Adjustable ¨ aims at the situation of uneven density.

Step 6 ¨ is recalculated. If the ´ is not out of range, step 2 would be turned to. If the ´ is out of range, step 7 would be turned to. And the ¨ could be calculated as $¨_i = (¨_{i¡1} + Sim_{min})=2¡ 0:1$.

Step 7 the clustering is ended. Its output form is $< Class_x; Plan_i; Plan_j; ¢¢¢; Plan_k >$. $Class_x$ is the identification code of the category. And $Plan_i$ is the corresponding text label of plan.

3. Experiment and Results Analysis

The corpus of this experiment is provided by Zhengzhou University of Light Industry R&D center (research and development center). According to the type of accident, plans are divided into four types which are natural calamity, accident disaster, public hygiene and social security. According to the administrative division, plans are divided into five types which are enterprise, prefecture, city, province and nation. According to the function of plans, plans are divided into four types which are special plan, comprehensive plan and disposal method.

3.1. Experiment Results

Plan compilation system is developed by JAVA. And this system is based on the pattern of B/S. Under the circumstance of 2.50 GHz CPU, 2.00 G memory, this system is tested. Tomcat server is used as background server and Oracle10g is used as database. 200 plans of different types are selected from corpus. The number of selected plans is shown as the following Table 2.

Table 2. Types Information Table

	Accidental classification/piece	Administrative classification/piece	Functional classification/piece
Natural calamities	73		
Accident disasters	62		
Public hygiene	24		
Social security	41		
National level		36	
Provincial level		44	
Municipal		54	
Prefectural level		25	
Enterprise level		41	
Special plan			72
Comprehensive plan			78
Disposal method			50
Total	200	200	200

According to the training set, the number of extractive keywords is set 10. And Minpts is set as 4 and the interval of ´ is [0.5,0.8]. The property value of “work summary” facet in plan Henan provincial forestry department forest fire special emergency plan is set to be needed to compile. And the type of accident classification is set as natural calamities. The type of administrative classification is set as provincial. The type of functional type is special plan.

DBSCAN algorithm and improved DBSCSAN algorithm are applied to the testing set. The results are shown as Table 3.

Table 3. Experimental Results Table

	N_{DBSCAN}	$N_{MDBSCAN}$	M_{DBSCAN}	$M_{MDBSCAN}$	C_{DBSCAN}	C_{DBSCAN}
Natural calamity	8	9	66	63	41	43
Provincial level	5	5	40	42	27	29
Special plan	7	7	68	60	44	45

Among Table 3, the number of reference plan group which are given by DBSCAN is N_{DBSCAN}. The number of reference plan group which are given by improved DBSCAN is $N_{MDBSCAN}$. The number of plans which are clustered by DBSCAN is C_{DBSCAN}. The number of plans which are clustered by improved DBSCAN is $C_{MDBSCAN}$.

3.2. Algorithm Analysis

Quality evaluation method of text clustering includes precision, call, entropy, fuzzy matrix and classification accuracy. In this paper, precision P and recall R are used as evaluation criterion of improved DBSCAN algorithm. The degree of merging similar text units and un-similar text units is represented by P and P reflects the evaluative criteria to different theme. The precision of clustering is higher, the contents are more gathered. The degree of merging similar text units into a category is represented by R and R reflects the recognition capability of same theme. The recall is higher, the similar units are more gathered, the occurrence of dividing similar units into different categories is less. P and R are united and taken measure as standard of evaluating clustering results in reference [10]. Therefore, P and R is defined as followed:

$$P = N_c = N_m \quad R = N_c = N_t$$

Among the formula, the number of plans which attributes to correct category is presented as N_c. The number of actual classified plans is presented as N_m. The number of plans which includes noise points in corresponding category.

By calculating the P and R of DBSCAN algorithm and improved algorithm in transverse classification, the average $P_{MDBSCAN}$ and $R_{MDBSCAN}$ is respectively 71.92%与 62.43%. The percentage of its results respectively is 8.21% and 2.22% higher than the results which applied DBSCAN algorithm. The results turn out improved algorithm is more suitable than traditional DBSCAN algorithm in the application of plan compilation.

● k-dist figure is established during the process of clustering. The first step is initialization of ¨. After areas of set is inquired repeatedly to realize search from one core point to every density-reachable points. Therefore, R3-tree is established to return all the objects in given area. The cost of above steps is a lot of memory and IO. Actually step1 is more vital for reason of that the realization of step 1 is the bottleneck of efficiency of time and space. As to the plan compilation system is based on the pattern of B/S. Ajax is used to read corresponding data table to SimList when the page of plan frame is pre-read. Data fitting is used to fit similarity curve to confirm the initialization of ¨ instead of the need of traditional way of establishment of k-dist figure. By this way, the efficiency of this system is improved.

- The result of clustering turns out that more noisy points are relatively removed under the circumstance of applied improved DBSCAN. And global ¨ is replaced with adjustable ˂ to ease the problem of wrong identification of boundary point. At the same time, the number of generating of category is more than traditional way. Due to ¨ is adjusted by the search results in last round. It is good for avoiding the problem of wrong merging categories as to smallness of margins. And the clustering effect is more apparent.

4. Conclusions

DBSCAN algorithm has the capacity of finding out any shape of category. Aiming at the specific problem of generating reference groups in plan compilation, DBSCAN algorithm is improved and expanded to improve the performance. Compare the effect of transverse reference groups which are generated by DBSCAN algorithm and improved DBSCAN algorithm. The results turn out improved DBSCAN algorithm is suitable applied to solve the problem. And it is valuable to provide intelligent scheme in plan compilation.

Acknowledgements

This work is supported by the National High Technology Development 863 Program of China (No. 2007AA01Z185) and doctorate foundation of Zhengzhou University of Light Industry (No.201203008).

References

[1] L. Zhiyong, "Text Mining Algorithm Based on Fuzzy Clustering", Computer Engineering, vol. 5, no. 35, **(2009)**.

[2] Z. Enla and H. Wenning, "Improved Track Clustering Algorithm Based on Density", Computer Engineering, vol. 5, no. 37, **(2011)**.

[3] L. Tiemin, "Emergency plan compilation guidelines", Labour Protection, vol. 2, no. 4, **(2004)**.

[4] S. Russel, "AI——a modern method", Posts and Telecom press, Beijijng, (**2010)**.

[5] L. Tiemin, "Design of the emergency plan system s concept", Journal of Safety Science and Technology, vol. 8, no. 7, **(2011)**.

[6] X. Wang, "Talking Simply about Several Methods of the Chinese Language Automatic Word Segmentation", Value Engineering, vol. 11, no. 13, **(2011)**.

[7] X. Li, "Research of the Text Subjective Question s Auto Remarking Algorithm Based on Word Segmentation Algorithm VSM", Computer Knowledge and Technology, vol. 7, no. 25, **(2011)**.

[8] Y. Yu and A. Zhou, "An Improved Algorithm of DBSCAN", Computer Technology and Development, vol. 21, no. 2, **(2011)**.

[9] Y. Cai and J. Yuan, "Text Clustering Based on Improved DBSCAN Algorithm", Computer Engineering, vol. 37, no. 12, **(2011)**.

[10] H. Suo and Y. Wang, "Reference-based k-means algorithm for document clustering", Computer Engineering and Design, vol. 30, no. 2, **(2009)**.

Authors

Wu Huaiguang

Wu Huaiguang received PhD degrees in computing from Wuhan University in 2011. Since 2011 he has been in the School of Computer and Communication Engineering at Zhengzhou University of Light Industry. His research interests include formal methods, software engineering, and algorithms.

Lin Qing

Lin Qing is a graduate student of School of Computer and Communication Engineering at Zhengzhou University of Light Industry. Her research interests include artificial intelligence, computer decision support system, computer software and theory and emergency management.

Jin Baohua

Jin Baohua is an associate professor of School of Computer and Communication Engineering at Zhengzhou University of Light Industry. His research interests include artificial intelligence, computer decision support system, computer software and theory and emergency management.

Qi Liu

BSc (2002), MSc (2006), PhD (2010) is an appointed Professor at the Nanjing University of Information Science and Technology. His research interests include context-awareness, data communication in MANET and WSN, smart home and energy efficiency. He also devotes time to WSN solutions on intelligent agriculture in the protected field. He is IEEE and ACM member.