Hindawi
Journal of Advanced Transportation
Volume 2023, Article ID 2570824, 13 pages
https://doi.org/10.1155/2023/2570824

Research Article

A Function Area Division Approach for Autonomous Transportation System Based on Text Similarity

Ke Huang[1], **Caiting Chen**[2], **Yao Xiao**[1] and **Ming Cai**[1]

[1]*School of Intelligent Systems Engineering, Shenzhen Campus of Sun Yat-sen University, Shenzhen 518107, China*
[2]*Urban Mobility Institute, Tongji University, Shanghai 201804, China*

Correspondence should be addressed to Ming Cai; caiming@mail.sysu.edu.cn

Received 19 August 2022; Revised 27 October 2022; Accepted 21 March 2023; Published 2 June 2023

Academic Editor: Yanyong Guo

Along with emerging technologies and increasing demands, autonomation has become a significant trend in current transportation systems. Within this context, the autonomous transportation system (ATS) framework hinges on functions that serve as fundamental units to support its operation. Recognizing the divisions among these function areas can enhance our understanding of their meanings and interrelationships. This study introduces a method for dividing function areas within the ATS framework, grounded in text similarity, to mitigate reliance on subjective experience. Precisely, this method quantifies the similarity between functions based on their textual descriptions, and implements hierarchical clustering to delineate them into distinct function areas. To validate the effectiveness of this proposed method, a case study analyzing a vehicle automatic driving scenario was conducted. The results demonstrate that our approach can efficiently divide function areas, producing clustering outcomes that possess superior accuracy and purity when juxtaposed with reference classifications. Consequently, this method has the potential to facilitate the formulation of function areas within ATS, thereby supporting the autonomous operation and construction of ATS. Moreover, its applicability extends beyond ATS, showing promise for other clustering problems that involve multiple texts, such as in text classification.

1. Introduction

Along with emerging technologies and increasing demands, autonomation has become a significant trend in current transportation systems. Against this background, the concept of autonomous transportation system (ATS) emerged [1], aiming to realize autonomous perception, autonomous learning, autonomous decision-making, and autonomous action for transportation systems.

The construction of the ATS depends on the guidance of the architecture framework, and function area division is an essential part of the ATS framework's research. Related concepts for the task can be explained as follows. The services are the applications and values that the system can provide for users. For example, a transportation system can provide users with services of "parking space management," "freight administration," "vehicle emergency response," etc. The functions are the processes and activities used to support services. For example, the realization of the "parking space management" service relies on functions such as "get personal driver request," "process vehicle location data," "determine dynamic parking lot state," and "output parking lot information to drivers." The function areas are the sets of functions with common data processing characteristics and application scope. The function area division contributes to organizing the functions of transportation systems and sorting out intra-area correlation and interarea coordination. Furthermore, it benefits in determining key modules of ATS development.

As the basis of the ATS framework's research, the traditional intelligent transportation system (ITS) frameworks have more than 20 years of history, including typical frameworks of the United States, the European Union, and China. The research on the ITS framework of the United States started first and has been continuously improved since 1993. The latest version 9.0 [2] was released in

2020 to adapt to the transportation reform for automatic driving. The ITS framework of the European Union has been studied since the 1990s, and it was updated to version 4.1 in 2011 [3]. The ITS framework of China was studied in the early 21st century. It has not been updated and developed since the completion of version 2.0 in 2005 [4]. The three ITS frameworks have affected the development of other countries' and cities' ITS frameworks [5–8], and the ITS frameworks are evaluated by some researchers [9–11]. Table 1 lists the function areas of ITS frameworks in the United States, the Europe Union, and China, and the contents of the three are similar. However, the three ITS frameworks do not explain and demonstrate the division logic of function areas, which often relies on expert experience. Jiang et al. [12] once tried to use rough sets to identify new function areas of the ITS framework, which was a preliminary study of function area division methods. Still, there have been no further or similar research and applications.

Based on the traditional ITS frameworks, many explorations of the new ITS frameworks and improved ITS subsystems have been carried out recently. Especially, with the rapid development of Internet of Vehicles technology, some related ITS frameworks appeared, such as the vehicle-to-vehicle-to-infrastructure framework [13], the vehicular ad hoc network (VANET) communication architecture [14], and the VANET architecture assisted by unmanned aerial vehicles [15]. In addition, some studies were devoted to improving subsystems of the ITS frameworks, including public transportation systems [16, 17], vehicle tracking systems [18, 19], ITS security systems [20–22], ITS information systems [23, 24], and ITS communication systems [25, 26]. However, most current research on the ITS frameworks combines limited emerging technologies or only focuses on the ITS subsystems. There are few in-depth and detailed types of studies similar to the three ITS frameworks mentioned previously.

Regarding the function area division, the new ITS frameworks' research has updated the functions' content. Still, they mostly rely on subjective construction methods and have not yet formed a clear and complete methodology. As the transportation systems become more complex and the functions become more abundant, an adaptive function area division method is urgent to adapt to the dynamic evolution of the transportation systems. Actually, each function has some short texts that embody function characteristics, which can be used to cluster functions with commonality into the same function areas. Therefore, dividing function area can be regarded as clustering the function texts here, and this task is generally called text clustering. Text clustering is to group similar texts from a set of texts [27] and has many useful algorithms, such as hierarchical clustering [28], k-means clustering [29], eigenspace-based fuzzy c-means clustering [30], and deep embedding [31]. In most text clustering algorithms, text similarity is a necessary step. The text similarity approach can measure the commonality between two texts, which is often used in text clustering [32], text classification [33], information retrieval [34], and document matching [35]. In product-service systems and web service discovery, text similarity is applied in service clustering research [36, 37], which is similar to our research problem. Inspired by their works, text similarity is also adopted in establishing the ATS function area division method.

In this paper, an ATS function area division method is proposed based on text similarity, transforming the function area division task into a short text clustering task. Based on the method, the function similarity is measured by texts, and the functions are clustered by hierarchical clustering. Therefore, the method can adaptively form function areas, helping reduce the dependence on subjective experience and improve the efficiency of function area division work.

This paper consists of five sections. Section 2 describes the steps of the function area division methods, including function text processing, function similarity calculating, function area dividing, and method performance evaluating. In Section 3, a case analysis for a vehicle automatic driving scenario is provided, and the method performance is evaluated. Finally, Section 4 concludes all the work and discusses possible future improvements.

2. Methods of Function Area Division

The function area division will cluster similar functions based on considering various aspects of the transportation system operation. At last, the functions within the same function area have high similarity degrees, and the functions among different function areas have low similarity degrees, as shown in Figure 1.

The technical route for function area division is shown in Figure 2.

The steps are as follows:

(1) Function text processing: The text of the function name is converted into a phrase set by phrase segmenting and stops word removal. The texts of three function attributes, including "function provider," "process object," and "service object," are converted into word sets directly.

(2) Function similarity calculating: The function name similarity and function attribute similarity are calculated based on the Jaccard coefficient. Then, the comprehensive similarity matrix of functions is obtained with a weighted average of the two similarities.

(3) Function area dividing: The final function area result is obtained by hierarchical clustering and the silhouette coefficient. The function areas are named based on analyzing the keywords of the function texts. In addition, the functions in every function area are reclassified with the "operation stage" attribute, and a three-level function list is finally formed.

(4) Method performance evaluating: A function set of a vehicle automatic driving scenario is constructed,

Table 1: Function areas of ITS frameworks in the United States, the Europe Union, and China.

Countries	Function areas
United States	(1) Manage traffic (2) Manage commercial vehicles (3) Provide vehicle monitoring and control (4) Manage transit (5) Manage emergency services (6) Provide driver and traveler services (7) Provide electronic payment services (8) Manage archived data (9) Manage maintenance and construction (10) Support secure transportation services
European Union	(1) Provide electronic payment facilities (2) Provide safety and emergency facilities (3) Manage traffic (4) Manage public transport operations (5) Provide support for host vehicle services (6) Provide traveler journey assistance (7) Provide support for law enforcement (8) Manage freight and fleet operations (9) Provide support for cooperative systems
China	(1) Manage traffic (2) Provide electronic payment services (3) Provide traffic information services (4) Manage emergency rescue (5) Manage passenger transportation (6) Manage freight (7) Manage urban public transport (8) Support intelligent highway and safety driving assist (9) Manage transportation infrastructures (10) Manage ITS data

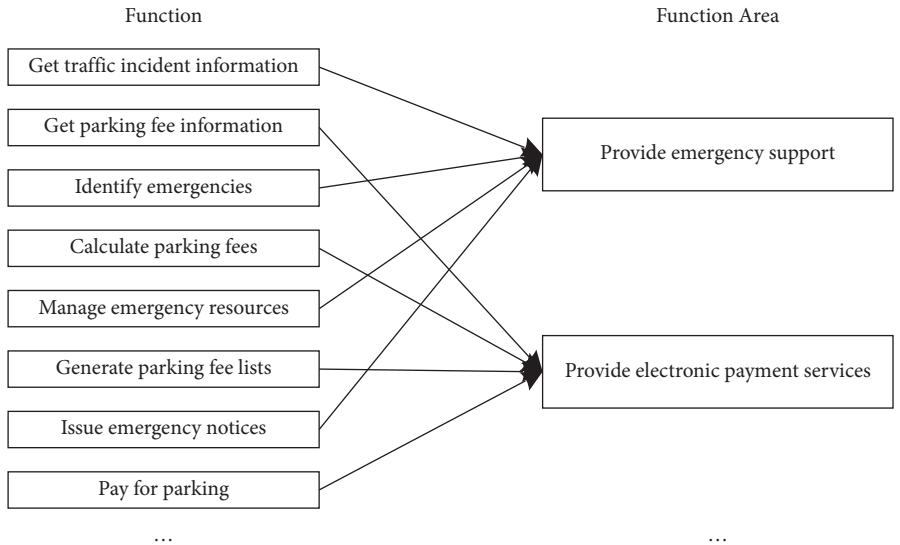

Figure 1: Schematic diagram of ATS function area division.

and the function areas are divided manually as reference classification. Based on the reference classification, the method performance is evaluated with accuracy and purity.

2.1. Function Text Processing. The text formats and processing of the ATS functions are introduced. In the ATS framework, a function is described by the function name and four attributes. The texts of the function name and three

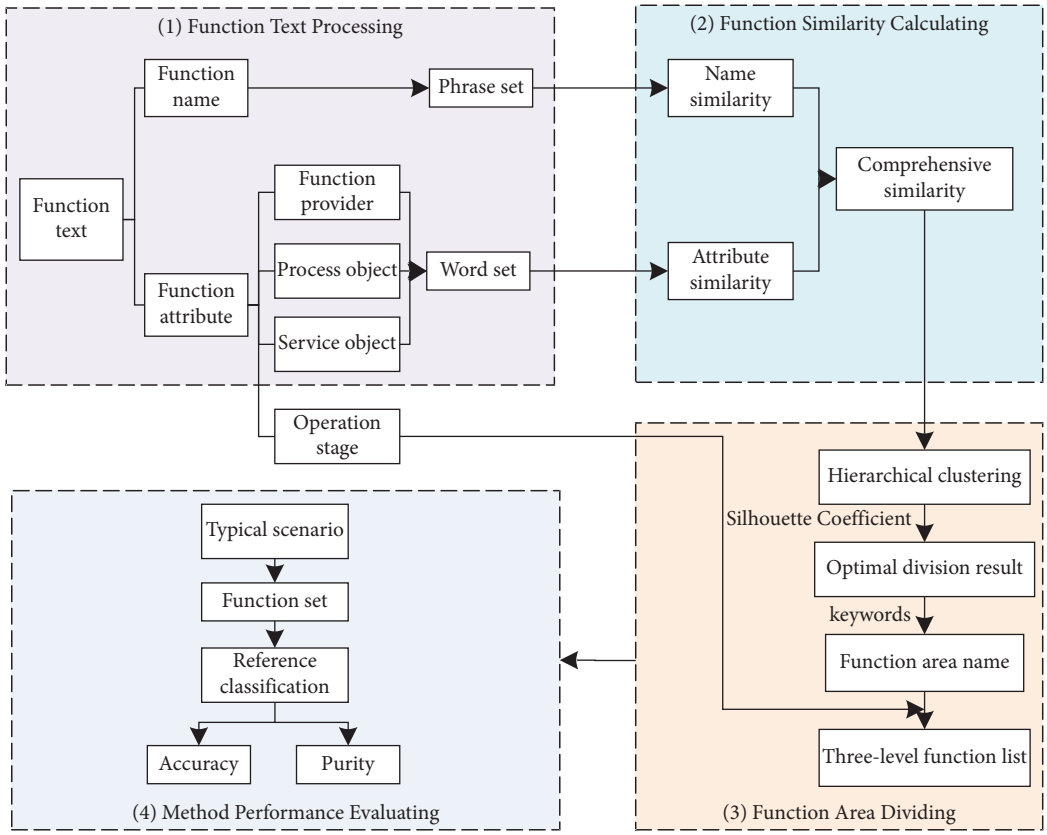

Figure 2: Technical route for function area division.

function attributes, including "function provider," "process object," and "service object," are processed for clustering functions. It should be noted that the function attribute "operation stage" is not used for clustering functions.

2.1.1. Function Name. The function name can summarize the function content and is a short Chinese text composed of "verb + noun or noun phrase," such as "dispatch emergency vehicles." Besides, extra information can be supplemented by parentheses, such as "provide traffic information query (traveler interface)."

Some auxiliary words, such as empty words and conjunctions, may exist in the function name text. Therefore, meaningful phrases must be extracted from the function name text for subsequent similarity calculation. The process of the function name text can be divided into the following two steps:

(1) Phrase segmenting: This work uses jieba, a mainstream Chinese phrase segmentation tool, to segment the function name texts into independent phrases, such as "monitor/passenger/anomaly/behavior," and each phrase is no more than three words.

(2) Stop word removal: In this work, stop words refer to meaningless symbols and redundant words. There are mainly two types: one is empty words, conjunctions, or other independent words after segmentation, and symbols, such as parentheses. Another is the verbs at the beginning of the text, such as "collect" and "process," which provide little help in distinguishing the function areas.

2.1.2. Function Attribute. The function attributes can describe the functional characteristics and embody the autonomous operation logic. Every function has four attributes:

(1) "Function provider": physical objects that provide the functions, including "user body," "system module," and "integration platform," such as vehicle-mounted equipment, infrastructures, and information platforms.

(2) "Process object": information objects used in the process of function realization, such as road network information and emergency events.

(3) "Service object": "user body" that can directly use functions, or "system module" and "integration platform" that directly use output results of functions, such as travelers, vehicle-mounted equipment, and information platforms.

(4) "Operation stage": system operation stages that can reflect the autonomous operation logic in the functions, including four stages of "perception," "learning," "decision-making," and "action".

The function attribute values are the specific contents of four function attributes, and the definition of all function attribute values can be found on the ATS website [38]. This research only studies functions whose function attributes have a single value.

The three function attributes, "function provider," "process object," and "service object," are used for clustering functions. The texts of their function attribute values are relatively simple, composed of short noun phrases without redundant components. Therefore, they can be directly converted to a word set for similarity calculation.

2.2. Function Similarity Calculating. The similarities of function names and function attributes are calculated first. Then, the comprehensive similarities between functions are calculated based on the two kinds of similarities.

The Jaccard coefficient can be used to calculate text similarity by measuring the overlap degree of phrases or words of two texts [37, 39]. Thus, the similarities of function names and attributes are measured based on the Jaccard coefficient.

2.2.1. Function Name Similarity. The function name similarity is calculated based on the phrase sets of two functions' names with the Jaccard coefficient:

$$r_{ij}^{(s)} = \frac{\left|W(s_i) \cap W(s_j)\right|}{\left|W(s_i) \cup W(s_j)\right|}, \tag{1}$$

where $r_{ij}^{(s)}$ is the similarity between the function i and the function j about the function name, and $W(s_i)$ is the phrase set of the function i about the function name.

2.2.2. Function Attribute Similarity. The function attribute similarity is calculated based on the word sets of two functions' attributes with the Jaccard coefficient:

$$r_{ij}^{(x_n)} = \frac{\left|C\left(t_i^{(x_n)}\right) \cap C\left(t_j^{(x_n)}\right)\right|}{\left|C\left(t_i^{(x_n)}\right) \cup C\left(t_j^{(x_n)}\right)\right|}, \tag{2}$$

where $r_{ij}^{(x_n)}$ is the similarity between the function i and the function j about function attribute x_n, and $C(t_i^{(x_n)})$ is the word set of the function i about function attribute x_n.

2.2.3. Comprehensive Similarity between Functions. The comprehensive similarity between two functions is obtained with a weighted average of the similarities of function names and function attributes:

$$r_{ij} = w_1 r_{ij}^{(s)} + w_2 r_{ij}^{(x_1)} + w_3 r_{ij}^{(x_2)} + w_4 r_{ij}^{(x_3)}, \tag{3}$$

where r_{ij} is the comprehensive similarity between the function i and the function j, $r_{ij}^{(x_1)}$, $r_{ij}^{(x_2)}$, and $r_{ij}^{(x_3)}$ are similarities of three function attributes: "function provider," "process object," and "service object." w_k is the weight of the k_{th} similarity, and $w_1 + w_2 + w_3 + w_4 = 1$.

The comprehensive similarity matrix of functions is

$$\mathbf{R} = \begin{bmatrix} r_{11} & r_{12} & \cdots & r_{1n} \\ r_{21} & r_{22} & \cdots & r_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ r_{n1} & r_{n2} & \cdots & r_{nn} \end{bmatrix}, \tag{4}$$

where $r_{ii} = 1$, $r_{ij} = r_{ji}$, and $0 \le r_{ij} \le 1$.

2.3. Function Area Dividing. All clustering results are obtained by aggregative hierarchical clustering, and the optimal clustering result is found by silhouette coefficient, which is the final function area result. Then, the function areas are named by analyzing the keywords of the function texts. At last, the functions in every function area are reclassified by the "operation stage" attribute, and a three-level function list is formed.

2.3.1. Function Clustering. Hierarchical clustering is a simple and effective unsupervised clustering method, which only needs the distances between samples. Specifically, this research adopts an agglomerative hierarchical clustering algorithm to cluster functions into function areas. The basic idea is to regard each function as an initial function area at the beginning and then continuously merge the two closest function areas until all functions are merged into one function area. The flow of function clustering by utilizing the agglomerative hierarchical clustering algorithm is shown in Figure 3.

In function clustering, the comprehensive similarity matrix should first be transformed into the comprehensive distance matrix, which is used as the initial distance between function areas. Then, during iteration, the distance between function areas is updated using the average sample distance between function areas:

$$d_{\text{avg}}(E_s, E_t) = \frac{1}{|E_s||E_t|} \sum_{i \in E_s} \sum_{j \in E_t} \text{dist}(i, j), \tag{5}$$

where $d_{\text{avg}}(E_s, E_t)$ is the average sample distance between function area E_s and function area E_t, and $\text{dist}(i, j)$ is the distance between function i and function j.

2.3.2. Optimal Division Result Determining. The hierarchical clustering can obtain all possible clustering results, whereas the optimal division result is what we need. Therefore, clustering performance should be evaluated to determine the optimal division result as the final function area division results.

The clustering performance is usually evaluated with external and internal indicators [40]. The external indicator compares the clustering result to the manual division result, whereas the internal indicator evaluates the clustering result directly. Compared with the external indicator, the internal

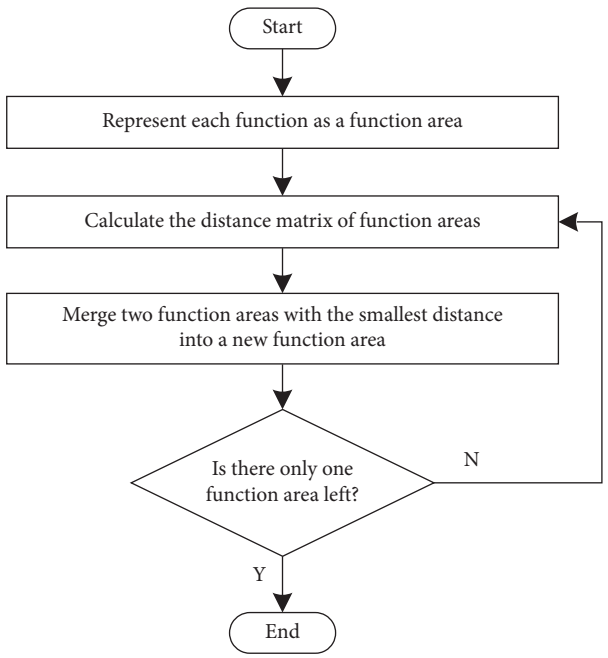

Figure 3: Flow of function clustering by the agglomerative hierarchical clustering algorithm.

indicator is more easily obtained and more adapted to the function area division work in different scenarios. Therefore, the internal indicator is selected to determine the optimal division result.

Silhouette coefficient [41] is a common internal indicator that considers both cluster cohesion and separation. Besides, it does not need to calculate the cluster center coordinates, so it is suitable for the clustering performance evaluation of our work with only sample similarity. The value of the silhouette coefficient ranges from −1 to 1, and a high value indicates that the intra-area similarity is high and the interarea similarity is low, representing a good clustering performance. Therefore, the clustering result with the maximum silhouette coefficient is the optimal division result.

Based on the function area division in the ITS framework, the maximum number of function areas can be set to 15. In addition, one function area that includes all functions is meaningless, so the minimum number of function areas is set to 2. The optimal division result is obtained by evaluating the silhouette coefficients of cluster numbers ranging from 2 to 15.

2.3.3. Function Area Naming. After obtaining the optimal division results, the function areas are named based on manual work because the number of the function area is no more than 15. For each function area, the three phrases with the highest frequency in the function name and the attribute value with the highest frequency in each function attribute are extracted as the keywords to provide references for naming manually.

The keywords and candidate names of function areas are concluded based on the function area research for the ITS framework, as shown in Table 2. The appropriate names can be directly selected according to the keywords.

2.3.4. Three-Level Function List Generating. Each function area has many functions, which is poor to display functions clearly, so the functions in all function areas are further categorized.

In each function area, the "operation stage" function attribute is used to reclassify the functions into "perception function," "learning function," "decision-making function," and "action function." As a result, a three-level function list is generated, whose structure is shown in Figure 4.

2.4. Method Performance Evaluating. The method performance of function area division needs to be evaluated. Thus, we construct a vehicle automatic driving scenario and obtain a function set of this scenario. Then, the function areas are divided manually as reference classification, and the method performance is evaluated by comparing the clustering result with the reference classification.

Two external indicators are used to assess the clustering performance compared with the reference classification: accuracy and purity [42].

2.4.1. Accuracy. The accuracy is the ratio of functions divided correctly. It indicates that the coherence between the clustering result and reference classification, and a value close to 1 represents the clustering performance is good. The accuracy is calculated with

$$\mathrm{ACC} = \frac{\left|F_{\mathrm{succ}}\right|}{|F|}, \tag{6}$$

where ACC is the accuracy, $|F_{\mathrm{succ}}|$ is the number of functions divided correctly, and $|F|$ is the total number of functions.

2.4.2. Purity. Since function distribution is not uniform in specific scenes, only the accuracy cannot reflect the clustering performance well. Therefore, the purity is introduced to help evaluate the clustering performance. It indicates the precision of clustering, and a value close to 1 represents the clustering performance is good.

The purity of each function area is calculated with

$$\mathrm{PUR}\left(C_k\right) = \frac{1}{\left|C_k\right|}\max\ \left(\left|C_k^s\right|\right), \tag{7}$$

where $\mathrm{PUR}(C_k)$ is the purity of the k_{th} function area, $|C_k|$ is the number of functions in the k_{th} function area, and $|C_k^s|$ is the number of functions belonging to the s_{th} function area of the reference classification.

The purity of the clustering result is calculated by the weighted average of the purities of all function areas

$$\mathrm{PUR} = \sum_{k\epsilon C}\frac{\left|C_k\right|}{|F|}\mathrm{PUR}\left(C_k\right), \tag{8}$$

Table 2: Keywords and candidate names of function areas.

Keywords	Candidate names
Vehicle	Provide vehicle control and safety Provide vehicle safety and aided driving
Traveler	Provide travel services
Nonmotorized traffic	Manage pedestrian and nonmotor vehicle Manage pedestrian and nonmotor vehicle safety
Weather environment	Monitor and manage environment
Road traffic	Manage traffic network Provide traffic management and planning
Freight	Manage traffic and freight
Bus operation, parking lot	Manage traffic operation Provide parking services Manage service facilities
Fee	Provide electronic payment services Provide electronic charging services
Infrastructure	Manage traffic facilities Manage infrastructures
Communication	Support communication
Emergency event	Provide emergency support Provide emergency rescue Provide emergency management/response
Safety	Provide traffic safety Provide vehicle safety Provide pedestrian and nonmotor vehicle safety Provide public safety
Maintenance	Manage maintenance and construction
Data	Manage data Manage information service

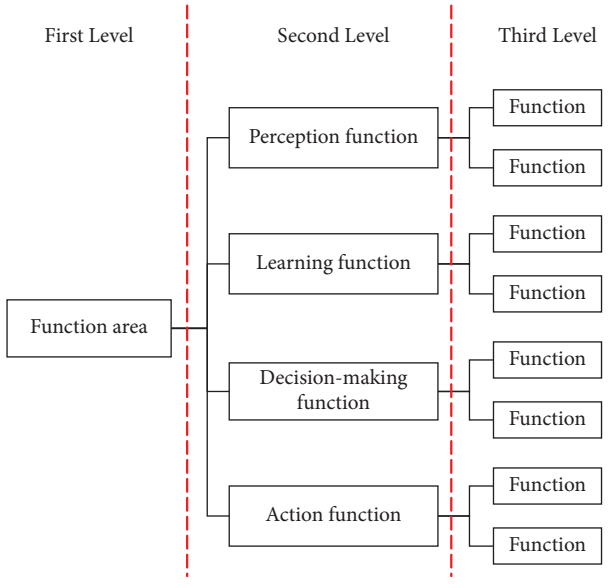

Figure 4: Structure of three-level function list.

where PUR is the purity of the clustering result, and $\mathrm{PUR}(C_k)$ is the purity of the k_{th} function area.

3. Scenario Verification and Result Discussion

A typical autonomous transportation scenario is constructed to verify the performance of the function area division method. First, a function set supporting this scenario is constructed. Then, the function areas are obtained by using the method. At last, the performance of the method is verified by comparing the clustering result with the reference classification through two external indicators: accuracy and purity.

3.1. Scenario Hypothesis and Function Set Construction. Vehicle automatic driving is a typical autonomous transportation scenario. In this scenario, we set an event that users want to drive from their homes to their workplaces on city roads. Before the travel, the users need to obtain travel information and plan routes. During travel, users need to drive cars and have safety requirements. Besides, they need to park their cars and pay traffic costs when arriving at the destination. Finally, after the travel, the system needs to provide travel evaluation and feedback to improve travel quality.

The completion of this event requires that the ATS framework has corresponding functions. As a result, a function set containing 174 ATS functions is constructed based on the autonomous operation logic (see Appendix I in Supplementary Materials (available here)), and the detailed contents of each function can be found in [38]. By referencing the traditional ITS framework, the 174 ATS functions are manually divided into 9 function areas as the reference classification (see Appendix I in Supplementary Materials (available here)). The function distribution of these function areas is shown in Table 3.

3.2. Function Area Division Result. Based on the method of Section 2, the functions can be clustered into function areas. The weights in (3) are set to $w_1 = 0.6$, $w_2 = 0.05$, $w_3 = 0.2$, and $w_4 = 0.15$, which are the optimal values by experiments.

The silhouette coefficients of cluster numbers ranging from 2 to 15 are calculated to judge the optimal cluster number, as shown in Figure 5. With the cluster number increasing, the silhouette coefficient increases first and then decreases. It reaches a maximum of 0.1989 when the cluster number is nine. Thus, the optimal cluster number is nine, the same as the function area number of the reference classification.

When the cluster number is nine, the accuracy and purity are both 0.8966, indicating that the function area division method can work well for function clustering.

Then, the keywords of the function name and three function attributes are extracted according to word frequency and combined with the candidate names in Table 2 to

Table 3: Function distribution of reference classification.

Function area	Autonomous operation logic				Function number
	Perception	Learning	Decision-making	Action	
Provide vehicle control and safety	14	14	13	13	54
Provide travel services	3	5	3	7	18
Provide electronic payment services	2	3	3	5	13
Monitor and manage environment	1	3	0	1	5
Manage traffic network	9	12	5	7	33
Manage traffic facilities	0	0	2	2	4
Support communication	0	0	0	2	2
Provide emergency support	7	7	6	9	29
Provide parking services	6	4	3	3	16
Total	42	48	35	49	174

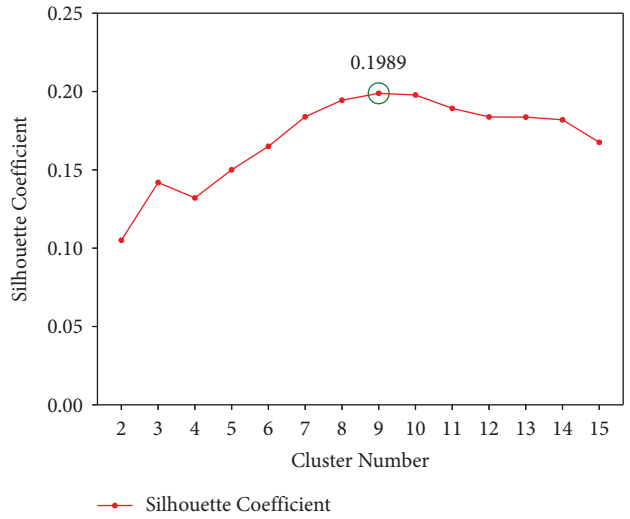

Figure 5: Silhouette coefficients of different cluster numbers.

name each function area manually. The keywords and recommended names of nine function areas obtained by clustering are shown in Table 4. Finally, the three-level function list is obtained based on the function attribute "operation stage" (see Appendix I in Supplementary Materials (available here)).

3.3. Analysis of Experiment Results. The effect of different text similarity weights and the clustering performance of optimal division results will be analyzed.

3.3.1. Effect Analysis of Text Similarity Weights. In (3), four weights need to be set when calculating the comprehensive similarity of functions. To explore the influence of weight values on clustering performance, we conduct the clustering experiments by changing weight values with a step size of 0.05 every time. The weight values and clustering performance of the top ten experiments by ranking the accuracy are shown in Table 5.

The accuracy and purity of the first seven experiments are the same, and most of the seven have nine function areas consistent with the reference classification. The last three experiments have more function areas than the reference classification, showing that the functions are divided into more details, resulting in higher purity and lower accuracy. The clustering result with high accuracy should be selected first because it is closer to the reference classification.

The relative relation of the four weights in Table 5 is $w_1 \geq w_3 > w_4 \geq w_2$, that is, the weights of the function name and "process object" are relatively large, while the weights of "function provider" and "service object" are relatively small. Specifically, the function name has the largest weight and $w_1 \geq 0.3$. It indicates that the semantic text information of the function name is rich to help most for function clustering. Among the three function attributes, the "process object" is the most important, the "function provider" is the least significant, and the "service object" plays a complementary role.

In Table 5, some experiments have different weights but the same clustering performance. In particular, experiment 1 and experiment 2 obtain different numbers of function areas, but their accuracies and purities are the same. The function distributions of the two experiments are shown in Figure 6, and the clustering results are close to the reference classification. However, the result of experiment 2 lacks the function area of "manage traffic facilities," and functions of this function area are mistakenly divided into the function area of "manage traffic network." In addition, some functions are incorrectly divided into the "monitor and manage environment" function area in experiment 1, while these functions are divided into correct function areas in experiment 2. Although experiment 1 and experiment 2 have different focuses on function clustering results, their correct function numbers are the same, resulting in their same accuracy and purity. In this scenario, both "manage traffic facilities" and "monitor and manage environment" have few functions. Because of the limited functions, it is difficult to perform an accurate performance comparison, and the two experiments can be considered similar clustering performances.

The clustering performance of the top 10 optimal experiments is not significantly different, especially since the index difference of the first four experiments is only 0.58%. When considering the number of function areas consistent with the reference classification preferentially, the clustering performance of experiment 1 is the best, followed by experiment 3 and experiment 4. Additionally, the two weights of experiment 3 are 0, which can simplify the calculation well without losing precision.

Table 4: Keywords and recommended names of nine function areas.

Function areas	Keywords				Recommended name
	Function name	Function provider	Process object	Service object	
C_1	Emergency, information, get	Emergency event management platform	Emergency event	Emergency event management platform	Provide emergency support
C_2	Vehicle, collect, information	Vehicle-mounted processing equipment	Vehicle operation control	On-board processing equipment	Provide vehicle control and safety
C_3	Provide, communication, interface	Vehicle-mounted interactive equipment	Communication information	Vehicle-mounted interactive equipment	Support communication
C_4	Road, electronic, indicator	Traffic facility management platform	Infrastructure information	Traffic facility management platform	Manage traffic facilities
C_5	Data, traffic, information	Road network management platform	Road network information	Road network management platform	Manage traffic network
C_6	Travel, generate, data	Traveler management platform	Travel demand information	Traveler management platform	Provide travel services
C_7	Information, weather, traveler	Vehicle-mounted interactive equipment	Weather environment information	Environmental climate information platform	Monitor and manage environment
C_8	Parking space, parking lot, information	Traffic facility management platform	Service location information	Traffic facility management platform	Provide parking services
C_9	Payment, parking fee, get	Charging equipment	Cost information	Vehicle-mounted payment equipment	Provide electronic payment services

Table 5: Weight values and clustering performance of the top ten experiments.

Rank	Function name (w_1)	Function provider (w_2)	Process object (w_3)	Service object (w_4)	Accuracy	Purity	Function area number
1	0.6	0.05	0.2	0.15	0.8966	0.8966	9
2	0.55	0.1	0.2	0.15	0.8966	0.8966	8
3	0.8	0	0.2	0	0.8908	0.8908	9
4	0.65	0.05	0.2	0.1	0.8908	0.8908	9
5	0.6	0.1	0.2	0.1	0.8793	0.8793	9
6	0.7	0.05	0.2	0.05	0.8736	0.8736	9
7	0.75	0.05	0.2	0	0.8678	0.8678	9
8	0.3	0.15	0.3	0.25	0.8506	0.9253	10
9	0.4	0.1	0.3	0.2	0.8448	0.9253	10
10	0.35	0.1	0.35	0.2	0.8448	0.9253	10

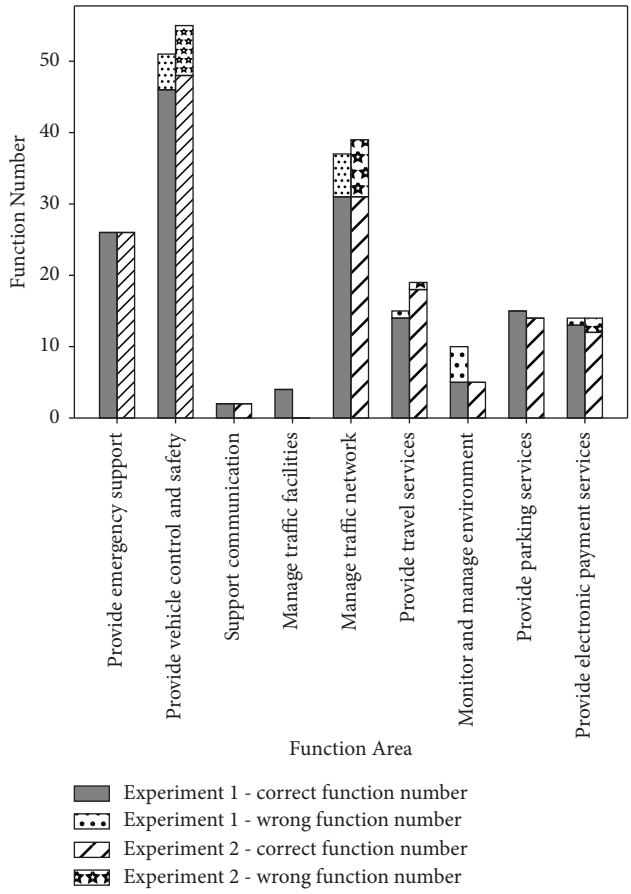

Figure 6: Clustering result distributions of experiment 1 and experiment 2.

In most experiments, the "function provider" weight is the smallest and even close to 0. Thus, we set this weight to 0 to explore the influence of the weights of the other two function attributes on the clustering performance.

Figure 7 shows the values of the accuracy and purity changing with the weight w_3 and w_4. On the whole, the accuracy and purity increase with w_3 increasing, and they can maintain a relatively stable high value in the range of $w_4 < 0.3$ and $w_3 \geq 0.2$. Besides, when $w_4 = 0$ and $w_3 = 0.2$, the purity reaches the highest value of 0.8908, which is the result of experiment 3 in Table 5. When $w_4 > w_3$, the accuracy and the purity both drastically decrease, whereas the purity is generally more significant than the accuracy, indicating that each function area is still relatively accurate. In short, the "process object" is the most critical influence factor and has the most weight.

3.3.2. Clustering Performance Analysis of Optimal Division Result. The accuracy and purity of the optimal division result (experiment 1) are both 0.8966, indicating that the function clustering precision is high and close to the results of the reference classification.

In the optimal division result, the purity of each function area relative to the reference classification is given in Table 6. Most of the function areas are divided relatively correctly; especially, the purities of C_1 (provide

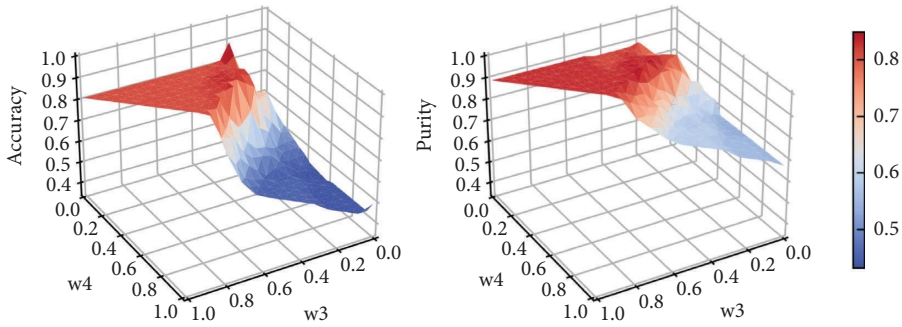

Figure 7: Influence of weights w_3 and w_4 on accuracy and purity.

Table 6: Purity of every function area.

Function area	Function number	Correct function number	Purity
C_1	26	26	1
C_2	51	46	0.9020
C_3	2	2	1
C_4	4	4	1
C_5	37	31	0.8378
C_6	15	14	0.9333
C_7	10	5	0.5000
C_8	15	15	1
C_9	14	13	0.9286
Total	174	156	0.8966

emergency support), C_3 (support communication), C_4 (manage traffic facilities), C_8 (provide parking services) are all equal to 1.

However, the clustering performance of C_7 (monitor and manage environment) is relatively poor. It is because the functions of the weather environment in this scenario are very few, and most of them are for travelers, causing it to confuse easily with the functions of C_6 (manage traveler services). If functions about the weather environment are added, the distinction between these two function areas might be improved.

In short, each function area is relatively independent, and the function similarity within the same function area is high, meeting the requirements of the ATS function area division. In addition, the keywords in the function text can show the characteristics of the function area to some extent, and the naming result can be consistent with the reference classification combined with the candidate name list. Therefore, the function clustering and naming methods proposed in this research are helpful to practical function area division.

4. Conclusion

A division approach that can adaptively divide the function areas is proposed based on text similarity for ATS framework research. Based on it, the function set of the vehicle automatic driving scenario is constructed, and the functions are clustered into function areas using the method. The experiment results show that the proposed method can effectively divide function areas, and the clustering results are relatively more accurate. Also, the text keywords can help to name function areas while reducing the dependence on subjective experience.

Even so, the ATS function area division research still has some limitations. First, the evaluation of the method performance depends on comparing the clustering result with the reference classification. Hence, the quality of the reference classification is vital, which still lacks inspection. Second, more realistic and complicated scenarios are necessary for the verification of the approach in the future. Last, the function areas are manually named here, while automatically generating function area names is worth studying, making it easy to verify a large number of scenarios. In the future, more efficient evaluation methods, more extensive scenario verification, and more appropriate function area name generation will be further studied.

The function area division method helps reduce the dependence on subjective experience and increases the efficiency of function area division work. In addition, the method is suitable for other ATS scenarios and clustering problems in other areas, such as text classification, whereas the clustering objects need to have multiple texts.

Data Availability

The data supporting the current study are available from the corresponding author upon request.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Acknowledgments

The work was funded by the National Key R&D Program of China (Grant no. 2020YFB1600400) and Shenzhen Science and Technology Program (Grant nos. 202206193000001, 20220817201129001).

Supplementary Materials

Appendix I. Table S1: Three-level function list. (*Supplementary Materials*)

References

[1] ATS Research Group, "Architecture of autonomous transportation system," 2021, http://auto-trans-sys.com/.

[2] United States Department of Transportation, *The National ITS Reference Architecture*, 2020,https://local.iteris.com/arc-it/index.html.

[3] FRAME Forum, "The FRAME architecture," 2020, https://frame-online.eu/.

[4] K. Zhang, T. Qi, D. Liu, C. Wang, R. He, and H. Liu, "The latest achievements of Chinese national ITS architecture (in Chinese)," *Journal of Transportation Systems Engineering and Information Technology*, vol. 5, no. 05, pp. 10–15, 2005.

[5] H. Borges, G. Knapp, and B. Eisenhart, "Development of Canadian architecture for intelligent transportation systems," *Artificial Intelligence and Intelligent Transportation Systems: Planning and Administration Transportation research record*, vol. 1774, pp. 80–89, 2001.

[6] X. M. Chen, Y. Lei, G. Jifu, Q. Yongshen, and G. Yanbin, "Development of beijing regional intelligent transportation system architecture," in *Proceedings of the 6th IEEE International Conference on Intelligent Transportation Systems*, pp. 560–565, Shanghai, China, October 2003.

[7] A. W. Sadek, R. Chamberlin, and P. R. Keating, "Use of the national architecture to develop an intelligent transportation systems strategic plan: case study for a medium-sized area," *Transportation Research Record*, vol. 1774, no. 1, pp. 71–79, 2001.

[8] R. Salazar-Cabrera and A. Pachon De La Cruz, "Design of urban mobility services for an intermediate city in a developing country, based on an intelligent transportation system architecture," in *Applied Computer Sciences in Engineering*, pp. 183–195, Springer International Publishing, Berlin, Germany, 2018.

[9] H. Vahidi and T. Sayed, "Using the Canadian ITS architecture for evaluating the safety benefits of intelligent transportation systems," *Canadian Journal of Civil Engineering*, vol. 30, no. 6, pp. 970–980, 2003.

[10] J. M. Golob, C. C. Stecher, and C. Felkins, "California statewide intelligent transportation systems plan evaluation: case study of conformity with national intelligent transportation systems architecture," *Transportation Research Record*, vol. 1826, no. 1, pp. 1–6, 2003.

[11] Y. Liu, J. Shi, and M. Jian, "Understanding visitors'responses to intelligent transportation system in a tourist city with a mixed ranked logit model," *Journal of Advanced Transportation*, vol. 2017, Article ID 8652053, 16 pages, 2017.

[12] C. Jiang, Q. Peng, K. Shi, X. Long, and F. Xu, "Rough set knowledge identification on the logic architecture of the national ITS of China (in Chinese)," *Presented at the 2007 Cross-Strait Symposium on ITS*, Tianjin University Press, Tianjin China, 2007.

[13] J. Miller, "Vehicle-to-vehicle-to-infrastructure (V2V2I) intelligent transportation system architecture," in *Proceedings of the IEEE Intelligent Vehicles Symposium*, pp. 1062–1067, Eindhoven, Netherlands, June 2008.

[14] S. Sousa, "A new approach on communications architectures for intelligent transportation systems," *Procedia Computer Science*, vol. 110, pp. 320–327, 2017.

[15] A. Raza, S. H. R. Bukhari, F. Aadil, and Z. Iqbal, "An UAV-assisted VANET architecture for intelligent transportation system in smart cities," *International Journal of Distributed Sensor Networks*, vol. 17, no. 7, 2021.

[16] J. Z. Wang and Z. J. Wang, "Architecture design of urban intelligent transportation using cloud computing," in *Proceedings of the 2nd International Conference on Materials and Products Manufacturing Technology (ICMPMT 2012)*, pp. 2549–2552, Guangzhou, China, December 2013.

[17] M. Barth and M. Todd, "Intelligent transportation system architecture for a multi-station shared vehicle system," in *Proceedings of the 3rd IEEE Intelligent Transportation Systems Conference (ITSC-2000)*, pp. 240–245, Dearborn, Michigan, USA, October 2000.

[18] G. Chen, H. Cao, M. Aafaque, and J. Chen, "Neuromorphic vision based multivehicle detection and tracking for intelligent transportation system," *Journal of Advanced Transportation*, vol. 2018, Article ID 4815383, 13 pages, 2018.

[19] R. Salazar-Cabrera, A. Pachon De La Cruz, and J. M. M. Molina, "Design of a public vehicle tracking service using long-range (LoRa) and intelligent transportation system architecture," *Journal of Information Technology Research*, vol. 14, no. 1, pp. 147–166, 2021.

[20] M. Mallegowda, V. Nete, and A. Kanavalli, "Intelligent transportation system based on the principles of service-oriented architecture," in *Proceedings of the 12th IEEE and IFIP International Conference on Wireless and Optical Communications Networks (WOCN)*, Bangalore, India, September 2015.

[21] Q. Chen, A. K. Sowan, and S. H. Xu, "Assoc Comp, "A safety and security architecture for reducing accidents in intelligent transportation systems," in *Proceedings of the 37th IEEE/ACM International Conference on Computer-Aided Design (ICCAD)*, California, CA, USA, November 2018.

[22] Y. R. B. Al-Mayouf, O. A. Mahdi, N. A. Taha, N. F. Abdullah, S. Khan, and M. Alam, "Accident management system based on vehicular network for an intelligent transportation system in urban environments," *Journal of Advanced Transportation*, vol. 2018, Article ID 6168981, 11 pages, 2018.

[23] M. K. Natvig and H. Westerheim, "National multimodal travel information - a strategy based on stakeholder involvement and intelligent transportation system architecture," *IET Intelligent Transport Systems*, vol. 1, no. 2, pp. 102–109, 2007.

[24] W. G. Li, Y. Yamashita, M. W. Koendjbiharie, R. C. D. Juca, and A. MacIver, "The development and implementation of the operation system and data bank for the intelligent transportation system - sitcuo," *Journal of Advanced Transportation*, vol. 38, no. 2, pp. 163–186, 2004.

[25] S. li, Y. Cheng, T. Y. Zhou, and P. Y. Liang, "The improved precoding method in the VLC-based intelligent transportation system," *Journal of Advanced Transportation*, vol. 2022, Article ID 5951389, 9 pages, 2022.

[26] S. Din, A. Paul, and A. Rehman, "5G-enabled hierarchical architecture for software-defined intelligent transportation system," *Computer Networks*, vol. 150, pp. 81–89, 2019.

[27] A. Subakti, H. Murfi, and N. Hariadi, "The performance of bert as data representation of text clustering," *Journal of Big Data*, vol. 9, no. 1, 2022.

[28] L. S. Lomakina, V. B. Rodionov, and A. S. Surkova, "Hierarchical clustering of text documents," *Automation and Remote Control*, vol. 75, no. 7, pp. 1309–1315, 2014.

[29] C. Xiong, Z. Hua, K. Lv, and X. Li, "An improved k-means text clustering algorithm by optimizing initial cluster centers," in

Proceedings of the 2016 7th International Conference on Cloud Computing and Big Data (CCBD), pp. 265–268, Macau, China, November 2016.

[30] H. Murfi, "The accuracy of fuzzy c-means in lower-dimensional space for topic detection," in *Proceedings of the Third International Conference on Smart Computing and Communications SmartCom*, pp. 321–334, Springer International Publishing, Tokyo, Japan, December 2018.

[31] J. Xie, R. Girshick, and A. Farhadi, "Unsupervised deep embedding for clustering analysis," in *Proceedings of the Presented at the 33rd International Conference on Machine Learning*, Proceedings of Machine Learning Research, New York, NY, USA, June 2016.

[32] S. Liu, X. Wu, and J. Chai, "A dynamic clustering method of hot topics based on user interaction and text similarity," in *Proceedings of the 2021 14th International Congress on Image and Signal Processing BioMedical Engineering and Informatics (CISP-BMEI)*, pp. 1–5, Shanghai, China, October 2021.

[33] G. S. Reddy and T. V. Rajinikanth, "A text similarity measure for document classification," *IADIS International Journal on Computer Science and Information Systems*, vol. 12, no. 1, pp. 14–25, 2017.

[34] H. Li and J. Xu, "Semantic matching in search," *Foundations and Trends® in Information Retrieval*, vol. 7, no. 5, pp. 343–469, 2014.

[35] H. Pham, M. T. Luong, and C. D. Manning, "Learning distributed representations for multilingual text sequences," in *Proceedings of the 1st Workshop on Vector Space Modeling for Natural Language Processing*, pp. 88–94, Denver, CO, USA, June 2015.

[36] Z. Guo, "Research on knowledge service matching based on attribute similarity and clustering (in Chinese)," *Modular Machine Tool & Automatic Manufacturing Technique*, vol. 09, pp. 171–174, 2020.

[37] B. Jiang, L. Ye, W. Pan, and J. Wang, "Service clustering based on the functional semantics of requirements (in Chinese)," *Chinese Journal of Computers*, vol. 41, no. 06, pp. 1035–1046, 2018.

[38] ATS Research Group, "ATS component - function," 2021, http://auto-trans-sys.com/func.

[39] M. A. Alksher, A. Azman, R. Yaakob, R. A. Kadir, A. Mohamed, and E. M. Alshari, "A review of methods for mining idea from text," in *Proceedings of the Third International Conference on Information Retrieval and Knowledge Management (CAMP)*, Malacca, Malaysia, August 2016.

[40] O. Arbelaitz, I. Gurrutxaga, J. Muguerza, J. M. Pérez, and I. Perona, "An extensive comparative study of cluster validity indices," *Pattern Recognition*, vol. 46, no. 1, pp. 243–256, 2013.

[41] P. J. Rousseeuw, "Silhouettes: a graphical aid to the interpretation and validation of cluster analysis," *Journal of Computational and Applied Mathematics*, vol. 20, pp. 53–65, 1987.

[42] J. G. Conrad, K. A. Al-Kofahi, Y. Zhao, and G. Karypis, "Effective document clustering for large heterogeneous law firm collections," in *Proceedings of the 10th International Conference on Artificial Intelligence and Law*, Bologna, Italy, June 2005.