IEEE Access
Multidisciplinary : Rapid Review : Open Access Journal

Received August 12, 2020, accepted August 29, 2020, date of publication September 3, 2020, date of current version September 21, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.3021421

A Text-Granulation Clustering Approach With Semantics for E-Commerce Intelligent Storage Allocation

QI LANG, XUEJUN PAN, (Member, IEEE), AND XIAODONG LIU
Faculty of Electronic Information and Electrical Engineering, School of Control Science and Engineering, Dalian University of Technology, Dalian 116024, China

Corresponding author: Xiaodong Liu (xdliuros@dlut.edu.cn)

This work was supported by the National Key Research and Development Program of China under Grant 2018YFB1702200.

ABSTRACT With the hot rise and increasing popularity of e-commerce and online shopping, exponentially growing orders require to be picked up timely. In the order-based warehouse picking process, more intelligent optimization methods are urgently demanded to meet the maturity of robot technology. Due to the wide variety of SKUs (stock-keeping units) in warehouses, direct analysis based on orders can hardly get strong correlations among SKUs. Since orders and SKUs usually are presented in text form files, this paper proposes a clustering optimization algorithm based on text information granules which makes full use of context information and semantics. Storage assignment is determined by clustering SKUs first and then performing correlation analysis on the clustering results according to orders. The algorithm can give the semantic description of each cluster so that the clustering process is interpretable and transparent. Extensive experiments on a distribution center of a clothing production enterprise indicate 40.5 to 57.9% improvement in the average order picking distance compared with the ABC classification.

INDEX TERMS Granular computing, EI algebra, text clustering, warehouse management system.

I. INTRODUCTION

A warehouse management system is a key element in the intelligence of electronic commerce. The ability of robots to efficiently and quickly collect a large number of orders from a large warehouse will be one of the core competencies of an e-commerce company. The objectives of warehouse optimization include minimizing the distance of picking orders and maximizing the space utilization of the warehouse [1]. A common form of order content demonstrates in table 1, each order contains several purchased items (e.g. goods). A recent survey of warehouses implied that more than 70% of the time was spent picking up orders [2]. To reduce the distance required for order picking, researchers proposed optimization methods from the following four ideas: carefully planned instruction picking path, batch orders, warehouse area division, and product allocation to the appropriate storage location [3].

In recent years, many order-based warehousing algorithms have been proposed. Chen *et al.* [4] proposed the ARBM method, which extracted the relationship between products by using association rules. ARBM method can shorten the travel distance by improving the batch processing efficiency of orders. However, Li *et al.* [5] pointed out that when the order quantity is small, or the warehouse is small, the optimization efficiency of ARBM will reduce. In the same year, Chen and Wu [6] proposed the approach ARIP, which depended a lot on whether there was an association between the orders. Chiang *et al.* [7] proposed DMSA to reduce the number of stops for picking one order. DMSA is only applicable to reallocate a portion of products [5]. Dijkstra and Roodbergen [8] proposed a dynamic programming (DP) storage strategy based on optimization to solve multi-aisle and multi-item problem. Azadeh *et al.* [9] presented

TABLE 1. Trading information.

order_number	item set
order_1	$\{sku_1, sku_3, sku_{16}, sku_{101}, \ldots\}$
order_2	$\{sku_7, sku_{12}, sku_{65}, sku_{98}, \ldots\}$
order_3	$\{sku_3, sku_6, sku_{11}, sku_{50}, \ldots\}$
$\vdots$	$\vdots$

The associate editor coordinating the review of this manuscript and approving it for publication was Biju Issac.

an algorithm based on the genetic algorithm to optimize the allocation of operators in a multi-product assembly shop. However, the full complement of robots has largely replaced human labor. Fontana and Nepomuceno [10] considered a 3D storage location-allocation model based on the ELECTRE TRI method, which realized the classification storage by defining the shelf level and shelf position of each item respectively. Zhou *et al*. [11] studied a warehouse layout with V-shaped and fishbone channels which removed the original constraint of the channel being a straight line. Other studies by conducting correlation analysis among products for storage location assignment can be found in [12] and [13]. The above algorithms optimize the process of intelligent warehouse pickup from many angles. However, the above algorithms do not make further use of the semantic information of goods, so the whole process is not interpretable. To sum up, this paper proposes a strategy for placing SKUs in large warehouses, and it has semantic meaning.

Most of the existing algorithms perform correlation analysis on SKUs directly based on orders. With the increasing variety of SKUs, the support degree of several SKUs appearing in one order at the same time will be almost submerged, and then correlation analysis can not get sufficient results. Meanwhile, the existing methods do not focus on exploring semantic information about text in the warehousing and picking process. Aiming at the order-based pickup problem of super large warehouses, this paper proposes a text-granulation clustering algorithm (TGCA) method to determine the storage location by clustering based on text information. Then the multilateral affinity among clustering results is calculated. Because of the variety of items in warehouses, order-based correlation analysis between single items becomes ineffective. Starting with a real-life warehouse, the TGCA algorithm proposed in this paper utilizes the basic text information such as commodity name, size, and color to cluster commodities, which not only improves the effectiveness of correlation, but also provides clustering description by EI algebra, facilitating human query and management. The goal of the TGCA model is not to optimize, but to provide a way to extend algorithms limited to small warehouses to large-scale problems that are more suited to actual needs.

The innovation of this paper are summarized as follows:

- A text-granulation clustering approach for intelligent storage allocation of large warehouse is proposed.
- The algorithm proposed in this paper is based on semantic learning, so the whole clustering process is interpretable.
- The use of EI algebra provides cluster descriptions for each cluster of SKUs.

Based on the above motivations and ideas, the structure of the paper is summarized as follows:

- A literature review of storage allocation approaches (section II)
- Definition of the storage assignment and theories used in the algorithm TGCA, including: granulation, EI algebra (section III).
- A detailed introduction of the TGCA algorithm (section IV).
- A case study for warehouse picking as well as experimental results (section V).
- Conclusions and future research plans (section VI).

II. RELATED WORK

Warehouse layout design, as an important part of the order pickup optimization process, has been widely valued by researchers. Zhang *et al*. [14] proposed a two-layer evolutionary algorithm to solve the problem of automatic warehouse layout design, which is suitable for large-scale warehouse logistics problems. Li *et al*. [5] proposed a new dynamic storage model that combined the product affinity with the ABC classification. However, this study only considers the two-sided affinity among products, whereas the multilateral affinity needs further study. De Koster *et al*. [15] optimized the warehousing and picking problem from the perspectives of storage assignment policies and layout design, to improve the overall speed of picking goods. Yang *et al.* [16] proposed an algorithm for clustering commodities with constraints on storage conditions. The principal component analysis is used to calculate the distance between clusters. Moshref-Javadi *et al.* [17] first clustered the goods in the warehouse and then discuss the different ways to locate the goods. The clustering methods include principal component analysis, singular value decomposition, and two-step clustering. However, in the process of locating the goods in the warehouse, the turnover rate of goods is not taken into account. Van Gils *et al*. [18] analyzed the relationships among storage, batch processing, partitioning, and routing and designed the warehouse as a whole to reduce orders pick time. In the paper [19], a warehouse picking algorithm combining storage allocation and travel distance estimation is proposed, which is effective for different routing strategies. Bozer *et al.* [20] compared two famous order picking systems mini-load system and Kiva system, and gave the advantages and limitations of each system in combination with the simulation model. Guo *et al*. [21] made a detailed comparison of four warehouse storage strategies which are storage zoning, random, full turnover-based, and class-based storage. A method of measuring similarity between commodities was proposed to construct a natural clustering model by Jane *et al.* [22]. Table 2 gives an overview of the algorithms of warehouse layout design. Different from the above algorithms, our TGCA algorithm proposes a clustering algorithm based on semantic learning for placing strategy of warehouse inventory goods, so that "related" goods can be placed closer together.

III. BACKGROUND

A. STORAGE PROBLEM DESCRIPTION

The core problem studied in this paper is the order-based pickup of the e-commerce platform. When the system

TABLE 2. Overview of algorithms of warehouse layout design.

Algorithm	Description
ARBM [4]	Extracting the relationship between products by using association rules
DSAP [5]	Combining the product affinity with the ABC classification
ARIP [6]	Mining the association relationship between orders through data mining and integer programming
DMSA [7]	Reducing the number of stops for picking one order
DP [8]	Using a dynamic programming storage strategy to solve multi-aisle and multi-item problem
MOGA [9]	Optimize the multi-product assembly shop based on the genetic algorithm
ELECTRE TRI [10]	Realize the classification storage by defining the shelf level and shelf position
V-shaped and fishbone [11]	Study a warehouse layout with V-shaped and fishbone channels
TLEA [14]	A two-layer evolutionary algorithm to solve the automatic warehouse layout design
A literature overview [15]	A literature overview of storage assignment policies and layout design
PCA cluster [16]	Clustering commodities with constraints on storage conditions. Principal component analysis is used to calculate the distance between clusters
Different clustering [17]	Clustering methods include principal component analysis, singular value decomposition, and two-step clustering
Relationships analysis [18]	Analyze the relationships among storage, batch processing, partitioning, and routing
Distance estimation [19]	Combining storage allocation and travel distance estimation, which is effective for different routing strategies
Miniload and Kiva [20]	Compare miniload system and Kiva system, and gave the advantages and limitations of each system
Comparison four strategies [21]	Compare four warehouse storage strategies: storage zoning, random, full turnover-based, and class-based storage
Natural clustering [22]	Measuring similarity between commodities to construct a natural clustering model

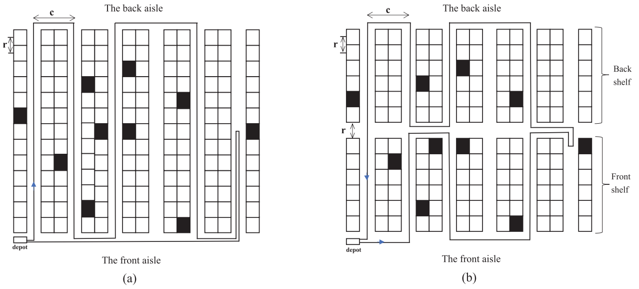


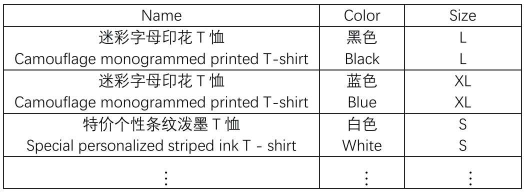
FIGURE 1. Schematic diagram of warehouse interior. The left (a) is a warehouse type that has only vertical channels, and the right (b) is a warehouse type that has both horizontal and vertical channels.

receives customers' orders, the robots need to go from an entrance to a warehouse to pick up the corresponding goods according to the requirements on orders. Robots can only pass through the existing horizontal and vertical channels in the warehouse. When robots get all the goods on an order, they will return to the warehouse entrance. We suppose the warehouse has n rows of shelves, each with m storage locations. In this way, the whole warehouse can store $m * n$ kinds of SKUs. Depot is located at the left of the front aisle. Figure 1 shows two usual warehouse picking diagrams for an order, with black squares representing items purchased in the order and white squares representing items not specified. Figure 1(a) shows the shelf arrangement with only vertical aisles, and figure 1(b) shows the multi-row shelf arrangement with both horizontal and vertical aisles. They are the two most common forms of storage shelves placement. The distance between the two storage locations and two parallel aisles are denoted by r and c, respectively. We assume that the robot can pick up items on the shelves on both sides in the middle of an aisle without shifting to the side, which is common in many other hypotheses in literatures (e.g., [23], [24]).

For the routing method, we consider the S-shape routing, Return routing, and Midpoint routing [18], [25]–[27].

- S-shape routing: from left to right, the robot traverses the aisle where the shelves need to be picked up. If it enters the current aisle from the front, it enters the next aisle from the back, and vice versa. If the shelves on either side of the aisle do not contain the items in the order, the robot will not enter the aisle (see figure 1).
- Return routing: each time, the robot enters the aisle from the front of the shelf, picks up the farthest item in the current column, and immediately returns to the front of the shelf.
- Midpoint routing: only the first and last aisle will be completed around the perimeter of the shelf. The shelves are divided in the middle into front and back sections.

The data used in this paper are from a real Chinese e-commerce history orders and SKUs warehouse (If it is to deal with English orders and SKUs, it degenerates to the problem without the word cutting process). In this practical problem, there are three attributes for each SKU, which are the name of commodity, size, and color and we use $sku_i = [sku_i^{(n)}, sku_i^{(s)}, sku_i^{(c)}]$ to represent them. Figure 2 illustrates an example of some SKUs. The proposed algorithm can widely



Name	Color	Size
迷彩字母印花T恤 Camouflage monogrammed printed T-shirt	黑色 Black	L L
迷彩字母印花T恤 Camouflage monogrammed printed T-shirt	蓝色 Blue	XL XL
特价个性条纹泼墨T恤 Special personalized striped ink T - shirt	白色 White	S S
⋮	⋮	⋮

FIGURE 2. Some Chinese specific examples of SKUs.

use in other e-commerce storage data, such as Amazon and Alibaba, and the attributes can also include brand, model, configuration, grade, pattern, packaging capacity, production date, shelf life, price, origin, etc. For different warehouses with different commodity types, our algorithm has similar principles in selecting commodity attributes. First, prioritize the attributes that are common to most goods. Such attributes are the basis of clustering. Secondly, attributes with semantic information should also be given priority, such as the name of the product, brand, scope of application, etc. In this way, the semantic richness of the cluster description can be guaranteed.

Each order covers the purchased items, as represented in table 1. An order may purchase only one item or more items at the same time. The significance of the proposed algorithm is to determine the position of each SKU on shelves to make the shortest path to pick up items according to a batch of orders.

B. RELEVANT THEORIES

In this part, we give a specific description of the definition and mathematical formulation of relevant theories.

1) GRANULAR COMPUTING

Granular computing (GrC) is a newly proposed research field that covers the world view and methodology of viewing the objective world with different granularity [29]. A granule is a block formed by some entities through indistinct, similar, adjacent, or functional relations. This process of processing information is called information granulation [30]. Information granules can be defined and analyzed in many formal frameworks such as fuzzy sets [31], rough sets [32], and concept lattice [33].

The granular layer refers to all particles with identical granularity. In this paper, text granularity is used to dynamically adjust the number of items contained in the granularity layer to cluster. In the algorithm proposed in this paper, the coarsest granularity is the cutting results of SKU name, color and size attribute, such as w_1 = casual, w_2 = T-shirt, w_3 = printed, w_4 = red, w_5 = XL, w_6 = dress, etc. Finer information granules such as $A_1 = w_1w_4w_2$ and $A_2 = w_3w_5w_6$, representing "casual red t-shirts" and "printed XL size dresses", respectively. The specific process will be described in detail in subsection IV-A

2) EI ALGEBRA

The most prominent advantage of a text-based clustering algorithm is the change of text information in the clustering process. However, it is unwise to obtain the cluster description by simply and directly combining all SKUs information particles belonging to the same cluster because it will generate a large amount of redundancy and the description will be extremely long. To obtain the most concise description of clustering, we introduce the relevant definitions and properties of EI algebra to solve this problem. EI algebra was first proposed by Liu [34] to deal with fuzzy set relations. It has also achieved good results in combination with particle calculation. The definitions of EI algebra are illustrated by the following examples:

$$\prod_{w\in A_1} w = w_1w_4w_2 \tag{1}$$

$$\gamma = \sum_{i\in\{1,2\}} (\prod_{w\in A_i} w) = w_1w_4w_2 + w_3w_5w_6 \tag{2}$$

where "+" denotes the disjunction of granules, A_i is same as mentioned in section III-B1, and γ represents "casual red t-shirts" or "printed XL size dresses". Our text granularity contains clear semantic information, and the coarse granularity is calculated by EI algebra to obtain a finer granularity with richer semantic information.

EI algebra also has the following axioms (natural language axioms) [35], [36]:

- Absorption law:

$$w_1w_4w_6 = w_1w_4w_1w_6,$$
$$w_1w_4 + w_1w_4 + w_3 = w_1w_4 + w_3,$$
$$w_1w_2 + w_1w_2w_8 + w_3 = w_1w_2 + w_3$$

- Commutative law:

$$w_5w_8 = w_8w_5,$$
$$w_5w_6 + w_5w_8 = w_5w_8 + w_5w_6$$

- Distributive law:

$$w_5w_6 + w_5w_8 = w_5(w_6 + w_8)$$

The strict mathematical definition of EI algebra is given as follows :

Let W is the set of some particles [37],
$EW^* = \{ \sum_{i\in I}(\prod_{w\in A_i} w) \mid A_i \subseteq W, i \in I, I$ is any indexing set$\}$
where $\sum_{i\in I} A_i$ is the sum in form, $\sum_{i\in I} A_i$ and EW represent the same set when these $\sum_{i\in I} A_i$ are summed by different orders.

Proposition 1: Let W be a set. The operators $\vee$ and $\wedge$ form a completely distributive lattice [38]:

For any $\sum_{i\in I}(\prod_{w\in A_i} w), \sum_{j\in J}(\prod_{w\in B_j} w) \in EW^*$,

$$\sum_{i\in I}(\prod_{w\in A_i} w) \vee \sum_{j\in J}(\prod_{w\in B_j} w) = \sum_{k\in I \bigsqcup J} (\prod_{w\in C_k} w)$$
$$= \sum_{i\in I}(\prod_{w\in A_i} w) + \sum_{j\in J}(\prod_{w\in B_j} w), \tag{3}$$

$$\sum_{i\in I}(\prod_{w\in A_i} w) \wedge \sum_{j\in J}(\prod_{w\in B_j} w) = \sum_{i\in I, j\in J} (\prod_{w\in A_i\sqcup B_j} w), \tag{4}$$

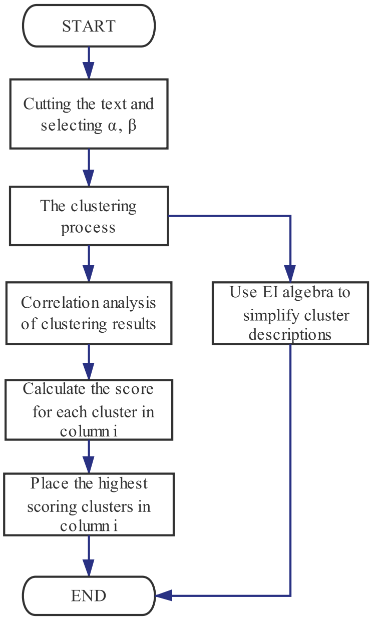

START
Cutting the text and selecting α, β
The clustering process
Correlation analysis of clustering results
Use EI algebra to simplify cluster descriptions
Calculate the score for each cluster in column i
Place the highest scoring clusters in column i
END

FIGURE 3. Flow chart of the dynamic text-granule clustering algorithm: where α and β respectively represent upper and lower bounds of cluster size.

where $I \bigsqcup J$ is the disjoin union of I and J and if $k \in I$, then $C_k = A_k$; if $k \in J$, then $C_k = B_k$.

For example: $A_1 = w_1w_3w_4$, $A_2 = w_3w_5w_6$ and $A_3 = w_4$,

$$A_1 \vee A_2 = w_1w_3w_4 + w_3w_5w_6 \tag{5}$$

$$A_1 \wedge A_2 = w_1w_3w_4w_5w_6 \tag{6}$$

$$\begin{aligned} A_1 \vee A_2 \wedge (A_3)' &= (w_1w_3w_4 + w_3w_5w_6) \wedge (w_4)' \\ &= w_3w_5w_6(w_4)' \end{aligned} \tag{7}$$

The negation of the concept $w \in W$ is represented by w' and $w \wedge w' = \emptyset$.

Relevant proof and further derivation can be seen in Kukkurainen and Paavo [39] and Liu *et al.* [40]. We will give concrete examples of the above two formulas in subsection IV-B.

IV. THE TEXT-GRANULATION CLUSTERING ALGORITHM FOR STORAGE ASSIGNMENT

In the following part, the order-based clustering optimization algorithm will be introduced in detail. Figure 3 illustrates the main steps of the clustering method. Input names, colors, and sizes of all SKUs and the coarse-level information granule are obtained through word cutting. These coarse granularities will be used as the basic units of EI algebra. The upper and lower limits of SKUs number contained in each cluster are determined as α and β, respectively, according to the size of the warehouse. The detail of the above operations and the clustering process will be described in subsection IV-A. The operation of using EI algebra to obtain the simplest description of each cluster is described in subsection IV-B. The rule for placing clusters in a warehouse is that each cluster places in the same row of shelves. Start from the first row and put the cluster with the highest score into the corresponding row shelf according to the scoring formula given in subsection IV-C.

A. TEXT PREPROCESS AND SKUs CLUSTERING BASED ON INFORMATION GRANULES

Due to the large quantity of SKUs and the dispersion of the order content, it is difficult to extract the effective commodity association rules directly between individual SKUs based on the orders. Therefore, the method of dividing SKUs into a uniform number of heaps for correlation analysis proposed. Clustering SKUs with the same attributes or varieties into the same category can achieve the purpose of heap separation and obtain comprehensible and manageable results. Thus, text information such as the name, size, and color of an SKU will serve as an important basis for clustering rules.

Define S as the set of all SKUs. For any given interval $[\alpha, \beta]$, where α and $\beta \in [1, N]$ are positive integers and N is the total number of SKUs. S will be divided into K ($K \leq N$) subsets $\{s_1, s_2, \ldots, s_K\}$ by clustering. These sets satisfy (8), (9) and (10):

$$S = s_1 \cup s_2 \cup \ldots \cup s_K \tag{8}$$

$$s_i \cap s_j = \emptyset, \quad i, j \in [1, K] \ \ and \ \ i \neq j \tag{9}$$

$$N(s_i) \in [\alpha, \beta], \quad i \in [1, K] \tag{10}$$

where $N(s_i)$ represents the number of SKUs in the subset s_i.

Firstly, text preprocessing refers to the cut names of SKUs by Jieba [1] to get the basic particles for clustering. At this point, each $sku_i^{(n)}$ is sliced into $g_i^{(n)} = [g_{i1}^{(n)}, g_{i2}^{(n)}, \ldots]$, where $g_{ij}^{(n)}$ means the jth word in the cut result of the name of the ith SKU. We also ranked each particle from highest to lowest according to its frequency of appearance in all names of SKUs. Then we cluster SKUs with the same result of name cutting as s_i. Next, according to the quantity of $N(s_i)$ which is the number of SKUs in the newly cluster, make the following judgment:

- If $\alpha \leq N(s_i) \leq \beta$: Save s_i as a new cluster;
- If $N(s_i) > \beta$: Add color attribute($s_i^{(c)}$) or size attribute($s_i^{(s)}$) to s_i (here we choose the attribute with high cross entropy of clustering results). Then update the clustering results, continue to judge the number of elements in the cluster. If both attributes are added and the number is still greater than β, the cluster will updated to the final result of clustering and put the excess on another cluster;
- If $N(s_i) < \alpha$: Remove the words with the lowest frequency in s_i and update the clustering results. Continue to judge the number of elements in the cluster.

Figure 4 is a flowchart for finding a cluster that conforms to the rules. Unclustered SKUs are continuously operated in the figure above until all SKUs belong to a single cluster, and then the clustering process ends. If both the color and size attributes are added and the value of $N(s_i)$ is still greater

[1] GitHub: http://github.com/fxsjy/jieba

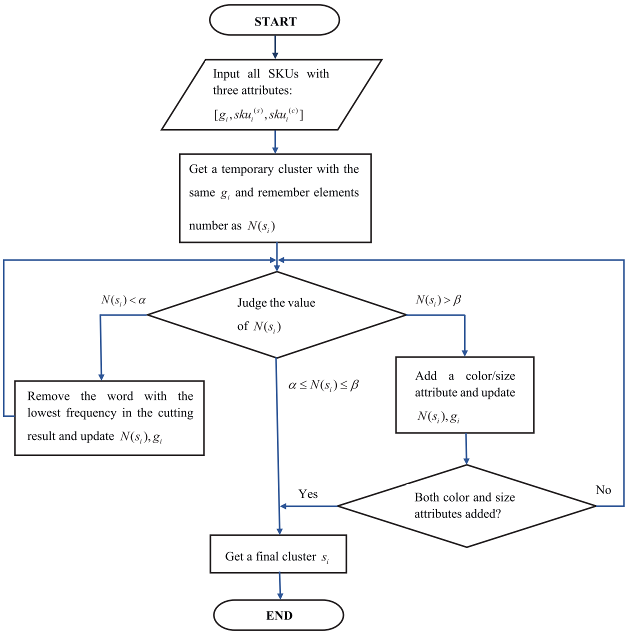

START
Input all SKUs with three attributes: $[g_i, sku_i^{(s)}, sku_i^{(c)}]$
Get a temporary cluster with the same g_i and remember elements number as $N(s_i)$
Judge the value of $N(s_i)$
$N(s_i) < \alpha$
$N(s_i) > \beta$
$\alpha \le N(s_i) \le \beta$
Remove the word with the lowest frequency in the cutting result and update $N(s_i), g_i$
Add a color/size attribute and update $N(s_i), g_i$
Both color and size attributes added?
Yes
No
Get a final cluster s_i
END

FIGURE 4. A concrete flowchart of the operation of identifying a cluster.

than β, extra SKUs will be placed at the end of shelves in the other few clusters. Algorithm 1 gives a first-round clustering process based on names granular. The next part describes the process of obtaining the cluster description.

Algorithm 1 TGCA Cluster

```
Input: SKUS = (name, color, size)
Output: applicability-dict, low-dict, high-dict

initialize high-dict, low-dict, title-num-dict
for SKU in SKUs: do
    name-list = Jiebacut(SKU[name])
    if name-list in keys of title-num-dict then
        value = value + 1
    else
        update title-num-dict[name-list] = 0
for key, value in title-num-dict: do
    if lower ≤ len(value) ≤ upper: then
        applicability-dict[key] = value
        if len(value)< lower: then
            low-dict[key] = value
        else
            high-dict[key] = value
```

B. UNIQUE CLUSTER DESCRIPTION BASED ON EI ALGEBRA

After getting clustering results, figure 5 represents SKUs contained in each cluster and common words of these SKUs. Each entry in figure 5 consists of two lines, with the actual Chinese text on the top and the English translation in brackets on the bottom. This also illustrates that the method proposed in this paper is not limited by language. Next, according to the common words of each cluster, an unique description of each cluster will be given through EI algebra.

There are two clusters in figure 5. Common words of each cluster can be seen as a description of each cluster. However, sku_1 can be shared by both s_1 and s_2. When the administrator retrieves s_2 using the description "printed T-shirt : red", sku_1 will also be retrieved. To solve the problem, we first denoted the description of s_1 as d_1 (all SKUs retrieved by d_1 belong to s_1). Then we use d_2 to retrieve SKUs in s_1 and if it retrieves an item, the new description of s_2 can be updated as follows:

$$d_2 = d_2 \wedge d_1' \tag{11}$$

Until d_i can only retrieve SKUs in s_i, stop updating d_i and simplify d_i by natural language axioms. In this way, all clusters get a single and simplified description.

Take the two clusters in figure 5 as an example, SKUs in s_2 cannot be retrieved as the first cluster s_1, otherwise, the SKU will be first extracted by s_1 in the clustering process. The first cluster description of s_1 defaults to $d_1 =$ [printed personalized T-shirt]. For cluster s_2, similar to (11), d_2 can be obtained by (5) and (6):

$$\begin{aligned}
d_2 &= d_2 * (d_1)' \\
&= [printed * (T - shirt) * red] \\
&\quad * [printed * personalized * (T - shirt)]' \\
&= [printed * (T - shirt) * red] \\
&\quad * [(printed)' + (personalized)' + (T - shirt)'] \\
&= printed * (T - shirt) * red * (personalized)'
\end{aligned}$$

where $()'$ denotes the negation of the contents and $word * (word)' = \emptyset$, d_2 means "printed red T-shirts but not personalized".

The above procedure ensures that only SKUs in the corresponding cluster can be retrieved for each cluster description, which will provide semantic assistance to the warehousing and refining process.

The cluster description obtained by the logical operation of information granules through EI algebra makes the warehousing, picking up and delivery process of e-commerce have description of text information, so that it is more convenient for people to manage and view the warehouse, SKUs, and orders.

C. RELEVANCE ANALYSIS AND LOCATION CRITERIA

As a simple and effective data analysis technique, correlation analysis can mine the correlation existing in large data and describe the law of some attributes appearing simultaneously in things [41], [42]. Here we use the classical Apriori algorithm to extract the correlation between clusters [43]. Table 3 implies the relationship between the order requirements and the s_i cluster. After correlation analysis, the parameters f_k and f_{kl} can be obtained, respectively representing the frequency of SKUs in s_k appear in all orders and the frequency of two SKUs in s_k and s_l appear together in all orders at the same time.

Common words in a cluster	s_1=[印花 个性 T 恤] (s_1=[printed personalized T-shirt])	s_2=[印花 T 恤 ：红色] (s_2=[printed T-shirt ：red])	…
Corresponding SKUs contained within the cluster	sku_1=' 印花个性 T 恤 ：红色: L ' (sku_1='printed personalized T-shirt : red : L ') sku_2=' 条纹印花个性 T 恤 : 蓝色: M ' (sku_2=' striped printed personalized T-shirt : blue : L ') sku_3=' 印花复古个性 T 恤 : 黄色: M ' (sku_3=' printed vintage style T-shirt : yellow : L ') ⋮	sku_4=' 夏季印花 T 恤 ：红色：S ' (sku_4=' summer printed T-shirt ：red ：S ') sku_5=' 夏季印花 T 恤 ：红色：M ' (sku_5=' summer printed T-shirt ：red ：M ') sku_6=' 男士印花 T 恤 ：红色：M ' (sku_6=' men's printed T-shirt ：red ：M ') ⋮	…

FIGURE 5. Clustering result diagram.

TABLE 3. Trading information with clusters.

order_number	Clusters included
order_1	$\{s_1, s_2, s_7, s_9, \ldots\}$
order_2	$\{s_2, s_7, s_{10}, s_{11}, \ldots\}$
order_3	$\{s_2, s_6, s_8, s_{10}, \ldots\}$
⋮	⋮

Determine the value of α and β according to the length of a row of shelves in the warehouse, and all SKUs will be clustered with quantities in $[\alpha, \beta]$ to ensure that each column can store a cluster of SKUs. For the type of shelves shown in figure 1(a), first place the most frequently purchased cluster of all orders in the first column on the left. For the type in figure 1(b), the first two clusters with the highest purchase frequency of all orders are placed in the front and back first columns on the far left. The following formula is given to calculate the degree to which cluster belongs to the *ith* column:

$$G_k^i = f_k + \sum_{l \in N_i} \frac{f_{kl}}{d_{kl}}, \tag{12}$$

where N_i represents the index set of clusters already placed in the warehouse. We assume the cluster s_l has been stored in the *jth* column, and if s_k will be stored in the *ith* column, d_{kl} can be calculated by:

$$d_{kl} = i - j \tag{13}$$

After calculating the score of each cluster in column i, the s_k with the highest score of G_k^i will be selected and put in the *jth* column (in the case of figure 1(b), the scores of the front and back shelves are calculated separately). Figure 6 shows the warehouse schematic diagram of correlation analysis according to clustering results. The cluster whose clustering result quantity is less than α will be placed in the remaining position at the end of each shelf.

Formula 12 is composed of two parts. The larger the value of f_k in the first part indicates that such items are purchased more often and should be placed on shelves closer to the depot. The other part of the formula measures the relevance of the current cluster to the SKUs already stored in the warehouse and takes into account the effect of shelf distance.

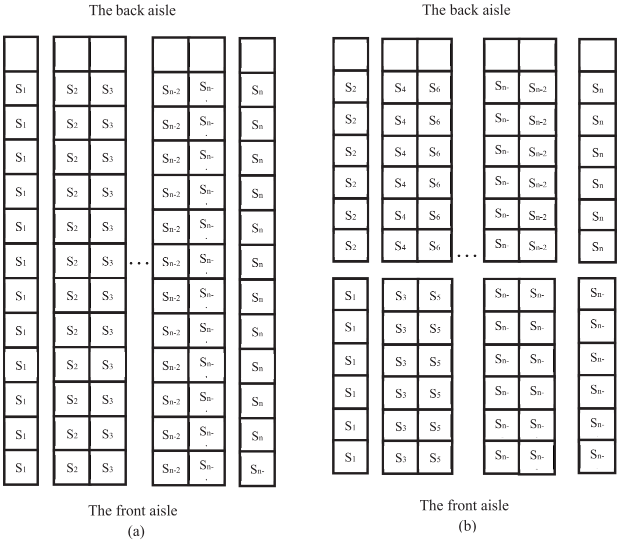


FIGURE 6. Warehouse layout diagram. (a) is the schematic diagram of goods placement with only longitudinal channels, and (b) is the goods placement with both transverse and longitudinal channels.

Cluster-based commodity storage methods also include within-aisle, across-aisle, and diagonal which are introduced in [21], [28]. We will further analyze the advantages and disadvantages of each method in combination with the experimental results in the comparative experiment.

V. EXPERIMENTAL SETUP

A. COMPARATIVE EXPERIMENTAL RESULTS

The total amount of actual data used in this paper includes 11,768 SKUs and the number of orders is 26,900. Each order contains an average of 2 items. Each order contains information about the name, color, and size of items purchased. In the experiment, we set r = 1 and c = 1, indicating that spacings between shelves are 1 *m*. We adopt the more common S-shape routing method as a basic routing method for order pickup. We also combine our TGCA algorithm with Return routing and Midpoint routing respectively for comparison experiments. For cluster-based storage placement, in addition to using formula 12, we also compared it with Across-aisle and Within-aisle methods.

To test the effectiveness of the proposed method on different scale warehouses, the following experiments conducted.

TABLE 4. Experimental design setup.

Parameter	Warehouse 1 (small size)	Warehouse 2 (medium size)	Warehouse 3 (mid-large)	Warehouse 4 (large size)
No. of storage locations	100	1500	7500	15000
No. of rows	10	150	75	1500
No. of SKUs	50	1000	5000	10000
No. of orders	149	1269	15630	23420
Warehouse width	5	10	20	40
Warehouse length	10	50	100	100

TABLE 5. Test results of average one order picking distance (m).

Models	Warehouse 1 (2, 5)	Warehouse 2 (30,60)	Warehouse 3 (20,40)	Warehouse 4 (50, 90)
ABC classification [5]	25.34	60.93	144.34	204.00
ARBM [4]	21.00	47.70	108.26	155.04
PCA-cluster [16]	16.75	36.78	89.20	134.73
TGCA+S-shape	13.86	25.63	85.89	120.65
TGCA+Return	**12.66**	**22.13**	**72.26**	**92.35**
TGCA+Midpoint	14.28	23.50	83.29	118.60

TABLE 6. Test results of average one order picking distance with different storage strategies (m).

	TGCA	TGCA+Across-aisle	TGCA+Within-aisle	TGCA+Diagonal
Warehouse 4	92.35	93.72	**87.64**	89.88

TABLE 7. The average order travel distance of one order varies with the values of α and β (total number of SKUs = 10000) (m).

	$\alpha = 30$	$\alpha = 40$	$\alpha = 50$	$\alpha = 60$	$\alpha = 70$	$\alpha = 80$
$\beta = \alpha + 10$	124.80	122.99	127.11	124.24	130.26	129.0
$\beta = \alpha + 20$	122.10	127.20	123.02	124.17	126.8	129.4
$\beta = \alpha + 30$	122.78	124.59	125.37	121.81	123.1	128.5
$\beta = \alpha + 40$	122.96	124.87	**120.65**	121.66	125.3	127.8
$\beta = \alpha + 50$	120.81	126.21	129.81	126.91	128.6	128.7

We designed four warehouses with sizes ranging from small to large in table 4 to evaluate the universality of the algorithm from multiple perspectives. Warehouses 1 and 2 belong to smaller warehouses, while warehouses 3 and 4 belong to medium and large warehouses. We set each robot completes 10 orders at a time and returns to the depot.

Table 5 indicates the results of three different algorithms and our TGCA algorithm with three different routing methods under different parameters. In the TGCA algorithm, the values of α and β are in parentheses following the warehouse name. Values in Table 5 represent the average distance the robot needs to travel for each order. Experimental results show that compared with the ABC classification algorithm, ARBM, and PCA-cluster, TGCA algorithms have the shortest pickup distance in all kinds of scale warehouse. Meanwhile, different routing methods also influence experimental results. In the four warehouses of different sizes, Return routing achieves the shortest distances.

Table 6 shows the results of the proposed TGCA algorithm in different assign clusters methods. All routing methods in the experiment adopt S-shape routing. Experimental results show that the Within-aisle has the best effect.

B. PARAMETER SELECTION

The parameters to be discussed experimentally are α and β, two important parameters of clustering, determine the structure of the warehouse. Table 7 illustrates the influence of values of α and β on the effectiveness of the proposed TGCA algorithm (this experiment is based on shelves placement in Figure 1(b)). The experiment is based on warehouse 4, with a total of 23,420 orders containing 10,000 SKUs. The bold number represents the minimum average distance traveled to take an order. Experimental results show that the selection of appropriate α and β can effectively reduce the picking distance. This paper only discusses values of α and β that are multiples of 10. A more detailed value may get a better result. For the actual data used in this paper, the best results are obtained when $\alpha = 40$ and $\beta = 60$. This means that the size of each cluster is [40, 60]. In the experimental results, the minimum distance of 120.65 is nearly 10 meters less than the maximum distance of 130.26, which means that the robot can save 10 meters for each order. That also indicates that adjusting the value of parameters α and β can effectively improve the algorithm results. When the number of orders is very large, the cost savings are quite objective. On the other hand, in the TGCA algorithm, even the most unsatisfactory experimental results reduce the distance by 36.15% compared with the ABC classification algorithm, which shows the validity of the TGCA algorithm.

VI. CONCLUSION

With the ubiquitous existence of big data, this paper proposes a strategy to reduce the cost of picking up orders based on text analysis. The proposed TGCA method clustered SKUs with the same information granules into the same cluster for further order-based correlation analysis. Meanwhile, EI algebra is used to give a unique cluster description of each cluster. We defined formula 12 to determine the placement of each cluster. It further extends the association analysis, where the location of each cluster is determined by the location of the other clusters.

The context-based clustering method can widely use in various industrial activities. Text description, along with the industrial activity process makes the process, with semantic information, thus letting the whole activity easier for people to understand and manage in improving efficiency.

REFERENCES

[1] H. L. Chan, A. Pang, and K.-W. Li, "Association rule based approach for improving operation efficiency in a randomized warehouse," in *Proc. Int. Conf. Ind. Eng. Oper. Manage.*, 2011, pp. 22–24.

[2] Z. Zhao, J. Fang, G. Q. Huang, and M. Zhang, "Location management of cloud forklifts in finished product warehouse," *Int. J. Intell. Syst.*, vol. 32, no. 4, pp. 342–370, Apr. 2017.

[3] K. J. Roodbergen and R. de Koster, "Routing order pickers in a warehouse with a middle aisle," *Eur. J. Oper. Res.*, vol. 133, no. 1, pp. 32–43, Aug. 2001.

[4] M.-C. Chen, C.-L. Huang, K.-Y. Chen, and H.-P. Wu, "Aggregation of orders in distribution centers using data mining," *Expert Syst. Appl.*, vol. 28, no. 3, pp. 453–460, Apr. 2005.

[5] J. Li, M. Moghaddam, and S. Y. Nof, "Dynamic storage assignment with product affinity and ABC classification—A case study," *Int. J. Adv. Manuf. Technol.*, vol. 84, nos. 9–12, pp. 1–16, 2015.

[6] M. Chen and H. Wu, "An association-based clustering approach to order batching considering customer demand patterns," *Omega*, vol. 33, no. 4, pp. 333–343, Aug. 2005.

[7] D. M.-H. Chiang, C.-P. Lin, and M.-C. Chen, "The adaptive approach for storage assignment by mining data of warehouse management system for distribution centres," *Enterprise Inf. Syst.*, vol. 5, no. 2, pp. 219–234, May 2011.

[8] A. S. Dijkstra and K. J. Roodbergen, "Exact route-length formulas and a storage location assignment heuristic for picker-to-parts warehouses," *Transp. Res. E, Logistics Transp. Rev.*, vol. 102, pp. 38–59, Jun. 2017.

[9] A. Azadeh, S. M. Asadzadeh, and S. Tadayoun, "Optimization of operator allocation in a large multi product assembly shop through unique integration of simulation and genetic algorithm," *Int. J. Adv. Manuf. Technol.*, vol. 76, nos. 1–4, pp. 471–486, Jan. 2015.

[10] M. E. Fontana and V. S. Nepomuceno, "Multi-criteria approach for products classification and their storage location assignment," *Int. J. Adv. Manuf. Technol.*, vol. 88, nos. 9–12, pp. 3205–3216, Feb. 2017.

[11] L. Zhou, J. Liu, X. Fan, D. Zhu, P. Wu, and N. Cao, "Design of V-type warehouse layout and picking path model based on Internet of Things," *IEEE Access*, vol. 7, pp. 58419–58428, 2019.

[12] F. Chen, H. Wang, Y. Xie, and C. Qi, "An ACO-based online routing method for multiple order pickers with congestion consideration in warehouse," *J. Intell. Manuf.*, vol. 27, no. 2, pp. 389–408, Apr. 2016.

[13] V. Giannikas, W. Lu, B. Robertson, and D. McFarlane, "An interventionist strategy for warehouse order picking: Evidence from two case studies," *Int. J. Prod. Econ.*, vol. 189, pp. 63–76, Jul. 2017.

[14] H. Zhang, Z. Guo, W. Zhang, H. Cai, C. Wang, Y. Yu, W. Li, and J. Wang, "Layout design for intelligent warehouse by evolution with fitness approximation," *IEEE Access*, vol. 7, pp. 166310–166317, 2019.

[15] R. de Koster, T. Le-Duc, and K. J. Roodbergen, "Design and control of warehouse order picking: A literature review," *Eur. J. Oper. Res.*, vol. 182, no. 2, pp. 481–501, Oct. 2007.

[16] C.-L. Yang and T. P. Q. Nguyen, "Constrained clustering method for class-based storage location assignment in warehouse," *Ind. Manage. Data Syst.*, vol. 116, no. 4, pp. 667–689, May 2016.

[17] M. Moshref-Javadi, and M. R. Lehto, "Material handling improvement in warehouses by parts clustering," *Int. J. Prod. Res.*, vol. 54, no. 2, pp. 1–15, 2016.

[18] T. van Gils, K. Ramaekers, K. Braekers, B. Depaire, and A. Caris, "Increasing order picking efficiency by integrating storage, batching, zone picking, and routing policy decisions," *Int. J. Prod. Econ.*, vol. 197, no. 4, pp. 243–261, Mar. 2018.

[19] D. Battini, M. Calzavara, A. Persona, and F. Sgarbossa, "Order picking system design: The storage assignment and travel distance estimation (SA&TDE) joint method," *Int. J. Prod. Res.*, vol. 53, no. 4, pp. 1077–1093, Feb. 2015.

[20] Y. A. Bozer and F. J. Aldarondo, "A simulation-based comparison of two goods-to-person order picking systems in an online retail setting," *Int. J. Prod. Res.*, vol. 56, no. 11, pp. 3838–3858, Jun. 2018.

[21] X. Guo, Y. Yu, and R. B. M. De Koster, "Impact of required storage space on storage policy performance in a unit-load warehouse," *Int. J. Prod. Res.*, vol. 54, no. 8, pp. 2405–2418, Apr. 2016.

[22] C.-C. Jane and Y.-W. Laih, "A clustering algorithm for item assignment in a synchronized zone order picking system," *Eur. J. Oper. Res.*, vol. 166, no. 2, pp. 489–496, Oct. 2005.

[23] C.-M. Chen, Y. Gong, R. B. M. de Koster, and J. A. E. E. van Nunen, "A flexible evaluative framework for order picking systems," *Prod. Oper. Manage.*, vol. 19, no. 1, pp. 70–82, Jan. 2010.

[24] S. S. Rao and G. K. Adil, "Optimal class boundaries, number of aisles, and pick list size for low-level order picking systems," *IIE Trans.*, vol. 45, no. 12, pp. 1309–1321, Dec. 2013.

[25] J. C.-H. Pan, P.-H. Shih, M.-H. Wu, and J.-H. Lin, "A storage assignment heuristic method based on genetic algorithm for a pick-and-pass warehousing system," *Comput. Ind. Eng.*, vol. 81, no. 3, pp. 1–13, Mar. 2015.

[26] K. J. Roodbergen and R. Koster, "Routing methods for warehouses with multiple cross aisles," *Int. J. Prod. Res.*, vol. 39, no. 9, pp. 1865–1883, Jan. 2001.

[27] A. Scholz, S. Henn, M. Stuhlmann, and G. Wäscher, "A new mathematical programming formulation for the single-picker routing problem," *Eur. J. Oper. Res.*, vol. 253, no. 1, pp. 68–84, Aug. 2016.

[28] Y. Yu, R. B. M. de Koster, and X. Guo, "Class-based storage with a finite number of items: Using more classes is not always better," *Prod. Oper. Manage.*, vol. 24, no. 8, pp. 1235–1247, Aug. 2015.

[29] T.-Y. Lin, "Granular computing on binary relations I: Data mining and neighborhood systems," in *Proc. 3rd Int. Conf. Rough Sets Current Trends Comput.*, 2002, pp. 296–299.

[30] J. Leng, Q. Chen, N. Mao, and P. Jiang, "Combining granular computing technique with deep learning for service planning under social manufacturing contexts," *Knowl.-Based Syst.*, vol. 143, no. 7, pp. 295–306, Mar. 2018.

[31] L. A. Zadeh, "Fuzzy sets," *Inf. Control*, vol. 8, no. 3, pp. 338–353, Jun. 1965.

[32] Z. Pawlak, "Rough sets," *Int. J. Comput. Inf. Sci.*, vol. 11, no. 5, pp. 341–356, Oct. 1982.

[33] J. Li, C. Mei, W. Xu, and Y. Qian, "Concept learning via granular computing: A cognitive viewpoint," *Inf. Sci.*, vol. 298, no. 3, pp. 447–467, Mar. 2015.

[34] X. Liu, "The structure of fuzzy Matrices," *J. Fuzzy Math.*, vol. 2, pp. 311–325, 1994.

[35] X. Liu, W. Jia, Y. Wang, H. Guo, Y. Ren, and Z. Li, "Knowledge discovery and semantic learning in the framework of axiomatic fuzzy set theory," *WIREs Data Mining Knowl. Discovery*, vol. 8, no. 5, p. e1268, Sep. 2018.

[36] L. Xiaodong and Z. Qingling, "The EI algebra representations of fuzzy concepts," in *Proc. 4th World Congr. Intell. Control Automat.*, Jun. 2002, vol. 4, no. 4, pp. 2968–2972.

[37] X. Liu and Q. Zhang, "The fuzzy cognitive maps based on AFS fuzzy logic," *Dyn. Continuous Discrete Impuls. Syst.*, vol. 11, no. 5, pp. 787–796, 2004.

[38] L. Xiaodong, Z. Kejiu, and H.-Z. Huang, "The representations of fuzzy concepts based on the fuzzy matrix theory and the AFS theory," in *Proc. IEEE Int. Symp. Intell. Control (ISIC)*, Oct. 2003, pp. 1006–1011.

[39] P. Kukkurainen, "Fuzzy logic and Zadeh algebra," *Adv. Pure Math.*, vol. 7, no. 7, pp. 350–365, 2017.

[40] X. Liu, T. Chai, W. Wang, and W. Liu, "Approaches to the representations and logic operations of fuzzy concepts in the framework of axiomatic fuzzy set theory I," *Inf. Sci.*, vol. 177, no. 4, pp. 1007–1026, Feb. 2007.

[41] R. Agrawal, T. Imieliński, and A. Swami, "Mining association rules between sets of items in large databases," in *Proc. ACM SIGMOD Int. Conf. Manage. Data (SIGMOD)*, 1993, pp. 207–216.

[42] S. Vijayarani and S. Sharmila, "Comparative analysis of association rule mining algorithms," in *Proc. Int. Conf. Inventive Comput. Technol. (ICICT)*, Aug. 2016, pp. 1–6.

[43] B. Christian and K. Rudolf, "Induction of association rules: Apriori implementation," in *Proc. 15th Conf. Comput. Statist.*, 2002, pp. 395–400.

QI LANG received the B.S. degree from Zhengzhou University and the M.S. degree from the Dalian University of Technology, China, where she is currently pursuing the Ph.D. degree with the Research Center of Information and Control. Her research interests include natural language processing and knowledge discovery.

XUEJUN PAN (Member, IEEE) received the B.S. and M.S. degrees from the Dalian University of Technology, Dalian, China, in 1989 and 1992, respectively, and the Ph.D. degree in information science from Northeastern University, Shenyang, China, in 1998.

He is currently an Associate Professor with the Research Center of Information and Control, Dalian University of Technology. His research interests include complex industrial process control and application and intelligent control.

XIAODONG LIU received the B.S. degree in mathematics from Northeastern Normal University, Changchun, China, in 1986, the M.S. degree in mathematics from Jilin University, Jilin, China, in 1989, and the Ph.D. degree in control theory and control engineering from Northeastern University, Shenyang, China, in 2003.

He was a Senior Visiting Scientist with the Department of Electrical and Computer Engineering, University of Alberta, Edmonton, AB, Canada, in 2003, and a Visiting Research Fellow with the Department of Computing, Curtin University of Technology, Perth, WA, Australia, in 2004. He is currently a Professor with the Research Center of Information and Control, Dalian University of Technology, Dalian, China, and a Guest Professor with the ARC Research Center of Excellence in PIMCE, Curtin University of Technology, Bentley, WA, Australia. He has proposed the axiomatic fuzzy sets (AFS) theory. He is the coauthor of three books. His research interests include algebra rings, combinatorics, topology molecular lattices, AFS theory and its applications, knowledge discovery and representations, data mining, pattern recognition, and hitch diagnoses, analysis, and design of intelligent control systems. Since 1993, he has been an American Mathematical Reviewer.