Extracting an optimal set of linguistic summaries using genetic algorithm combined with greedy strategy

Lan Pham-Thi[1], Ho Nguyen-Cat[2,3], Phong Pham-Dinh[4]
[1] Faculty of Information Technology, Hanoi National University of Education
[2] Theoretical and Applied Research Institute, Duy Tan University
[3]Faculty of Information Technology, Duy Tan University
[4]Faculty of Information Technology, University of Transport and Communications
Correspondence: Lan Pham-Thi, ptlan@hnue.edu.vn
Communication: received 25 December 2020, revised 26 January 2021, accepted 31 January 2021
Digital Object Identifier: 10.32913/mic-ict-research.v2020.n2.949

***Abstract*: The goal of extracting linguistic data summaries is to produce summary sentences expressed in natural language which represent knowledge hidden in numerical dataset. At the most general level, human users can get a very large number of linguistic summaries. In this paper, we propose a model of genetic algorithm combined with greedy strategy to extract an optimal set of linguistic summaries based on the evaluation measures of goodness and diversity of the set of linguistic summaries. The experimental results on creep dataset have demonstrated the outperformance of the proposed model of genetic algorithm combined with greedy strategy in comparison with the existing genetic algorithm models in extracting linguistic summaries from data.**

Keywords: *Linguistic data summary, hedge algebras, linguistic frame of cognition, genetic algorithm, greedy strategy.*

I. INTRODUCTION

Technology is more and more developed, and data is obtained easily, and their quantity is also increasing rapidly. Therefore, data mining methods are constantly being developed to help people exploit information and knowledge hidden in giant data warehouses. Among those methods, linguistic data summarization is considered a research branch of data mining that has many useful practical applications [1]. The outputs of linguistic data summarization are summary sentences expressed in natural language in a given sentence structure, denoted by *LS* (Linguistic Summary). Each *LS* represents knowledge about the real-world objects stored in a given dataset. The form of knowledge represented in sentences in natural language is easy to understand for every human user. We study the structures of the *LS* used in most studies on linguistic data summarization, which are the sentences with linguistic quantifier proposed by Yager [2]: "*Q y are S*" or "*Q F y are S*" [2–14]. For example, "*Very few* (*Q*) sales of printers (*y*) is with *high* commission (*S*)" [10], "*Most* (*Q*) hospitals (*y*) with *very high* average hospital stay (*F*) have *very low* computer (*S*)" [8].

Human users read the linguistic summaries to understand information and knowledge in the dataset through the semantics of the words '*very few*', '*most*', '*high*', '*very low*', '*very high*' in those linguistic summaries. The linguistic quantifier *Q* represents a proportion satisfying the conclusion *S* with respect to all objects in the dataset in the first sentence sample, or it represents the objects in the group satisfying filter criterion *F* in the second sentence sample.

In the fuzzy set theory approach to extract the linguistic summary from the numerical dataset, the summaries with linguistic quantifier are considered fuzzy propositions expressing knowledge about the objects stored in the dataset. Therefore, the quality of each summary sentence is evaluated at least by a validity measure or a truth value measure. The validity measure formula uses the membership functions of the fuzzy sets representing the semantics of linguistic terms in a sentence like *'very few', 'most', 'high', 'very low', 'very high'* in the above examples. When given a dataset, the results of a linguistic data summarization algorithm are the linguistic summaries with the validity measure greater than a given threshold.

Table I shows a classification of the generality degrees in increasing the order of linguistic data summarization of Kacpryzk and Zadrożny [4]. In which, $S^{structure}$ is the structure of the summarizer *S* as "SALARY = *x*" (*x* is a linguistic term), S^{value} is a linguistic term in the summarizer

S. The degree 5 is the most general degree, when all three components Q, F, S are completely undefined in terms of attributes as well as linguistic terms, then the linguistic data summarization is equivalent to extracting fuzzy rules. Degree 5 poses the big challenge of high computational volume and the large number of linguistic summaries mined from data. However, human users can discover interesting relationships and knowledge hidden in the dataset. The studies in [8, 15–19] applied genetic algorithm to find an optimal set of linguistic summaries. Therefore, human users need to define the constraints and a quality evaluation function for the set of linguistic summaries based on the user's needs. The genetic algorithm will search for an optimal set of linguistic summaries from a set of very large number of linguistic summaries.

TABLE I
CLASSIFICATION OF LINGUISTIC SUMMARY DEGREE

Degree	Given	Search	Note
1	S	Q	Simple summarization by fuzzy query
2	SF	Q	Conditional summarization by fuzzy query
3	$Q\ S^{structure}$	S^{value}	Simple summarization towards determining value
4	$Q\ S^{structure}\ F$	S^{value}	Conditional summarization towards determining value
5	Nothing	$S\ F\ Q$	General fuzzy rules

In the studies on extracting the optimal set of linguistic summaries from the databases by genetic algorithm models [17, 19], in addition to the basic genetic operators (selection, crossover and mutation), the author applied two specific operators. The first one is *Propositions Improver operator* to improve the quality of the linguistic summaries, and the second one is *Cleaning operator* to substitute all propositions in the chromosomes having $T = 0$ with others randomly generated propositions. These two operators were applied into the model of genetic algorithm Hybird-GA to obtain better results than the basic genetic algorithm. However, the experimental results show that there are still limitations of those two operators as follows:

• The *Propositions Improver operator* is applied to substitute an existing linguistic summary with the one selected by a local search based on the best first strategy towards better validity. In the experimental results, there are still 3 out of 30 linguistic summaries with the value of $T < 0.8$, which reduces the evaluation of the goodness of the set of linguistic summaries. This result may be due to the linguistic quantifier set having only five linguistic terms *'none', 'few', 'half', 'much', 'most'*. The fuzzy sets representing the semantics of those five linguistic terms form a strong partition on the universe of discourse, so it is possible to generate a linguistic summary with validity value in the vicinity of 0.5.

• In the experimental results, there is still one out of 30 linguistic summaries with the value of $T = 0$. That is, the *Cleaning operator* has not removed all linguistic summaries having $T = 0$. The linguistic summaries have the value of $T = 0$ because there is not any record in the dataset that satisfies the filter criterion F.

To overcome the aforementioned limitations, we propose a genetic algorithm model that combines the greedy strategy in randomly generating linguistic summaries. The ideas are given as follows:

• We can extend the set of quantifier terms by adding more specificity terms according to the Linguistic Frame of Cognition (LFoC) design method based on the methodology of the hedge algebras as examined in [20]. Since the fuzzy set structure which represents the semantics of the quantifier words [20] is multi-granularity, the higher the specificity level of the quantifier word is set, the higher the chance it is to obtain the linguistic summaries with validity's values closer to 1.

• We only randomly generate the filter criterion F, which is the structure of the summarizer S. We then use the greedy strategy to determine the linguistic term in S and Q such that the validity's value of T and the semantic order of Q are as great as possible. This strategy is applied towards generating favorite linguistic summaries, i.e., increasing the quality measure of the linguistic summaries.

• The linguistic summaries with the value of $T = 0$ are the ones without any record in the dataset that satisfies the filter criterion F. Therefore, in order not show such linguistic summaries, we propose using the support measure $supp(F)$ to evaluate the cardinality of records satisfying filter criterion F. A new linguistic summary is only generated when $supp(F)$ is greater than a given threshold.

To demonstrate the effectiveness of the above mentioned proposals, we perform experiments on the dataset of *creep* and compare the results in the study [19].

The rest of the paper is organized as follows: Section II presents the related issues such as the linguistic summary with linguistic quantifier, the application of genetic algorithm to extract the optimal set of linguistic summaries, and the problem of constructing the fuzzy sets representing semantics of linguistic terms; Section III presents a new proposal of a genetic algorithm model that combines greedy strategy to generate an optimal set of linguistic summaries; Section IV shows the experimental results of one creep dataset and comparative analysis to demonstrate the effectiveness of the new proposals; Some conclusions are presented in section V.

II. SOME FUNDAMENTAL KNOWLEDGE

1. Linguistic summaries with quantifier word

Yager [2] proposes using fuzzy propositions in the structure with linguistic quantifiers expressed in natural language to represent summary information extracted from the dataset. In this section, we briefly present some concepts and notations of the problem of linguistic data summarization.

Let $Y = \{y_1, y_2, \ldots, y_n\}$ be the set of objects (records) in the dataset such as the set of customers of a bank; $A = \{A_1, A_2, \ldots, A_m\}$ is the set of attributes needed to consider objects in the set Y such as AGE, SALARY, MARITAL, etc. We denote $A_i(y_j)$ as attribute value A_i of the object y_j. The dataset is given by the set D = $\{\{A_1(y_1), A_2(y_1), \ldots, A_m(y_1)\}, \ldots, \{A_1(y_n), A_2(y_n), \ldots, A_m(y_n)\}\}$ which is the input of the problem of linguistic data summarization. The output is the linguistic summaries with the linguistic quantifiers having the general structure as follows:

$$Q\ y \text{ are } S \tag{1}$$

$$Q\ F\ y \text{ are } S \tag{2}$$

where:

- *The summarizer* S is an evaluation expressed by a linguistic word in the word domain corresponding to an attribute. For example, AGE = '*young*', SALARY = '*very high*', etc.
- *Linguistic quantifier* Q is a linguistic word representing the proportion of records that satisfies the summarizer S in the entire dataset D like in sentence form (1) or in the object group that satisfies the filter criterion F like the summary sentence of the form (2). For example, '*very few*', '*a half*', '*most*', etc.
- *Validity value* T is a value in the normalized interval [0, 1] evaluating the validity of the linguistic summaries. The value of T is considered the truth value of the fuzzy proposition with linguistic quantifier.
- *Filter criterion* F is optional to define a subgroup of objects in the set of objects Y considered in the linguistic summaries. For example, a fuzzy filter criterion in the form of AGE = '*young*', i.e., only considers the objects in the age group '*young'*.

In a general form, the filter criterion F and the summarizer S are the association of multiple linguistic predicates connected by the connector words AND/OR. Each linguistic predicate is identified by a pair of attribute – linguistic term, such as "AGE = '*young*'", "SALARY =' *high*'", etc. The linguistic summaries in the form of (1), (2) are considered fuzzy propositions with linguistic quantifier. To calculate the truth value of these propositions, it is necessary to design the fuzzy sets representing semantics of the linguistic terms in those propositions. Assume that the semantics of the linguistic terms in the components Q, F, S are represented by the fuzzy sets with respective membership functions μ_Q, μ_F and μ_S. The truth value T is calculated by the formula of Zadeh [21] for the fuzzy proposition with the linguistic quantifier as follows:

$$T(Q\ y \text{ are } S) = \mu_Q\left[\frac{1}{n}\sum_{i=1}^{n}\mu_S\left(y_i\right)\right] \tag{3}$$

$$T(Q\ F y \text{ are } S) = \mu_Q\left[\frac{\sum\limits_{i=1}^{n}\left(\mu_F\left(y_i\right) \wedge \mu_S\left(y_i\right)\right)}{\sum\limits_{i=1}^{n}\mu_F\left(y_i\right)}\right] \tag{4}$$

The truth value T is the basic measure used to evaluate the quality of the linguistic summaries. Therefore, only the linguistic summaries with the truth value of T greater than a given threshold δ are extracted, for example, δ = 0.85 [17] or δ = 0.8 [18]. These are considered linguistic summaries representing information and knowledge hidden in the dataset. In addition, some other evaluation measures are proposed in [4, 22] such as imprecision, covering, focus, and appropriateness. The formulas for calculating these evaluation measures are also based on the membership functions of the fuzzy sets that represent the semantics of linguistic terms appeared in the linguistic summaries.

2. The genetic algorithm extracts the optimal set of linguistic summaries from a given dataset

In several studies of extracting the sets of linguistic summaries from relational databases, we are interested in the research of Donis-Diaz et al. in [17, 19], where the optimal set of linguistic summaries is selected based on *goodness* and *diversity*. The authors coded each linguistic summary into a gene, and each individual is a set of linguistic summaries. In addition to basic genetic operators, the *Propositions improver* operator is added to create an enhanced genetic models [17].

The *Propositions improver* operator uses the neighborhood search technique to replace an existing linguistic summary by a better one. In [19], the *Cleaning operator* is used to replace the linguistic summaries with the truth value of T = 0 by the other randomly generated ones. Two operators - *Cleaning* and *Improver* - are added to the genetic algorithm to make a more efficient hybrid genetic model than the basic genetic algorithms.

Donis-Diaz et al. [17] evaluate the goodness of a good linguistic summary by the value of goodness Gn calculated by formula (5), where T_1, T_2, T_3, T_4 are the truth, imprecision, covering and appropriateness, respectively. Doniz-Diaz et al. in [19] evaluate the goodness of a linguistic summary according to formula (6). In both studies, the goodness of a set of linguistic summaries Gd are calculated by the average of the goodness of the linguistic summaries in the set by formula (7), where l is the number of linguistic summaries in the set.

$$Gn = 0.4 \times T_1^{St} + 0.1 \times T_2 + 0.25 \times T_3 + 0.25 \times T_4 \quad (5)$$

$$Gn = T \times St(Q) \quad (6)$$

$$Gd = \frac{\sum_{i=1}^{l} Gn_i}{l} \quad (7)$$

where $T_1^{St} = T_1 \times St(Q)$ shows the concept of *Linguistic Strength*, and $St(Q)$ is the weight of the linguistic quantifier Q pre-specified based on the priority evaluation of the linguistic quantifiers. Specifically, the parameters used in [17, 19] are St (Most) = 1, St (Much) = 0.75, St (Half) = 0.20, St (Some) = 0.15, St (Few) = 0.05. Thus, the larger proportion the linguistic quantifier represents, the greater the weight.

The diversity of a set of linguistic summaries is calculated by formula (8) in both papers [17, 19], where C is the number of clusters, and l is the number of linguistic summaries in the set.

$$De = \frac{C}{l} \quad (8)$$

C is the number of clusters when clustering linguistic summaries based on the similarity function L as follows:

$$L(p1, p2) = \begin{cases} Yes & if \sum_{k=0}^{m} H(p1_k, p2_k) < 2 \\ No & in\ other\ case \end{cases} \quad (9)$$

The two linguistic summaries $p1$ and $p2$ extracted from the database comprise m attributes represented by a numeric vector consisting of $(m + 1)$ elements. The elements $p1_0$ and $p2_0$ are indexes of the linguistic quantifier Q in $Dom(Q)$, the elements $p1_i$, $p2_i$ are indexes of the linguistic terms in $Dom(A_i)$ of the vector representing the linguistic summaries $p1$, $p2$ ($Dom(A_i)$ - the linguistic domain of the attribute A_i). If the attribute A_i is not present in a linguistic summary, the i^{th} element in the vector representing the linguistic summary takes the value 0. When the result of the function $L(p1, p2)$ is '*yes*', two linguistic summaries $p1$ and $p2$ are similar. The function $H(p1_k, p2_k)$ is calculated by formula (10) to compare the k^{th} element in two vectors. The k^{th} element is different (the value of function $H(p1_k, p2_k) = 1$) when: (1) $p1_k = 0$ and $p2_k \neq 0$; $p1_k \neq 0$ and $p2_k = 0$ (the attribute A_k is present in one linguistic summary, not in the remaining one); (2) the attribute A_k is present in both linguistic summaries, but the index of the two terms are different. Two indices of the terms in the same $Dom(A_k)$ are considered distinct when they are in two positions in ascending semantic order greater than 20% of the number of words in $Dom(A_k)$. For example, if $Dom(A_k)$ = {*'very low', 'low', 'little low', 'medium', 'little high', 'high', 'very high'*}, the term '*low*' at position 2 and the term '*medium*' at position 4 have their distance |2 - 4| > 20%*7 = 1.4. Therefore, two terms '*low*' and '*medium*' are considered distinct.

$$H(p1_k, p2_k) = \begin{cases} 1 & \begin{array}{l} if\ |p1_k - p2_k| > \\ \quad round(0.2 * size(Dom(A_k)) or \\ if\ p1_k = 0\ and\ p2_k \neq 0\ or \\ if\ p1_k \neq 0\ and\ p2_k = 0 \end{array} \\ 0 & in\ other\ case \end{cases} \quad (10)$$

From that, the fitness function Fit of an individual, which corresponds to a set of linguistic summaries, is the weighted sum of two measures: Gd (the goodness of a set of linguistic summaries) and De (diversity) according to the formula as follows:

$$Fit = m_g Gd + m_d De \quad (11)$$

where: m_g, m_d are the weights of two measures Gd and De satisfying the condition $m_g + m_d$ =1. The authors in [19] select m_g = 0.7, m_d = 0.3, i.e., the goodness of the set of linguistic summaries is weighted more than 2 times the diversity of the entire linguistic summaries.

3. Fuzzy set based semantics representation of terms ensures the interpretability of the content of the linguistic summaries

a) The interpretability of the content of the linguistic summaries

In the studies on linguistic data summarization based on fuzzy set theory, the word domain of each attribute is usually limited to 7 ± 2 words, the fuzzy sets representing their semantics are often in the form of strong and uniformly distributed partitions. Figure 1 is an example of the fuzzy set design for the attributes of the patient database in the paper of Almeida et al. [9]. Figure 1(a) includes five fuzzy sets

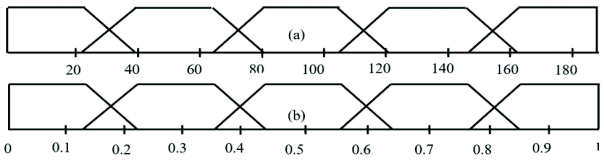

Figure 1. Examples of fuzzy set representation that form strong partitions on the reference domain.

representing the semantics of five terms '*very low*', '*low*', '*medium*', '*high*' and '*very high*' of the attribute "Heard rate". Figure 1(b) includes five fuzzy sets representing the semantics of five quantifier words '*very few*', '*few*', '*half*', '*most*' and '*almost all*'. The linguistic summary extraction algorithm handles directly the membership functions of the fuzzy sets, so they play a decisive role in the output of the algorithm.

Assume that when considering a linguistic summary, denoted by *LS*, extracted from database *D*, the information content assigned to *LS* from the output of the linguistic summary extraction algorithm *M* interacting on the fuzzy sets representing the semantics of linguistic terms in the *LS* is $Cont_{M,D}(fs\_REP(LS))$. The human user interprets the linguistic summary *LS* as a sentence in natural language and receives the information as $Cont_D(LS)$. Determining the conditions to ensure $Cont_{M,D}(fs\_REP(LS)) = Cont_D(LS)$ is considered the problem of examining the interpretability of content of the linguistic summary. However, in the existing studies, the linguistic terms are only considered as linguistic labels assigned to the fuzzy sets designed based on the intuition of the designer of the algorithm *M* and the number of them is limited by 7 ± 2. Thus, when there is no formal formalism to ensure the linkage between the semantics of the linguistic terms and the designed fuzzy sets, it will not ensure that the human user can correctly interpret the content of the linguistic summaries received from the linguistic summary extraction algorithm. To overcome this limitation, in the study [20], the authors have proposed a method of designing fuzzy set structure to ensure the correct semantic representation of terms based on the formalism of hedge algebras theory. In which, the semantics represented by the fuzzy set is designed from qualitative semantics, preserving inherent semantic relationships of the terms in the entire term domain $Dom(A)$ of each attribute *A*. The summary of the method in [20] is as follows:

• Step 1: Determine the syntactic structure, the qualitative semantics of the set of terms $Dom(A)$ of each attribute *A* by a hedge algebras structure. A subset of those terms forms a Linguistic Frame of Cognition (LFoC) $\mathcal{F}_A$ for each attribute *A*.

• Step 2: Based on the inherent semantics of terms in $Dom(A)$, the multi-semantic structure based on the semantic order and the generality - specificity relationships in $Dom(A)$ and $\mathcal{F}_A$ is discovered. Simultaneously, $\mathcal{F}_A$ is scalable in the system development process in practice.

• Step 3: Propose a procedure to design the trapezoidal fuzzy sets from the independent fuzzy parameter set of the hedge algebras structure in Step 1. The trapezoidal fuzzy sets form a structure preserving the multi-semantic structure discovered in Step 2. Simultaneously, the trapezoidal fuzzy set structure of $\mathcal{F}_A$ is also scalable.

b) Multi-semantic structure and the scability of linguistic frame of cognition of each attribute

Based on the viewpoint of hedge algebras, each term domain of the attribute *A*, denoted by $Dom(A)$, is an algebraic structure based on the semantic order relationship (notation ≤) between words. The elements in $Dom(A)$ are induced from two primary terms (considered generator elements) $c^- \leq c^+$, e.g, when considering the attribute *AGE*, c^- = '*young*' and c^+ = '*old*'. The linguistic hedges such as '*very', 'little'*, etc. are used to generate new terms with the semantic order determined based on the semantic change tendency to when the hedges act on the primary terms. We have '*very young*' ≤ '*young*' ≤ '*little young*', but '*little old*' ≤ '*old*' ≤ '*very old*'. Denote the set of hedges by *H*. In general, each term in $Dom(A)$ has the string representation of the form ωc^*, where ω is a sequence of hedges (i.e., $\omega \in H^*$). The length of ωc^* is $|\omega| + 1$, for example, the length of '*very very young*' is 3. In $Dom(A)$, there are 3 constant elements corresponding to the smallest, the neutral and the largest in semantic order. Denote the constant elements are $\boldsymbol{0}, W, \boldsymbol{1}$ respectively; where $\boldsymbol{0} \leq W \leq \boldsymbol{1}$. Since the semantics of the terms need to be determined in the context of the entire $Dom(A)$, the authors in [23] introduce the definition of Linguistic Frame of Cognition (LFoC) of an attribute *A* as follows:

Definition 1: An LFoC of attribute *A*, denoted by $\mathcal{F}_A$, is a set of terms that satisfy the following conditions [23]:

(i) $\{\boldsymbol{0}, c^-, W, c^+, \boldsymbol{1}\} \subseteq \mathcal{F}_A$

(ii) $hx \in \mathcal{F}_A \Rightarrow (\forall h' \in H,\ h'x \in \mathcal{F}_A\ )$

(iii) $x \in \mathcal{F}_A$ và $x = hx'$ $(h \in H) \Rightarrow x' \in \mathcal{F}_A$

From the definition of $\mathcal{F}_A$, each LFoC of an attribute *A* has the form $\mathcal{F}_{A,\kappa} = X_{(\kappa)}$, where $X_{(\kappa)}$ is the set of all words with a length smaller than κ. Then, κ determines the specificity of LFoC. For simplicity, we denote $\mathcal{F}_{A,\kappa}$ as $\mathcal{F}_\kappa$ when *A* is clearly specified.

Based on inherent semantics of the linguistic terms, in $Dom(A) = X$ and $\mathcal{F}_\kappa$ there are the whole semantic order relation ≤ and the generality – specificity partial relation $\mathcal{G}$ between the terms. Two of those semantic relations in conjunction with *X* and $\mathcal{F}_\kappa$ form the multi-semantic

structure $S_{\leq,\mathcal{G}} = (X, \leq, \mathcal{G})$ and $F_{\leq,\mathcal{G}} = (\mathcal{F}_\kappa, \leq, \mathcal{G})$, respectively. Furthermore, when it is necessary to extend LFoC by adding more terms with higher speficity level, we increase the level of κ in LFoC. The semantic order relation and the generality - specificity relation of the existing terms have not been changed. That is, their semantics are not changed when LFoC grows. This scalability is in line with the requirements set out in practice when human users use the set of linguistic terms of attributes in real world perception.

c) Construct the computational semantics to ensure the interpretability and scalability of the linguistic frame of cognition of each attribute

Trapezoidal fuzzy sets are used to represent semantics of linguistic terms in most studies of linguistic data summarization. In the enlarged hedge algebras [24], the interval quantifying mapping value is f: $Dom(A) \rightarrow P[0,1]$ ($P[0,1]$ – which are all subsets of [0, 1]) which allow us to define an interval semantics core of the terms.

The interval semantics core of term x is used as the small base of the trapezoid representing the semantics of x. In [20], the authors propose a procedure for constructing trapezoidal fuzzy sets of terms in $\mathcal{F}_{A,\kappa}$ from the independent semantic parameter set of an enlarged hedge algebra. The trapezoidal fuzzy sets form a multi-granularity structure as illustrated in Figure 2 for a LFoC $\mathcal{F}_{A,3}$, consisting of 3 levels, the terms are induced from the hedge algebras structure with the hedge set $H = \{L\ (\text{'}little\text{'}), V(\text{'}very\text{'})\}$.

The trapezoid set $\mathrm{T}(\mathcal{F}_\kappa)$ is designed to form a multi-granularity structure that preserves the multi-semantic structure of $F_{\leq,\mathcal{G}} = (\mathcal{F}_\kappa, \leq, \mathcal{G})$ and is scalable as LFoC $\mathcal{F}_\kappa$. Hence, they are considered isomorphic images of $\mathcal{F}_\kappa$. These conclusions have been stated into theorems and proved in [20]. We summarize them as follows:

- *Preserving two semantic structural relations*: two inherent semantic-based relations of terms are the semantic order relation $\leq$ and the generality - specificity $\mathcal{G}$. These relations are preserved when mapping from the set of words in $\mathcal{F}_\kappa$ to the set of trapezoid $\mathrm{T}(\mathcal{F}_\kappa)$. We denote $Tr(x)$ as a trapezoid corresponding to the term x. If $x, y \in \mathcal{F}_\kappa$ and $x \leq y$, then $Tr(x) \leq Tr(y)$ because the small base of $Tr(x)$ is at the left of the small base of $Tr(y)$ and at least one of the two end-points of the large base of $Tr(x)$ is smaller than the corresponding one of $Tr(y)$. If $x = hy$ ($h \in H$) (x has the specificity greater than y), then the large base of $Tr(x)$ lies in the inner large base of $Tr(y)$. This is easily seen in the illustration in Figure 2.

- *Scalability*: When LFoC is extended, it means increasing the specificity κ, that is, adding all words that have the specificity level $\kappa + 1$ to the LFoC that has the current specificity level κ. The practical requirement requires that the newly added words do not change the semantics of words already present in the LFoC. With the multi-granularity structure shown in Figure 2, the design of fuzzy sets of words with the specificity level $\kappa + 1$ will not change the fuzzy sets based semantics of words at the specificity level $\kappa + 1$. This is not possible if the fuzzy sets that are based on semantics of LFoC form strong partitions as shown in Figure 1.

Based on the formalism of the hedge algebras theory, we will use the fuzzy sets designed as the proposed method as in paper [20] for the proposed linguistic summary extraction algorithm in Section III and the experiments in Section IV to ensure the interpretability of the content of the linguistic summaries. This is an advantage of the fuzzy set design to ensure the interpretability of the linguistic summary content in this paper, which is not considered in the studies on extracting the optimal set of linguistic summaries by applying the variant of genetic algorithms.

III. GENETIC ALGORITHM COMBINED WITH GREEDY STRATEGY FOR EXTRACTING THE OPTIMAL SET OF LINGUISTIC SUMMARIES

In this study, we propose a genetic algorithm model combined with the greedy strategy to find an optimal set of linguistic summaries extracted from a relational database. The criteria for selecting the optimal set of linguistic summaries in studies [17, 19] are the combination of the goodness and the diversity of the linguistic summary set. This greedy idea is demonstrated in the Random-Greedy-LS procedure to extract an optimal linguistic summary as described in Subsection III.2 below. This procedure is used in genetic algorithm to model genetic algorithm Greedy-GA to extract an optimal set of linguistic summaries as presented in Subsection III.3.

1. Determine the LFoC and the fuzzy sets based semantics of linguistic terms

We use the method of determining the set of linguistic terms for each attribute and designing the trapezoidal fuzzy sets based semantics described in Subsection II.3 to ensure the interpretability of the information content of the linguistic summaries. Firstly, it is necessary to define the syntactic and qualitative semantics of terms in the LFoC $\mathcal{F}_{A,\kappa}$ of attribute A. Information needed include:

- Two generator terms $c^-$and c^+, three term constants $\mathbf{0}$, W, $\mathbf{1}$.
- The linguistic hedges in H and the relative sign table between hedges.

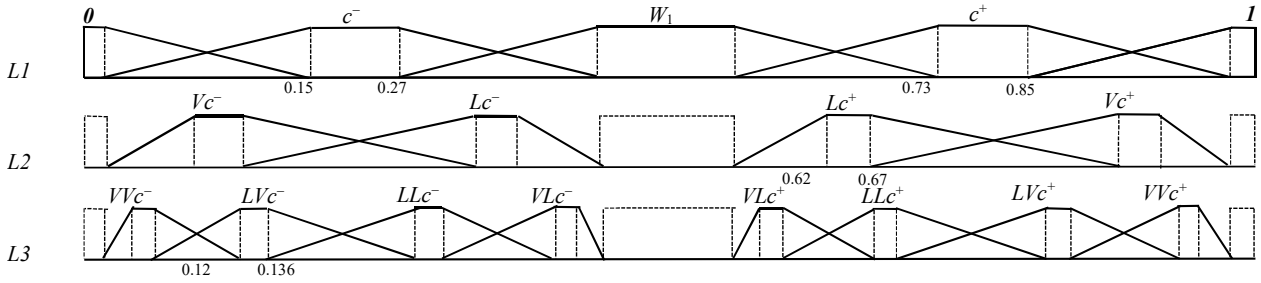

Figure 2. Multi-granularity representing trapezoidal fuzzy sets of terms of an LFoC with 3 levels.

- The speficity level of $\mathcal{F}_{A,\kappa}$, i.e., the value of κ (the maximum length of terms in $\mathcal{F}_A$)

Secondly, to design all trapezoidal fuzzy sets representing the semantics of words in $\mathcal{F}_{A,\kappa}$, it is necessary to provide a set of independent semantic parameter values of the enlarged hedge algebra for attribute A that includes:

- Four fuzzines measure values of three term constants and a generator term c^+ such that they satisfy the constraint: $fm(\mathbf{0}) + fm(c^-) + fm(W) + fm(c^+) + fm(\mathbf{1}) = 1$.
- Fuzziness parameter values of the hedges in H satisfying the constraint $\sum_{h\in H}^{\mu}(h) + \mu(h_0) = 1$, where h is the artificial hedge h_0 which is used to induce the semantics core of the terms.

The set of hedges used in this paper includes a negative hedge '*little*' and a positive hedge '*very*'. In case the specificity level $\kappa = 2$, LFoC has 9 words, i.e., $X_{(2)} = \{\mathbf{0}, Vc^-, c^-, Lc^-, W, Lc^+, c^+, Vc^+, \mathbf{1}\}$. In case the specificity level $\kappa = 3$, LFoC has 17 words, i.e., $X_{(2)} \cup \{VVc^-, LVc^-, LLc^-, VLc^-, VLc^+, LLc^+, LVc^+, VVc^+\}$.

The set of linguistic quantifiers Q plays an important role in extracting linguistic summaries. When the body of the linguistic summary has been specified, that is, the filter criterion F and the summarizer S have been specified, the validity value of the linguistic summary depends on the selection of the linguistic quantifier Q. Since the fuzzy sets are structured in the form of multi-granularity as shown in Figure 2, the set of linguistic quantifier Q consisting of more terms at a greater level of specificity will increase the chance of selecting the linguistic quantifier which allows us to obtain the linguistic summary with higher validity value T.

2. Generate the optimal linguistic summaries based on greedy strategy

Consider the example of an extended linguistic summary in (2): "Q y that (AGE = '*young*' AND SCORE_TEST = '*very high*') are "SALARY = x". In this example, we determine a group of objects satisfying the filter criterion F (AGE = '*young*' AND SCORE_TEST = '*very high*'), the structure of the summarizer S is "SALARY = x", where x is a term in LFoC $\mathcal{F}_{SALARY}$. For each $x \in \mathcal{F}_{SALARY}$, we determine a linguistic summary body, then calculate the value of $r(x)$, which is the proportion of objects satisfying the summarizer "SALARY = x" in the group of objects satisfying the filter condition in the following formula:

$$r(x) = \frac{\sum_{i=1}^{n}(\mu_{young}(o_i) \wedge \mu_{very\_high}(o_i) \wedge \mu_x(o_i))}{\sum_{i=1}^{n}(\mu_{young}(o_i) \wedge \mu_{very\_high}o_i))} \quad (12)$$

where, $o_i \in \mathcal{D} = \{o_1, o_2, \ldots, o_n\}$ is the i^{th} record in the database. From now on, the notation o is used instead of y to avoid confusion with the notation y of the linguistic terms.

Then, choose the term $Q(x)$ such that the validity value $T(x) = \mu_Q(r(x))$ reaches the maximum value. That is, choose the linguistic quantifier that best represents the proportion $r(x)$. When there are many linguistic quantifiers producing the same greatest $T(x)$ value, choose the term with the greatest semantic order.

When $r(x)$ is bigger, $Q(x)$ is chosen with greater semantic order, that is, if $r(x_1) < r(x_2)$ then $Q(x_1) \leq Q(x_2)$. From the formula for evaluating the goodness of a linguistic summary, [19] $Gn = T.St(Q)$, where $St(Q)$ is the priority weight of the terms Q satisfying the condition $Q_1 < Q_2$, then $St(Q_1) < St(Q_2)$. Therefore, if $x^* \in \mathcal{F}_{SALARY}$ then $r(x^*)$ reaches the maximum value, which means there is the largest number of objects in the group satisfying the filter criterion F that SALARY is at x^*, then $Q^* \in Dom(Q)$ is chosen so that the linguistic summary "Q^*y that (AGE = '*young*' AND SCORE_TEST = '*very high*') are SALARY = x^*" has the highest priority $St(Q^*)$ of Q^*.

Referring to the fuzzy association rule, the value of $r(x)$ is the confident measure of the fuzzy association rule of the form "IF (AGE = '*young*' AND SCORE_TEST = '*very*

high') THEN SALARY = x". Therefore, the idea of selecting x^* as in the above example gives a linguistic summary where the summarizer shows the most common property of the attribute SALARY of the object group determined by the filter criterion (AGE = '*young*' AND SCORE_TEST = '*very high*'). When the filter criterion F is completely determined (including the attribute and the corresponding linguistic value), the structure S is determined (the attribute is determined, the term is undefined), we only give out a linguistic summary with the linguistic summarizer x^* that satisfies the following conditions:

(C1): $r(x^*)$ reaches maximum value;

(C2): the validity value T is maximal;

(C3): linguistic quantifier Q^* has maximum order.

Such greedy strategy will make the goodness Gn of each linguistic summary increase, which simultaneously makes the goodness Gn of the whole set of linguistic summaries increase. Moreover, when each group of subjects satisfies the filter criterion F, giving out only one conclusion will also increase the diversity of the set of linguistic summary as evaluated by formula (8). Thus, the quality of the set of linguistic summaries evaluated by formula (11) will also increase.

In the process of extracting the optimal set of linguistic summaries by genetic algorithm, the study [19] used *Cleaning operator* to replace the linguistic summaries with the validity of T = 0 by the other random linguistic summaries. The linguistic summaries with T = 0 correspond to the ones whose filter criterion F includes many AND clauses of linguistic predicates, so there is not any record in the database satisfying the filter criterion F.

The experimental results in [19] show that after an average of 10 runs of the Hybird-GA genetic model using *Cleaning* and *Improver* operators, there is still a linguistic summary with $T = 0$ in the optimal set of linguistic sentences which has 30 sentences in total. In order to not show such linguistic summaries, we propose using the support measure $supp(F) = \sum_{i=1}^{n} \mu_F(y_i)/n$ (n is the number of records in the database) to evaluate the cardinality of the object group satisfying the filter criterion F. We only generate the linguistic summary with the criterion F when the measure support $supp(F)$ is greater than a given threshold β. Thus, the optimal set of linguistic summaries will include summarizers about groups of objects whose cardinality is greater than a threshold β or has the diversity greater than a given threshold.

We use the linguistic summary pattern shown in [20] as follows:

"Qos are $o(E_s)$," and "Qos that are $o(F_q)$ is $o(E_s)$" (13)

where $o(E_s)$ = "$o(A_{s1})$ is/has x_{s1} AND ... AND $o(A_{sm})$ is/has x_{sm}" is the summarizer corresponding to S in (1) and (2); $o(F_q)$ = "$o(A_{q1})$ is/has x_{q1} AND ... AND $o(A_{qh})$ is/has x_{qh}" is the filter criterion in the linguistic summary corresponding to F in (2); A_{ij} is an attribute and x_{ij} is a linguistic term in LFoC $\mathcal{F}_{Aij}$.

From the above analysis, general greedy strategy is applied to select a linguistic summary that is implemented according to the following idea:

- *Step 1*: Randomly generate the filter criterion $o(F_q)$ (include corresponding attribute and linguistic term). Calculate the support measure of $o(F_q)$ by the formula $supp(o(F_q)) = \sum_{i=1}^{n} \mu_{F_q}(y_i)/n$ (n is the number of records in the database). If $supp(o(F_q)) > \beta$, $o(F_q)$ is accepted, go to Step 2. Otherwise, randomly generate another filter criterion $o(F_q)$.

- *Step 2*: Randomly select the attributes in $o(E_s)$ with the given amount, scan the term combinations in LFoC of the attributes $o(E_s)$ to find term combination where the expression $r = \frac{\sum_{o(Fq)}^{\mu} \wedge \mu_{o(Es)}}{\sum_{o(Fq)}^{\mu}}$ reaches the maximum value.

- *Step 3: Select a linguistic quantifier* Q^* in LFoC $\mathcal{F}_Q$ such that the value $T = \mu_{Q^*}(r)$ reaches maximum value. If there are many terms Q^* that make T maximized, select term Q^* with the greatest semantic order.

Step 1 selects the filter criterion $o(F_q)$ satisfying the threshold $supp(o(F_q)) > \beta$, and step 2 selects the term in the summarizer $o(E_s)$ that is the most popular for the object group satisfying $o(F_q)$ according to the defined structure. Step 3 selects the term Q to have the greatest T and the greatest $St(Q)$ (corresponding to the greatest semantic order of Q). It results in the obtained linguistic summary towards larger goodness measure of Gn in the linguistic summaries with the same $o(F_q)$ and the same structure $o(E_s)$.

The procedure for generating linguistic summaries using greedy strategy is described as Algorithm 1.

3. Genetic algorithm combined with greedy strategy for extracting the optimal set of linguistic summaries

a) The object coding in genetic algorithm

Each gene represents a linguistic summary including the following components:

- The filter criterion $o(F_q)$: includes pairs of (ql_i, vq_i), where ql_i is the index of the attribute in the attribute list of the database, and vq_i is the index of the term in LFoC of the attribute at the index ql_i

- The summarizer $o(E_s)$: is similar to the filter criterion F, which includes the pairs of (sm_i, vs_i).

- The linguistic quantifier is the index q_i of linguistic quantifier in LFoC $\mathcal{F}_Q$.

Algorithm 1: Procedure Random-Greedy-LS.

Inputs: Database D, LFoC $\mathcal{F}_A$ và T($\mathcal{F}_A$), where A is an attribute of D, the summary pattern "Qos that are $o(F_q)$ is $o(E_s)$", threshold β.
Outputs: A linguistic summary LS satisfying $supp(o(F_q)) \geq \beta$, the maximum value $r = \frac{\sum^{\mu}_{o(Fq)} \wedge \mu_{o(Es)}}{\sum^{\mu}_{o(Fq)}}$, the maximum value T, Q with the greatest semantic order in the linguistic summaries having the same $o(F_q)$ and $o(E_s)$.
begin
  do
    $o(F_q) \leftarrow$ Random_List((A_{q1}, x_{q1}), …, (A_{qh}, x_{qh}));
  while $supp(F_q) \geq \beta$;
  Random_List(A_{s1}, …, A_{sm});
  $(x_{s1}, \ldots, x_{sm}) \leftarrow (x_{s1}, \ldots, x_{sm}) \in \mathcal{F}_{A_{s1}} \times \ldots \times \mathcal{F}_{A_{sm}}$ AND maximize $r = \frac{\sum^{\mu}_{o(Fq)} \wedge \mu_{o(Es)}}{\sum^{\mu}_{o(Fq)}}$
  $Q \leftarrow Q \in \mathcal{F}_Q$ AND maximize $\mu_Q(r)$ AND maximize $St(Q)$
  return "Qos that are $o(F_q)$ is $o(E_s)$";
end

Algorithm 2: The scheme of genetic algorithm combined with greedy strategy.

for i = 1 to $size\_generation(P)$ **do**
  Add(P, Random-Greedy-LS);
end
evaluate(P);
while *termination criterion not satisfied* **do**
  createEmpty(P');
  add(P', selectElitistChromosomes(P));
  while *P' is not full* ***do*** **do**
    Parent ← select(P);
    Children ← crossover(*Parent*);
    add(P', *Children*);
    evaluate(*Children*);
    end;
  end
  while *mutate criterion satisfied* **do**
    individual ← chooseRandom(P');
    Replace_Genes(*individual*, Random-Greedy-LS);
    end;
  end
  $P \leftarrow P'$;
  end;
end

- The truth value T of linguistic summary.

Each individual (chromosome) represents a set of linguistic summaries consisting of many different genes. Each generation consists of many different individuals.

TABLE II
AN ILLUSTRATION OF THE STRUCTURE OF A GENE REPRESENTS A LINGUISTIC SUMMARY.

q_i	(ql_1, vq_1)	(ql_2, vq_2)	….		(sm_1, vs_1)	(sm_2, vs_2)	…	T

b) The fitness function

The fitness function *Fit* of each individual that represents a set of linguistic summaries is a weighted aggregate measure of two measure: the goodness of the linguistic summary *Gd* in formula (7), and the diversity of the set of linguistic summaries *De* in formula (8). The value of fitness function *Fit* is a value in the interval [0, 1] calculated by formula (11). The best individual in the last generation is selected as the solution to the problem when that individual has the greatest value of *Fit*.

The weight of linguistic quantifier $St(Q)$ in the linguistic summary as in studies [17, 19] is applied to the linguistic quantifiers with the specificity level 1 in the set {**0**, *few, a half, many,* **1**}. The linguistic quantifiers with the specificity level 2 and level 3 are assigned the weighted value such that: if two linguistic quantifiers Q and Q' satisfy the condition $Q < Q'$, then $St(Q) < St(Q')$.

c) The genetic operators

The basic genetic operators are used as follows:

- Selection operator: with each evolution, a proportion of the best individuals (the fitness value *Fit* is greatest) in the current generation are selected for the next generation.

- The crossover operator: one point crossover operator is applied to two randomly selected individuals to generate two offspring. The crossover operator swaps genes between two individuals, i.e. swaps two linguistic summaries between two sets of linguistic summaries. Crossover operator changes the diversity measure of the sets of linguistic summaries, not the goodness of each linguistic summary, but the goodness of the set of linguistic summaries.

- The mutation operator: a small proportion (usually around 0.05) alters some of the genes in a randomly selected individual with a newly randomly generated gene, i.e., replacing some linguistic summaries in the set of summaries. The mutation operator changes both the goodness *Gd* and the diversity *De* of the set of linguistic summaries.

The genetic algorithm scheme combined greedy Greedy-GA strategy is shown in Algorithm 2. In which, the procedure Random-Greedy-LS (as shown in Algorithm 1) is used to generate a linguistic summary using the greedy strategy as presented in Section III.2.

In the genetic algorithm model Greedy-GA proposed in this paper, at the initial generation step, all linguistic summaries (the genes of individuals) are generated by the procedure using the greedy strategy Random-Greedy-LS. In the process of applying the genetic operators to the evolutionary cycles, the selection and crossover operators do not change the linguistic summaries. They just swap the linguistic summaries between the different sets of linguistic

summaries. The mutation operator replaces some linguistic summaries with new ones which are also generated by the procedure Random-Greedy-LS. Thus, all linguistic summaries in the entire executing process of algorithm Greedy-GA are generated by the procedure Random-Greedy-LS. As analyzed in Section III.2, these linguistic summaries tend to increase the fitness function values of the individuals. That is, the results of the proposed algorithm Greedy-GA will give a set of better linguistic summaries according to the adaptability evaluated by the fitness function *Fit* as in formula (11).

IV. EXPERIMENT

In the experiment, we implement the proposed genetic algorithm that combines greedy strategy model Greedy-GA as presented in section III. To demonstrate the advantages of the new proposal in Greedy-GA, the database *creep* is used. The summary patterns and the genetic algorithm parameters are the same as in the model hybrid-GA in study [19] to compare and evaluate the results of extracting the optimal set of linguistic summaries.

1. Database and sentence patterns

The database used in the experiment is *creep* of steel annealing as in the study of Donis-Diaz [19]. The database includes 2066 records and 30 attributes. In which, the attribute *CREEP* represents the strength of steel. There are 19 attributes of chemicals in steel and 6 attributes of temperature. The linguistic summaries are extracted in the form of sentences as in [19], as follows:

- The filter criterion F is a combination of pairs of (*att, val*), each pair (*att, val*) representing a linguistic predicate, where *att* is an attribute in 19 attributes of chemicals or 6 attributes of temperature. The authors in [19] have shown that when the filter criterion F has more than 6 pairs (*att, val*), there is almost no record that satisfies the filter criterion F. Therefore, in this experiment, F will be a combination of no more than 6 pairs of (*att, val*).

- The summarizer S is in the form of 'CREEP= x', where x is a term in LFoC $\mathcal{F}_{CREEP}$.

2. Linguistic of cognition of the attributes and the linguistic quantifiers Q

We use a simple hedge algebra structure which includes two generator terms, 3 term constants, a negative hedge '*little*', and a positive hedge '*very*'. The linguistic frame of cognition of the attributes is $\mathcal{F}_{A,3}$, which consists of 3 specificity levels and 17 linguistic terms. Fuzzy sets representing semantics of terms are represented by the multi-granularity structure shown in Figure 2.

The attribute *CREEP* has a value domain of [13, 550], and the studies in [5, 19] have shown that values between 330 and 550 are considered ideal. The authors used 9 trapezoidal fuzzy sets to represent semantics of 9 terms in *Dom*(*CREEP*), in which the fuzzy set representing the term '*ideal*' (the term has the greatest semantic order in *Dom*(*CREEP*)) has a small base in the interval from 330 to 550, the remaining 8 fuzzy sets are distributed uniformly in the interval from 13 to 330. Thus, we select the following set of parameters for the attribute *CREEP* as follows: *fm*(***0***) = 0.0195; *fm*(*low*) = 0.2832; *fm*(*medium*) = 0.0273; *fm*(*high*) = 0.2793; *fm*(***1***) = 0.3906; $\mu(L) = 0.4$; $\mu(h_0) = 0.25$; $\mu(V) = 0.35$. Then, the trapezoid represents semantics of the term ***1*** (the word with the greatest semantic order in LFoC $\mathcal{F}_{CREEP}$) has a small base coinciding with the small base of the trapezoid representing semantics of the term '*ideal*' in [5, 19]. The trapezoids representing the semantics of the term ***0***, *low*, *medium*, *high,* form a uniform partitions in the interval from 13 to 330 of the reference domain.

The trapezoidal fuzzy sets represent the semantics of terms of the attributes of time, temperature, and chemicals in [19] that form the strong partitions on their reference domains. Therefore, we choose a balanced fuzziness parameter value set for those attributes as follows: *fm*(***0***) = 0.03; *fm*(*low*) = 0.42; *fm*(*W*) = 0.1; *fm*(*high*) = 0.42; *fm*(***1***) = 0.03; $\mu(L) = 0.4$; $\mu(h_0) = 0.25$; $\mu(V) = 0.35$.

The set of fuzziness parameter value of the set of linguistic quantifier Q is determined as follows: *fm*(***0***) = 0.03; *fm*(*few*) = 0.42; *fm*(*a half*) = 0.1; *fm*(*many*) = 0.42; *fm*(***1***) = 0.03; $\mu(L) = 0.4$; $\mu(h_0) = 0.25$; $\mu(V) = 0.35$.

In this experiment, we use the linguistic frame of cognition with the specificity of 3. That is, there are 17 terms in LFoC for each attribute in the database creep and the LFoC of the linguistic quantifier Q. The number of terms of 17 is more than twice the number of terms of the attributes in studies [5, 19].

3. The parameters of genetic algorithm

The parameters of genetic algorithms are selected as in study [19]. Specifically, the number of linguistic summaries in each summary set is 30, corresponding to the 30 genes in each individual. The number of individuals in each generation is 20, and the number of iterations is 100. The selection rate is 0.15, and the mutation rate is 0.1. The fitness function *Fit* evaluates each individual in equation (11) with parameter values $m_g = 0.7$, $m_d = 0.3$.

4. Experimental results

Figure 3 shows the change of the best value of the fitness function *Fit* of the best individual in each generation

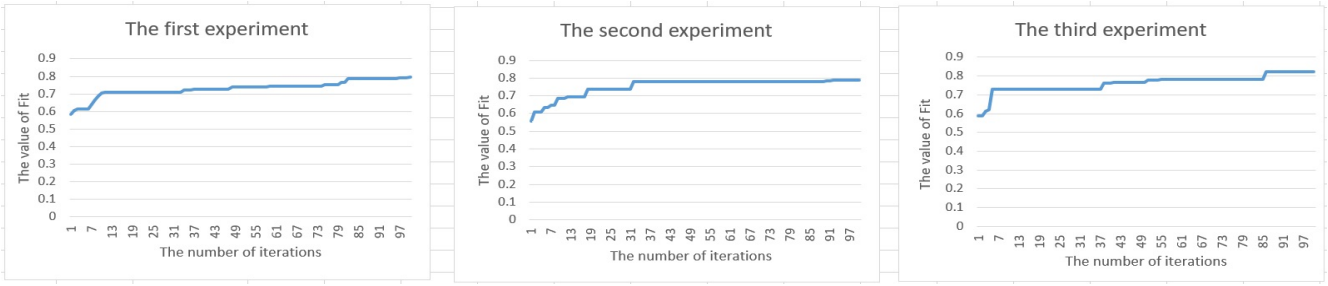

Figure 3. Fitness function value Fit of the best individual in the population in over 100 iterations.

TABLE III
COMPARISON OF THE PERFORMANCE RESULTS OF THE GREEDY-GA MODEL IN THIS STUDY WITH THE HYBRID-GA MODEL IN [19].

Model GA	**Fitness function value *Fit***	**The average of truth values *T***	**The number of summaries with *Q* > *a half***	**The number of summaries with *T* > *0.8***	**The number of summaries with *T* = 0**
Hybrid-GA .[19]	0.6659	0.9139	17.8	27.0	1.0
Greedy-GA	0.7905	0.9951	18.8	30	0

over each loop. From there, it shows that this value has increased gradually and will converge to a value at the last iterations. It proves that the results reflect the evolution through iterations.

Table III shows the comparison between the results of the algorithm Greedy-GA proposed in this paper and those of the algorithm Hybird-GA in the paper of Donis-Diaz et al. [19] on the evaluation function value *Fit*, the average of the validity values T of the linguistic summaries, the number of linguistic summaries with linguistic quantifier with semantic order greater than '*a half*', the number of linguistic summaries with the truth value $T > 0.8$, and the number of linguistic summaries with truth value $T = 0$ (corresponding to the case that there is not any records satisfying condition in F). The model Hybrid GA has been evaluated to be better than the basic model GA (Classical-GA) and the basic model GA combined with the *Cleaning operator* (Classcial + Cleaning-GA) to remove summaries with the truth value $T = 0$. Table III shows that the model Greedy-GA in this study has some advantages in comparison with the model Hybrid-GA:

• The optimal set of linguistic summaries has the greater fitness function value *Fit*. It proves that Greedy-GA gives a better optimal solution.

• There are more summaries which have linguistic quantifiers with semantic order greater than '*a half*'. This is the result of using the greedy strategy for selecting linguistic quantifiers with semantic order as large as possible in linguistic summaries with the same filter criterion F.

• The number of linguistic summaries with truth value T > 0.8 in our experiment reaches the maximum of 30, higher than the result of 27 summaries in [19]. This is because we use the linguistic quantifier term set with 17 terms, and the trapezoids representing the semantics of the linguistic quantifiers form the fuzzy partitions in the form of multi-granularity structure. This proves the advantage of the trapezoidal semantic representation designed based on the hedge algebras theory in [20] and the meaning of the scalability of LFoC in practical applications. Specifically, increasing the number of linguistic quantifiers by using more terms with a larger level of specificity will increase the ability to represent by quantifier term for any proportion in the interval of [0, 1]. The experimental results show that when LFoC of Q consists of 3 levels, it has the ability to select quantifier term for linguistic summaries with truth values greater than 0.8.

• There is no linguistic summary with the truth value $T = 0$. As analyzed at the end of Section III, all linguistic summaries in the execution process of genetic algorithm are generated by the procedure Random-Greedy-LS. Because we have used the condition $supp(F) > 0.1$ for the support measure in the procedure Random-Greedy-LS, there will be no summary with $T = 0$ in the execution of the algorithm Greedy-GA.

V. CONCLUSIONS

The extraction of knowledge represented in words in natural language from numerical datasets is a current trend. In this paper, we proposed a genetic algorithm model combined with the greedy strategy Greedy-GA to extract the optimal set of linguistic summaries based on the evaluation of the goodness and diversity of the set of linguistic summaries. The use of greedy strategy led to the generation of linguistic summaries with high goodness level and the increase of diversity in the set of linguistic summaries. Therefore, the Greedy-GA has better efficiency

when applied to extracting the optimal set of linguistic summaries compared to some existing genetic models. One difference from the existing genetic models to extract the optimal set of linguistic summaries is that we utilize hedge algebras methodology for designing the fuzzy sets based on semantics of linguistic terms. This ensures the interpretability of the content of the linguistic summaries, and that the set of terms consists of many high specificity terms also increase the quality of the set of linguistic summaries. The experimental results of dataset *creep* have proved the effectiveness of fuzzy set design method based on hedge algebras methodology and the model of genetic algorithm combined with greedy strategy in comparison with their counterparts.

REFERENCES

[1] S. Mitra, S. K. Pal, and P. Mitra, "Data mining in soft computing framework: a survey", *IEEE transactions on neural networks*, vol. 13, no. 1, pp. 3-14, 2002.

[2] R. R. Yager, "A new approach to the summarization of data", *Information Sciences*, vol. 28, no. 1, pp. 69-86, 1982.

[3] J. Kacprzyk, R. R. Yager, and S. Zadrożny, "A fuzzy logic based approach to linguistic summaries of databases", *International Journal of Applied Mathematics and Computer Science*, vol. 10, no. 4, pp. 813-834, 2000.

[4] J. Kacprzyk and S. Zadrożny, "Linguistic database summaries and their protoforms: towards natural language based knowledge discovery tools", *Information Sciences*, vol. 173, no. 4, pp. 281-304, 2005.

[5] C. A. Donis-Díaz, R. Bello-Pérez, and E. V. Morales, "Using Linguistic Data Summarization in the study of creep data for the design of new steels", *11th International Conference on Intelligent Systems Design and Applications (ISDA)*, 2011, pp. 160-165: IEEE.

[6] A. Wilbik, J. Keller, and J. C. Bezdek, "Generation of prototypes from sets of linguistic summaries", *IEEE International Conference on Fuzzy Systems (FUZZ-IEEE)*, 2012, pp. 1-8: IEEE.

[7] J. Kacprzyk and S. Zadrożny, "Linguistic summarization of the contents of Web server logs via the Ordered Weighted Averaging (OWA) operators", *Fuzzy Sets and Systems*, vol. 285, pp. 182-198, 2016.

[8] T. Altintop, R. R. Yager, D. Akay, F. E. Boran, and M. Ünal, "Fuzzy Linguistic Summarization with Genetic Algorithm: An Application with Operational and Financial Healthcare Data", *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, vol. 25, no. 04, pp. 599-620, 2017.

[9] R. J. Almeida, M.-J. Lesot, B. Bouchon-Meunier, U. Kaymak, and G. Moyse, "Linguistic summaries of categorical time series for septic shock patient data", *IEEE International Conference on Fuzzy Systems (FUZZ)*, 2013, pp. 1-8: IEEE.

[10] J. Kacprzyk and R. R. Yager, "Linguistic summaries of data using fuzzy logic", *International Journal of General System*, vol. 30, no. 2, pp. 133-154, 2001.

[11] M. D. Peláez-Aguilera, M. Espinilla, M. R. Fernández Olmo, and J. Medina, "Fuzzy linguistic protoforms to summarize heart rate streams of patients with ischemic heart disease", *Complexity*, vol. 2019, 2019.

[12] A. Duraj, P. S. Szczepaniak, and L. Chomatek, "Intelligent Detection of Information Outliers Using Linguistic Summaries with Non-monotonic Quantifiers", *International Conference on Information Processing and Management of Uncertainty in Knowledge-Based Systems*, 2020, pp. 787-799: Springer.

[13] A. Jain, M. Popescu, J. Keller, M. Rantz, and B. Markway, "Linguistic summarization of in-home sensor data", *Journal of biomedical informatics*, vol. 96, p. 103240, 2019.

[14] A. Wilbik, I. Vanderfeesten, D. Bergmans, S. Heines, and W. van Mook, "Linguistic summaries for compliance analysis of a glucose management clinical protocol", *IEEE International Conference on Fuzzy Systems (FUZZ-IEEE)*, 2018, pp. 1-7: IEEE.

[15] J. Kacprzyk, A. Wilbik, and S. Zadrożny, "Using a genetic algorithm to derive a linguistic summary of trends in numerical time series", *International Symposium on Evolving Fuzzy Systems*, 2006, pp. 137-142: IEEE.

[16] F. E. Boran and D. Akay, "A generic method for the evaluation of interval type-2 fuzzy linguistic summaries", *IEEE transactions on cybernetics*, vol. 44, no. 9, pp. 1632-1645, 2013.

[17] C. A. Donis-Díaz, R. Bello-Pérez, and J. Kacprzyk, "Linguistic data summarization using an enhanced genetic algorithm", *Czasopismo Techniczne*, vol. 2013, no. Automatyka Zeszyt 2 AC (10) 2013, pp. 3-12, 2014.

[18] R. Castillo Ortega, N. Marín, D. Sánchez, and A. G. Tettamanzi, "Linguistic summarization of time series data using genetic algorithms", *EUSFLAT*, 2011, vol. 1, no. 1, pp. 416-423: Atlantis Press.

[19] C. A. Donis-Díaz, A. G. Muro, R. Bello-Pérez, and E. V. Morales, "A hybrid model of genetic algorithm with local search to discover linguistic data summaries from creep data", *Expert Systems with Applications*, vol. 41, no. 4, pp. 2035-2042, 2014.

[20] C. H. Nguyen, T. L. Pham, T. N. Nguyen, C. H. Ho, T. A. Nguyen, "The linguistic summarization and the interpretability, scalability of fuzzy representations of multilevel semantic structures of word-domains", *Microprocessors and Microsystems*, vol. 81, Article 103641, 2021.

[21] L. A. Zadeh, "A computational approach to fuzzy quantifiers in natural languages", *Computers & Mathematics with applications*, vol. 9, no. 1, pp. 149-184, 1983.

[22] A. Wilbik, "Linguistic summaries of time series using fuzzy sets and their application for performance analysis of investment funds", *Ph. D. dissertation, Syst. Res. Inst.*, Polish Academy Sci., 2010.

[23] C. H. Nguyen, V. T. Hoang, and V. L. Nguyen, "A discussion on interpretability of linguistic rule based systems and its application to solve regression problems", *Knowledge-Based Systems*, vol. 88, pp. 107-133, 2015.

[24] C. H. Nguyen, T. S. Tran, and D. P. Pham, "Modeling of a semantics core of linguistic terms based on an extension of hedge algebra semantics and its application", *Knowledge-Based Systems*, vol. 67, pp. 244-262, 2014.

Lan Pham-Thi was born in Hanoi, Vietnam, in 1984. She received a B.S. in informatics teaching from Hanoi National University of Education in 2006 and an M.S. in computer science from Hanoi National University of Education in 2008. She is currently pursuing a PhD in computer science at the Graduate University of Science and Technology, Vietnam Academy of Science and Technology. Since 2008, she has been working as a lecturer in the Faculty of Information Technology, Hanoi National University of Education. Her research interests include data structure and algorithms, applications of hedge algebra, and approaches to extracting linguistic summaries.
Email: ptlan@hnue.edu.vn

Ho Nguyen-Cat was born in Hanoi, Vietnam, in 1941. He received a B.S. degree in mathematics from the University of Hanoi, the former VNU University of Science of Vietnam National University, Hanoi; a Dr. degree in mathematics and mechanics from the Warsaw University, Warsaw, Poland, in 1971; and a Dr. Sc. degree in mathematics, cybernetics and computer science from the Dresden University of Technology, Dresden, Germany, in 1987. He is now a researcher at the Institute of Theoretical and Applied Research, Duy Tan University in Da Nang and Hanoi. He is the author of more than 90 articles, including more than 40 articles published in international journals and conferences. His research interests include fuzzy logic, fuzzy databases, and computing with words, especially hedge algebras as a mathematical basis for directly handling linguistic words and their applications.
Email: ncatho@gmail.com

Phong Pham-Dinh received a Master degree in Information Technology and a Doctor of Philosophy degree in Computer Science from University of Engineering and Technology, Vietnam National University, Hanoi in 2011 and 2018, respectively. Now, he is a lecturer at the University of Transport and Communications. His research interests include hedge algebras, fuzzy systems, soft computing, data mining, and machine learning.
Email: phongpd@utc.edu.vn