ANALES DE LA
Academia de Ciencias de Cuba

CIENCIAS TÉCNICAS

Artículo original de investigación

Ecosistema *software* para el aprendizaje y toma de decisiones basados en sumarización lingüística de datos

Iliana Pérez Pupo [1] https://orcid.org/0000-0003-1433-0601
Pedro Y. Piñero Pérez [1] https://orcid.org/0000-0002-7635-8290
Rafael Bello Pérez [2] https://orcid.org/0000-0001-5567-2638
Roberto García Vacacela [3] https://orcid.org/0000-0002-1834-6806

[1] *Grupo Inteligencia Artificial para un Desarrollo Sostenible. La Habana, Cuba*
[2] *Centro de Investigaciones de la Informática, Universidad Central Marta Abreu de Las Villas. Santa Clara, Cuba*
[3] *Universidad Católica Santiago de Guayaquil. Guayaquil, Ecuador*

**Autor para la correspondencia:* iliperezpupo@gmail.com

Editor

Lisset González Navarro
Academia de Ciencias de Cuba.
La Habana, Cuba

Traductor

Darwin A. Arduengo García
Academia de Ciencias de Cuba.
La Habana, Cuba

RESUMEN

Introducción: El aumento del volumen de los datos en disímiles escenarios de toma de decisiones incrementa la necesidad de técnicas para el descubrimiento de dependencias no triviales ocultas en los datos. Se identifica a la sumarización lingüística de datos, como una de las ramas de la inteligencia artificial que permite generar resúmenes lingüísticos con aplicación en diferentes áreas. **Objetivo:** Desarrollar un ecosistema para el aprendizaje y la toma de decisiones, basado en técnicas de sumarización lingüística de datos bajo un enfoque multilingüe mejorando la eficiencia de los sistemas existentes. **Métodos:** Se aplican método histórico lógico que permitió la identificación de las principales oportunidades de mejora de los métodos existentes. Se proponen nuevos algoritmos y métodos para la construcción de resúmenes lingüísticos. Finalmente, se aplican métodos empíricos y se validan los resultados combinando técnicas de triangulación de datos y métodos. **Resultados:** Se obtienen 4 nuevos algoritmos que mejoran la eficiencia y la eficacia de otros algoritmos reportados en la bibliografía. Los algoritmos propuestos se destacan por las facilidades para el trabajo multilingüe, el empleo de técnicas de neutrosofía y el tratamiento de variables correlacionadas. **Conclusiones:** A partir de las comparaciones se evidencia la superioridad del algoritmo *RST_LDS,* se demuestra la complementariedad y la importancia de combinar diferentes técnicas en el descubrimiento de los resúmenes lingüísticos. Los nuevos indicadores para la evaluación de resúmenes lingüísticos mejoran el tratamiento de la indeterminación y la falsedad complementando los indicadores reportados en la bibliografía.

Palabras clave: sumarización lingüística de datos; resúmenes lingüísticos; soft computing; toma de decisiones

Premio Nacional Anual de la Academia de Ciencias de Cuba, 2022

Software ecosystem for learning and decision making based on linguistic summarization of data

ABSTRACT

Introduction: The increase in data volume in dissimilar decision-making scenarios increases the need for techniques for the discovery of non-trivial dependencies hidden in the data. Linguistic data summaries are identified as one of the branches of artificial intelligence that allow generating linguistic summaries with application in different areas. **Objective:** To develop an ecosystem for learning and decision-making, based on linguistic data summary techniques under a multilingual approach, improving the efficiency of existing systems. **Methods**: It is applied the logical historical method that allows the identification of the main opportunities for improvement of existing methods. They are proposed new algorithms and methods for the construction of linguistic summaries. Finally, they are applied empirical methods, and the results are validated by combining data triangulation techniques and methods. **Results:** As a result of the research, four new algorithms were obtained that improved the efficiency and effectiveness of other algorithms reported in the literature. The proposed algorithms stand out for their facilities for multilingual work, the use of neutrosophic techniques, and the treatment of correlated variables. **Conclusions:** From the comparisons it is concluded that the RST_LDS algorithm demonstrates the complementarity and importance of combining different techniques in the discovery of linguistic summaries. The new indicators for the evaluation of linguistic summaries improve the treatment of indeterminacy and falsity by complementing the indicators reported in the bibliography.

Keywords: linguistic data summarization; linguistic summaries; soft computing; decision-making

INTRODUCCIÓN

Uno de los problemas latentes en los procesos de toma de decisiones no estructuradas está asociado al descubrimiento de patrones de comportamiento en datos provenientes de sistemas de información. [1,2,3,4] Por lo general estos datos y sus relaciones no son entendibles a simple vista por los decisores, lo cual dificulta los procesos de toma de decisiones. [5] En este contexto se identifica, a la sumarización lingüística de datos (SLD), como una de las ramas de la inteligencia artificial, que permite el descubrimiento de las dependencias no triviales ocultas en los datos. [3,5,6,7] Las técnicas de la SLD permiten generar resúmenes lingüísticos legibles y entendibles que describan el comportamiento de los datos facilitando la toma de decisiones. [8,9,10,11,12,13,14]

Sin embargo, en la bibliografía consultada se identifican algunas problemáticas, por ejemplo, los entornos de toma de decisiones donde se han aplicado estas técnicas se caracterizan por ser entornos bajo incertidumbre; [15] no obstante, las técnicas reportadas en la bibliografía solo analizan el grado de certeza en la información obviando el tratamiento de la indeterminación y la falsedad. Además, la mayoría de los algoritmos en la bibliografía se concentran en la construcción de resúmenes a partir de datos frecuentes, pero no los de baja frecuencia. Esto afecta su aplicabilidad en la minería de datos anómalos. [16] A pesar de la alta aplicabilidad de la gestión de proyectos en disímiles sectores de la sociedad, hay pocas investigaciones sobre sumarización lingüística de datos en este contexto, lo cual es una de las motivaciones de esta investigación. [17]

Respecto al análisis de la eficacia de los algoritmos para la generación de resúmenes la mayoría de los trabajos, se basan en el tratamiento simple de cadenas en idioma inglés. [2,18,19] Esta situación afecta la eficacia de los algoritmos para la generación de los resúmenes en múltiples idiomas. En general las limitaciones principales de los trabajos reportados en la bibliografía presentan las siguientes situaciones. Se identifican en la bibliografía 3 enfoques fundamentales para la generación de resúmenes lingüísticos: extensiones a lenguajes de consultas, la generación de resúmenes a partir de reglas y la generación de resúmenes usando metaheurísticas; [20,6,21,22] pero generalmente los algoritmos sobre estos enfoques no explotan la información asociada a la correlación entre las variables, elemento que afecta la eficacia en la generación de resúmenes. Además, la mayoría de los algoritmos no facilitan la identificación de situaciones con poca ocurrencia en la base de datos, afectando la detección de datos anómalos. [23]

Además, las metaheurísticas reportadas en la bibliografía tienen alto costo computacional, elemento que influye negativamente en la aplicabilidad de estos algoritmos en entornos reales de toma de decisiones. [23] Así mismo, los indicadores más empleados para la evaluación de los resúmenes son los descritos en el trabajo realizado por Janusz Kacprzyk y Sławomir Zadrożny publicado en el año 2010: T1 grado de verdad, T2 grado de imprecisión, T3 grado de cobertura, T4 grado de adecuación y T5 longitud de un resumen. [24] En el cálculo de estos la mayoría utiliza el grado de pertinencia de los objetos considerando el cero como umbral y no consideran la indeterminación o la falsedad. [11]

En este contexto, esta investigación se propone como objetivo desarrollar un ecosistema para el aprendizaje y la toma de decisiones, basado en técnicas de sumarización lingüística de datos y su hibridación con otras tecnologías emergentes para mejorar la eficacia y la eficiencia en la construcción y evaluación de los resúmenes lingüísticos y su aplicación en entornos de toma de decisiones bajo un enfoque multilingüe.

MÉTODOS

En la investigación se aplican los métodos: Histórico-lógico a partir de aplicar este método se construye un marco teórico referencial para realizar una revisión sistemática sobre métodos para la generación y evaluación de resúmenes lingüísticos de datos, se identifican las oportunidades de mejora de los métodos reportados en la bibliografía. El hipotético deductivo se utiliza en el transcurso de la investigación; la hipótesis es resuelta siguiendo métodos fundamentados científicamente y luego se realizan pruebas estadísticas para demostrar la validez de los resultados.

Para evaluar la variable dependiente y las independientes se diseñan conjuntos de pruebas por cada dimensión: Se compara el comportamiento de los indicadores de evaluación tradicionales con las nuevas extensiones propuestas considerando las medidas: "capacidad para el tratamiento de la incertidumbre" y "comportamiento en los escenarios de prueba". Se evalúa la variable dependiente respecto a la dimensión "calidad de los resúmenes", "cubrimiento de diferentes situaciones en la base de datos" y "fortaleza de las dependencias descubiertas". Se comparan los algoritmos propuestos con diferentes algoritmos reportados en la bibliografía aplicando técnicas de triangulación de datos en 12 bases de datos.

Análisis de las dimensiones: "eficiencia de los algoritmos" y "desempeño integral de los algoritmos". En la primera se comparan los algoritmos propuestos con otros reportados en la bibliografía respecto a la complejidad computacional y al tiempo de ejecución. En la segunda se analiza el desempeño integral de los algoritmos al aplicarles el *test* de *ranking* Page´L Trent Test.

Se evalúa la variable dependiente en la dimensión "aplicabilidad de los algoritmos propuestos en la toma de decisiones". Se analizan los indicadores "interpretabilidad de los resúmenes" y "facilidad para la toma de decisiones" a partir de aplicar una técnica de análisis multicriterio; y se realiza el análisis de la variable dependiente respecto a la dimensión "enfoque multilingüe"

En las pruebas de comparación de los algoritmos, se emplean pruebas estadísticas bien fundamentadas, para una significación de 0,05 e intervalos de confianza del 99 % cumpliendo los siguientes pasos:

Paso 1. Se aplica prueba de normalidad (empleando los *test* de Shapiro-Wilk y Kolmogorov-Smirnov) y análisis descriptivo.

Paso 2. En caso de que las muestras cumplan con la distribución normal, se aplican pruebas paramétricas empleando T-Student para 2 muestras relacionadas.

Paso 3. En caso de que las muestras no cumplan con la distribución normal, se aplican pruebas no paramétricas empleando Friedman para *n* muestras relacionadas.

Paso 4. Si Friedman indica que no hay diferencias significativas en las *n* muestras relacionadas, entonces se concluye que no hay diferencias significativas.

Paso 5. Si Friedman indica que hay diferencias significativas en las *n* muestras relacionadas, entonces se aplican pruebas *post-hoc test* Wilcoxon para 2 muestras relacionadas.

RESULTADOS

Los resultados de la investigación se concretan en las siguientes direcciones que constituyen epígrafes de esta sección.

- Se proponen los algoritmos RST_LDS basados en conjuntos aproximados, PCA_LDS que se apoyan en componentes principales, LPA_LDS basados en grafos probabilísticos que generan resúmenes lingüísticos bajo un enfoque multilingüe;
- Pérez Pupo junto a otros autores generan resúmenes a partir de datos anómalos, y definen el algoritmo LDS_Outliers; [25,26,27]
- se proponen nuevas extensiones para evaluar la calidad de los resúmenes complementando a los indicadores existentes; mejorando el tratamiento de la indeterminación;
- desarrollo de un repositorio de datos para investigaciones que incluye 18 bases de datos sobre gestión de proyectos, 2 sobre auditorías y control interno y 2 sobre información médica; [17]
- introducción de los resultados en un ecosistema de ***software*** para la ayuda a la toma de decisiones en la gestión de proyectos de I+D+i y de inversión.

Notación básica empleada en los algoritmos:

- U: conjunto de datos (dataset) formado por n objetos, cada objeto representa una fila del dataset U tal que n = |U|,

- H: conjunto de atributos que describen a los objetos en U tal que p = |H|,
- ALV: conjunto de variables lingüísticas que describen a los atributos H,
- LVQ: variable lingüística para los cuantificadores de los resúmenes,
- LNC: lenguaje natural controlado empleado, que incluye la gramática,
- LNCGrammar y el diccionario LNCDictionary,
- $T = T_{tradicional} \cup T_e$, donde $T_{tradicional}$: indicadores tradicionales de evaluación de calidad de los resúmenes lingüísticos, [28]
- Te: indicadores de extensión propuestos en la investigación;
- α: umbral de pertenencia,
- CandidateSummaries: listado de resúmenes candidatos,
- percentil_component: percentil para seleccionar las componentes relevantes,
- percentil_elbow: percentil para seleccionar las variables relevantes,
- support, confidence: medidas de calidad para evaluar las reglas de asociación,
- *SI* = (*U, A* ∪ D): sistema de información,
- A, D: conjuntos de atributos,
- αk1, αk2: alfa cortes,
- k_1, k_2: límites inferior y superior de dependencia en grado k,
- C: conjunto de semillas centroides,
- distancia (*d, P*): función de distancia desde d al conjunto de puntos P,
- percentil: percentil usado para la determinación de datos anómalos.

Los resúmenes lingüísticos que construyen los algoritmos propuestos constituyen objetos con los siguientes atributos:

- *Q*: cuantificador, conjunto borroso con universo de discurso en el intervalo [0,1];
- *R*: calificador o filtro, atributo que determina un subconjunto borroso del objeto *yi*, ejemplo "joven" para el atributo "edad";
- *S*: sumarizador, atributo con un valor lingüístico definido en el dominio del atributo *Aj*, ejemplo "bajo salario" para el atributo "salario";

Las salidas de los algoritmos incluyen una etapa de humanización (algoritmo 1) que emplea lenguajes naturales controlados para expresar los resúmenes en los idiomas: español, inglés, japonés y árabe. Los resúmenes generados cumplen con las protoformas "Qy' s are S" y "QRy' s are S", [29] por ejemplo:

- La mayoría de los trabajadores tienen bajo salario (sin filtro ´*R*´);
- La mayoría de los trabajadores jóvenes tienen bajo salario (con filtro ´*R*´).

Algoritmo 1 humanize_summaries (CandidateSummaries, ALV, LVQ, LNC)

Entradas

CandidateSummaries: listado de resúmenes candidatos

A_{LV}, LV_Q, LNC: parámetros descritos en la sección 2.1 sobre notaciones.

Inicio:

Paso 1. *summaries = [ ]*

Paso 2 *para cada P* (Q, *R*, *S*, *T*) *en CandidateSummaries*

Paso 3 *translations = [ ]*

Paso 4 *para cada element de P (Q, R, S, T)*

Paso 5 *element_translation = search_dictionary (element, ALV, LVQ , LNCDictionary)*

Paso 6 *translations ← element_translation*

Fin del paso 4

Paso 7 *sentence = generate_sentence (translations, LNCGrammar)*

Paso 8 *summaries ← sentence*

Fin del paso 2

Paso 9 *Return summaries*

Algoritmo PCA_LDS, basado en análisis de componentes principales

Este algoritmo genera resúmenes combinando los componentes principales PCA y reglas de asociación (Algoritmo 2). [30,31]

Algoritmo 2 PCA_LDS

Entradas:

U, ALV, LVQ, T, LNC, α : parámetros definidos en la sección 2.1 sobre notaciones.

percentil_component: valor de percentil que permite seleccionar las componentes relevantes.

percentil_elbow: valor de percentil empleado para seleccionar las variables relevantes en el contexto de una componente.

support, confidence: representan las medidas de calidad soporte y confianza respectivamente, para evaluar las reglas de asociación.

Inicio

Rules = ∅

components = do_pca(U)

UB = transform_fuzzy (U, ALV)

μ = calculate_percentil (components, percentil_component)

choosed_components = choose_components (components, μ)

por cada i-ésima componente principal en choosed_components hacer:

Paso 6.1 *Velbow = calculate_percentil (componentsi, percentil_elbow)*

Paso 6.2 *ACi = get_attributes (componentsi, Velbow)*

Paso 6.3 *UCi = get_data (U, ACi)*

Paso 6.4 *rulesCi = do_rules (UCi, support, confidence)*

Paso 6.5 *Rules ← rulesCi*

Fin del ciclo iniciado en el paso 6

Paso 7 *Candidates=generate_summaries_from_rules (U, Rules, ALV, LVQ, T)*

Paso 8 *calculate_quantifiers (Candidates, LVQ)*

Paso 9 *calculate_T (Candidates, α)*

Paso 10 *summaries = humanize_summaries (candidates, ALV, LVQ, LNC)*

Paso 11 *ranking_summaries = sort_summaries (summaries)*

Paso 12 *return ranking_summaries*

En el paso 2 se aplica el algoritmo PCA y se generan los componentes principales. En el paso 4, se calcula el umbral μ a partir del cual se seleccionan los componentes cuyo valor de fortaleza sea mayor que α permitiendo la reducción del espacio de búsqueda. [32] Entre los pasos 6-12 se generan reglas de asociación a partir de las componentes seleccionadas. [33,34] Luego del paso 9 se tiene el conjunto de resúmenes candidatos, cada resumen cumple con la estructura *P(Q, R, S,T)*. Finalmente, en el paso 10 se generan los resúmenes usando el Algoritmo 1. La complejidad computacional del PCA_LDS es *O* (*max*[$P2^{p+1}$, n^2 *P*, *n m* p^2]).

Algoritmo *LPA_LDS* basado en gráficos probabilísticos

En este algoritmo se descubren las relaciones más fuertes entre las variables empleando el aprendizaje de grafos probabilísticos, [26,35] reduciendo el espacio de búsqueda. Finalmente, los resúmenes lingüísticos candidatos son generados en múltiples idiomas ver algoritmo 3.

Este algoritmo es capaz de trabajar con datos heterogéneos pero tiene como limitante que solo descubre relaciones de primer, segundo y tercer orden, porque el consumo de espacio y tiempo aumenta exponencialmente con respecto al orden de las relaciones.

Algoritmo 3 LPA_LDS

Entradas

U, ALV, LVQ, T, LNC, α : parámetros definidos en la sección 2.1 sobre notaciones.

Inicio:

Paso 1 Candidates = ∅

Paso 2 *UB = transform_fuzzy (U, ALV)*

Paso 3 *prob_graph = do_probabilistic_graph (UB)*

Paso 4 *Por cada uno de los prob_graph i árboles en prob_graph hacer*

Paso 4.1 *Si prob_graph i tiene más de un vértice*

Paso 4.2 *candidate_summariesi = do_candidate_from_branchs (prob_graph i)*

Paso 4.3 *candidates* ← *candidate_summariesi*

Fin de la condicional iniciada en paso 4.1

Fin del ciclo iniciado en el paso 4

Paso 5 *CandidateSummaries = generate_summaries (U, Candidates, ALV, LVQ, T)*

Paso 6 *calculate_quantifiers (CandidateSummaries, LVQ)*

Paso 7 *calculate_T (CandidateSummaries, α)*

Paso 8 *summaries = humanize_summaries (CandidateSummaries, ALV, LVQ, LNC)*

Paso 9 *ranking_summaries = sort_summaries (Summaries)*

Paso 10 *Return ranking_summaries*

En el paso 2 se transforman los datos de entrada empleando variables lingüísticas y luego se aplica un algoritmo para el aprendizaje del modelo probabilístico, que mejor se aproxime al comportamiento de los datos, generando un poliárbol. [36] En el paso 4, por cada árbol se genera un objeto candidato, donde el nodo raíz del árbol es identificado como el atributo sumarizador y los nodos ramas son identificados como filtros. Finalmente, entre los pasos del 6-10 se generan los resúmenes lingüísticos. La complejidad del algoritmo *LPA_LDS es O*(n^3).

Algoritmo *RST_LDS* basado en conjuntos aproximados

El algoritmo 4 *RST_LDS* permite el trabajo con datos heterogéneos. Aplica la dependencia en grado k y otros conceptos de la teoría de conjuntos aproximados para la generación de resúmenes lingüísticos. Bajo este principio es posible identificar relaciones entre variables con muy baja frecuencia de aparición en los datos, lo que permite que también pueda ser empleado para identificar datos anómalos. [16]

Algoritmo 4 *RST_LDS*

Entradas:

U, ALV, LVQ, T, LNC, α : parámetros definidos en la sección 2.1 sobre notaciones.

Sistema de información SI = (U, A ∪ *D)*

A, D: atributos del sistema de información,

α_{k1}*: alfa corte* k_1*, establece el límite superior de la dependencia en grado k en la generación de resúmenes,*

α_{k2}*: alfa corte* k_2*, establece el límite inferior de la dependencia en grado k en la generación de resúmenes.*

Inicio

Paso 1 *transformar SI = (U, A* ∪ *D) en un dataset lingüístico considerando los conjuntos borrosos y los principios de máxima membresía,*

Paso 2 *A_ItemSet = attributes_to_sets(A* ∪ *D)*

Paso 3 *D_ItemSet = A_ItemSet,*

Paso 4 *stackset.push(A_ItemSet),*

Paso 5 *CS = { }*

Paso 6 *mientras Stackset no esté vacío,*

Paso 6.1 *A_ItemSet = Stackset.pop,*

Paso 6.2 *Por cada B* ⊆ *A_ItemSet,*

Paso 6.3 *Por cada X* ⊆ *D_ItemSet: B* ∩ *X =* ∅*,*

Paso 6.4

Paso 6.5 *por cada* $O_t(B,X) \in POS_B(D)$,
Paso 6.6 *si* $(k(B, X) \geq \alpha_{k1})$ *entonces,*
Paso 6.7 $CS = CS \cup O_t(B,X)$,
Paso 6.8 *sino,*
Paso 6.9 *si* $(k(B, X) \geq \alpha_{k2})$ *entonces,*
Paso 6.10 *Stackset.push({*$B \cup X$*})*
fin del condicional iniciado en el paso 6.9,
fin del condicional iniciado en el paso 6.6,
fin del ciclo iniciado en el paso 6.5,
fin del ciclo iniciado en el paso 6.3,
fin del ciclo iniciado en el paso 6.2

$$POS_B(D) = \bigcup_{X \in U/D} B_*(X),\ k(X,B) = \frac{|POS_B(D)|}{|U|},\ B \Rightarrow_k D,\ k(B, X) \geq \alpha_{k2}$$

fin del ciclo iniciado en el paso 6.

Paso 7 *CandidateSummaries = generate_summaries_from_candidates (U, CS, ALV, LVQ, T),*
Paso 8 *calculate_quantifiers (CandidateSummaries, LVQ),*
Paso 9 *calculate_T (CandidateSummaries, α),*
Paso 10 *summaries = humanize_summaries (CandidateSummaries, ALV, LVQ, LNC),*
Paso 11 *ranking_summaries = sort_summaries (summaries),*
Paso 12 *return ranking_summaries,*

El paso 1 transforma el sistema de información SI en un conjunto de datos lingüísticos considerando el principio de máxima membresía en los conjuntos borrosos. El paso 6 constituye el corazón de este algoritmo, en los pasos 6.2 y 6.3 se crean pares de conjuntos (*B, X*) que representan posibles filtros y sumarizadores respectivamente. En el paso 6.4 se calcula la región positiva $POS_B(D)$ mientras que la condición $K(B, X) \geq \alpha_{k2}$ ayuda a encontrar relaciones entre atributos de dependencia parcial, considerando los conjuntos de atributos que tienen un nivel mínimo de relación. En el paso 6.6 se identifican las relaciones de dependencia mayor que el umbral α_{k1}, en el contexto O_t, con los cuales se construyen los resúmenes candidatos. En los pasos del 8-12 se aplica el algoritmo 1 para generar los resúmenes lingüísticos. La complejidad del algoritmo *RST_LDS es O* $(max[n^2 p, n\ m\ p^2])$.

Algoritmo LDS_Outliers, generación de resúmenes lingüísticos a partir de datos anómalos

En esta investigación se proponen 2 enfoques para la generación de resúmenes lingüísticos a partir de datos anómalos. [37] El primer enfoque está centrado en la generación de resúmenes lingüísticos que permitan identificar situaciones poco frecuentes donde se pueden aplicar los algoritmos *RST_LDS* y *LPA_LDS*. El segundo enfoque primero descubre datos anómalos que son empleados para construir los resúmenes (algoritmo 5 *LDS_Outliers)*.

Algoritmo 5: LDS_outliers

Entradas:
U, ALV, LVQ, T, LNC, α: parámetros definidos en la sección 2.1 sobre notaciones,
C: conjunto de semillas centroides,
Distancia(d,P): función de distancia desde d al conjunto de puntos P,
percentil: percentil usado para la determinación de datos anómalos,
α_{k1}*: alfa corte* k_1*, establece límite superior de la dependencia en grado k en la generación de resúmenes,*
α_{k2}*: alfa corte* k_2*, establece límite inferior de la dependencia en grado k en la generación de resúmenes.*
Inicio
Paso 1 Out = U // se inicializa con todos los datos
Paso 2 clusters = Cluster (U, centers = C)
Paso 3 centers = clusters.centers
Paso 4 Para cada clusteri en clusters, hacer
Paso 4.1 B0 = Calculate_threshold(clustersi)
Paso 4.2 O_i *= out_centers (Bo, clusteri)*
Paso 4.3 Out = Out $\cap O_i$
Fin del ciclo paso 4
Paso 5 Out = Ranking_outlier(Out, percentil)
Paso 6 CandidateSummaries = Aplicar RST_LDS(Out, A_{LV}, α_{k1}, α_{k2}, *LVQ, T, α)*
Paso 7 CandidateSummaries = empower_summaries (CandidateSummaries, α, Out_i*)*
Paso 8 Summaries = filter_with_experts(CandidateSummaries)
Paso 9 Return Summaries

En el paso 2 se construyen *C* clústeres a partir de los datos, este algoritmo podría aplicarse con diferentes métodos de agrupación. En el paso 4, por cada clúster se calcula el umbral B0, basado en el concepto "conjunto B_0 compacto" definido en un artículo de Shulcloper del 2009. [38] En el paso 4.2 se detecta, para cada clúster, el conjunto de datos que tienen comportamiento anómalo. En el paso 5 los objetos identificados son ordenados en un *ranking*, los elementos más alejados de los centros tienen mayor probabilidad de ser datos anómalos. En el paso 8 se produce un proceso de aprendizaje activo y se identifican los anómalos, con los que se generan los resúmenes. La complejidad computacional del algoritmo 5 está acotada superiormente por *O(n3)*.

Nuevos métodos de evaluación de los resúmenes lingüísticos

En la investigación se proponen nuevos indicadores para la evaluación de resúmenes que emplean elementos de los conjuntos aproximados y la teoría neutrosófica. Estos nuevos

indicadores complementan los reportados en la bibliografía. [9,28,39] Se propone el indicador Te_{1a} y Te_{1b} (como complementos al T1) son medidas de certeza, pero incorporan un *α*-corte que flexibilizar el umbral de pertenencia de los objetos a un resumen (ecuaciones 1, 2, 3, 4, 5 y 6).

μQ es el conjunto borroso que representa al cuantificador del resumen.Donde representa la evaluación del objeto yi en la función de pertenencia de los conjuntos borrosos de R y representa la evaluación del objeto yi en la función de pertenencia de los conjuntos borrosos de S_j.

Se proponen indicadores que complementan al T_2 basados en la neutrosofía: Te_{2a} combina la pertenencia, indeterminación y la no pertenencia (ecuación 7, 8, 9, 10 y 11), Te_{2b} mide el grado de no pertenencia (ecuación 12), Te_{2c} mide el grado de indeterminación (ecuación 13) y el Te_{2d} combina la no pertenencia y la indeterminación (ecuación 14). [24]

Se propone Te_3 (como reemplazo al indicador tradicional T_3) (ecuación 15). [24]

Se propone Te_4 que complementa al T_4 incluye elementos de neutrosofía (ecuación 16). [24]

Se propone reemplazar el indicador T_5 con Te_5 mejorando la calidad en el cálculo de la longitud del resumen (ecuación 17). [24]

La validación se organiza en 2 grupos de pruebas: uno asociado a la comparación de los nuevos indicadores propuestos con los tradicionales, y el segundo compara los algoritmos propuestos *(RST_LDS, LPA_LDS, PCA_LDS, Outlier_LDS)* con los reportados en la bibliografía *(GA_LDS, Apriori_LDS, ACO_LDS);* empleando los indicadores tradicionales, los extendidos, la "fortaleza de las relaciones que descubren" y el tiempo de ejecución. Posteriormente se analiza el desempeño global de los algoritmos. Finalmente, se demuestra la aplicabilidad de la propuesta a través de un caso de estudio. [21,23,25,26,27,40]

En las comparaciones se utilizaron 7 bases de datos sobre gestión de proyectos: 2 bases de datos sobre auditorías suministradas

$$\mathrm{Te}_{1a} = \mu_Q\left(\frac{|SUMMARY_{\alpha*}(Y)|}{|SUMMARY^{*}(Y)|}\right), \text{ tal que } \mathrm{Te}_{1a} \in [0,1] \quad (1)$$

$$SUMMARY_{\alpha*}(Y) = \{ y_i \in Y : \mu_{SUMMARY}(y_i) \geq \alpha \} \quad (2)$$

$$SUMMARY^{*}(Y) = \{ y_i \in Y : \mu_{SUMMARY}(y_i) \geq 0 \} \quad (3)$$

$$\mu_{SUMMARY}(y_i) = \mathrm{TNorm}_{j\in\{1,2,\ldots,m\}}\left(\mu_{S_j}(y_i)\right) : y_i \in Y \quad (4)$$

$$\mu_{SUMMARY}(y_i) = \mathrm{TNorm}_{\substack{k\in\{1,2,\ldots,p\}\\ j\in\{1,2,\ldots,m\}}}\left(\mu_{R_k}(y_i), \mu_{S_j}(y_i)\right) : y_i \in Y \quad (5)$$

$$\mathrm{Te}_{1b} = \frac{\sum_{y_i \in SUMMARY_{\alpha*}(Y)} \mu_{SUMMARY}(y_i)}{|SUMMARY_{\alpha*}(Y)|}, \text{ tal que } \mathrm{Te}_{1b} \in [0,1] \quad (6)$$

$$\mathrm{Te}_{2a} = \frac{\sum_{y_i \in SUMMARY_{\alpha*}(Y)} \left(\frac{1-\mu_{SUMMARY}(y_i)+\tau_{SUMMARY}(y_i)+\sigma_{SUMMARY}(y_i)}{3}\right)}{|SUMMARY_{\alpha*}(Y)|} \quad (7)$$

$$\text{tal que } \mathrm{Te}_{2a} \in [0,1]$$

$$\tau_{SUMMARY}(y_i) = \mathrm{TNorm}_{j\in\{1,2,\ldots,m\}}\left(\tau_{S_j}(y_i)\right) \quad (8)$$

$$\tau_{SUMMARY}(y_i) = \mathrm{TNorm}_{\substack{k\in\{1,2,\ldots,p\}\\ j\in\{1,2,\ldots,m\}}}\left(\tau_{R_k}(y_i), \tau_{S_j}(y_i)\right) \quad (9)$$

$$\sigma_{SUMMARY}(y_i) = \mathrm{TNorm}_{j\in\{1,2,\ldots,m\}}\left(\sigma_{S_j}(y_i)\right) \quad (10)$$

$$\sigma_{SUMMARY}(y_i) = \mathrm{TNorm}_{\substack{k\in\{1,2,\ldots,p\}\\ j\in\{1,2,\ldots,m\}}}\left(\sigma_{R_k}(y_i), \sigma_{S_j}(y_i)\right) \quad (11)$$

$$\mathrm{Te}_{2b} = \frac{\sum_{y_i \in SUMMARY_{\alpha*}(Y)} \left(\frac{\tau_{SUMMARY}(y_i) + \sigma_{SUMMARY}(y_i)}{2} \right)}{|SUMMARY_{\alpha*}(Y)|} \quad (12)$$

$$\text{tal que } \mathrm{Te}_{2b} \in [0,1]$$

$$\mathrm{Te}_{2c} = \frac{\sum_{y_i \in SUMMARY_{\alpha*}(Y)} \tau_{SUMMARY}(y_i)}{|SUMMARY_{\alpha*}(Y)|}, \text{ tal que } \mathrm{Te}_{2c} \in [0,1] \quad (13)$$

$$\mathrm{Te}_{2d} = \frac{\sum_{y_i \in SUMMARY_{\alpha*}(Y)} \sigma_{SUMMARY}(y_i)}{|SUMMARY_{\alpha*}(Y)|}, \text{ tal que } \mathrm{Te}_{2d} \in [0,1] \quad (14)$$

$$\mathrm{Te}_{3} = \frac{|SUMMARY_{\alpha*}(Y)|}{n}, \text{ tal que } \mathrm{Te}_{3} \in [0,1] \quad (15)$$

$$\mathrm{Te}_{4} = \frac{\sum_{y_i \in SUMMARY_{\alpha*}(Y)} \left(\frac{\left(1 + \mu_{SUMMARY}(y_i) - \tau_{SUMMARY}(y_i) - \sigma_{SUMMARY}(y_i)\right)}{2} \right)}{|SUMMARY^{*}(Y)|} \quad (16)$$

$$\text{tal que } \mathrm{Te}_{4} \in [0,1]$$

$$Te_5 = e^{-k(\bar{x}-b)^2}, \text{ tal que } Te_5 \in [0,1],\ k = 0.125 \text{ y } b = 2.5 \quad (17)$$

por la Contraloría General de la República de los años 2017-2018; 3 bases de datos sobre toma de decisiones médicas, 2 de ellas sobre embarazadas cardiópatas y la tercera sobre COVID-19. [41,17]

Comparación de los indicadores tradicionales y extendidos

Para esta comparación se simularon 29 escenarios que representan clases de pruebas abarcando las posibles combinaciones de cuantificador-filtro-sumarizador.

Comparación de T_1 y Te_{1a}

La figura 1 grafica estos 2 indicadores ante resúmenes del tipo *QRy's are S*, se demuestra que en el 86,2 % de las clases de pruebas el Te_{1a} tiene mejor comportamiento que el T_1.

Comparación de T_3 y Te_3

La figura 2, evidencia que el indicador T_3 prácticamente no refleja variaciones, lo cual no es un buen comportamiento.

En general, el Te_3 extendido es más estricto que el tradicional y logra diferenciar mejor los escenarios representados por las clases de pruebas.

Comparación de T_4 y Te_4

En la figura 3 se observa que el Te_4 tiene mayor poder discriminatorio de las situaciones representadas en mayoría de las clases de pruebas respecto al T_4.

Comparación de T_5 y Te_5

En la figura 4 se observa que el T_5 tradicional (curva azul) cae abruptamente, asignando alta puntuación a resúmenes de tamaño 1 y bajos valores a los de tamaño 2 y 3 (longitudes más deseadas). Mientras que el Te_5 extendido asigna valores altos a resúmenes que involucran hasta 4 variables, y la caída de la curva es más suave. Se concluye que el Te_5 extendido tiene mejor comportamiento que el T_5.

Comparación de algoritmos considerando indicadores de calidad

En esta comparación se realizan pruebas de normalidad usando el *test* Shapiro-Wills. En dependencia de los resultados, se comparan los algoritmos usando pruebas paramétricas o no paramétricas según corresponda, de forma tal que los algoritmos ubicados en el grupo A, reportan resultados significativamente mejores que el resto, los ubicados en el grupo B son significativamente mejores que los del C y así sucesivamente. El análisis comparativo evidenció que los algoritmos *RST_LDS*, *LPA_LDS* y *GA_LDS*, reportan mejores resultados respecto a los indicadores tradicionales, mientras que *ACO_LDS* y *Apriori_LDS* son los de peores resultados.

La comparación de la tabla 1 evidencia que, a modo general, los algoritmos propuestos reportan mejores resultados que los tomados de la bibliografía. También se compararon las corridas de los algoritmos respecto al cubrimiento de diferentes situaciones en la base de datos. De esta forma se mide la capacidad de los algoritmos para generar resúmenes lingüísticos diversos. En este criterio, el algoritmo de mejores resultados fue *ACO_LDS* seguido por *RST_LDS*.

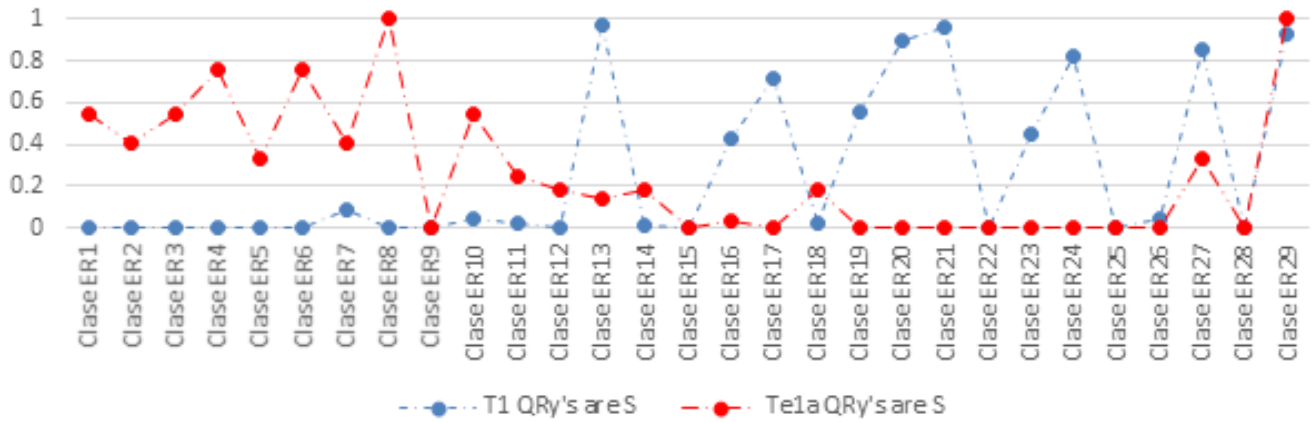

Fig. 1. Comportamiento del T_1 y Te_{1a} en resúmenes con filtro

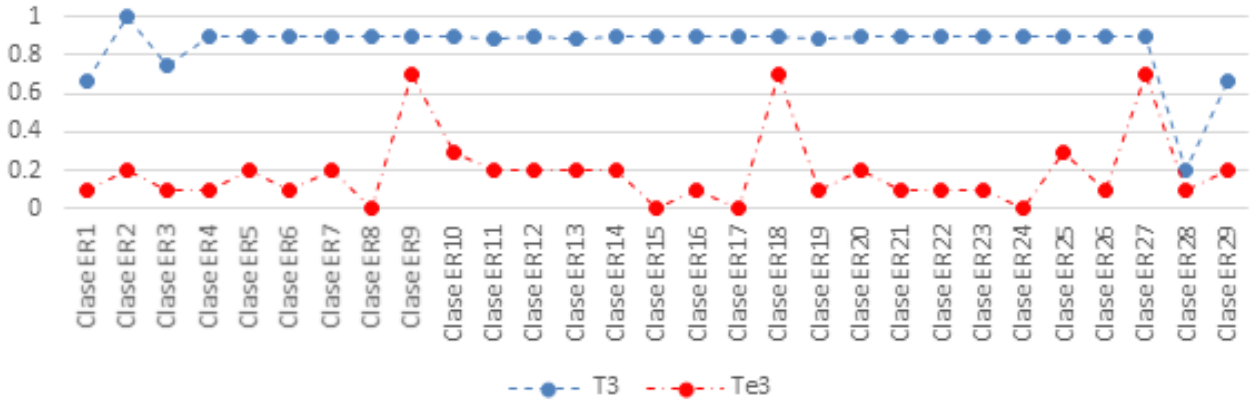

Fig. 2. Comportamiento del T_3 y Te_3

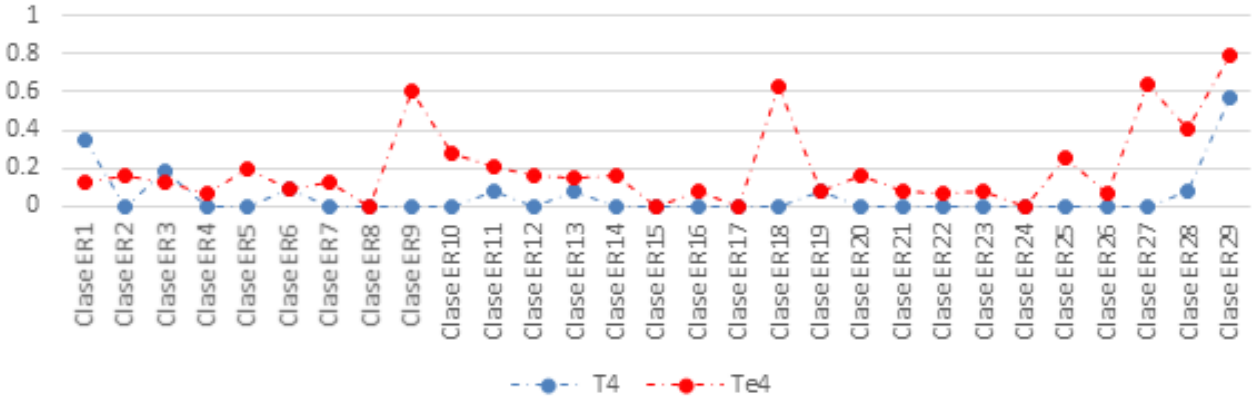

Fig. 3. Comportamiento del T_4 y Te_4

La fortaleza de las dependencias descubiertas es una medida que asigna mayor puntuación a los algoritmos que generen resúmenes cuyos cuantificadores son "la mayoría", "casi la totalidad", "muchos". En este criterio, los de mejores resultados fueron los algoritmos *RST_LDS* y *LPA_LDS*. [23]

Comparación de los algoritmos respecto a la eficiencia

La tabla 2 evidencia que los algoritmos propuestos mejoran la eficiencia respecto a los reportados en la bibliografía empleados en la experimentación. El de mejor eficiencia fue *RST_LDS*, mientras que los basados en metaheurísticas son los menos eficientes.

Tamaño del resumen	T_5 Tradicional	Te_5 Extendido
1	1	0.755
2	0.5	0.969
3	0.25	0.969
4	0.125	0.755
5	0.063	0.458
6	0.031	0.216
7	0.016	0.08
8	0.008	0.023

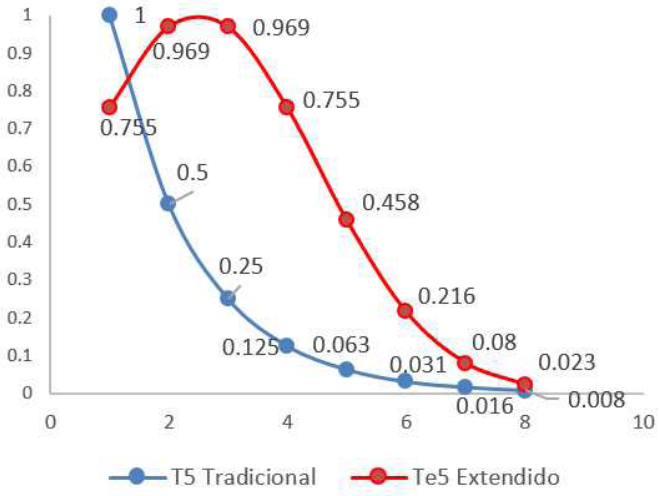

Fig. 4. Curvas del T_5 (tradicional) y Te_5 (extendido)

donde:

n: tamaño de la base de datos. *p:* cantidad de variables,

k: tamaño medio de los resúmenes, *k < p. g:* generaciones del algoritmo,

d: número de individuos,

z: cantidad de resúmenes por individuos.

Nota: El asterisco (*) refiere los algoritmos propuestos en esta investigación.

Desempeño global de los algoritmos

Para el análisis del desempeño global de los algoritmos se utilizó el método *Page's L Test,* implementado en la herramienta *Statex,* [42] el cual generó el siguiente *ranking* global: *RST_LDS, PCA_LDS, apriori_LDS, GA_LDS, LPA_LDS y ACO_LDS.* De este análisis se concluye que el algoritmo *RST_LDS* es el de mejor comportamiento global y el *ACO_LDS* el peor.

Validación de la aplicabilidad de la propuesta para la toma de decisiones

En esta sección se demuestra la aplicabilidad en la toma de decisiones. Para ello se aplican los algoritmos propuestos en 2 escenarios: para esta prueba los algoritmos *RST_LDS, LPA_LDS* y *PCA_LDS* fueron integrados en un metaalgoritmo *Hybrid_LDS* e incorporados al ecosistema propuesto cuadro

Tabla 1. Comparación de los algoritmos respecto a los indicadores extendidos de evaluación

Indicadores	Grupo A	Grupo B	Grupos C-D
Te_{1a}, Te_{1b}	RST_LDS Apriori_LDS PCA_LDS LPA_LDS	ACO_LDS	GA_LDS
Te_3	PCA_LDS	Apriori_LDS	RST_LDS (C) ACO_LDS (C) LPA_LDS (D) GA_LDS (D)
Te_4	PCA_LDS Apriori_LDS GA_LDS RST_LDS LPA_LDS	ACO_LDS	

Tabla 2. Complejidad de los algoritmos

Algoritmos	Complejidad	Media t. (∀ BD)
* RST_LDS (A)	O(n2·p)	2,11
* PCA_LDS (B)	O(max(p·2p+1, n2·p))	43,99
Apriori_LDS (B)	O(2p+1)	45,94
* LPA_LDS (B)	O(n3)	52,76
GA_LDS (C)	O(kgnd)	1212,62
ACO_LDS (D)	O(zgnd2)	19692,03

de mando para la toma de decisiones (*businessredmine*). [43] Con este metaalgoritmo se aprovechan las potencialidades de cada algoritmo generando un único conjunto de resúmenes lingüísticos sin repetición, y en cada caso de estudio se refleja la complementariedad respecto a la coincidencia y exclusividad de los resúmenes generados por los algoritmos combinados.

Escenario uno: tratamiento médico a embarazadas cardiópatas

En este escenario se dispone de 2 bases de datos asociadas a embarazadas cardiópatas, provistas por el Hospital Ramón González Coro. A continuación, se muestran algunos ejemplos de resúmenes sobre los casos de embarazadas recogidas en el estudio.

Sobre embarazadas con enfermedades congénitas:

"Muy pocos registros en la base de datos reportan que en el 100,0 % de las veces, las embarazadas cardiópatas que tienen como tratamiento aplicar cierre percutáneo tienen como tratamiento no aplicar cuartoplastia".

"La mayoría de los registros en la base de datos reportan que en el 96,0 % de las veces las embarazadas cardiópatas que tienen ausencia de fracción de eyección del ventrículo izquierdo (FEVI) tienen ausencia de soplo sistólico tricúspide".

Sobre embarazadas con enfermedades valvulares:

"Algunos registros en la base de datos reportan que en el 100,0 % de las veces, las embarazadas cardiópatas que tienen presencia de primer ruido cardíaco fuerte tienen como tratamiento no aplicar heparina".

"Casi la totalidad de los registros en la base de datos reportan que en el 100,0 % de las veces las embarazadas cardiópatas que tienen presencia de soplo sistólico aórtico y ausencia de soplo diastólico aórtico tienen como pronóstico parto vaginal".

Escenario dos: toma de decisiones en gestión de proyectos

En este escenario se dispone de datos recopilados por el sistema de información para la gestión de proyectos entre los años 2010-2020, [44,45] ofrecidos por el Centro de Consultoría y Desarrollo de Arquitecturas Empresariales. Se obtuvieron resúmenes que cubren diferentes áreas de gestión:

Respecto al desempeño de los recursos humanos los decisores consideraron que los resúmenes obtenidos permitieron descubrir relaciones entre los rasgos de la personalidad y los roles principales de un proyecto de software. También consideran que los resúmenes brindan información relevante para apoyar la toma de decisiones en el proceso de adquisición de los recursos humanos. A continuación, se muestran 2 ejemplos de resúmenes.

"La mayoría de los especialistas con rol de programador y con alto rendimiento son poco comprensivos y poco tolerantes".

"En casi la totalidad de los especialistas con rol de calidad y con alto rendimiento son moderadamente creativos y muy afectivos".

Además, se obtuvieron resúmenes que facilitaron la toma de decisiones durante los cortes de proyectos, como por ejemplo "Pocos proyectos que tienen aproximadamente el 50 % de los recursos humanos con baja competencia tienen perfecto el tiempo trabajado". A partir de este resumen y otros similares los decisores notaron la alta dependencia de las competencias laborales con la eficiencia en el trabajo e identificaron como medidas: elevar las competencias técnicas de los recursos humanos y reordenar los equipos de la organización. Otros ejemplos de resúmenes que facilitaron la toma de decisiones durante los cortes de proyectos son:

"Aproximadamente el 50 % de los proyectos que tienen alta cantidad de recursos humanos evaluados de mal tienen perfecto el tiempo planificado".

"Muchos proyectos que tienen perfecto el tiempo planificado tienen alta cantidad de recursos humanos evaluados de mal".

Estos 2 resúmenes evidenciaron que existían proyectos donde los recursos humanos están falseando los datos de la gestión del tiempo. Como medida se propuso revisar los proyectos afectados y corregir la dificultad en la generación de la información.

En el caso de "Aproximadamente el 50 % de los proyectos anómalos tienen muy alta cantidad de tareas planificadas para los recursos humanos" se evidencia sobrecarga de trabajo; para disminuirla, se tomaron como medidas priorizar las tareas, reajustar los tiempos, revisar qué tareas se pueden paralelizar y subcontratar otros recursos humanos.

Los resúmenes "Cercano al 33 % de los proyectos anómalos tienen muy alta cantidad de recursos materiales planificados" y "Cercano al 33 % de los proyectos anómalos que tienen alto plan de recursos materiales tienen alta tarifa horaria del equipamiento contratado" evidencian proyectos con sobreplanificación de recursos materiales. Como medidas aplicadas están la selección de alternativas más económicas, revisión de contratos y ajuste de la gestión logística.

A partir del análisis de estos resúmenes lingüísticos sobre elementos anómalos durante la planificación de proyectos se identifica que, en su mayoría, se refieren a la sobreestimación de los recursos humanos en las tareas del proyecto por lo tanto, incurren en costos más altos por usar más recursos de lo planeado. La detección de estos tipos de resúmenes lingüísticos ayuda a los gerentes de proyectos a corregir errores en la programación del proyecto y detectar el costo excesivo.

Resumen del análisis multicriterio

En ambos escenarios se aplicó análisis multicriterio donde participaron especialistas en cada área, quienes evaluaron los resúmenes respecto a 10 criterios para medir la "interpretabilidad" y la "facilidad para la toma de decisiones". Ambos criterios fueron evaluados por los especialistas obteniéndose altos valores de calificación y baja dispersión en las opiniones. Se aplicó el coeficiente de correlación *RWG* para el análisis de la concordancia entre expertos, [46] cuyo valor fue 0,97 y 0,95 en los escenarios 1 y 2 respectivamente, demostrando la concordancia. Se demuestra que los resúmenes generados por los algoritmos propuestos son fácilmente interpretables por los especialistas y útiles para la toma de decisiones.

DISCUSIÓN

Como elementos novedosos de la investigación se señala el desarrollo de nuevas extensiones de evaluación de calidad de resúmenes complementando a los existentes; incorporando elementos de teoría de conjuntos aproximados y teoría neutrosófica. Las extensiones propuestas complementan a los tradicionales respecto al "tratamiento de la incertidumbre" al mejorar la medición de los grados de certeza, indeterminación y falsedad. Además, el uso del α-corte y las aproximaciones superior e inferior facilitan la aplicabilidad de los indicadores en diferentes dominios de aplicación.

Otro resultado logrado en la presente investigación es la publicación de una biblioteca de algoritmos de sumarización lingüística combinando técnicas de computación emergente como neutrosofía, conjuntos aproximados, aprendizaje de reglas de asociación y grafos probabilísticos, mejorando la eficacia y eficiencia de los algoritmos existentes. La tabla 3 evidencia esta mejora respecto a algunos aspectos identificados como limitaciones en la problemática de la investigación, para ello se comparan los nuevos algoritmos con otros reportados en bibliografía. [21,23,25,26,27,40]

Otro resultado notorio de la investigación es el uso de lenguajes naturales controlados para la generación de resúmenes en múltiples idiomas, lo cual permite la internacionalización y uso en diferentes zonas geográficas. Este enfoque multilingüe queda validado usando métodos multicriterio de expertos en idiomas inglés, japonés y árabe, considerando los criterios de gramática, sencillez, legibilidad y ortografía. Como resultado de esta validación, en los 3 idiomas la mayoría de los criterios fueron evaluados de "alto" o "muy alto", los criterios mejores evaluados globalmente fueron "sencillez" y "ortografía", aunque el árabe y el japonés resultaron mejor evaluados que el inglés. Se identifica como línea abierta de la investigación continuar trabajando en la legibilidad de los resúmenes lingüísticos.

El ecosistema de *software* para la ayuda a la toma de decisiones con cuadro de mando es sin duda un resultado de impacto en los escenarios de toma de decisiones. En este ecosistema fueron incluidos los algoritmos propuestos en la investigación para la gestión de proyectos de I+D+i en: Empresa Xetid, [47] Empresa ETECSA, [48] Fundación de la Universidad de la Habana, [49] Parque Científico Tecnológico de la Habana, [50] Red Colaborativa de Ingeniería y Gestión de Proyectos, [51] Red de Centros Productivos de la Universidad de las Ciencias Informáticas (UCI). [52] Finalmente fue construido un repositorio de datos para investigaciones que incluye 18 bases de datos sobre gestión de proyectos, 2 sobre auditorías y 2 sobre información clínica. [17]

Conclusiones

Esta investigación aporta 4 nuevos algoritmos para la construcción de resúmenes lingüísticos bajo un enfoque multilingüe mejorando en este aspecto los algoritmos reportados en la bibliografía permitiendo generar resúmenes en español, inglés, japonés y árabe. A partir de las comparaciones se concluye que el algoritmo *RST_LDS* obtiene los mejores resulta-

Tabla 3. Comparación de los nuevos algoritmos (los que tienen el asterisco) con otros reportados en bibliografía

Algoritmos	Tratamiento correlación entre variables	Umbral variable (granularidad)	Emplea neutrosofía	Multilingüe	Identificación datos anómalos
* RST_LDS	X	X	X	X	X
* LPA_LDS	X	X	X	X	X
* PCA_LDS	X	X	X	X	
* Outlier_LDS		X	X	X	X
GA_LDS					
Apriori_LDS					

dos globales y que en general los algoritmos propuestos superaron a los reportados en la bibliografía.

Los nuevos indicadores para la evaluación de resúmenes lingüísticos mejoran el tratamiento de la indeterminación y la falsedad complementando a los indicadores reportados en la bibliografía. Fue probada la aplicabilidad de los algoritmos en diferentes entornos de toma de decisiones médicas y de gestión de proyectos. Se demostró que los resúmenes generados por los algoritmos propuestos son fácilmente interpretables por los especialistas y útiles para la toma de decisiones.

Los resultados obtenidos se encuentran completamente introducidos en más de 6 escenarios reales asociados a la toma de decisiones en la gestión de proyectos de I+D+i. Se recomienda continuar aplicando los métodos de sumarización lingüística de datos en otros escenarios para la ayuda a la toma de decisiones, explotando su flexibilidad para ser combinados con otras técnicas.

REFERENCIAS BIBLIOGRÁFICAS

1. Peláez-Aguilera MD, Espinilla M, Fernández MR, Medina J. Fuzzy Linguistic Protoforms to Summarize Heart Rate Streams of Patients with Ischemic Heart Disease. Hindawi. 2019;2019:11.
2. Dijkman R, Wilbik A. Linguistic summarization of event logs–A practical approach. Information Systems. 2017 jul;67:114-25.
3. Amghar D, Chikh AmineM. Extracting a Linguistic Summary from a Medical Database. International Journal of Intelligent Systems and Applications. 2018 dic;10(12):16-26.
4. Kaczmarek-Majer K, Hryniewicz O, Dominiak M, Święcicki Ł. Personalized linguistic summaries in smartphone-based monitoring of bipolar disorder patients. In Atlantis Press; 2019 [cited 2019 dic 11]. Disponible en: https://www.atlantis-press.com/proceedings/eusflat-19/125914826
5. Wilbik A, Barreto D, Backus G. On Relevance of Linguistic Summaries-A Case Study from the Agro-Food Domain. In: International Conference on Information Processing and Management of Uncertainty in Knowledge-Based Systems. Springer; 2020.289-300p.
6. Wu D, Mendel JM. Linguistic summarization using IF–THEN rules and interval type-2 fuzzy sets. IEEE Transactions on Fuzzy Systems. 2010;19(1):136-51.
7. Igde EY, Aydoğan S, Boran FE, Akay D. Linguistic Summarization of Structured Patent Data. In: International Scholarly and Scientific Research & Innovation. 2017;9(11):1062-5.
8. Kacprzyk J, Zadrożny S. Linguistic summarization of the contents of Web server logs via the Ordered Weighted Averaging (OWA) operators. Fuzzy Sets and Systems. 2016;285:182-98.
9. Kacprzyk J, Zadrożny S. On a fuzzy querying and data mining interface. Kybernetika. 2000;36(6):657-70.
10. Yager RR. On Linguistic Summaries of Data. Knowledge Discovery in Databases. 1991;378-89.
11. Pérez Pupo I, Piñero Pérez PY, Martín N, Bello Pérez RE. Tendencias en la sumarización lingüística de datos. Revista cubana de transformación digital. 2021;2(1):79-101.
12. Kacprzyk J, Strykowski P. Linguistic summaries of sales data at a computer retailer via fuzzy logic and a genetic algorithm. Proceedings of the 1999 Congress on Evolutionary Computation-CEC99 (Cat No 99TH8406). 1999;2:937-43.
13. Kacprzyk J, Zadrożny S. Fuzzy logic-based linguistic summaries of time series: a powerful tool for discovering knowledge on time varying processes and systems under imprecision. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery. 2016;6(1):37-46.
14. Wilbik A, Vanderfeesten I, Bergmans D, Heines S, Mook W van. Linguistic Summaries for Compliance Analysis of a Glucose Management Clinical Protocol. In: 2018 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE). 2018.1-7p.
15. García JAL, Peña AB, Piñero Pérez PY, Pérez RB. Project Control and Computational Intelligence: Trends and Challenges. International Journal of Computational Intelligence Systems. 2017;10(1):320-35.
16. Castro Aguilar GF. Modelo para el aseguramiento de ingresos en organizaciones orientadas a proyectos basado en minería de datos anómalos [Internet] Tesis Doctoral. La Habana, Cuba: Universidad de las Ciencias Informáticas; 2017 [citado 2021 abr 2]. Disponible en: https://repositorio.uci.cu/jspui/handle/123456789/7933

17. Piñero P, Pupo I, Rivero Hechavarría CC, Rojas C, Sosa R, Torres S. Repositorio de datos para investigaciones en gestión de proyectos. Revista Cubana de Ciencias Informáticas. 2019;13(1):176-91.
18. Ramos-Soto A, Martin-Rodillab P. Enriching linguistic descriptions of data: A framework for composite protoforms. Fuzzy Sets and Systems. 2019;26.
19. Marín N, Sánchez D. On generating linguistic descriptions of time series. Fuzzy Sets and Systems. 2016;285:6-30.
20. Kacprzyk J, Zadrożny S. Fuzzy linguistic data summaries as a human consistent, user adaptable solution to data mining. In: Gabrys B, Leiviskä K, Strackeljan J, editors. Do Smart Adaptive Systems Exist? Best Practice for Selection and Combination of Intelligent Methods [Internet]. Berlin, Heidelberg: Springer; 2005 [citado 2020 ene 10]. 321-40 p. (Studies in Fuzziness and Soft Computing). Disponible en: https://doi.org/10.1007/3-540-32374-0_16
21. Donis-Díaz CA, Muro AG, Bello-Pérez R, Morales EV. A hybrid model of genetic algorithm with local search to discover linguistic data summaries from creep data. Expert Systems with Applications. 2014 mar;41(4, Part 2):2035-42.
22. Donis-Díaz CA, Bello R, Kacprzyk J, others. Linguistic data summarization using an enhanced genetic algorithm. Czasopismo Techniczne. 2014;2013(Automatyka Zeszyt 2 AC (10) 2013:3-12.
23. Donis-Díaz CA, Bello R, Kacprzyk J. Using Ant Colony Optimization and Genetic Algorithms for the Linguistic Summarization of Creep Data. In: Angelov P, Atanassov KT, Doukovska L, Hadjiski M, Jotsov V, Kacprzyk J, ***et al.,*** editors. Intelligent Systems'2014. Cham: Springer International Publishing; 2015. 81-92 p. (Advances in Intelligent Systems and Computing).
24. Kacprzyk J, Zadrożny S. Linguistic data summarization: a high scalability through the use of natural language? In: Scalable Fuzzy Algorithms for Data Management and Analysis: Methods and Design. IGI Global; 2010.214-37 p.
25. Pérez Pupo I, Piñero Pérez PY, Bello Pérez RE, García Vacacela RC, Piñero Ramírez PE, Piñero Ramírez CM. Aplicaciones de la sumarización lingüística de datos en la toma de decisiones en gestión de proyectos. In: V Conferencia Internacional en Ciencias Computacionales e Informáticas (CICCI' 2020). Informática CICCI´2020; 2020.
26. Pérez Pupo I, Santos Acosta O, Bello Pérez RE, Piñero Pérez P. Algorithms for linguistic data summarization, help in decision-making in project-oriented organizations. In: XXII Ibero-American Conference on Software Engineering, CIbSE 2019. Springer; 2019. 633-40 p.
27. Pérez Pupo I, Piñero Pérez PY, Vacacela RG, Bello R, Acuña LA. Discovering Fails in Software Projects Planning Based on Linguistic Summaries. International Joint Conference on Rough Sets Lecture Notes in Computer Science ISSN 0302-9743, Springer, 12179 LNAI, ISBN 978-3-030-52704-4 [Internet]. Lecture Notes in Computer Science, Springer; 2020. 365-75p. (12179 LNAI). Disponible en: https://link.springer.com/content/pdf/10.1007%2F978-3-030-52705-1_27.pdf
28. Zadeh LA. A computational approach to fuzzy quantifiers in natural languages. Computers & Mathematics with applications. 1983;9(1):149-84.
29. Kacprzyk J, Zadrożny S. Linguistic database summaries and their protoforms: towards natural language based knowledge discovery tools. Information Sciences. 2005;173(4):281-304.
30. Lasisi A, Attoh-Okine N. Principal components analysis and track quality index: A machine learning approach. Transportation Research Part C: Emerging Technologies. 2018 Jun 1;91:230-48.
31. Dunteman GH. Principal Components Analysis. SAGE, ISBN: 978-0-8039-3104-6; 1989. 98 p.
32. Naik G. Advances in Principal Component Analysis. Research and Development. Springer. 2018. 252 p.
33. Wasilewska A. Apriori algorithm. Lecture Notes, accessed [Internet]. 2007.
34. Medina JE, Hernández J, Hernández R, Pérez A, Hechavarría. A, González R. Generación de conjuntos de ítems y reglas de asociación [Internet]. Cenatav; 2007. (Serie Gris). Report No.: 2143.
35. Chow C, Liu C. Approximating discrete probability distributions with dependence trees. IEEE Transactions on Information Theory. 1968 May;14(3):462-7.
36. Soto M. A single connected factorized distribution algorithm and its cost of evaluation. Tesis de Doctorado. Universidad de La Habana, Cuba: Instituto de Cibernética, Matemática y Física. Departamento Matemática Interdisciplinaria; 2003.
37. Aggarwal CC, Sathe S. Outlier Ensembles: An Introduction. Springer; 2017. 288 p.
38. Ruiz-Shulcloper J. Reconocimiento lógico combinatorio de patrones: teoría y aplicaciones. PhD Thesis. Universidad Central Marta Abreu de Las Villas, Santa Clara: Centro de Investigaciones de Tecnologías de Avanzadas CENATAV; 2009.
39. Yager RR. A new approach to the summarization of data. Information Sciences. 1982;28(1):69-86.
40. Wilbik A, Kaymak U, Dijkman RM. A method for improving the generation of linguistic summaries. In: 2017 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE). 2017. 1-6 p.
41. Pérez Pupo I. Repositorio de investigaciones en gestión de proyectos [Internet]. 2021. Disponible en: https://gespro.uci.cu/projects/repositorio-de-investigaciones-en-gestion-de-proyectos/dmsf
42. Mandel I. Expert system for Statistics STATEX – 30 years after [Internet]. Rochester, NY: Social Science Research Network; 2020 jul [cited 2021 mar 29]. Report No.: ID 3664990. Disponible en: https://papers.ssrn.com/abstract=3664990
43. Piñero Pérez P, Pérez Pupo I, Piñero Cruz P. Software "Suite Business-Redmine Sistema para la toma de decisiones empresariales y la gestión de proyectos 19.05". Registrado en el Centro Nacional de Derecho de Autor de Cuba (CENDA); No. de registro 4076-12-2019, 2019.
44. Pérez Pupo I, García Vacacela R, Piñero Pérez P, Sadeq G, Peña Abreu M. Experiencias en el uso de técnicas de softcomputing en la evaluación de proyectos de software. Revista Investigación Operacional. 2020;41(1):106-17.
45. López P. Procedimiento para la aplicación de test de personalidad como apoyo a la gestión de recursos humanos en proyectos informáticos. Tesis de maestría defendida en Universidad de las Ciencias Informáticas, La Habana, Cuba, 2017.
46. Benavente Reche AP. Medidas de acuerdo y de sesgo entre jueces [PhD Thesis]. Facultad de Psicología, Universidad de Murcia; 2009. Diponible en: https://digitum.um.es/digitum/bitstream/10201/35117/1/TAPBR.pdf
47. Gestión de Proyectos XETID [Internet]. Dirección Integrada de Proyectos de gestión, desarrollo e innovación de la Xetid. 2021. Disponible en: https://proyectos.xetid.cu/

48. Dirección de Desarrollo e Investigación de la Empresa de Telecomunicaciones de Cuba (Etecsa). Gestión y dirección de los proyectos de desarrollo e investigación de ETECSA. 2021. Disponible en: https://gespro.etecsa.cu dentro de la red interna de ETECSA.
49. Proyectos FUNDACIÓN UH. Ecosistema para la Gestión de la Innovación en la Fundación de la Universidad de la Habana. 2021. Disponible en: https://fundacionuh.xutil.cu/ dentro de la red interna de la Institución.
50. Solución de Inteligencia de Negocios [Internet]. Parque Científico Tecnológico de la Habana (PCT 3CE). 2021. Disponible en: https://cm.3ce.cu/
51. Red Colaborativa de Ingeniería y Gestión de Proyectos [Internet]. Gestión y Dirección de Proyectos de Innovación-Desarrollo-Investigación en Gestión de Proyectos y Transformación Digital. 2020 [citado 2022 nov 21]. Disponible en: https://gespro.uci.cu/
52. Plataforma para Dirección Integrada de Proyectos, Universidad de las Ciencias Informáticas. Ecosistema de Gestión de Proyectos. 2020 [citado 2022 Nov 21]. Disponible en: https://gp.prod.uci.cu/

Recibido: 07/06/2024
Aprobado: 25/06/2024

Conflictos de intereses

Los autores declaran que no existen conflictos de intereses entre ellos, ni con la investigación presentada.

Contribuciones de los autores

Conceptualización: Iliana Pérez, Pedro Piñero
Curación de datos: Iliana Pérez
Análisis formal: Iliana Pérez, Pedro Piñero, Rafael Bello
Adquisición de fondos: no se emplearon fondos adicionales.
Investigación: Iliana Pérez, Pedro Piñero, Rafael Bello, Roberto García Vacacela
Metodología: Iliana Pérez, Pedro Piñero
Administración del proyecto: Iliana Pérez
Recursos: Pedro Piñero.
Software: Iliana Pérez, Pedro Piñero
Supervisión: Pedro Piñero, Rafael Bello
Validación: Iliana Pérez, Pedro Piñero, Rafael Bello, Roberto García Vacacela
Visualización: Iliana Pérez, Pedro Piñero
Redacción-borrador original: Iliana Pérez
Redacción-revisión y edición: Iliana Pérez, Pedro Piñero, Rafael Bello

Financiamientos

No se utilizó financiamiento específico para realizar la investigación presentada.

Cómo citar este artículo

Pérez Pupo I, Piñero Pérez PY, Bello Pérez R, García Vacacela R. Ecosistema ***software*** para el aprendizaje y toma de decisiones basados en sumarización lingüística de datos. An Acad Cienc Cuba [internet] 2024 [citado en día, mes y año];14(2):e1606. Disponible en: http://www.revistaccuba.cu/index.php/revacc/article/view/1606