REVISTA CUBANA
DE TRANSFORMACIÓN DIGITAL

RNPS 2487 • ISSN 2708-3411
Vol. 2 • Nro. 1 • enero-marzo 2021 • págs. 79-101

ARTÍCULO DE REVISIÓN

Tendencias en la sumarización lingüística de datos

Linguistic Data Summarization and Overview

Iliana Pérez Pupo
iliperezpupo@gmail.com, iperez@uci.cu • http://orcid.org/0000-0003-1433-0601

Pedro Yobanis Piñero Pérez
pppyob@gmail.com, ppp@uci.cu • http://orcid.org/0000-0002-7635-8290

Nayma Martín Amaro
nayma@uci.cu • http://orcid.org/0000-0002-1003-7224

UNIVERSIDAD DE LAS CIENCIAS INFORMÁTICAS

Rafael Esteban Bello Pérez
rbellop@uclv.edu.cu • http://orcid.org/0001-5567-2638

UNIVERSIDAD CENTRAL "MARTA ABREU" DE LAS VILLAS, CUBA

Recibido: 2020-12-23 • Aceptado: 2021-01-30

RESUMEN

Las técnicas de sumarización lingüística de datos han surgido para ayudar a descubrir relaciones complejas entre variables y presentar la información en lenguaje natural. En el desarrollo de estas técnicas se combinan la inteligencia artificial, la estadística, el aprendizaje automático, entre otras áreas del conocimiento humano. Esta investigación tiene como objetivo realizar un marco teórico referencial de la temática que permita a los investigadores analizar las tendencias en los métodos de generación de resúmenes lingüísticos de datos, las estrategias de validación empleadas en las investigaciones y las principales áreas de aplicación. Como conclusiones se identifica la necesidad de mejora de los métodos de validación, las técnicas emergentes en la generación de resúmenes y la posibilidad del empleo de los resúmenes en problemas de predicción entre otros.

Palabras clave: descubrimiento de conocimiento; resúmenes lingüísticos de datos; sumarización lingüística de datos.

ABSTRACT

Linguistic data summarization techniques have emerged to help discover complex relationships between variables and present information in natural language. In the development of these techniques, artificial intelligence, statistics, machine learning, among other areas of human knowledge, are combined. This research aims to carry out a study of the state of the art of the subject that allows researchers to analyze the evolution and trends in its development. Trends in linguistic summaries generation methods, validation strategies used in researches, and main areas of application are analyzed. As conclusions, the need to improve validation methods, emerging techniques in the generation of summaries and the possibility of using summaries in prediction problems, among others, are identified.

Keywords: *knowledge discovery; linguistic data summarization; linguistic summaries.*

INTRODUCCIÓN

Uno de los problemas latentes en los procesos de toma de decisiones está asociado al descubrimiento de patrones de comportamiento y su interpretabilidad. Este problema se incrementa con el rápido crecimiento de los datos generados por los sistemas de información. Por lo general, los datos tomados directamente de los sistemas de información no son legibles ni entendibles a simple vista por los decisores, lo que dificulta el proceso de toma de decisiones. Surge así el llamado dilema de "datos ricos, información pobre" presente en la toma de decisiones (Wu & Mendel, 2010). En este sentido constituye objeto de interés para los decisores la existencia de herramientas que permitan el descubrimiento de las dependencias no triviales ocultas en los datos.

En general, las técnicas de descubrimiento de conocimiento a partir de los datos se han clasificado en dos clases: técnicas predictivas y las técnicas descriptivas. El objetivo de las técnicas predictivas es predecir el valor de un atributo específico (conocido como variable dependiente) basado en los valores de otros atributos; mientras que el objetivo de las técnicas descriptivas es explorar patrones (correlaciones, tendencias, grupos, trayectorias, o anomalías) que resumen las relaciones subyacentes en los datos (Janusz Kacprzyk & Zadrożny, 2000).

En este contexto surge la sumarización lingüística de datos (SLD, también conocido como LDS del inglés *linguistic data summarization)* como una de las técnicas de descubrimiento de conocimiento descriptivo, con un enfoque interesante y prometedor para producir resú-

menes a partir de una base de datos que utiliza lenguaje natural (Yager, 1991). Autores como Yager y Zadeh fueron pioneros en el desarrollo de la sumarización lingüística de datos. Sus publicaciones marcaron pautas en la conceptualización y el desarrollo posterior de la sumarización lingüística de datos como técnica en el campo de la computación blanda (*soft computing)* (Ronald R Yager, 1982; Zadeh, 1983). Diferentes autores han caracterizado esta técnica como se muestra a continuación.

Kacprzyk (1999) plantea que los resúmenes lingüísticos están en línea con los nuevos paradigmas de la computación con palabras, y pueden captar muy bien la esencia de los datos. Posteriormente, Zadrożny identificó que la esencia de esta técnica es que un conjunto de datos se puede resumir lingüísticamente con respecto a un atributo o atributos seleccionados, mediante proposiciones cuantificadas lingüísticamente (Kacprzyk & Zadrożny, 2005).

Boran, Akay & Yager (2016) plantean que la SLD es una de las poderosas técnicas de descubrimiento de conocimiento descriptivo de un gran conjunto de datos, capaz de extraer conocimiento potencial, útil y abstracto de datos tanto numéricos como categóricos. El objetivo es descubrir los patrones que resuman las relaciones existentes entre atributos en una gran base de datos con representaciones conceptuales más abstractas.

Por su parte, para Anna Wilbik, el objetivo de los resúmenes lingüísticos es proporcionar una descripción general rápida, pues facilitan la comprensión de grandes cantidades de datos al describir las principales propiedades de los datos en forma lingüística (Wilbik & Dijkman, 2016). Mientras que otros autores identifican a los resúmenes lingüísticos como un intento de describir mediante expresiones lingüísticas, patrones que emergen en los datos (Eciolaza, Pereira-FariñA, & Trivino, 2013).

En la bibliografía consultada se identifican pocos trabajos de revisión asociados a esta temática. Por ejemplo, autores como Boran y colaboradores (2016) se han centrado en los métodos de generación y evaluación de resúmenes y Marín y Sánchez (2016) muestran una colección de ejemplos de resúmenes generados por diferentes autores entre el 2001 hasta el 2016. Mientras que Hudec, Bednárová, & Holzinger (2018) en sus investigaciones se centran en sintetizar las protoformas utilizadas, que abarcan el análisis de diversos artículos, desde los de Ronald R. Yager (1982) hasta más recientes en el 2013. Mientras que los autores Ramos-Soto y Martin-Rodillab (2019) exponen una tabla en la que se lista los tipos de protoformas más empleadas entre el 2006 hasta el 2016. Estas revisiones se centran solo en aspectos específicos de la teoría, no aplican técnicas de revisiones sistemáticas o de meta-análisis y no logran caracterizar completamente a cada uno de los trabajos que referencian.

A partir de esta situación problemática se plantea como objetivo de este trabajo construir un marco teórico referencial a través de una revisión terciaria que permita analizar la evolución y las tendencias en el desarrollo de la sumarización lingüística de datos y sus aplicaciones. El trabajo se encuentra organizado de la siguiente forma: la segunda sección Metodología explica el diseño experimental empleado, la sección Desarrollo muestra los resultados del análisis y finalmente se presentan las Conclusiones.

METODOLOGÍA

El estudio que se presenta constituye una investigación exploratoria (Hernández-Sampieri & Torres, 2018), en la cual se desarrolla el siguiente diseño de experimentos:

Paso 1: Diseño de un protocolo de revisión sistemática que permite construir un marco teórico referencial asociado a la temática.

Paso 2: Análisis a profundidad de la bibliografía consultada por cada una de las categorías establecidas.

Paso 3: Aplicación de técnicas de estadística descriptiva y meta-análisis para la presentación de los resultados de la investigación.

Paso 4: Identificar líneas abiertas de investigación.

En relación con lo plateado, se diseñó un protocolo de revisión sistemática (Littell, Corcoran, & Pillai, 2008) y se analizó la evolución histórica de la sumarización lingüística en diferentes etapas: publicaciones antes del 2000, publicaciones entre el 2000-2015 y a partir del 2015 hasta el 2020.

Protocolo de revisión sistemática aplicado:

1. Definición del objeto de estudio, campo de acción y objetivo de la investigación.
   a). Objeto de investigación: sumarización lingüística de datos
   b). Objetivo general: construir un marco teórico referencial asociado a los métodos relacionados con la sumarización lingüística de datos, tendencias, y aplicaciones en la gestión de proyectos.
   c). Campo de investigación: estructuras, métodos y aplicaciones de generación de resúmenes lingüísticos de datos
2. Definición de un gestor bibliográfico.
3. Definir fuentes de información académica para el desarrollo de la revisión: *Semantic Scholar, Google Scholar* y *Scopus* y otros meta-buscadores basados en ciencia abierta.
4. Definición del siguiente conjunto de frases claves para la realización de las búsquedas: *linguistic data summarization, linguistic data summaries, linguistic summarization, linguistic summaries.*
5. Definir las metas del análisis bibliométrico en forma de preguntas de investigación y criterios de inclusión-exclusión:
   a). ¿Cómo ha sido la tendencia de las publicaciones por año?
   b). ¿Cuáles son los principales autores?
   c). ¿Cuáles son las afiliaciones y países de los principales autores?
   d). ¿Cómo se distribuyen las publicaciones considerando los tipos de documentos en: artículos, libros, tesis y memorias de conferencias o congresos?
   e). Exclusión de trabajos publicados en espacios con poco nivel de arbitraje.
   f). Exclusión de trabajos de minería de textos no asociados al uso de técnicas de sumarización lingüística de datos.

6. Clasificar y filtrar las publicaciones en el siguiente conjunto de categorías:
    - Clásicas: se refiere a publicaciones pioneras en sumarización lingüística de datos, en las que se exponen los principios fundamentales que marcan pastas en la teoría.
    - Extensiones a las teorías: se refiere a publicaciones que extienden la teoría planteada en las publicaciones entendidas como clásicas; no marcan pautas que cambien significativamente los métodos propuestos con anterioridad, aunque si desarrollan aportes al conocimiento.
    - Resultados de aplicación: publicaciones que se concentran en el empleo de la teoría existente en escenarios prácticos concretos.
    - Revisiones terciarias: se refiere a artículos de revisión de las tendencias y la evolución en la temática en cuestión.
7. Sintetizar las principales tendencias.

Como segundo paso en la metodología empleada se realiza un análisis detallado de cada uno de los trabajos y se caracterizan respecto a los siguientes elementos:

- Estructura de los resúmenes que se generan (protoformas).
- Métodos o técnicas para la generación de los resúmenes lingüísticos de datos.
- Principales técnicas y métodos de validación empleados en las investigaciones.
- Áreas de aplicación de la propuesta.

Luego se aplican de técnicas de estadística descriptiva y el meta-análisis para la presentación de los resultados de la investigación (Littell, *et al.,* 2008). Finalmente se analizan y proponen líneas abiertas en el objeto de estudio analizado.

DESARROLLO

A partir del estudio realizado se identifica que autores europeos y norteamericanos fueron pioneros en el desarrollo de la teoría en la década del 80, mientras que en la actualidad se observa una mayor dispersión de los trabajos respecto a áreas geográficas (ver figura 1).

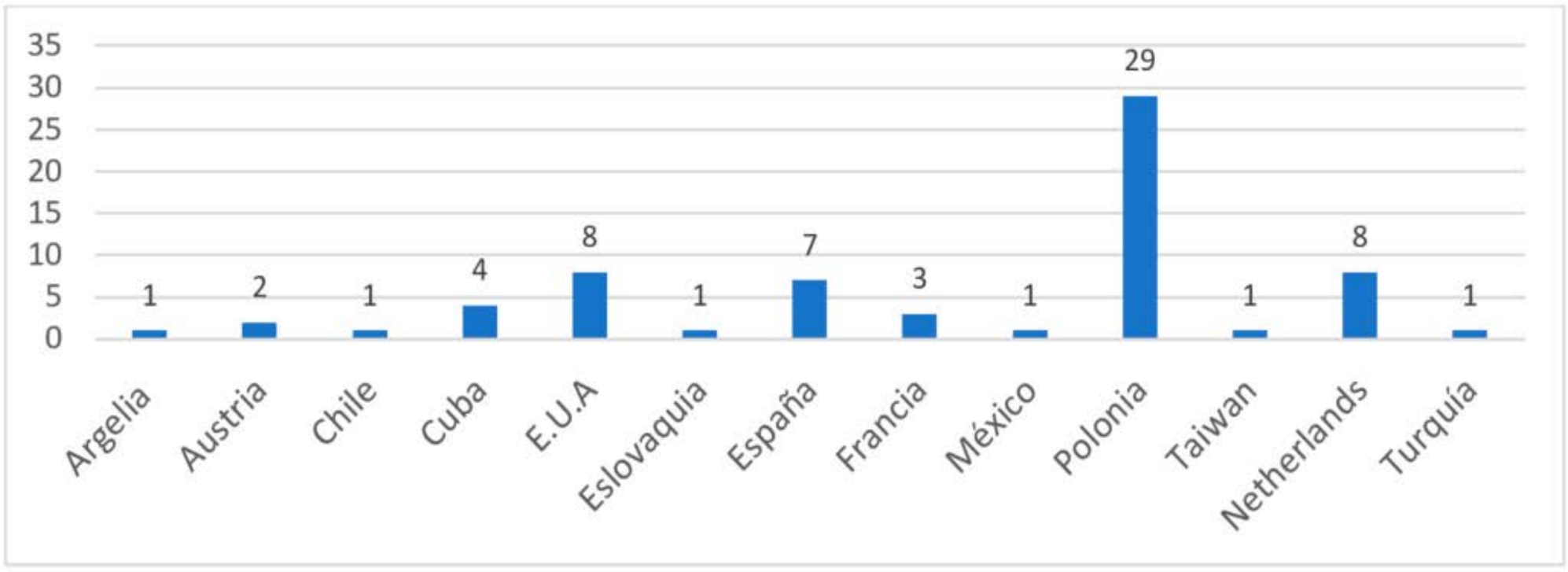

Figura 1. Publicaciones por países de la bibliografía consultada.

Una mejor caracterización de la bibliografía consultada para este trabajo se muestra a continuación:

- La mayoría de las investigaciones consultadas son artículos en revistas, se identifica que en esta área temática hay pocas tesis de doctorado u otro tipo publicadas (figura 2).
- La mayoría de las publicaciones consultadas están indexadas en *Scopus* y *Web Of Science* (WOS) (ver figura 3).
- Las fuentes de información fundamentales y las editoriales principales para las investigaciones consultadas fueron *IEEE, Elsevier* y *Springer* respectivamente (figura 4).

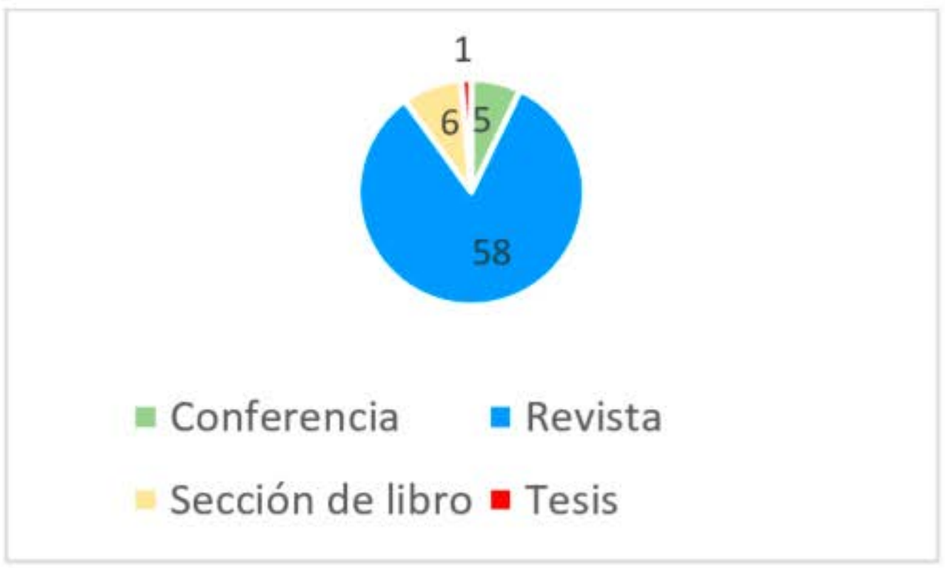

Figura 2. Bibliografía por tipo de publicación.

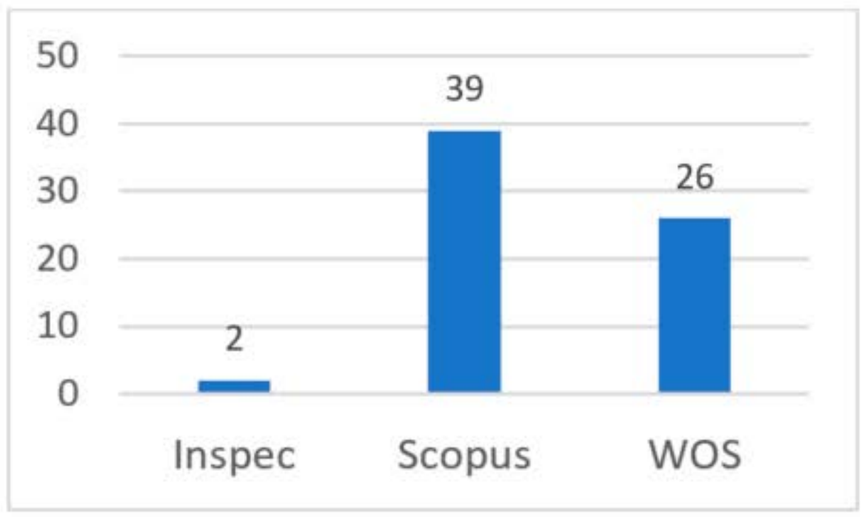

Figura 3. Indexado de las publicaciones en revistas y conferencias analizadas.

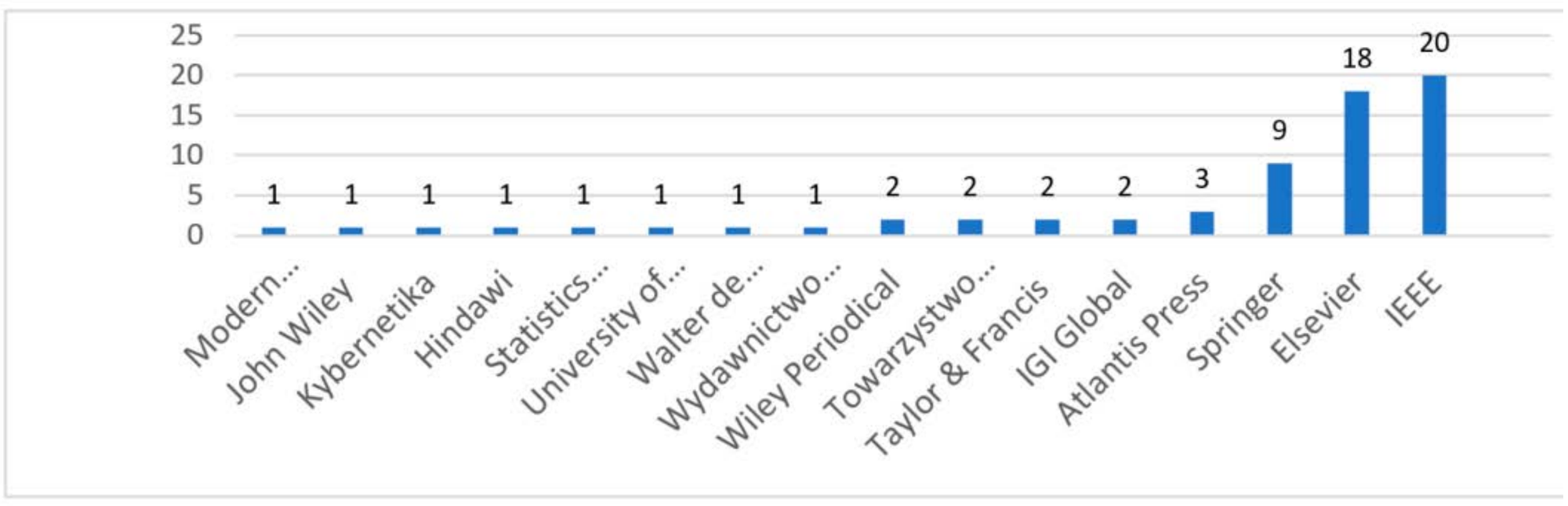

Figura 4. Fuentes y editoriales principales de la bibliografía analizada en profundidad.

En la bibliografía consultada se identifica un conjunto de elementos que caracterizan o forman parte de la notación establecida en esta área temática (Yager, 1982; Donis-Díaz, Bello, & Kacprzyk, 2015; Wu & Mendel, 2010):

- *D:* base de datos, por ejemplo, de trabajadores, o proyectos.
- $Y = \{y_1, ..., y_n\}$, conjunto de objetos (registros) en la base de datos.
- $A = \{A_1, ..., A_m\}$, conjunto de atributos (variable difusa) que caracterizan a los objetos Y, Salario, edad.
- $A_j(y_i)$*:* denota el valor del atributo A_j para el objeto y_i, por ejemplo, “Bajo salario” para el atributo “salario”.
- *Q:* cuantificador, es un conjunto borroso con el universo de discurso en el intervalo [0, 1] que expresa una cantidad, por ejemplo, “La mayoría”, “el 60 %” o “Más de la mitad”.

- *R:* Calificador o filtro, es otro atributo con un valor lingüístico (predicado borroso) que determina un subconjunto borroso del objeto y_i, por ejemplo, "joven" para el atributo edad.
- *S:* Sumarizador, es un atributo con un valor lingüístico (predicado borroso) definido en el dominio del atributo A_j, por ejemplo, "Bajo salario" para el atributo "salario".
- *T:* Grado de verdad (validez) del resumen, es un número del intervalo [0, 1] que evalúa el grado de verdad del resumen; generalmente, solo resúmenes con un alto valor de la *T* son de interés.

Las sintaxis más empleadas para la construcción de los resúmenes son las siguientes (Janusz Kacprzyk & Zadrożny, 2016a; Donis-Díaz, et al., 2015; Wilbik, Kaymak, & Dijkman, 2017):

- *Qy's are S:* para resúmenes lingüísticos sin filtro, en los que la declaración tiene la forma: "*Q* objetos son *S*".
  Ejemplo: "La mayoría de los trabajadores son puntuales".
- *QRy's are S:* para resúmenes lingüísticos con filtro, en los que la declaración tiene la forma: "*Q* objetos *R* son *S*", donde *R* es un filtro que puede estar conformado por uno o varios atributos que califican al objeto.
  Ejemplo: "La mayoría de los trabajadores jóvenes son impuntuales".

Las estructuras de los resúmenes lingüísticos se han conceptualizado en protoformas, la siguiente subsección resume las tendencias en este sentido.

Evolución y las tendencias en las protoformas para la construcción de resúmenes

Las protoformas que más se han empleado en la bibliografía para la construcción de los resúmenes lingüísticos siguen la estructura propuesta por Zadeh (2002) y extendidas posteriormente por Kacprzyk y Zadrożny (2005). La tabla 1 sintetiza los elementos principales de estas protoformas.

Tabla 1. Clasificación de la sumarización lingüística (Janusz Kacprzyk & Zadrozny, 2009).

Tipo	Protoforma	Conocido	Duda	Comentarios
0	*QRy's* are *S*	Todo	T	Resúmenes condicionales a través de consultas ad-hoc
1	*Qy's* are *S*	S	Q	Resúmenes simples a través de consultas ad-hoc
2	*QRy's* are *S*	S R	Q	Resúmenes condicionales a través de consultas ad-hoc
3	*Qy's* are *S*	Q $S^{Estructura}$	S^{Valor}	Resúmenes sencillos orientados a valores
4	*QRy's* are *S*	Q $S^{Estructura}$, R	S^{Valor}	Resúmenes orientados a valores condicionales
5	*QRy's* are S	Nada	Q R S	Reglas difusas generales

En estas protoformas se establecen jerarquías en las que se transita por los niveles de mayor a menor abstracción y se caracterizan por los siguientes elementos:

- El término $S^{Estructura}$ significa que en el resumen es conocido cuáles variables conforman el mismo; mientras que S^{Valor} denota que se desconoce el valor del sumarizador.
- Protoforma 0: responde a la estructura *QRy's are S.* Es la de menor nivel de abstracción, pues se asume que todos los elementos que conforman el resumen son conocidos. Lo que

se desea conocer es su grado de verdad *T* (Zadeh, 1983) (Janusz Kacprzyk & Zadrożny, 2005).

- Protoforma 1: estructura *Qy's are S,* se conoce el sumarizador (*S*), y se quiere descubrir el cuantificador (*Q*). En el caso de la aplicación de las protoformas 1 y 2, pueden generar resúmenes a partir de sentencias SQL y sus extensiones basadas en FQUERY (Janusz Kacprzyk & Zadrożny, 2005).
- Protoforma 2: estructura *QRy's are S,* se conocen el sumarizador (*S*) y el filtro (*R*), se desea descubrir el cuantificador (*Q*). Al igual que en las protoformas 0 y 1, estos resúmenes pueden ser obtenidos a través de consultas a la base de datos, aunque se requiere un poco más de esfuerzo respecto a la protoforma 1.
- Protoforma 3: estructura *Qy's are S,* se conocen el cuantificador (*Q*) y la estructura del resumen, se quiere descubrir el sumarizador (*S*).
- Protoforma 4: estructura *QRy's are S,* se conocen el cuantificador (*Q*) y la estructura del resumen, se quiere descubrir el sumarizador (*S*) y el filtro (*R*).
- Protoforma 5: estructura *QRy's are S.* Es el de mayor nivel de abstracción, pues no se conoce ningún elemento y por tanto se trata de descubrir todo. La sumarización en este caso, puede implicar un alto consumo de tiempo debido al tamaño del espacio de búsqueda; pero los resúmenes podrían ser más interesantes. Estos son los resúmenes más complejos y son el objetivo principal de este trabajo.

En los resúmenes lingüísticos los tipos de protoformas más empleados responden a: protoformas clásicas, protoformas de series de tiempo y protoformas temporales (Hudec, *et al.*, 2018).

Las protoformas clásicas sumarizan los atributos de todo el conjunto de datos o las relaciones entre ellos (Yager, 1982; Kacprzyk & Zadrożny, 2005). Estos resúmenes son de la estructura *Qy's are S* y *QRy's are S,* como se mencionó anteriormente. Un ejemplo de resumen con este tipo de protoforma es: "La mayoría de las casas tienen alto consumo de gas" y "La mayoría de las casas viejas tienen alto consumo de gas", respectivamente. Estas protoformas han sido usadas en disímiles escenarios de aplicación, con aplicaciones en la gestión de glucosa en la Unidad de Cuidados Intensivos en los Países Bajos; un ejemplo de resumen obtenido es "Todos los bolos de insulinas administrados son más altos que el protocolo" (Wilbik, Vanderfeesten, Bergmans, Heines, & Mook, 2018a).

En la bibliografía consultada se encuentran investigaciones que particularizan las protoformas para problemas específicos (Hudec, et al., 2018). Por ejemplo, en algunas investigaciones relacionadas con series de tiempo se introducen protoformas con la estructura: *Q Bs are A* Q_T *time,* donde Q_T es un cuantificador aplicado al atributo tiempo. Algunos resúmenes que siguen estas protoformas son: "la mayoría de las tendencias del tópico B son de variabilidad baja" y "cerca de la mitad de los negocios pequeños tienen tiempo de respuesta pequeño la mayor parte del tiempo".

Existen algunas extensiones asociadas a secuencia de eventos o actividades, (Wilbik & Dijkman, 2016) podemos encontrar resúmenes como "cuando el tiempo de procesamiento

era largo, el caso contenía una secuencia como *'abcdefgh' "*. Otros ejemplos de estos autores se presenta en el análisis de logs asociados a registros de procesos de apelación de la municipalidad de Dutch, (Wilbik, *et al.*, 2017) en los que se obtienen resúmenes tales como "muchos casos tienen tiempo de operación corto".

Otras extensiones a las protoformas se presentan en escenarios de aplicación en los que se quiera incorporar épocas o fenómenos temporales, donde la bibliografía consultada emplea protoformas que no usan cuantificadores lingüísticos. Este tipo de resúmenes tienen la estructura *P, datos son A,* donde *P* es un término temporal y *A* es un término lingüístico borroso. Por ejemplo, "regularmente en otoño, las precipitaciones con altas".

También, podemos encontrar modelos híbridos en que se introducen protoformas temporales simples como "T_t entre todos los segmentos, *Q* es/son *P*", para el análisis de logs en servidores web (Janusz Kacprzyk & Zadrożny, 2016b). En este caso se generan resúmenes tales como "Recientemente, entre todos los segmentos, la mayoría están aumentando lentamente"; y protoformas temporales extendidas "T_t entre todos los segmentos *R, Q* es/son *P*", y generan resúmenes como "Inicialmente, entre todos los segmentos cortos, la mayoría aumentan lentamente" siendo T_t un término temporal.

Existen otras variantes de protoformas relacionadas con datos sobre el clima (Ramos-Soto & Martin-Rodillab, 2019), en que los autores establecen protoformas con la estructura: *Q Xs are TEMP* y *Q Ts are TEMP.* En esa investigación los autores establecen que *Q* es el cuantificador, *Xs* es el referencial (temperatura máxima diaria, en un mes dado de una locación determinada), *TEMP* es uno de los términos lingüísticos de la variable lingüística "Temperatura" y *Ts* se refiere a una de las subdivisiones semanales del mes. Los autores Ramos-Soto y Martin-Rodillab combinan estos dos tipos de protoformas con los términos, "pero" y "especialmente" para establecer sentencias de contraste y de énfasis, respectivamente. A continuación, se muestran diferentes ejemplos de ambos casos:

Ejemplos de protoforma de contraste:

- "Muchos valores en Ancares son normales", pero
- "La mayoría de los valores de la tercera semana en Ancares son calientes"
- "La mayoría de los valores de la cuarta semana en Ancares son fríos"

Ejemplo de protoforma de énfasis:

- "Muchos valores en Pontevedra-Campolongo son calientes", especialmente
- "Muchos valores de la segunda semana en Pontevedra-Campolongo son muy calientes"

Otras extensiones a las protoformas que proponen en la bibliografía se concentran en el uso de múltiples sumarizadores en que se generan resúmenes como "Muchos casos tienen actividades Confirmar recepción, Registrar apelación y persona A involucrada" (Wilbik, et al., 2017a).

En las investigaciones consultadas, generalmente los autores construyen los resúmenes para lenguajes específicos. Se identifica en este sentido como una línea abierta a la investigación la posibilidad de representación de un mismo resumen con diferentes lenguajes naturales controlados (Kuhn, 2014) lo que facilita la internacionalización de los resultados.

Métodos o técnicas para la generación de los resúmenes lingüísticos de datos

En la bibliografía consultada se identifican diferentes tendencias para la construcción de resúmenes lingüísticos entre las que se destacan:

- Construcción de resúmenes a partir de técnicas de estadística descriptiva (Boran, *et al.,* 2016; Khedidja, Allel, & Mohand, 2020).
- Generación de resúmenes lingüísticos a partir de lenguaje de consultas, inicialmente desarrollados por J. Kacprzyk, S. Zadrożny y P. Strykowskien (Janusz Kacprzyk, 1999).
- Generación de resúmenes lingüísticos a partir de reglas de asociación (Janusz Kacprzyk & Zadrożny, 2003) y de reglas de producción (Dubois & Prade, 1992).
- Generación de resúmenes lingüísticos a partir de meta-heurísticas, los primeros trabajos se generan por George y Srikant (1996).
- Generación de resúmenes a partir de técnicas de agrupamiento (Wilbik & Dijkman, 2016), y combinaciones con problemas de generación de resúmenes a partir de datos anómalos (*outliers)* (Pérez, Piñero, Vacacela, Bello, & Acuña, 2020).
- Generación de resúmenes a partir conjuntos aproximados (Pérez, Piñero, Bello, Acuña, & Vacacela, 2020).

Enfoque estadístico y enfoque basado en consultas

Las técnicas de estadística descriptiva generalmente son empleadas para la construcción de resúmenes basados en las protoformas 1 y 2 (Boran, *et al.,* 2016). Estas técnicas se caracterizan generalmente por generar resúmenes cuantitativos y con una limitada capacidad de descubrimiento de nuevo conocimiento (Khedidja, et al., 2020).

Otro ejemplo del enfoque basado en estadística descriptiva se encuentra en Kaczmarek-Majer, Hryniewicz, Dominiak, & Święcicki (2019), en que se presentan resultados preliminares sobre resúmenes lingüísticos aplicados a la duración media diaria de las llamadas salientes basadas en teléfonos inteligentes. En este trabajo los algoritmos fueron aplicados a pacientes con trastorno bipolar. El estudio tuvo como objetivo el desarrollo de una herramienta de diagnóstico basada en el uso de teléfonos inteligentes para la detección y predicción de los primeros síntomas de un cambio de fase. Los resúmenes lingüísticos obtenidos en esta investigación son personalizados, basados en la protoforma corta que toma la forma "Entre todos los objetos que pertenecen al paciente *A, Q* son *P*", donde *Q* es el cuantificador y *P* es el sumarizador.

Como ventaja se identifica el procesamiento de resúmenes lingüísticos personalizados en lugar de números reales, para así evitar el almacenamiento de datos confidenciales sobre los registros históricos del paciente que son procesados con el empleo estadígrafos.

En la generación de resúmenes lingüísticos a partir de lenguaje de consultas se destaca el Paquete FQUERY desarrollado por Kacprzyk y Zadrożny (1999, 2000). Estos autores utilizan consultas difusas que extienden al SQL para la construcción de resúmenes con las protoformas descritas en la tabla 1, es un paquete integrado al gestor *Microsoft Access* (Janusz Kacprzyk, Zadrożny, & Dziedzic, 2014; Janusz Kacprzyk & Zadrożny, 2016b). Otros ejemplos son propuestos por Janusz Kacprzyk & Zadrożny (1995) y Rasmussen & Yager (1999) en que

se generan resúmenes lingüísticos a partir de un lenguaje de consulta llamado *SummarySQL.* Este enfoque tiene como ventaja la flexibilidad y versatilidad de los lenguajes de consultas como los basados en SQL, algunos trabajos representativos de este enfoque se relacionan a continuación.

Tanto el enfoque estadístico como el basado en consultas generalmente trabajan con datos numéricos, no tienen en cuenta la correlación entre las variables, se basan simplemente en una observación de los objetos, identifican cosas explícitas, pero no implícitas, por lo que pueden presentar dificultades para identificar outliers.

Enfoque basado en reglas

El enfoque de generación de resúmenes a partir de reglas tiene su base en técnicas de construcción de reglas de producción como los árboles de decisión y otros (Wu & Mendel, 2011). En este sentido se destacan autores como Dubois y Prade (1992), Wilbik y colaboradores (2017), Janusz Kacprzyk & Zadrożny (2003), Wu, Mendel y Joo (2010).

En Dubois y Prade (1992) se generan reglas de inferencia gradual de la forma "Cuanto más *X* es *F,* más *Y* es *G*", que expresan un cambio progresivo del grado en que la entidad *Y* satisface la propiedad gradual *G* cuando se modifica el grado en que la entidad *X* satisface la propiedad gradual *F.* Luego las reglas son transformadas empleando conjuntos borrosos hasta convertirlas en resúmenes lingüísticos.

Kacprzyk y Zadrożny (2003) y Wu y colaboradores (2010) proponen la generación de resúmenes lingüísticos a partir de reglas de asociación y reglas de producción respectivamente. En estos trabajos se logra encontrar resúmenes en las protoformas más complejas. Wilbik, Kaymak y Dijkman (2017) proponen un método para la generación de resúmenes lingüísticos inspirados en el algoritmo *Apriori.* Estos autores asumen que se dan los posibles cuantificadores y etiquetas lingüísticas para los atributos del conjunto de datos a analizar. Aplican la poda de los resúmenes y para el caso de los resúmenes extendidos, actualizan el filtro.

En Pérez y colaboradores (2018) se presenta una propuesta basada en la generación de reglas de asociación que luego son refinadas, que toma como base el soporte y la confianza de las reglas, y son transformadas posteriormente en resúmenes lingüísticos. Esta propuesta se presenta combinada con otros enfoques que consideran el comportamiento estadístico de los datos.

Por su parte, Smits y colaboradores (2018) combinan diferentes estrategias y aprovechan el conocimiento estadístico asociado a la distribución de los datos. Luego calculan la cardinalidad relativa en el caso de un atributo numérico, y la moda en el caso de valores categóricos para finalmente aplicar un algoritmo basado en los principios del algoritmo Apriori. Se centran en generar un resumen lingüístico para el cual se calcula su cardinalidad y posteriormente identificar el cuantificador que mejor lo describe.

Como principal limitante de los enfoques basados en reglas se señala con frecuencia que se apoyan en la construcción de un grupo de reglas candidatas, que luego son transformadas en resúmenes lingüísticos y generan un número elevado de resúmenes que son eliminados en

una segunda etapa. Este enfoque, generalmente no explotan suficientemente la información asociada a la correlación de los datos en el problema en cuestión.

Enfoque basado en meta-heurísticas

Otro enfoque en la construcción de resúmenes lingüísticos se ha basado en el uso de meta-heurísticas, se destacan autores como George y Srikant (1996), seguidos por J. Kacprzyk y P. Strykowski (1999), Donis-Diaz, Bello y colaboradores (2014). Bajo este enfoque los algoritmos genéticos han sido la técnica más utilizada.

Donis-Díaz y colaboradores (2014) proponen un modelo híbrido de algoritmo genético (AG) con un escalador de colinas para la identificación de resúmenes lingüísticos. Estos autores experimentan además con diferentes meta-heurísticas de inteligencia colectiva como las colonias de hormigas (ACO, siglas en inglés) (Donis-Díaz, *et al.*, 2015). Como debilidad de este enfoque se le señala que no aprovecha la información asociada a las correlaciones entre los datos. Además, en ambos trabajos los autores establecen un algoritmo de generación de resúmenes con un alto costo computacional que se aprecia a partir de la cantidad de evaluaciones de resúmenes lingüísticos que debe realizar este algoritmo.

Enfoque basado en agrupamientos y descubrimiento de *outliers*

En el enfoque basado en el agrupamiento de los datos se destacan trabajos como los de Wilbik & Dijkman (2016), en los que se construyen matrices de partición y emplean los algoritmos Single Linkage (SL) y Non-Euclidean Relational Fuzzy CMeans (NERFCM). Otros trabajos se basan en la construcción de un árbol a partir de considerar todas las combinaciones de filtros y sumarizadores y luego podan las ramas que representa una combinación con bajo soporte en la base de datos (Dijkman & Wilbik, 2017). La naturaleza de estos trabajos, implica un alto costo computacional durante el proceso de búsqueda.

Los métodos basados en este enfoque generalmente construyen los *clusters* y luego por cada grupo generan un resumen lingüístico que explica el comportamiento del mismo. Entre las principales limitantes de este enfoque se identifica que generalmente dependen de parámetros que acotan la cantidad de grupos a construir. En otros casos las estrategias basadas en distancias tienden a formar hiper-esferas y no logran identificar grupos que describan otras topologías y centran el análisis en valores numéricos que limitan el trabajo con datos heterogéneos.

Duraj, Szczepaniak, & Chomatek (2020) y Duraj, Szczepaniak, & Ochelska-Mierzejewska (2016) son ejemplos de generación de resúmenes a partir de datos anómalos (*outliers),* en estos casos, para el proceso de construcción de los resúmenes lingüísticos. Duraj y colaboradores hacen alusión al método basado en construcción exhaustiva de todas las posibles combinaciones de cuantificadores y sumarizadores predefinidos; por lo que estas propuestas no usan la heurística u otros métodos que simplifiquen la complejidad de búsqueda en la construcción de los resúmenes.

Otros autores han publicado diferentes trabajos asociados al descubrimiento de *outliers* que combinan diferentes técnicas, entre las que destacan los resúmenes lingüísticos (Pérez, Piñero, Vacacela, *et al.,* 2020; Aguilar, *et al.,* 2016; Castro, *et al.,* 2016).

La figura 5 muestra un resumen de las principales tendencias en los métodos de generación y su evolución en el tiempo.

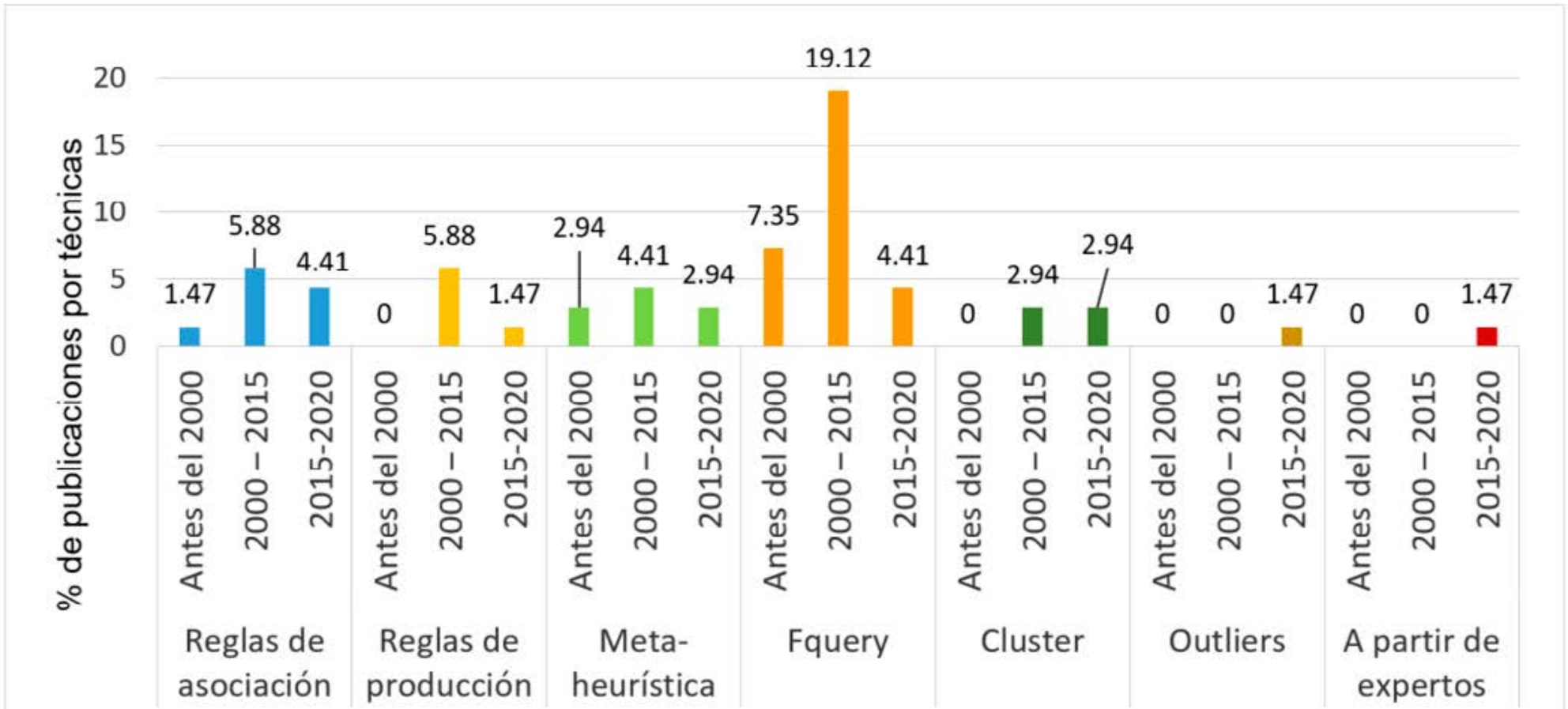

Figura 5. Tendencia de los métodos o técnicas para generar resúmenes lingüísticos de datos.

Como resumen de este análisis se identifica que antes del 2015 los métodos más utilizados son los basados en consultas difusas, una de las ventajas de este enfoque es las facilidades que aporta el uso de lenguajes de consultas. Mientras que en el período 2015-2020 hay mayor diversidad de métodos sin diferencias significativas en la frecuencia de uso de estos. Este comportamiento muestra la inclusión de nuevo métodos basados en otras teorías y que aportan resultados prometedores con potencialidades para el uso de los resúmenes en otras problemáticas asociadas a problemas de predicción entre otros.

Principales técnicas y métodos de validación empleados en las investigaciones

En la bibliografía estudiada también se analizaron cuáles fueron las principales estrategias de validación empleados en las investigaciones sobre sumarización lingüística de datos (Figura 6).

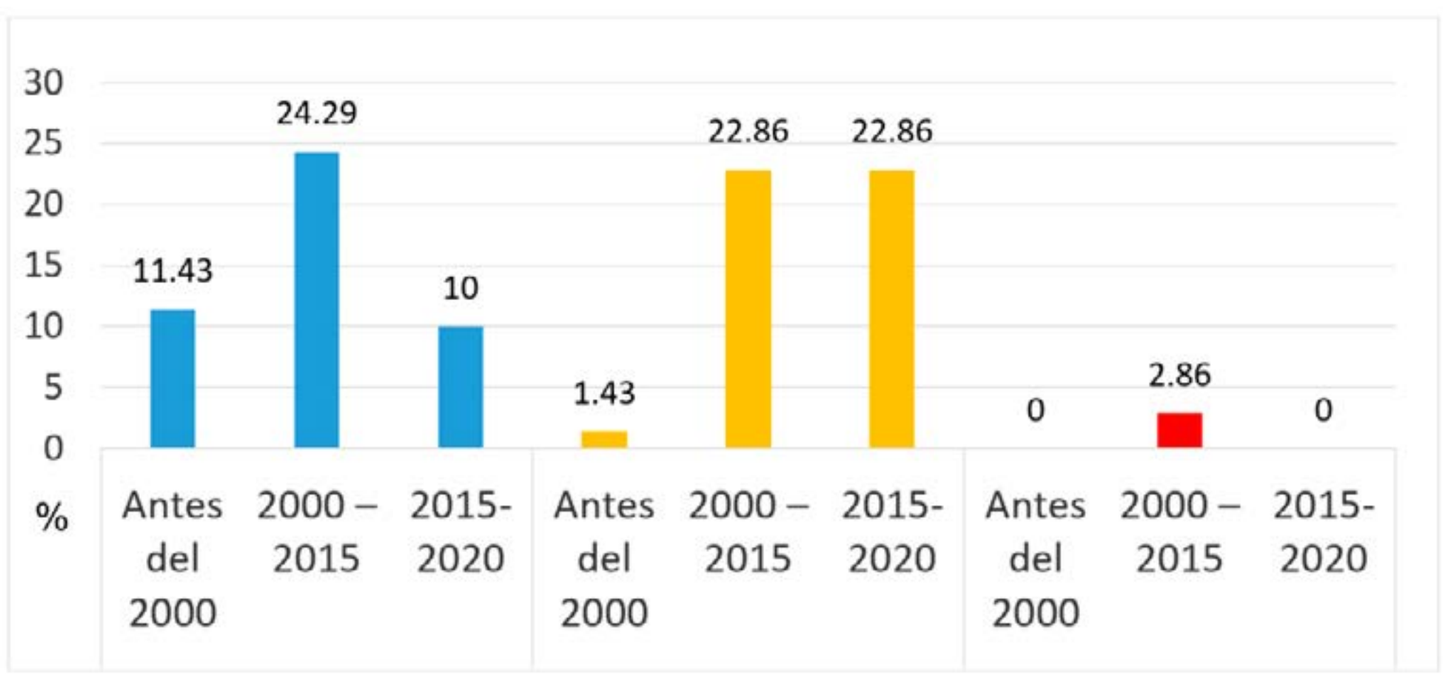

Figura 6. Estrategias de validación empleadas en la bibliografía estudiada.

A partir de este análisis se identifica que antes del 2015, la mayoría de las investigaciones solo se validaban a partir de la observación y el análisis cualitativo simple en casos de estudio. Mientras que, en los últimos cinco años, existe un uso amplio de técnicas básicas de estadística descriptiva aplicada sobre los indicadores de calidad de los resúmenes (Kaczmarek-Majer, *et al.,* 2019; Peláez-Aguilera, *et al.,* 2019; Amghar & Chikh, 2018). En esta temática muy pocos estudios aplican técnicas estadísticas de rigor basadas en test paramétricos y no paramétricos (Donis-Díaz, *et al.,* 2014, 2015).

Respecto a la validez externa se realiza un análisis en el cual se observa como cada una de las investigaciones consultadas hace uso de alguna de las siguientes estrategias de validación: triangulación de datos, triangulación de métodos, triangulación de expertos y estudio de casos (ver figura 7).

Respecto a la triangulación de datos solo en los últimos años se encuentran investigaciones que lo aplican, se muestran a continuación algunas de estas investigaciones:

- En Ramos-Soto & Martin-Rodillab (2019) utilizan datos de 15 estaciones meteorológicas de Galicia, en Amghar & Chikh (2018) se aplica la investigación a la base de datos *Pima Indian Diabetes Dataset* y *Wisconsin Breast Cancer Dataset* (WBCD) ambas del *UCI Repository of Machine Learning* (Dua & Graff, 2017).
- Kacprzyk y Zadrożny (2016a) en el análisis de cotizaciones de fondos de inversión en Bolsas de Valores de Varsovia, Zagreb y Moscú, y a registros de *logs* de servidores web.
- En Rojas Valenzuela (2018) se generan resúmenes a partir del nivel de polución de diferentes distritos de una ciudad.
- En Smits y colaboradores (2018) se aplica a datos sobre vuelos comerciales en Estados Enidos entre1987-2008, y a descripciones de carros de segunda mano en que se considera el número de puertas, precio, kilometraje, año, caballo de fuerza y nuevo precio inicial.
- En Khedidja y colaboradores (2020) se aplica a datos sobre vuelos reales, y datos sobre ciudad inteligente extraído del proyecto *neOCampus* de la universidad Toulouse en Francia.

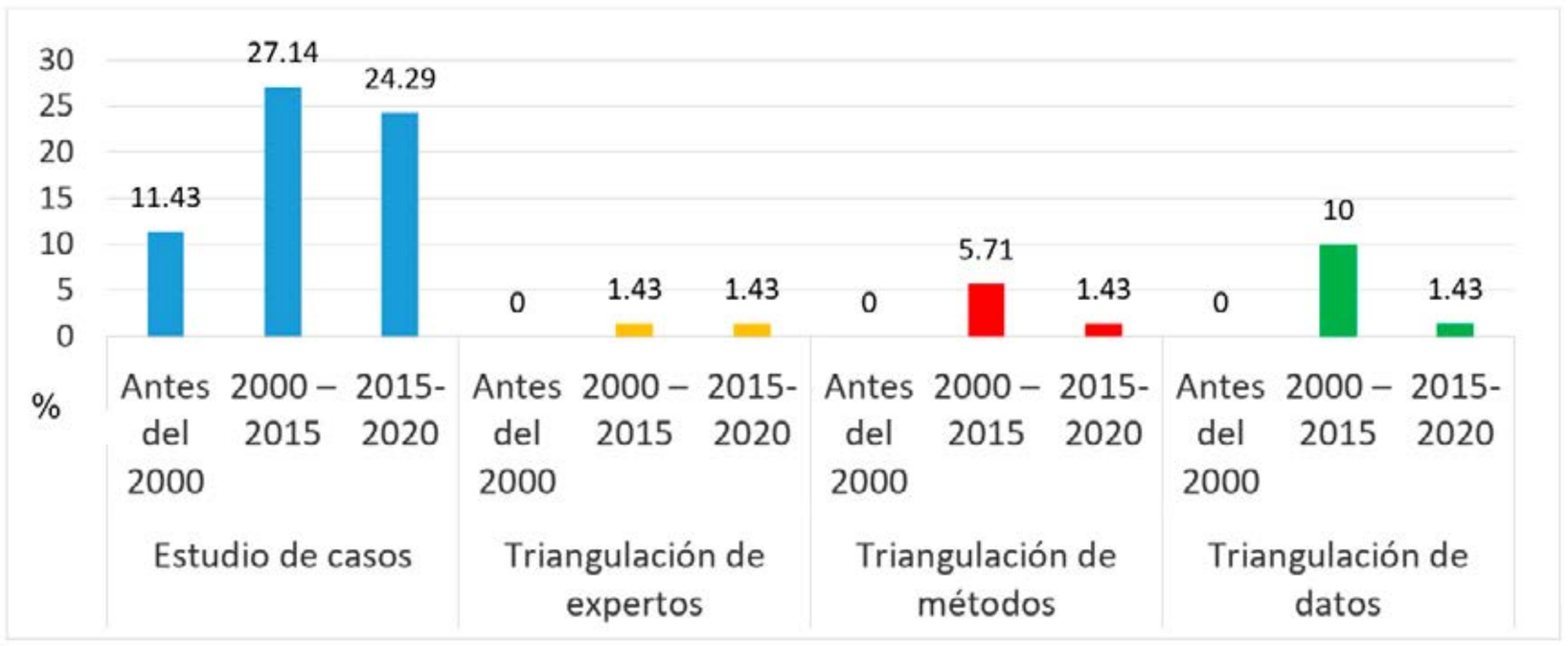

Figura 7. Alcance de la validación empleada en la bibliografía estudiada.

Algunos autores han aplicado triangulación de métodos en sus investigaciones, entre ellos están Donis-Díaz y colaboradores (2014), quienes realizan una comparación de los resultados

utilizando tres variantes de modelos de algoritmos genéticos: el clásico, clásico modificado y el híbrido. Otro ejemplo en el que se realiza una comparación respecto al tiempo de ejecución entre dos algoritmos de sumarización aplicados a dos bases de datos se puede encontrar en Khedidja y colaboradores (2020).

Es muy reducido el número de trabajos publicados que apliquen triangulación de expertos. En Rojas Valenzuela (2018) se aplican encuestas para la validación de los resultados.

En esta área temática las técnicas de validación que han predominado en las investigaciones son las basadas en el estudio de casos. Constituye una oportunidad de mejora el desarrollo de investigaciones que comparen con profundidad los diferentes métodos de generación de resúmenes.

Áreas de aplicación de los resúmenes lingüísticos

El resumen lingüístico se ha aplicado en numerosos dominios y en diferentes tipos de aplicación, ya sea aplicación descriptiva, como para la predicción o clasificación. La figura 8 resume el análisis realizado en este sentido.

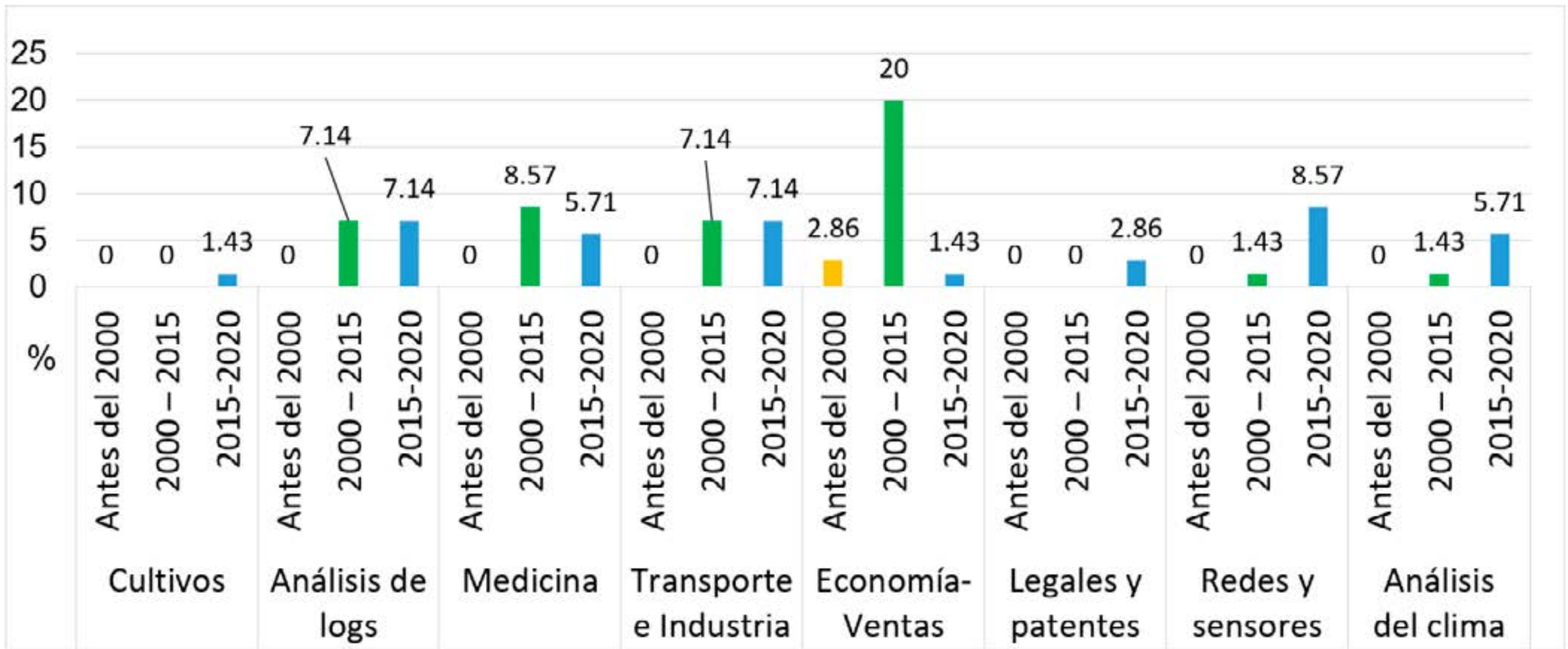

Figura 8. Análisis de publicaciones respecto a los entornos de aplicación.

A partir del análisis realizado se identifica que el área de mayor cantidad de aplicaciones en el periodo 2000 al 2015 fue la asociada a los sectores financiero y económico. Sin embargo, la tendencia en los últimos cinco años se concentra en la medicina, dispositivos inteligentes y redes de sensores.

En la bibliografía consultada se identifican diferentes aplicaciones de los resúmenes lingüísticos en problemas de toma de decisiones como se muestra a continuación:

- Economía y ventas (Piñero, Pérez Pupo, García Vacacela, & Toscanini, 2020).
- Registros de *logs* (Janusz Kacprzyk & Zadrożny, 2016b; Janusz Kacprzyk & Zadrożny, 2016a; Wilbik & Dijkman, 2016; Dijkman & Wilbik, 2017; Kaymak & Wilbik, 2017; Wilbik, Gilsing, Turetken, Ozkan, & Grefen, 2020).
- Redes de dispositivos y sensores (Jain, *et al.,* 2019; Jain, Keller, & Bezdek, 2016; Khedidja, *et al.,* 2020; Díaz-Hermida & Vidal, 2018; Peláez-Aguilera, *et al.,* 2019; Kaczmarek-Majer *et al.,* 2019; Genç, Akay, Boran, & Yager, 2020).

- Sector de la salud: en gestión de glucosa (Wilbik, *et al.,* 2018a), enfermedades del corazón (Peláez-Aguilera, *et al.,* 2019), actividad física (Sanchez-Valdes, Alvarez-Alvarez, & Trivino, 2016), diabetes y cáncer (Amghar & Chikh, 2018), entre otros.
- En la meteorología (Khedidja, *et al.,* 2020; Ramos-Soto & Martin-Rodillab, 2019; Heble-Lahera, *et al.,* 2020).
- En la agricultura (Hudec, *et al.,* 2018; Rojas Valenzuela, 2018) y la industria (Donis-Díaz, *et al.,* 2014).
- En gestión de tráfico y movilidad urbana (Degtiarev & Remnev, 2016) (Gilsing, Wilbik, Grefen, Turetken, & Ozkan, 2020).
- En gestión de proyectos (Pupo, *et al.,* 2020), (Pérez Pupo, *et al.,* (Pérez, Piñero, Vacacela, *et al.,* 2020), (Pérez, Piñero, Bello, *et al.,* 2020)
- Temas legales (Wilbik, *et al.,* 2017[a]; Dijkman & Wilbik, 2017), gestión de patentes (Igde, Aydoğan, Boran, & Akay, 2017) y gestión del consenso (Janusz Kacprzyk & Zadrożny, 2018; Janusz Kacprzyk & Zadrożny, 2016c).

En menor medida se han empleado resúmenes en problemas de predicción o clasificación como se muestra en (Chiang, Chow, & Wang, 2000; Amghar & Chikh, 2018; Chiang, *et al.,* 2000; J. Kacprzyk, Yager, & Merigo, 2019).

A partir de la metodología aplicada como parte del diseño de experimentos y el análisis detallado de las diferentes fuentes bibliográficas, se establece en esta investigación que la sumarización lingüística de datos es un "área de conocimiento interdisciplinar, que forma parte del *soft computing,* e involucra un conjunto de áreas de conocimiento entre las que se destacan la lógica borrosa, el aprendizaje automático, la estadística, las bases de datos y otras. Tiene como objetivo la ayuda a la toma de decisiones a partir de la construcción automática de resúmenes en lenguaje natural que reflejen el conocimiento implícito y las relaciones intrínsecas que existen entre las variables en bases de datos heterogéneas".

CONCLUSIONES

En este trabajo se ha realizado una revisión del estado del arte de la sumarización lingüística de datos desde las siguientes perspectivas: la notación y la estructura de los resúmenes, las técnicas de generación, las estrategias de validación de las investigaciones y las áreas de aplicación. Se identifica que existe consenso en los elementos que componen los resúmenes y la notación empleada en las investigaciones.

Respecto a la estructura de los resúmenes se han definido diferentes protoformas; las más generalizadas son las propuestas por Kacprzyk y Zadrożny. No obstante, cada escenario en particular tiene sus características y esta situación ha motivado a los investigadores a desarrollar diferentes extensiones. Resulta particularmente interesante el desarrollo de extensiones que permiten la integración de múltiples resúmenes en un único análisis, lo cual facilita la toma de decisiones. Una línea abierta a la investigación en este campo se ubica en la posibilidad de representación de un mismo resumen lingüístico que emplea

diferentes lenguajes naturales controlados, elemento que facilitaría la internacionalización de los resultados.

Las técnicas de generación de resúmenes son diversas. Se destacan antes del 2015 el uso de métodos basados en consultas que explotan las facilidades de los lenguajes de consultas de bases de datos y combinan estos con otras técnicas. En últimos años existe mayor variedad en los métodos de generación de resúmenes lingüísticos, se ha incrementado el uso de técnicas asociadas al aprendizaje automático y la minería de datos. Estas nuevas tendencias incorporan a los algoritmos mayor robustez, pero se identifica que quedan líneas abiertas a la investigación a partir de la posibilidad de hibridación de diferentes técnicas y una mejor explotación de información asociada a los datos, tales como la correlación entre las variables o aspectos específicos de los dominios de aplicación.

Respecto a la validación de las investigaciones publicadas se identifica que hay pocos trabajos que aplican las técnicas de triangulación de datos, triangulación de métodos u otras que ayuden a verificar la capacidad de generalización de las investigaciones. Constituye una oportunidad de mejora el desarrollo de investigaciones que comparen con profundidad los diferentes métodos de generación de resúmenes y permitan identificar las potencialidades de estos para diferentes escenarios.

Los resúmenes lingüísticos de datos han sido empleados generalmente de forma descriptiva. Las áreas de aplicación más trabajadas son la toma de decisiones en la gestión económica y las ventas desde etapas tempranas del desarrollo de la teoría. No obstante, a partir del 2016 hay un incremento significativo de aplicaciones en la medicina, del análisis de datos de dispositivos inteligentes y en el análisis de redes de sensores como parte del internet de las cosas. Se identifica como una línea abierta a la investigación el uso, en mayor medida, de las técnicas de sumarización lingüística de datos en la solución de problemas de predicción y en el desarrollo de sistemas conversacionales, como parte de procesos de ayuda a la toma de decisiones.

Los resultados obtenidos pueden ser empleados en procesos cuyas dinámicas que definen su comportamiento no están bien determinadas existiendo relaciones difusas entre ellas. La robustez y capacidad de generalización del algoritmo propuesto, permite su utilización en situaciones donde existe poco conocimiento de las relaciones causales entre variables, o incluso a donde no exista conocimiento experto alguno para definir un modelo inicial a partir del cual parta el modelado.

El uso del algoritmo de aprendizaje propuesto, posibilitaría resultados superiores en investigaciones previas. En la predicción de parámetros de crecimiento y desarrollo de hortalizas (Madruga, *et al.,* 2019) hubo una afectación en el modelado por la limitación en la cantidad de datos adquiridos, así como incertidumbre en algunas relaciones entre variables y valores de los pesos iniciales. Estos aspectos pueden ser resueltos a partir de la utilización del algoritmo de aprendizaje utilizando AR.

REFERENCIAS

Aguilar, C., Fernando, G., Pérez Pupo, I., Pérez, P., Martínez, N., y Crúz Castillo, Y. (2016). Aplicación de la minería de datos anómalos en organizaciones orientadas a proyectos. *Revista Cubana de Ciencias Informáticas, 10,* 195-209.

Amghar, D., & Chikh, A. M. (2018). Extracting a Linguistic Summary from a Medical Database. *International Journal of Intelligent Systems and Applications, 10*(12), 16-26. https://doi.org/10.5815/ijisa.2018.12.02

Boran, F. E., Akay, D., & Yager, R. R. (2016). An overview of methods for linguistic summarization with fuzzy sets. *Expert Systems with Applications, 61,* 356-377. https://doi.org/10.1016/j.eswa.2016.05.044

Castro, G. F., Pérez, I., Piñero, P., Torres, S., Vásquez, M., Hidalgo, J., & Vera-Lucio, N. (2016). *Platform for Project Evaluation Based on Soft-Computing Techniques* (pp. 226-240). Springer, Cham. Retrieved from https://link.springer.com/chapter/10.1007/978-3-319-48024-4_18

Chiang, D.-A., Chow, L. R., & Wang, Y.-F. (2000). Mining time series data by a fuzzy linguistic summary system. *Fuzzy Sets and Systems, 112*(3), 419-432. https://doi.org/10.1016/S0165-0114(98)00003-7

Degtiarev, K. Y., & Remnev, N. V. (2016). Linguistic resumes in software engineering: the case of trend summarization in mobile crash reporting systems. *Procedia Computer Science, 102,* 121-128. https://doi.org/210.1016/j.procs.2016.09.378

Díaz-Hermida, F., & Vidal, J. C. (2018). *Fuzzy quantification for linguistic data analysis and data mining.* Retrieved from https://arxiv.org/abs/1807.07389v1

Dijkman, R., & Wilbik, A. (2017). Linguistic summarization of event logs – A practical approach. *Information Systems, 67,* 114-125. https://doi.org/10.1016/j.is.2017.03.009

Donis-Díaz, C. A., Muro, A. G., Bello-Pérez, R., & Morales, E. V. (2014). A hybrid model of genetic algorithm with local search to discover linguistic data summaries from creep data. *Expert Systems with Applications, 41*(4, Part 2), 2035-2042. https://doi.org/10.1016/j.eswa.2013.09.002

Donis-Díaz, Carlos A., Bello, R., & Kacprzyk, J. (2015). Using Ant Colony Optimization and Genetic Algorithms for the Linguistic Summarization of Creep Data. In P. Angelov, K. T. Atanassov, L. Doukovska, M. Hadjiski, V. Jotsov, J. Kacprzyk, … S. Zadrożny (Eds.), *Intelligent Systems'2014* (pp. 81-92). Cham: Springer International Publishing. https://doi.org/10.1007/978-3-319-11313-5_8

Dua, D., & Graff, C. (2017). *UCI Machine Learning Repository.* University of California, Irvine, School of Information and Computer Sciences. Retrieved from http://archive.ics.uci.edu/ml

Dubois, D., & Prade, H. (1992). Gradual inference rules in approximate reasoning. *Information Sciences, 61*(1), 103-122. https://doi.org/10.1016/0020-0255(92)90035-7

Duraj, A., Szczepaniak, P. S., & Chomatek, L. (2020). Intelligent Detection of Information Outliers Using Linguistic Summaries with Non-monotonic Quantifiers. In *International Conference on Information Processing and Management of Uncertainty in Knowledge-Based Systems* (pp. 787-799). Springer. https://doi.org/10.1007/978-3-030-50153-2_58

Duraj, A., Szczepaniak, P. S., & Ochelska-Mierzejewska, J. (2016). Detection of Outlier Information Using Linguistic Summarization. In T. Andreasen, H. Christiansen, J. Kacprzyk, H. Larsen, G. Pasi, O. Pivert, … S. Zadrożny (Eds.), *Flexible Query Answering Systems*

2015 (pp. 101-113). Cham: Springer International Publishing. https://doi.org/10.1007/978-3-319-26154-6_8

Eciolaza, L., Pereira-FariñA, M., & Trivino, G. (2013). Automatic linguistic reporting in driving simulation environments. *Applied Soft Computing, 13*(9), 3956-3967. https://doi.org/10.1016/j.asoc.2012.09.007

Genç, S., Akay, D., Boran, F. E., & Yager, R. R. (2020). Linguistic summarization of fuzzy social and economic networks: an application on the international trade network. *Soft Computing, 24*(2), 1511-1527. https://doi.org/10.1007/s00500-019-03982-9

George, R., & Srikant, R. (1996). Data summarization using genetic algorithms and fuzzy logic. *Genetic Algorithms and Soft Computing,* 599-611.

Gilsing, R., Wilbik, A., Grefen, P., Turetken, O., & Ozkan, B. (2020). A Formal Basis for Business Model Evaluation with Linguistic Summaries. In *Enterprise, Business-Process and Information Systems Modeling* (pp. 428-442). Springer. https://doi.org/10.1007/978-3-030-49418-6_29

Heble-Lahera, C., Cascallar-Fuentes, A., Ramos-Soto, A., & Diz, A. B. (2020). Empirical study of fuzzy quantification models for linguistic descriptions of meteorological data. In *2020 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE)* (pp. 1-7). IEEE. https://doi.org/10.1109/FUZZ48607.2020.9177716

Hernández-Sampieri, R., & Torres, C. P. M. (2018). *Metodología de la investigación* (Vol. 4). McGraw-Hill Interamericana México^ eD. F DF.

Hudec, M., Bednárová, E., & Holzinger, A. (2018). Augmenting Statistical Data Dissemination by Short Quantified Sentences of Natural Language. *Journal of Official Statistics, 34*(4), 981-1010. https://doi.org/10.2478/jos-2018-0048

Igde, E. Y., Aydoğan, S., Boran, F. E., & Akay, D. (2017). Linguistic Summarization of Structured Patent Data.

Jain, A., Keller, J. M., & Bezdek, J. C. (2016). Quantitative and qualitative comparison of periodic sensor data. In *2016 IEEE-embs international conference on biomedical and health informatics (bhi)* (pp. 37-40). IEEE. https://doi.org/10.1109/BHI.2016.7455829

Jain, A., Popescu, M., Keller, J., Rantz, M., & Markway, B. (2019). Linguistic summarization of in-home sensor data. *Journal of Biomedical Informatics, 96,* 103240. https://doi.org/10.1016/j.jbi.2019.103240

Kacprzyk, J. (1999). Fuzzy logic for linguistic summarization of databases. In *FUZZ-IEEE'99. 1999 IEEE International Fuzzy Systems. Conference Proceedings (Cat. No.99CH36315)* (Vol. 2, pp. 813-818 vol.2). https://doi.org/10.1109/FUZZY.1999.793053

Kacprzyk, J, & Strykowski, P. (1999). 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), 2,* 937-943. https://doi.org/10.1109/CEC.1999.782523

Kacprzyk, J., Yager, R. R., & Merigo, J. M. (2019). Towards Human-Centric Aggregation via Ordered Weighted Aggregation Operators and Linguistic Data Summaries: A New Pers-

pective on Zadeh's Inspirations. *IEEE Computational Intelligence Magazine, 14*(1), 16-30. https://doi.org/10.1109/MCI.2018.2881641

Kacprzyk, J., & Zadrożny, S. (2005). Fuzzy linguistic data summaries as a human consistent, user adaptable solution to data mining. In B. Gabrys, K. Leiviskä, & J. Strackeljan (Eds.), *Do Smart Adaptive Systems Exist? Best Practice for Selection and Combination of Intelligent Methods* (pp. 321-340). Berlin, Heidelberg: Springer. https://doi.org/10.1007/3-540-32374-0_16

Kacprzyk, Janusz. (1999). An interactive fuzzy logic approach to linguistic data summaries. In *18th International Conference of the North American Fuzzy Information Processing Society-NAFIPS (Cat. No. 99TH8397)* (pp. 595-599). IEEE. https://doi.org/10.1109/NAFIPS.1999.781763

Kacprzyk, Janusz, & Zadrożny, S. (1995). Fquery for Access: Fuzzy Querying for a Windows-Based DBMS. In *Fuzziness in database management systems* (Vol. 5, pp. 415-433). Springer. Retrieved from https://link.springer.com/chapter/10.1007/978-3-7908-1897-0_18

Kacprzyk, Janusz, & Zadrożny, S. (2000). On a fuzzy querying and data mining interface. *Kybernetika, 36*(6), 657-670.

Kacprzyk, Janusz, & Zadrożny, S. (2003). Linguistic summarization of data sets using association rules. In *Fuzzy Systems, 2003. FUZZ'03. The 12th IEEE International Conference on* (Vol. 1, pp. 702-707). IEEE.

Kacprzyk, Janusz, & Zadrożny, S. (2005). Linguistic database summaries and their protoforms: towards natural language based knowledge discovery tools. *Information Sciences, 173*(4), 281-304. https://doi.org/10.1016/j.ins.2005.03.002

Kacprzyk, Janusz, & Zadrożny, S. (2009). Linguistic database summaries using fuzzy logic, towards a human-consistent data mining tool, (20), 10.

Kacprzyk, Janusz, & Zadrożny, S. (2016a). 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, 6*(1), 37-46. https://doi.org/10.1002/widm.1175

Kacprzyk, Janusz, & Zadrożny, S. (2016b). Linguistic summarization of the contents of Web server logs via the Ordered Weighted Averaging (OWA) operators. *Fuzzy Sets and Systems, 285,* 182-198. https://doi.org/10.1016/j.fss.2015.07.020

Kacprzyk, Janusz, & Zadrożny, S. (2016c). On a fairness type approach to consensus reaching support under fuzziness via linguistic summaries. In *2016 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE)* (pp. 1999-2006). https://doi.org/10.1109/FUZZ-IEEE.2016.7737937

Kacprzyk, Janusz, & Zadrożny, S. (2018). Reaching Consensus in a Group of Agents: Supporting a Moderator Run Process via Linguistic Summaries. In *Soft Computing Applications for Group Decision-making and Consensus Modeling* (pp. 465-485). Springer.

Kacprzyk, Janusz, Zadrożny, S., & Dziedzic, M. (2014). A Novel View of Bipolarity in Linguistic Data Summaries. In L. T. Kóczy, C. R. Pozna, & J. Kacprzyk (Eds.), *Issues and Challen-*

ges of Intelligent Systems and Computational Intelligence (pp. 215-229). Cham: Springer International Publishing. https://doi.org/10.1007/978-3-319-03206-1_16

Kaczmarek-Majer, K., Hryniewicz, O., Dominiak, M., & Święcicki, Ł. (2019). *Personalized linguistic summaries in smartphone-based monitoring of bipolar disorder patients.* Atlantis Press. https://doi.org/10.2991/eusflat-19.2019.56

Khedidja, B., Allel, H., & Mohand, L. (2020). Data Summarization for Sensor Data Management: Towards Computational-Intelligence-Based Approaches. *International Journal of Computing and Digital Systems, 9*(5), 825-833. https://doi.org/10.12785/ijcds/090505

Kuhn, T. (2014). A Survey and Classification of Controlled Natural Languajes. *Computational Lingusitic, 40*(1), 121-170. http://dx.doi.org/10.1162/COLI_a_00168

Littell, J. H., Corcoran, J., & Pillai, V. (2008). *Systematic reviews and meta-analysis.* Oxford University Press.

Marín, N., & Sánchez, D. (2016). On generating linguistic descriptions of time series. *Fuzzy Sets and Systems, 285,* 6-30. https://doi.org/10.1016/j.fss.2015.04.014

Peláez-Aguilera, M. D., Espinilla, M., Fernández, M. R., & Medina, J. (2019). Fuzzy Linguistic Protoforms to Summarize Heart Rate Streams of Patients with Ischemic Heart Disease. *Hindawi, 2019,* 11. https://doi.org/0.1155/2019/2694126

Pérez, I., Piñero, P. Y., Bello, R., Acuña, L. A., & Vacacela, R. G. (2020). Linguistic Summaries Generation with Hybridization Method Based on Rough and Fuzzy Sets. In *International Joint Conference on Rough Sets* (pp. 385-397). Springer. https://doi.org/10.1007/978-3-030-52705-1_29

Pérez, I., Piñero, P. Y., Vacacela, R. G., Bello, R., & Acuña, L. A. (2020). Discovering Fails in Software Projects Planning Based on Linguistic Summaries. In *International Joint Conference on Rough Sets* (pp. 365-375). Springer. https://doi.org/10.1007/978-3-030-52705-1_27

Pérez, I., Santos, O., García, R., Piñero, P., & Ramírez, E. C. (2018). Descubrimiento de resúmenes lingüísticos para ayuda a la toma decisiones en gestión de proyecto. *Revista Cubana de Ciencias Informáticas, 12,* 163-175.

Pérez Pupo, I., Villavicencio, N., Piñero, P., García Vacacela, R., & García Sánchez, R. (2020). PROERP Ecosistema de software para la toma de decisiones en gestión de proyectos. In *Experiencias Iberoamericanas de Ingeniería de Proyectos* (p. 899). Guayaquil, Ecuador: Universidad Católica de Santiago de Guayaquil.

Piñero, P., Pérez Pupo, I., García Vacacela, R., & Toscanini, P. (2020). *Caracterización de los estándares de gestión de proyectos y su impacto en la gestión económico financiera de las organizaciones orientadas a proyectos.* Guayaquil, Ecuador: Universidad Católica de Santiago de Guayaquil.

Pupo, I. P., Santos Acosta, O., Piñero, P., García Vacacela, R., & Alvarado, L. (2020). Descubrimiento de errores en la planificación de proyectos basado en resúmenes linguísticos. In *Experiencias Iberoamericanas de Ingeniería de Proyectos* (pp. 867-876). Guayaquil, Ecuador: Universidad Católica de Santiago de Guayaquil.

Ramos-Soto, A., & Martin-Rodillab, P. (2019). Enriching linguistic descriptions of data: A framework for composite protoforms. *Fuzzy Sets and Systems, 26.* https://doi.org/10.1016/j.fss.2019.11.013

Rasmussen, D., & Yager, R. R. (1999). Finding fuzzy and gradual functional dependencies with SummarySQL. *Fuzzy Sets and Systems, 106*(2), 131-142. https://doi.org/10.1016/S0165-0114(97)00268-6

Rojas Valenzuela, Á. R. (2018). *Resúmenes lingüísticos para riego de cultivos* (Tesis). Universidad Técnica Federico Santa María, Departamento de Informática, Santiago, Chile. Retrieved from https://repositorio.usm.cl

Sanchez-Valdes, D., Alvarez-Alvarez, A., & Trivino, G. (2016). Dynamic linguistic descriptions of time series applied to self-track the physical activity. *Fuzzy Sets and Systems, 285,* 162-181. https://doi.org/10.1016/j.fss.2015.06.018

Smits, G., Nerzic, P., Pivert, O., & Lesot, M.-J. (2018). Efficient Generation of Reliable Estimated Linguistic Summaries. In *2018 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE)* (pp. 1-8). https://doi.org/10.1109/FUZZ-IEEE.2018.8491604

Wilbik, A., & Dijkman, R. M. (2016). On the generation of useful linguistic summaries of sequences. In *2016 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE)* (pp. 555-562). https://doi.org/10.1109/FUZZ-IEEE.2016.7737736

Wilbik, A., Gilsing, R., Turetken, O., Ozkan, B., & Grefen, P. (2020). Intentional linguistic summaries for collaborative business model radars. In *2020 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE)* (pp. 1-7). IEEE. https://doi.org/10.1109/FUZZ48607.2020.9177587

Wilbik, A., Kaymak, U., & Dijkman, R. M. (2017). A method for improving the generation of linguistic summaries. In *2017 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE)* (pp. 1-6). https://doi.org/10.1109/FUZZ-IEEE.2017.8015752

Wilbik, A., Vanderfeesten, I., Bergmans, D., Heines, S., & Mook, W. van. (2018). Linguistic Summaries for Compliance Analysis of a Glucose Management Clinical Protocol. In *2018 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE)* (pp. 1-7). https://doi.org/10.1109/FUZZ-IEEE.2018.8491449

Wu, D., & Mendel, J. M. (2010). Linguistic summarization using IF–THEN rules and interval type-2 fuzzy sets. *IEEE Transactions on Fuzzy Systems, 19*(1), 136-151. https://doi.org/10.1109/TFUZZ.2010.2088128

Wu, D., & Mendel, J. M. (2011). Linguistic summarization using IF–THEN rules and interval type-2 fuzzy sets. *IEEE Transactions on Fuzzy Systems, 19*(1), 136-151.

Wu, D., Mendel, J. M., & Joo, J. (2010). Linguistic summarization using if-then rules. In *Fuzzy Systems (FUZZ), 2010 IEEE International Conference on* (pp. 1-8). IEEE.

Yager, R. R. (1991). On Linguistic Summaries of Data. *Knowledge Discovery in Databases,* 378-389.

Yager, Ronald R. (1982). A new approach to the summarization of data. *Information Sciences, 28*(1), 69-86. https://doi.org/10.1016/0020-0255(82)90033-0

Zadeh, L. A. (1983). A computational approach to fuzzy quantifiers in natural languages. *Computers & Mathematics with Applications, 9*(1), 149-184.

Zadeh, L. A. (2002). A prototype-centered approach to adding deduction capability to search engines–the concept of protoform. In *Intelligent Systems, 2002. Proceedings. 2002 First International IEEE Symposium* (Vol. 1, pp. 2-3). IEEE. https://doi.org/10.1109/IS.2002.1044219