Centro de Investigación en Métodos de Producción de Software

Informe Técnico / Technical Report

Hacia la Caracterización de las Aplicaciones Big Data

Ana C. Marcén, Carlos Cetina y Óscar Pastor

Ref. #:	**PROS-TR-2019-I**
Title:	**Hacia la Caracterización de las Aplicaciones Big Data**
Author (s):	**Ana C. Marcén, Carlos Cetina, and Óscar Pastor**
Corresponding autor (s):	**acmarcen@pros.usj.es** **ccetina@usj.es** **opastor@pros.usj.es**
Document versión number:	**1**
Final version:	**-**
Release date:	**-**
Key words:	**Aplicaciones Big Data, Modelado Conceptual, Ontología**

Hacia la Caracterización de las Aplicaciones Big Data

Ana C. Marcén[1,2], Carlos Cetina[2], and Óscar Pastor[1]

[1] Centro de Investigación en Métodos de Producción de Software, Universitat Politècnica de València, acmarcen, opastor@pros.upv.es
[2] SVIT Research Group, Universidad San Jorge, acmarcen, ccetina@usj.es

Abstract. Mediante este artículo proponemos el uso de una ontología como base esencial para construir aplicaciones Big Data determinando qué se entiende por aplicación Big Data, cuáles son sus características particulares o de qué depende que una aplicación sea o no sea Big Data. Como primer paso hacia la caracterización de las aplicaciones Big Data, este artículo presenta diversas técnicas actuales Big Data de manera homogeneizada. Y posteriormente, se propone una ontología inicial basada tanto en los antecedentes del término Big Data, como en la homogeneización de las técnicas realizada.

Keywords: Aplicaciones Big Data, Modelado Conceptual, Ontología

1 Introducción

En la actualidad, existen millones de aplicaciones de todo tipo en el mercado: aplicaciones móviles, aplicaciones web, aplicaciones de entretenimiento, aplicaciones de negocios, y en los últimos años proliferan cada vez con mayor frecuencia las aplicaciones Big Data [29].

Sin embargo, la carencia de un consenso para definir el término Big Data da lugar a ambigüedad e incluso contrariedades que afectan a la comprensión y el desarrollo de nuevas aplicaciones Big Data [13]. La finalidad de este artículo es dar solución al desafío de carácter conceptual que surge a la hora de desarrollar aplicaciones y servicios basados en Big Data en entornos industriales.

Para alcanzar el consenso ontológico necesario que determine con precisión el soporte conceptual de las nociones utilizadas en las aplicaciones Big Data, este articulo presenta de manera homogeneizada las técnicas Big Data usadas más frecuentemente en aplicaciones para el tratamiento de textos. Posteriormente, dichas técnicas se han empleado como base para proponer una ontología que sirva para caracterizar las aplicaciones Big Data.

Las secciones de este artículo se estructuran de la siguiente manera: La Sección 2 presenta una visión general del concepto Big Data. Posteriormente, la Sección 3 caracteriza las principales técnicas Big Data para el tratamiento de texto. En la sección 4, se presenta la ontología para la caracterización de las aplicaciones Big Data. Finalmente, en las Secciones 5 y 6, se discute y concluye el trabajo.

2 Background

En esta primera sección, se presenta una visión en conjunto de las aplicaciones Big Data, qué son y qué proceso siguen para alcanzar su objetivo. En el primer apartado, se recopilan varias definiciones de Big Data pertenecientes al sector industrial. Mediante dichas definiciones, se obtiene la visión en conjunto de qué se entiende por Big Data, al mismo tiempo que se plantean discusiones respecto a dicho término. Y el segundo apartado se centra en los pasos que lleva a cabo una aplicación Big Data para lograr su cometido, es decir, en el proceso que siguen las aplicaciones Big Data.

2.1 Definición de Big Data

Mediante las siguientes definiciones se pretende identificar qué características capacitan a una aplicación como Big Data y la diferencia del resto de aplicaciones no Big Data. Todas estas definiciones provienen de empresas técnologicas reconocidas.

Big data is being generated by everything around us at all times. Every digital process and social media exchange produces it. Systems, sensors and mobile devices transmit it. Big data is arriving from multiple sources at an alarming velocity, volume and variety. To extract meaningful value from big data, you need optimal processing power, analytics capabilities and skills.

IBM [24]

Big data describes a holistic information management strategy that includes and integrates many new types of data and data management alongside traditional data. Big data has also been defined by the four Vs: Volume (The amount of data), Velocity(The fast rate at which data is received and perhaps acted upon), Variety (New unstructured data types), and Value((Data has intrinsic value—but it must be discovered).

Oracle [33]

Big data is a term that describes the large volume of data – both structured and unstructured – that inundates a business on a day-to-day basis. But it's not the amount of data that's important. It's what organizations do with the data that matters. Big data can be analyzed for insights that lead to better decisions and strategic business moves.

SAS [38]

Por lo tanto, algunas definiciones se basan en el principio de la tres V's: volumen, velocidad y variedad, modificándolo incluyendo otras características cómo veracidad o valor. Este principio fue enunciado por primera vez por Doug Laney [12] cómo forma de explotar los retos en el manejo de datos teniendo en cuenta esas tres dimensiones volumen, velocidad y variedad. El volumen hace referencia

a la gran cantidad de los datos, la velocidad hace referencia a la rapidez con que son creados, almacenados y procesados, y la variedad hace referencia a los distintos tipos de datos: estructurados, semi-estructurados o no-estructurados. Partiendo de estas tres características, algunas definiciones incluyen otras nuevas como valor que hace referencia a la importancia que tienen ciertos datos sobre otros dentro del gran volumen de datos generados.

También, existen definiciones que reconocen el volumen de datos, la velocidad o su variedad como características, pero Big Data es definida en función del impacto del conocimiento generado en la sociedad [30]. Y otras definiciones ponen en duda que el volumen de los datos sea una característica relevante para definir Big Data aunque actualmente los datos disponibles sean muchos [7].

Partiendo que Big Data es un término que surgió en el campo empresarial, e incluso después de varios años sigue evolucionando a medida que las soluciones Big Data proliferan, algunos investigadores han intentado asentar y contextualizar el término en base a las diversas definiciones del mismo [42][10][16]. Pero, al igual que en el entorno empresarial la importancia del volumen de los datos, la velocidad, la variedad, el valor, la veracidad u otras características para definir y caracterizar el término Big Data está en discusión.

2.2 Proceso Big Data

Teniendo en cuenta las definiciones anteriores, podemos deducir que una aplicación Big Data tendrá como entrada unos datos (de mayor o menor tamaño, estructurados de alguna manera o sin estructurar) y cómo salida se generará cierto conocimiento. Pero hasta el momento desconocemos el proceso para lograr alcanzar ese conocimiento. Por esta razón, a partir de diferentes artículos académicos dentro del campo [3][9][19][25][35], se ha generalizado la Fig. 1 que muestra los siguientes pasos para los procesos Big Data:

A. Raw Data. La entrada de un proceso Big Data está formada por **datos** que provienen de diversos recursos como son las redes sociales, los dispositivos wearables o los blogs. Además, dichos datos pueden ser de diversos tipos como texto, audio, o video.

B. Clasificación. Los datos pueden ser almacenados en diversas **fuentes** o procesados directamente. Por lo tanto, los datos pueden ser almacenados en una fuente u otra, o incluso ser procesados en tiempo real en función de diversos criterios.

C. Almacenamiento. A la hora de procesar los datos se seleccionan una o más fuentes de las que extraer sus datos. Los datos pueden provenir tanto de fuentes internas como las bases de datos de una empresa, como de fuentes externas como Internet. Por lo tanto, es necesario conocer donde se encuentran almacenados los datos a procesar, y tener en cuenta que pueden estar almacenados en más de una fuente.

D. Preparación de Datos. Los datos pueden ser preparados para darles un **formato** concreto para aplicar cierta técnica o extraer de ellos cierta característica (por ejemplo, la frecuencia de una señal).

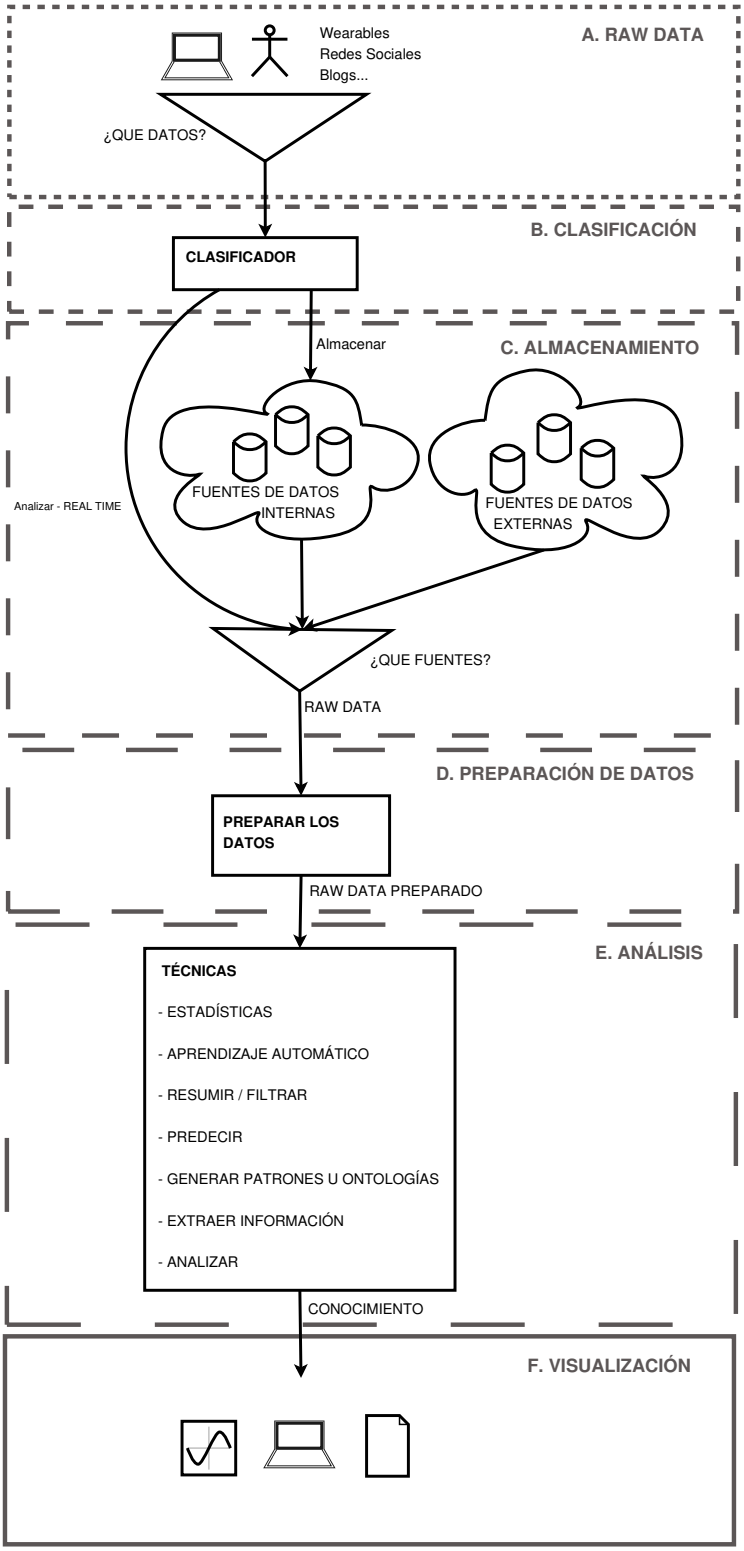

Fig. 1. Visión en conjunto de un Proceso Big Data.

E. Análisis. Se emplean **técnicas** Big Data en función del **conocimiento** a obtener. Analizar, extraer información, predecir, clasificar, filtrar o resumir, generar patrones u ontologías, hacer estadística o aprender de manera automática son algunas de las operaciones.

F. Visualización. El conocimiento generado se visualiza mediante reportes, representaciones gráficas o herramientas de visualización.

Teniendo en cuenta este proceso, es posible deducir algunos conceptos relevantes de Big Data y relacionarlos entre sí como base para una ontología que defina de manera precisa las aplicaciones Big Data. Pero mientras que la mayoría de estos conceptos son fácilmente entendibles, el concepto 'Técnica Big Data' puede resultar complejo de conceptualizar teniendo en cuenta la gran variedad de técnicas que existen. Por esta razón el primer paso es la caracterización de las técnicas Big Data.

3 Caracterización de las Técnicas Big Data

Hoy en día, existe un gran número de técnicas Big Data que pueden ser empleadas para el tratamiento de diversos tipos de datos como textos, audios o videos [17]. Dado el gran volumen de técnicas existentes, este artículo se enfoca como punto de partida en las principales técnicas Big Data para el tratamiento de texto.

Partiendo de estas técnicas, en este apartado se lleva a cabo una homogeneización de dichas técnicas como base para proponer una ontología para caracterizar las aplicaciones Big Data. Para llevar a cabo dicha homogeneización, en primer lugar, se han agrupado las técnicas teniendo en cuenta cómo se genera el conocimiento a partir de unos datos: Question Answering, Information Extraction, Text Summarization, Clustering, Sentiment Analytics, o Predictive Analytics. Por ejemplo, teniendo en cuenta sinopsis de diferentes películas el primer grupo de técnicas (Question Answering) sería capaz de contestar a preguntas como ¿Quien es el personaje principal?, el segundo grupo de técnicas (Information Extraction) sería capaz de extraer conocimiento como el nombre de los actores o las fechas de estreno, el tercer grupo de técnicas (Text Summarization) sería capaz de resumir dichas sinopsis, el cuarto grupo de técnicas (Clustering) sería capaz de clasificar las sinopsis teniendo en cuenta algún criterio como el año de estreno o el tipo de película, el quinto grupo de técnicas (Sentiment Analytics) sería capaz de evaluar las películas en función de los votos de los usuarios, y el sexto grupo de técnicas (Predictive Analytics) sería capaz de predecir qué tipo de películas serán las más vistas en los próximos años teniendo en cuenta cuales son las más vistas en la actualidad.

En segundo lugar, dichas técnicas son clasificadas en función de su enfoque, es decir, en función de cómo una técnica lleva a cabo su cometido. Por ejemplo, existen técnicas Big Data que generan resúmenes extrayendo las frases más relevantes de un texto, mientras que otras técnicas emplean la semántica para escribir un resumen cuyas frases no forman parte del texto original.

Y en tercer lugar, para cada técnica se han descrito los datos de entrada y el conocimiento que genera. Tanto los nombres de las agrupaciones, como los nombres de los enfoques y las técnicas son términos empleados frecuentemente en artículos académicos para definir dichos conceptos.

A continuación, se presenta una tabla por cada grupo de técnicas teniendo en cuenta los diferentes enfoques posibles, las técnicas Big Data correspondientes a dicho enfoque, los datos de entrada y el conocimiento de salida. Se incluyen también algunos ejemplos que ilustran de manera sencilla tanto los datos de entrada como el conocimiento de salida para facilitar su comprensión.

– QUESTION ANSWERING: Responder a una pregunta. Partiendo de una pregunta formulada en lenguaje natural, la técnicas pertenecientes a Question Answering buscan la respuesta a dicha pregunta dentro de un texto. [37][23][27][22]

Table 1. Técnicas para responder preguntas

ENFOQUE	TÉCNICAS	ENTRADA: DATOS	SALIDA: CONOCIMIENTO
INFORMATION RETRIEVAL (IR): Extrae la respuesta a una pregunta filtrando, buscando, relacionando y creando un ranking.	I. Indexing and searching techniques	(1) Texto original y (2)Pregunta y (3) Historico con respuestas a otras preguntas	Respuesta a la pregunta
	Ej.	(1) "Amazon tiene millones de productos.", (2) ¿Cuántos productos tiene amazon? y (3) "¿Cúantos productos vende amazon? Vende miles".	Tiene millones de productos.
KNOWLEDGE: Crea una representación semántica de la pregunta para poder responderla.	I. Knowledge-Based, II. Métodos Supervisados y III. Métodos Semi-Supervisados	(1) Texto original y (2)Pregunta y (3) Historico o patrones	Respuesta a la pregunta
	Ej. (Hand-built)	(1) "Amazon tiene millones de productos.", (2) ¿Cuántos productos tiene amazon? y (3) "X vende Y,X tiene Y".	Tiene millones de productos.

– INFORMATION EXTRACTION (IE): Extraer información estructurada a partir de documentos de texto con información no estructura o semi estructurada.[5][40][26][43][4][11][28][15]

Table 2. Técnicas para la extracción de información

ENFOQUE	TÉCNICAS	ENTRADA: DATOS	SALIDA: CONOCIMIENTO
NAMED ENTITY RECOGNITION (NER): Extrae todas las Named Entities (Campos como organización, persona, localización o fecha) de un texto.	I. Heurísticas	(1) Texto original y (2) Patrones con gramáticas	Valores para: persona, organización o fecha.
	Ej.	(1) "Contacte con Luis en luis@gmail.com" y (2) "§+@§+" (Patrón)	Email: luis@gmail.com
	II. Técnicas de Aprendizaje Supervisado	(1) Texto original y (2) Historico (Conjunto de datos para generar un clasificador o patrones)	Valores para: persona, organización o fecha.
	Ej. (Support Vector Machine)	(1) "Contacte con Luis en luis@gmail.com" y (2) "Jesús NOMBRE jesus@gmail.com"	Email: luis@gmail.com, Nombre: Luis
	III. Otros: Bootstrapping, Listados o Tesauros.		
RELATION EXTRACTION (RE): Extrae las relaciones existentes entre las Named Entities (NacidoEn, FundadoPor, SituadoEn).	I. Knowledge-Based Methods	(1) Texto original y (2) Tabla con patrones	Tabla de relaciones
	Ej. (Hand-built extraction)	(1) "Luis nació en Zaragoza" y (2) "X nació en Y, X es Y"	(Persona, Origen), (Luis, Zaragoza)
	Ej. (Support Vector Machine)	(1) Texto Original y (2) Historico	Relaciones entre dos entidades
		(1) "Luis nació en Zaragoza" y (2) "Olga nació en Alemania, María es de España"	(Persona, Origen), (Luis, Zaragoza)
	III. Métodos semi-supervisados	(1) Texto Original, (2) Historico pequeño y (3) Algún patron relevante.	Relaciones entre dos entidades
	Ej. (Bootstrapping)	(1) "Luis es de Zaragoza", (2) "María es de España, Álvaro es Belga." y (3) "X es Y".	(Persona, Origen), (Luis, Zaragoza)

– TEXT SUMMARIZATION: Resume el texto en otro texto de menor longitud pero que contiene las partes más importantes del original. [1][20][21][18][2]

Table 3. Técnicas para resumir texto

ENFOQUE	TÉCNICAS	ENTRADA: DATOS	SALIDA: CONOCIMIENTO
EXTRACTIVE APPROACH: Extrae las unidades de texto (frases, párrafos) más relevantes para formar un resumen del mismo.	I. Term frequency – Inverse document frequency	(1) Texto original	Resumen
	Ej. (Si el tf-idf es "resumen")	"El resumen es un escrito que sintetiza las ideas principales de un texto. Su extensión puede variar, pero no suele superar el 25% de la extensión del original."	"El resumen es un escrito que sintetiza las ideas principales de un texto."
ABSTRACTIVE APPROACH: Escribe un resumen del texto teniendo en cuenta la relevancia de las unidades y su semántica.	I. Structured Based Approach (Basado en plantillas, patrones o árboles)	(1) Texto Original y (2) Esquema o estructura del texto	Resumen
	Ej. (Template)	(1) Texto del ejemplo anterior y (2) "Definición: ¡Un documento es un escrito¿, Tamaño: ¡Su extensión puede variar¿"	- "El resumen es un escrito, su extensión es inferior a la texto original."
	II. Semantic Based Approach (Basado en sistemas de procesamiento de lenguaje natural)	(1) Texto original y (2) Histórico	Resumen

– CLUSTERING: Agrupa un conjunto de objetos de tal manera que los objetos en el mismo grupo (cluster) son similares entre sí en un sentido u otro.[1][39] Como es posible observar en la tabla 4, las técnicas para agrupar no están divididas según su enfoque o no se han encontrado descripciones de los enfoques en artículos académicos.

Table 4. Técnicas para agrupar

ENFOQUE	TÉCNICAS	ENTRADA: DATOS	SALIDA: CONOCIMIENTO
HIERARCHICAL BASED: Genera una jerarquía entre los clusters.	I. Agglomerative (De abajo hacia arriba) y II. Divisive (De arriba hacia abajo)	(1) Texto original Y (2) Características para establecer la similaridad	Grupos
	Ej.(Agglomerative hierarchical clustering)	(1) Varias noticias y (2) Temática (Cultura, Economía o General)	Artículos culturales, Artículos de economía y Artículos generales.
PARTITIONING BASED: Genera tantos clusters cómo se indiquen.	I. K-means	(1) Texto original y (2) Número de clusters a generar	Grupos
	Ej.	(1) Varias noticias y (2) 2 clusters	Artículos antiguos y Artículos recientes.

– SENTIMENT ANALYSIS/ OPNION MINING: Identificar y extraer información subjetiva.[8][34][14][6]

Table 5. Técnicas para evaluar opiniones

ENFOQUE	TÉCNICAS	ENTRADA: DATOS	SALIDA: CONOCIMIENTO
DOCUMENT LEVEL, SENTENCE LEVEL y ASPECT LEVEL	I. KNOWLEDGE BASED (Basado en palabras claves)	(1) Texto original Y (2) Palabras clave, patrones o reglas	Opinión positiva, negativa o neutral
	Ej.	(1) "Me gusta mi móvil" y (2) "Contento, aburrido, gustar, odiar"	Positiva.
	II. STATISTICAL BASED (Basado en aprendizaje automático)	(1) Texto original y (2) Histórico	Opinión positiva, negativa o neutral
	Ej. (Support Vector Machine)	(1) "Mi móvil es genial" y (2) "Es aburrido, negativa; Es genial, positiva."	Positiva.

– PREDICTIVE ANALYTICS: Engloba una gran variedad de técnicas estadísticas de modelado predictivo, aprendizaje automático y minería de datos que realizan predicciones o determinan eventos desconocidos basándose en datos actuales e históricos.[32][36][31]

Table 6. Técnicas para predecir

ENFOQUE	TÉCNICAS	ENTRADA: DATOS	SALIDA: CONOCIMIENTO
REGRESSION MODELS: Emplea ecuaciones matemáticas para representar las interacciones entre la variables a considerar.	I. Linear regression, II. Logistic regression	(1) Texto original Y (2) Histórico	Predicción
	Ej. (Linear Regression)	(1"Lunes - Día de cielos despejados en gran parte del país. Martes - Predominará el sol, con algunas nubes en el Cantábrico." y (2) Tiempo en Zaragoza: "Viernes – Soleado con alguna nubes. Sábado – Soleado. Domingo – Soleado por la mañana, son nubes por la tarde"	En Zaragoza el tiempo el miércoles será soleado con alguna nube.
MACHINE LEARNING, DECISION MODELS	I. Neural Networks, II.Support Vector Machine, III. Naïve Bayes	(1) Texto original y (2) Histórico	Predicción
	Ej. (Support Vector Machine)	(1) "Lunes - Día de cielos despejados en gran parte del país. Martes - Predominará el sol, con algunas nubes en el Cantábrico." y (2) Tiempo en Zaragoza: "Viernes, Sol con nubes; Sábado, Sol; Domingo, Sol con nubes"	En Zaragoza el tiempo el miércoles será "Sol con nubes".

4 Ontología

En esta sección, se describe una ontología inicial para identificar y definir de manera precisa los conceptos esenciales de una aplicación Big Data tomando como punto de partida la caracterización de las técnicas Big Data realizada en la sección anterior. Mediante las siguientes definiciones, se describen los conceptos que conformarían la ontología fundacional que se propone en este artículo.

Conocimiento: Resultado que se busca a partir de una Aplicación Big Data, y cuya finalidad es el manejo y compresión de los datos.

Datos: Conjunto de información donde se encuentra integrado el conocimiento que se pretende obtener.

Características: Los datos poseen ciertas características en cuanto a su volumen, su variedad (estructurados, semi-estructurados, no-estructurados), su velocidad de creación, almacenamiento o procesado, su valor o relevancia para obtener cierto conocimiento, o su veracidad.

Fuentes: Los datos pueden ser almacenados en diversas fuentes, por lo tanto para obtener un conocimiento es posible que sea necesario manejar varias fuentes a la vez.

Técnicas: Están formadas por un conjunto de procedimientos cuya finalidad es lograr obtener cierto conocimiento basándose en datos.

Enfoque: Las técnicas pueden clasificarse teniendo en cuenta el punto de vista que siguen para manejar datos y obtener un conocimiento.

Formato: Cada técnica maneja los datos ordenándolos de una manera específica para facilitar su compresión y análisis. Por lo tanto, tanto los datos de entrada cómo el conocimiento que se genera serán consecuentes con dicho formato.

Estas podrían ser definiciones básicas de los principales conceptos involucrados a la hora de crear Aplicaciones Big Data. Sus relaciones están gráficamente representadas en el siguiente diagrama:

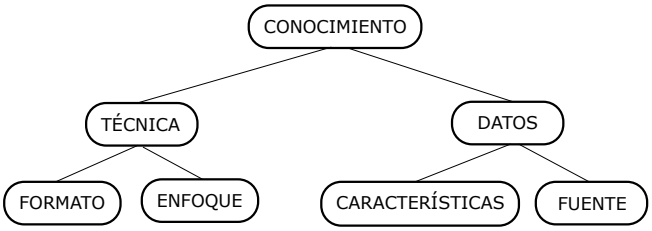

Fig. 2. Modelo Conceptual de las Aplicaciones Big Data

Una Aplicación Big Data sirve para obtener el conocimiento que el usuario quiere o necesita, a partir de ciertos datos y empleando técnicas específicas. Dichos datos provienen de diversas fuentes (bases de datos, tiempo real o documentos) y poseen ciertas características que dificultan su manejo y comprensión

cómo su volumen, su variedad o su velocidad de creación. Por lo que es necesario aplicar técnicas específicas que mediante ciertos enfoques cómo utilización de ecuaciones matemáticas, extracción de palabras clave, aprendizaje automático o jerarquías, y mediante la descripción de los datos con un formato específico consiguen obtener el conocimiento deseado.

5 Discusión

La ontología definida en la sección anterior se basa en el estudio previo de las aplicaciones Big Data para texto, pero existen también aplicaciones Big Data que son empleadas en otras áreas como las aplicaciones de análisis de información genética en medicina, o las aplicaciones de tratamiento de las imágenes o video para reconocer objetos, o las aplicaciones para procesar el sonido y la voz para ejecutar instrucciones o comandos. La principal ventaja de la ontología definida es que es tan general que todas estas aplicaciones podrían basarse en dicha ontología. Por ejemplo, el conocimiento en una aplicación de análisis de información genética estaría formado por los patrones obtenidos a partir de la información genética de las personas, la información genética serían los datos que se recopilan de diferentes personas o bases de datos (fuentes) y cuyo volumen y complejidad (características) dificultan ver a simple vista sus similitudes y diferencias para generar los patrones manualmente. Una de las técnicas más empleadas para reconocimiento de patrones (enfoque) en análisis de información genética [41] es Support Vector Machine cuyo formato tanto de los datos de entrada como del conocimiento que se genera está formado por vectores de características.

6 Conclusión

Como parte de este trabajo, hemos presentado una homogeneización de las principales técnicas Big Data empleadas para el tratamiento de textos como base de una ontología que identifica y define los conceptos esenciales de las Aplicaciones Big Data. Proponemos emplear esta ontología como primer paso para alcanzar el consenso ontológico necesario para que el soporte conceptual preciso de las nociones utilizadas en Big Data esté perfectamente determinado. Además, nuestro siguiente paso implica el enriquecimiento de esta ontología para otorgarle mayor precisión.

References

1. Aggarwal, C.C., Zhai, C.: Mining text data. Springer Science & Business Media (2012)
2. Atif, K., Naomie, S.: A review on abstractive summarization methods. Journal of Theoretical and Applied Information Technology 59(1) (2014)
3. Aufaure, M.A., Chiky, R., Curé, O., Khrouf, H., Kepeklian, G.: From business intelligence to semantic data stream management. Future Generation Computer Systems (2015)

4. Bach, N., Badaskar, S.: A survey on relation extraction. Language Technologies Institute, Carnegie Mellon University (2007)
5. Bird, S., Klein, E., Loper, E.: Natural language processing with Python. " O'Reilly Media, Inc." (2009)
6. Boiy, E., Moens, M.F.: A machine learning approach to sentiment analysis in multilingual web texts. Information retrieval 12(5), 526–558 (2009)
7. Boyd, D., Crawford, K.: Six provocations for big data. In: A decade in internet time: Symposium on the dynamics of the internet and society. pp. 1–17. Oxford Internet Institute (2011)
8. Cambria, E., Schuller, B., Liu, B., Wang, H., Havasi, C.: Statistical approaches to concept-level sentiment analysis. IEEE Intelligent Systems 3(28), 6–9 (2013)
9. Ceri, S., Della Valle, E., Pedreschi, D., Trasarti, R.: Mega-modeling for big data analytics. In: International Conference on Conceptual Modeling. pp. 1–15. Springer (2012)
10. Chen, M., Mao, S., Liu, Y.: Big data: a survey. Mobile Networks and Applications 19(2), 171–209 (2014)
11. Culotta, A., McCallum, A., Betz, J.: Integrating probabilistic extraction models and data mining to discover relations and patterns in text. In: Proceedings of the main conference on Human Language Technology Conference of the North American Chapter of the Association of Computational Linguistics. pp. 296–303. Association for Computational Linguistics (2006)
12. D, L.: 3-d data management: controlling data volume, velocity and variety. meta group research note, 6 february. https://blogs.gartner.com/doug-laney/files/2012/01/ad949-3D-Data-Management-Controlling-Data-Volume-Velocity-and-Variety.pdf (2001), [Website; Accessed on 25-10-2016]
13. De Mauro, A., Greco, M., Grimaldi, M., Giannakopoulos, G., Sakas, D.P., Kyriaki-Manessi, D.: What is big data? a consensual definition and a review of key research topics. In: AIP conference proceedings. vol. 1644, pp. 97–104. AIP (2015)
14. Farra, N., Challita, E., Assi, R.A., Hajj, H.: Sentence-level and document-level sentiment mining for arabic texts. In: 2010 IEEE International Conference on Data Mining Workshops. pp. 1114–1119. IEEE (2010)
15. Feldman, R., Fresko, M., Kinar, Y., Lindell, Y., Liphstat, O., Rajman, M., Schler, Y., Zamir, O.: Text mining at the term level. In: European Symposium on Principles of Data Mining and Knowledge Discovery. pp. 65–73. Springer (1998)
16. Gandomi, A., Haider, M.: Beyond the hype: Big data concepts, methods, and analytics. International Journal of Information Management 35(2), 137–144 (2015)
17. Gibert, K., Sànchez-Marrè, M., Codina, V.: Choosing the right data mining technique: classification of methods and intelligent recommendation (2010)
18. Greenbacker, C.F.: Towards a framework for abstractive summarization of multimodal documents. In: Proceedings of the ACL 2011 Student Session. pp. 75–80. Association for Computational Linguistics (2011)
19. Gu, Y., Storey, V.C., Woo, C.C.: Conceptual modeling for financial investment with text mining. In: International Conference on Conceptual Modeling. pp. 528–535. Springer (2015)
20. Gupta, V., Lehal, G.S.: A survey of text summarization extractive techniques. Journal of emerging technologies in web intelligence 2(3), 258–268 (2010)
21. Harabagiu, S.M., Lacatusu, F.: Generating single and multi-document summaries with gistexter. In: Document Understanding Conferences. pp. 40–45 (2002)
22. Hermjakob, U., Hovy, E.H., Lin, C.Y.: Knowledge-based question answering. In: Proceedings of the Sixth World Multiconference on Systems, Cybernetics, and Informatics (SCI-2002) (2000)

23. Hirschman, L., Gaizauskas, R.: Natural language question answering: the view from here. natural language engineering 7(04), 275–300 (2001)
24. IBM: What is big data? Disponible en: https://www.ibm.com/big-data/us/en/, [Website; Accessed on 23-10-2016]
25. Jayapandian, C., Chen, C.H., Dabir, A., Lhatoo, S., Zhang, G.Q., Sahoo, S.S.: Domain ontology as conceptual model for big data management: Application in biomedical informatics. In: International Conference on Conceptual Modeling. pp. 144–157. Springer (2014)
26. Jiang, J.: Information extraction from text. In: Mining text data, pp. 11–41. Springer (2012)
27. Jurafsky, D.: Speech & language processing. Pearson Education India (2000)
28. Konstantinova, N.: Review of relation extraction methods: What is new out there? In: International Conference on Analysis of Images, Social Networks and Texts_x000D_. pp. 15–28. Springer (2014)
29. Lambán, L., Martín-Mateos, F., Rubio, J., Ruiz-Reina, J.: Towards a verifiable topology of data. EACA 2016 p. 113
30. Mayer-Schönberger, V., Cukier, K.: Big data: A revolution that will transform how we live, work, and think. Houghton Mifflin Harcourt (2013)
31. Mitra, T., Gilbert, E.: The language that gets people to give: Phrases that predict success on kickstarter. In: Proceedings of the 17th ACM conference on Computer supported cooperative work & social computing. pp. 49–61. ACM (2014)
32. Nguyen, D., Smith, N.A., Rosé, C.P.: Author age prediction from text using linear regression. In: Proceedings of the 5th ACL-HLT Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities. pp. 115–123. Association for Computational Linguistics (2011)
33. Oracle: What is big data? https://www.oracle.com/es/big-data/index.html, [Website; Accessed on 23-10-2016]
34. Pang, B., Lee, L.: Opinion mining and sentiment analysis. Foundations and trends in information retrieval 2(1-2), 1–135 (2008)
35. Peral, J., Ferrández, A., Tardío, R., Maté, A., de Gregorio, E.: Energy consumption prediction by using an integrated multidimensional modeling approach and data mining techniques with big data. In: International Conference on Conceptual Modeling. pp. 45–54. Springer (2014)
36. Radford, A.: Recurrent neural networks for text analysis in open data science conference. http://es.slideshare.net/odsc/alec-radfordodsc-presentation (2015), [Website; Accessed on -10-2016]
37. Roshdi, A., Roohparvar, A.: Review: Information retrieval techniques and applications
38. SAS: Big data - what it is and why it matters. http://www.sas.com/en-us/insights/big-data/what-is-big-data.html, [Website; Accessed on 29-10-2016]
39. Steinbach, M., Karypis, G., Kumar, V., et al.: A comparison of document clustering techniques. In: KDD workshop on text mining. vol. 400, pp. 525–526. Boston (2000)
40. Urbano, J., Morato, J., Marrero, M., Sánchez-Cuadrado, S.: Recuperación y acceso a la información. OpenCourseWare, University Carlos III of Madrid (2010)
41. Valafar, F.: Pattern recognition techniques in microarray data analysis. Annals of the New York Academy of Sciences 980(1), 41–64 (2002)
42. Ward, J.S., Barker, A.: Undefined by data: a survey of big data definitions. arXiv preprint arXiv:1309.5821 (2013)
43. Zettlemoyer, L.: Relation extraction. https://goo.gl/Z7NdJp, [Website; Accessed on 12-09-2016]