TRABAJO FIN DE MÁSTER

MÁSTER EN CIENCIA DE DATOS E INGENIERÍA DE COMPUTADORES

Análisis de tendencias en Big Data

Autor
José Ángel Díaz García

Directoras
María José Martín Bautista
María Dolores Ruiz Jiménez

Escuela Técnica Superior de Ingenierías Informática y de Telecomunicación

—

Granada, Abril de 2019

Análisis de tendencias con Big Data

Palabras clave: Reglas de Asociación, Reglas de Asociacion Difusas, Big Data, Minería de textos, Twitter

Resumen:

Trend Analysis with Big Data

Keywords: Association rules, Fuzzy Association Rules, Big Data, Text mining, Twitter

Abstract:

Agradecimientos

Este último año podríamos catalogarlo como uno de los más interesantes que probablemente la vida me depare, al menos en cuanto a lo profesional e intelectual se refiere. Compaginar los últimos meses de máster profesional, con mi incorporación como personal de investigación a la Universidad de Granada y la posterior matriculación en el Máster de Ciencia de Datos, no ha sido nada fácil y a muy seguro no habría sido posible sin esas personas que siempre han estado ahí. Por ello no se me ocurre mejor manera de comenzar a escribir este proyecto que con los agradecimientos.

En primer lugar, es a mi familia, concretamente a mi madre y abuela pero sobre todo a mi pareja, Rocio, a quien va dirigida esta sección ya que son las personas más cercanas y que más han tenido que soportar mis momentos de estrés y porque no decirlo, mal humor.

En segundo lugar, a mis compañeros de piso y amigos, Luis, Nourdin, Rosa y Alberto con los que tan buenas tardes de evasión he pasado, así como a mis compañeros del máster profesional en Ingeniería Informática, junto a los cuales las clases y largas prácticas del máster fueron superadas de la mejor forma posible.

Por último agradecer a los tutores del proyecto, María José Martín Bautista y María Dolores Ruiz Jiménez el haberme dado la oportunidad de trabajar y desarrollarme intelectualmente con grandes profesionales en el sector de informática y la investigación como ellas.

Índice general

Índice de figuras

Índice de tablas

Capítulo 1

Introducción

Actualmente nadie debería sorprenderse cuando escuche que vivimos en la *sociedad de la información*, concepto acuñado para referenciar a una sociedad cambiante y donde la manipulación de datos e información juega un papel más que relevante en las actividades sociales, culturales y sobre todo, económicas. El tratamiento de estos datos puede suponer una ardua labor, más aún cuando el volumen de éstos es tan grande que los paradigmas para su procesado deben migrar hacia nuevas vertientes y aún más cuando estos datos provienen de fuentes tan dispares como nuestras tendencias en la compra diaria, el uso que le damos a una tarjeta de crédito o a una red social... Es por ello, que fruto de la necesidad del análisis y la obtención de información de estos datos en especie desestructurados y aparentemente carentes de significado, surgen técnicas y herramientas capaces de procesar y obtener información útil y relevante.

Como hemos mencionado anteriormente, las redes sociales son grandes factorías de datos. Datos que una vez procesados pueden servir de ayuda para comprender temas relevantes de la sociedad actual, o incluso desvelar patrones aparentemente ocultos en los hábitos de comportamiento de usuarios que pueden ser de ayuda en procesos de toma de decisiones o para diversos estudios posteriores. A este proceso se le denomina minería de redes sociales o *social media mining* y es una de las vertientes de estudio sobre la que más se investiga actualmente dentro del ámbito de la minería de datos.

A continuación, haremos una breve introducción a la motivación, para continuar con la puntualización de los objetivos principales del proyecto de fin de

máster y concluiremos la sección dando al lector una idea de la organización final de la memoria.

1.1. Motivación

Vivimos en un mundo aparentemente obsesionado por etiquetar, clasificar y buscar relaciones entre todo lo que nos rodea. El buscar relaciones entre distintos factores como por ejemplo asociar la existencia de un tipo determinado de nubes con una probabilidad más alta de lluvias, es algo innato del ser humano desde tiempos inmemoriales e inherente a la totalidad de los ámbitos de estudio habidos y por haber a lo largo de la historia.

Las técnicas de minería de datos y extracción de conocimiento, tales como reglas de asociación, clustering o modelos de clasificación entre otras, no son muy distintas al menos en el concepto general de la búsqueda de relaciones en cualquier ámbito o problema. Pese a que estas técnicas están presentes en casi todas las vertientes de estudio y desarrollo con las que los seres humanos actualmente trabajan, hay ciertos problemas o enfoques en los que destacan notablemente y en los cuales son herramientas esenciales. Estos problemas son tales como la detección de comunidades [1], la realización de diversos estudios y herramientas enfocados al marketing [4] en pequeñas y grandes compañías, la elaboración de modelos predictivos en ámbitos financieros o de seguros [5] y por supuesto la minería de redes sociales o el análisis de sentimientos [2] [3].

Estos últimos campos, actualmente se han convertido en una de las vertientes más estudiadas, dado su interés para comprender los hábitos de los usuarios desde una perspectiva de análisis más fiable incluso a preguntar de forma particular a las personas, cuyas respuestas pueden estar sesgadas por el estudio en cuestión. Es en la minería de redes sociales, junto con las técnicas anteriormente introducidas y que veremos con más detalle en los puntos siguientes, donde surge lo que conocemos como análisis de tendencias o **minería de opiniones**. Objeto de estudio en el que se trata de comprender o analizar comportamientos, actividades y opiniones, por ejemplo, de consumidores de cierto producto o usuarios de cierta red social. El fin de estas técnicas es por tanto la extracción de conocimiento útil que pueda traducirse en ventajas competitivas en el proceso de toma de decisiones de una pequeña

o gran compañía, sin olvidar claro está las connotaciones científicas y áreas de estudio que se pueden desarrollar en el proceso.

Este ámbito, que aúna, técnicas de minería de datos, redes sociales y en cierta medida Big Data, es relativamente nuevo, debido sin duda alguna a la novedad que las redes sociales ofrecen. Por poner algún ejemplo, Twitter fue fundada en el año 2006 y Facebook en el 2005, lo que nos da una media de unos 11 años de vida en las redes sociales más famosas, antiguas y usadas. Por otro lado, debemos tener en cuenta que su implantación y comercialización en la sociedad no tuvo lugar el mismo día de fundación por lo que su 'edad' sería aún menor.

Si dejamos apartado el 'problema' de la reciente novedad de las redes sociales, y nos centramos en los aspectos puramente informáticos del proyecto (BigData y minería de opiniones), también tienen un notable carácter de novedad. El BigData por su parte, es uno de los más recientes avances de la computación a gran escala, haciendo que nos encontremos aún en los albores de la explotación de esta tecnología. Por su parte, la minería de opiniones, íntimamente ligada a la minería de redes sociales y la aparición de estas por tanto, promueve un gran interés tanto en los aspectos empresariales y comerciales de la sociedad como en ámbitos relacionados con la investigación.

La novedad del estudio de estas técnicas, hace que haya pocos trabajos previos completamente relacionados con el ámbito de estudio, pero también hace que actualmente sea una de las áreas de investigación que más interés suscita entre la comunidad científica, dada la creciente importancia que las redes sociales digitales están tomando en casi la totalidad de las acciones y tareas de nuestro día a día.

1.2. Objetivos del proyecto

En el presente proyecto de fin de máster podemos encontrar un objetivo principal del cual posteriormente se desgranarán objetivos secundarios. El objetivo principal del proyecto será obtener tendencias y patrones de opinión que los usuarios de una red social puedan tener sobre distintos temas o personajes como por ejemplo políticos o artistas. Estas tendencias, serán obtenidas por un modelo de minería de datos basado en técnicas de aprendizaje no supervisado y minería de textos que se nutrirá de datos provenientes

de Twitter. Este objetivo a su vez podemos descomponerlo en objetivos con menos granularidad, que serían los siguientes:

- Obtención de información sobre la minería de redes sociales, el análisis de tendencias y las técnicas de Big Data aplicadas en estos campos anteriormente.
- Estudio del estado del arte en el campo del análisis de tendencias y la minería de opiniones con técnicas no dirigidas.
- Extracción de los datos provenientes de la red social Twitter para analizar y aplicar las técnicas desarrolladas.
- Almacenamiento y procesado de los datos usando técnicas de minería de textos y Big Data.
- Aplicación de técnicas de minería de datos descriptiva para obtener patrones interesantes en los datos.
- Pruebas y experimentación.
- Análisis de resultados y comparación de los mismos con posibles eventos políticos y sociales.

1.3. Organización de la memoria

Tras el estudio del problema e introducción al tema visto en este punto, los siguientes capítulos se centran en el estudio del estado del arte de la materia y finalmente en el desarrollo de la solución aportada. En el siguiente capítulo podemos encontrar detalladamente la planificación seguida durante la elaboración del proyecto, así como el estudio de los recursos empleados; tras este capítulo encontramos una serie de capítulos donde estudiamos el estado del arte del uso de reglas de asociación en minería de redes sociales y por supuesto, en conjunción con los paradigmas y arquitectura propuestos usando el BigData. En la parte central del proyecto se entrará en detalle en la solución aportada, así como en el estudio y la documentación de las técnicas usadas para la limpieza, integración y visualización de los datos. Se concluirá con un estudio de los resultados y las vías futuras que la elaboración de este proyecto abre.

Capítulo 2

Planificación del proyecto

Una correcta planificación puede suponer el éxito o rotundo fracaso del proyecto en cualquier ámbito o disciplina aplicable. Si esta disciplina es a su vez la ingeniería en cualquiera de sus vertientes, la necesidad de una correcta planificación se acentúa aún más llegando a convertirse en una de las partes cruciales y más importantes del proyecto en sí. En ciencia de datos, esta parte no es menos importante, ya que una correcta planificación temporal que ayude al científico de datos, a distribuir su tiempo y esfuerzos entre las distintas partes que integran un proyecto de análisis de datos (integración, preprocesado, minería de datos) puede ser de vital importancia de cara al triunfo o el fracaso del proyecto.

En este capítulo haremos un resumen de la planificación del proyecto versando este en los recursos software, humano y hardware empleados así como de la planificación temporal seguida por el mismo.

2.1. Gestión de recursos

En esta sección se hará un repaso por los recursos utilizados, siendo estos como vimos en la introducción del capítulo, tres categorías bien diferenciadas, recursos de personal, hardware y software los cuales son a su vez los tres pilares clave de un proyecto de ingeniería informática, a pesar de que en este caso que nos compete esté más enfocado al ámbito de investigación que al del

diseño y desarrollo de un producto final, como sería el caso de un proyecto íntegro de ingeniería del software.

2.1.1. Personal

El personal del proyecto radica exclusivamente en el autor José Ángel Díaz García, encargado de todas las partes del mismo, bajo la supervisión de los tutores.

2.1.2. Hardware

Elemento	Características
Procesador	2,6 GHz Intel Core i5
Memoria Ram	8 GB 1600 MHz DDR3
Disco duro	SATA SSD de 120 GB

Tabla 2.1: Especificaciones técnicas de la máquina personal usada.

Además de la máquina personal se ha utilizado un cluster de procesado de datos para el proceso de *named entity recognition (NER)*, que es sin duda el más costoso computacionalmente hablando. Este cluster está formado por cuatro máquinas con las especificaciones técnicas que podemos ver en la tabla 3.3.1.

Elemento	Características
Procesador	Intel Xeon E5-2665
Memoria Ram	32 GB
Nucleos	8

Tabla 2.2: Especificaciones técnicas del cluster.

2.1.3. Software

El software utilizado es en su práctica totalidad software libre, siendo el restante software propietario cuyas licencias vienen incluidas en el sistema operativo de la máquina usada siendo este OS X . El software usado es:

- **TeXShop**: procesador de textos basado en Latex usado para elaborar la documentación del presente proyecto.
- **Scrapy**: Librería de Python que ofrece un *framework* para la creación de *web crawlers*.
- **Twitter**: Red social de microblogging.
- **MongoDB**: Base datos noSQL usada como almacén persistente de los datos.
- **RStudio**: Entorno de Desarrollo en R donde se ha realizado la mayor parte del proceso del proyecto.
- **RSpark**: Librería para R que ofrece grandes ventajas a la hora de procesar grandes cantidades de datos bajo este lenguaje de programación.

2.2. Planificación temporal

La parte más importante de esta sección radica en la planificación temporal seguida en los meses de trabajo que el proyecto ha ocupado, siendo este elaborado continuamente etapa a etapa.

1. **Obtención de información y estudio del tema**: La primera parte del proyecto consistió en la obtención de información acerca de la minería de opiniones y de las reglas de asociación así como de la aplicación de estas en el ámbito de la minería de redes sociales y más concretamente en Twitter. En este primer proceso de recopilación de información también se estudiaron temas más genéricos dentro del Big Data y la minería de datos con el fin de tener una visión global de las herramientas y técnicas a estudiar y usar en el problema. Esta etapa aunque ha sido continua, tuvo especial importancia desde mediados de noviembre de 2016 a finales de diciembre de ese mismo año.

2. **Estudio del estado del arte**: Tras obtener buena cantidad de información y comprender el problema a resolver, se realizó un estudio exhaustivo del estado del arte de la materia así como a comenzar a desarrollar los primeros capítulos de la memoria en cuestión. Esta etapa tuvo lugar desde finales de diciembre de 2016 hasta finalizar el proyecto debido a que se ha realizado un estudio continuo de los nuevos trabajos que iban apareciendo sobre la temática.

3. **Selección de herramientas**: Una vez fijado Twitter como medio objetivo, se llevó a cabo una investigación sobre las herramientas más oportunas para la obtención de los tuits de la red social. Esta etapa tuvo lugar entre final de junio y principio de julio de 2017.

4. **Obtención del dataset**: Para poder comenzar a hacer pruebas y desarrollar el sistema basado en reglas, una vez elegida la herramienta, se comenzó a obtener datos de la red social durante unos días ininterrumpidamente para tener un conjunto de entrenamiento suficiente. Esta tarea tomo lugar a mediados de julio de 2017.

5. **Carga y preprocesado de los datos**: Una vez obtenidos los datos y almacenados en MongoDB se hizo necesaria su carga y limpieza, esta tarea no es trivial ya que necesitó de técnicas de procesado del lenguaje natural y aplicaciones de Big Data para poder trabajar con un volumen de datos muy elevado en una máquina estándar como es el caso. Esta tarea fue llevada a cabo entre los meses de julio y octubre de 2017.

6. **Limpieza de datos**: Dado que partimos de un problema no supervisado, donde los datos carecían de filtrado alguno, esta fue una de las etapas que más tiempo tomó. Tras la aplicación de técnicas básicas de limpieza en minería de textos, se aplicaron técnicas experimentales de procesamiento del lenguaje natural para filtrar los datos y poder poner el foco del problema en aquel subconjunto de datos que hace referencia a personas. Dado el volumen de datos esta etapa precisó el uso de un cluster de procesado así como de técnicas de programación paralela y concurrente que podríamos enmarcar como Big Data. Esta tarea fue llevada a cabo entre octubre y diciembre de 2017.

7. **Análisis exploratorio de datos**: Sobre el dataset final, se han realizado gráficos y estudios estadísticos básicos con el fin de conocer y

entender mejor la naturaleza de los mismos. Esta tarea fue llevada a cabo durante el mes de diciembre de 2017.

8. **Análisis de sentimientos**: Sobre los datos, se aplicaron técnicas de análisis de sentimientos para poder realizar gráficos que nos ayudaran a discernir qué palabras o expresiones estaban relacionadas en nuestro dataset con sentimientos para en el paso de obtención de reglas de asociación poder polarizar en cierta medida las mismas, o tener al menos, otro enfoque subjetivo de éstas, pudiendo así desambiguar en cierta medida las mismas. Esta tarea fue llevada a cabo durante el mes de diciembre de 2017.

9. **Reglas de asociación y experimentación**: Con los datos limpios y estudiados, se obtienen un conjunto de reglas de asociación sobre la temática y se experimenta sobre el mismo obteniendo distintos conjuntos en función de itemsets frecuentes, así como de la variación de los parámetros de confianza y soporte en las reglas. Esta tarea comprendió los meses de diciembre de 2017 y enero de 2018.

10. **Elaboración de la memoria**: La memoria ha constado de una elaboración continua, ya que continuamente se han ido añadiendo y refinando capítulos en función de cómo se avanzaba en el proceso de desarrollo y experimentación. Los meses que ha comprendido su elaboración, han sido por tanto desde primeros de febrero de 2017 hasta enero de 2018.

2.3. Costes

Tras el análisis de los recursos empleados y la planificación temporal seguida, es menester estimar los costes del proyecto en el supuesto caso de su implantación en una empresa o grupo de investigación. Esta estimación de costes está realizada en función de dos verticales, los gastos de personal y los gastos de ejecución.

Personal

Como hemos descrito en la sección 2.1.1 el personal radica en un solo investigador y la dedicación total atendiendo a la carga lectiva del proyecto

final de máster, estaría en torno a los 3 meses a jornada completa. Teniendo en cuenta una estimación de unos 2000 euros brutos al mes tendríamos un total de 6000 euros en gastos de personal.

Ejecución

En esta categoría encontramos los gastos de adquisición del material inventariable así como los gastos del material fungible. Como inventariable, tenemos los equipos descritos en la tabla 2.1.2, es decir, el equipo personal y el cluster de procesado cuyo coste en función del período de amortización (precio por uso dividido entre tiempo de amortización) puede verse en la tabla 2.3.

Unidad	Precio	Periodo Amortización	Duración proyecto	Total
Mac pro 2,6GHz Intel Core i5	1300	2 años	1,5 meses	81,25
Cluster	6000	6	1semana	30

Tabla 2.3: Detalle de costes inventariables.

Si atendemos a los costes fungibles, es decir, aquellos no inventariables como el material de oficina, podríamos estimarlos en unos 100 euros.

Resumen de gastos

En la tabla 2.4 podemos ver un total de los gastos del proyecto en función a lo descrito en esta sección.

Gastos elegibles	Total
Personal	6000
Costes inventariables	111,25
Costes fungibles	100
TOTAL	6211,25

Tabla 2.4: Resumen final de costes.

Capítulo 3

Marco de trabajo

Antes de comenzar a abordar el estado del arte del análisis de tendencias o *minería de opiniones*, y más concretamente su aplicación en el ámbito de la web 2.0, es necesario introducir algunos conceptos teóricos que nos permitan comprender mejor los conceptos de los trabajos de los que discutiremos en el capítulo **??**. Sobre ello, hablaremos en este capítulo.

3.1. Minería de medios sociales digitales

La reciente incursión de las redes sociales digitales en nuestro mundo han cambiado el paradigma de trabajo, económico y social de la sociedad. Dada su importancia, diversos sectores y ámbitos de estudio han puesto el punto de mira en el estudio de estos nuevos paradigmas sociales. La minería de datos es uno de los campos que estudia los medios sociales digitales originando una nueva vertiente de la misma denominada como **minería de medios sociales**.

La **minería de medios sociales**, acorde a P. Gundecha [6], comprende el proceso de representar, analizar y extraer de datos provenientes de medios sociales patrones con significado y valor. La minería de medios sociales es por tanto un campo multidisciplinar y su alcance puede ser dividido en los siguientes ámbitos de aplicación:

- **Análisis de comunidades**: Por medio de teoría de grafos, se obtienen comunidades dentro de nuestra población objetivo. Estos pueden ser usuarios con similares intereses, gustos o preferencias.
- **Sistemas de Recomendaciones colaborativos** : Se basa en la hipótesis en que usuarios similares tendrán gustos similares por lo que se pueden afinar los sistemas de recomendación teniendo estos factores en cuenta.
- **Estudios de Influencia**: Se basan en la obtención de la influencia de marcas o personas en determinados sectores.
- **Difusión de la información**: En un mundo saturado de información como el actual, saber de qué manera tendremos que difundirla para llegar a un mayor número de personas es un factor decisivo. Esto es lo que estudia este área dentro de la minería de medios sociales.
- **Privacidad, seguridad y veracidad**: Este punto se centra en la verificación automática de cuentas falsas, identificación de fuentes de spam así como de la identificación de la veracidad de información o identificación de problemas de violación de privacidad.
- **Opinion mining**: Este punto, es uno de los más estudiados en **minería de medios sociales**, podemos encontrarlo junto al análisis de sentimientos aunque como veremos en el punto siguiente hay ligeras diferencias. Dada la relevancia de cara al presente trabajo ampliaremos este concepto en la sección siguiente.

3.2. Minería de opiniones

La minería de opiniones, conocida en el ámbito internacional como *opinion mining*, es una vertiente al alza dentro de la famosa minería de textos y tiene su raíz por tanto en las técnicas de procesamiento de lenguaje natural. Si analizamos la web o las publicaciones en redes sociales, encontraremos cientos de miles de *reviews* o posts de personas acerca de un producto o marca, el potencial de analizar la finalidad de esta opinión, ver si es una critica constructiva, si se promueve el producto o si simplemente lo crítica puede suponer una gran ventaja competitiva para las empresas y marcas,

por ello, son más las que cada vez usan estas técnicas en sus procesos de vigilancia tecnológica u obtención del *feedback* del consumidor.

Como todas las especializaciones o vertientes dentro del área de la minería de textos, en *opinion mining* tratamos por tanto de obtener información relevante y de valor a partir de textos, como los que hemos mencionado anteriormente, blogs, tweets o diversas redes sociales, de ahí que sea estudiada dentro del proceso de *social media mining* descrito anteriormente, ya que podríamos decir que una técnica complementa a la otra. Pero, ¿qué es una opinión? Acorde a la definición dada por Liu en [7], una opinión es una quíntupla compuesta de los siguientes elementos:

1. **Entidad**: Puede ser un objeto, persona, servicio, lugar sobre el que se emite la opinión.
2. **Emisor**: Entidad que emite la opinión.
3. **Aspecto**: Es un aspecto que se valora sobre la **entidad** en cuestión.
4. **Orientación**: Puede ser positiva, negativa o neutra.
5. **Momento temporal**: Corresponde al momento en que la opinión se emite, ya que mismos **emisores, entidades y aspectos** podrán cambiar de **orientación** en momentos distintos, por lo que es un registro importante a tener en cuenta.

Pese a que aún no hemos entrado en el estudio de las redes sociales ni de la **anatomía** de un tweet, estos serán la fuente y la unidad mínima de información en nuestro proyecto. En el punto 3.5.2 trazaremos un claro paralelismo entre esta definición y los tweets en concreto.

La minería de opiniones, se centrará por tanto en obtener de textos que podrán provenir de diferentes fuentes, *aspectos* de opinión, esto difiere en cierta medida del proceso de *análisis de sentimientos* [8] [9] que se centra desde un enfoque mayormente supervisado en la clasificación de estas entidades textuales acorde a sentimientos u orientación. Analizando estos *aspectos* y sus implicaciones sobre su *entidad* relacionada, podremos obtener por tanto ventajas muy relevantes como por ejemplo saber qué opinan los consumidores de una marca en concreto, posicionar productos u obtener análisis de confianza entre otras muchas aplicaciones.

En el presente proyecto, obviaremos la rama supervisada, para centrarnos en el enfoque no supervisado dentro del campo de la *minería de opiniones*, en el que no conocemos las clases o *etiquetas* a priori, aunque sí que se incluye una cierta polarización básica sobre las opiniones, ya que es información relevante en el conjunto del proceso. Las técnicas de aprendizaje no supervisado que estudiaremos, son el *clustering* y las reglas de asociación, ambas las trataremos en el siguiente punto.

3.3. Técnicas de aprendizaje no supervisado

Pese a la gran relación que existe entre las técnicas mencionadas anteriormente, estas difieren en factores tan dispares como su utilización, su aplicación o la información que aportan sobre un problema. El análisis de estos factores nos ayudará a elegir la técnica o el conjunto de técnicas adecuadas para cada problema concreto, es decir, podemos partir de enfoques diferentes que se apoyen y retroalimenten mutuamente. Cabe destacar y diferenciar las técnicas más relevantes aplicadas a los problemas inherentes al estudio del análisis de tendencias, para así poder comprender mejor los siguientes capítulos y nuestro problema en cuestión.

3.3.1. Clustering

Las técnicas de clustering, se basan en la obtención de grupos o clases en función de un determinado conjunto de muestras o población, sin conocer a priori estas clases. Las técnicas de clustering, están enmarcadas dentro del aprendizaje no supervisado y basan la obtención de estos grupos y clases en dos factores como pueden ser la distancia o la similitud. De todos los problemas mencionados anteriormente estas técnicas son muy utilizadas en estudios relativos al marketing y estudios sociales, donde es relevante obtener agrupaciones. Un ejemplo concreto sería discernir entre los distintos tipos de cliente que compran regularmente en un supermercado, para poder ofrecer ofertas concretas en función del grupo de manera que estas sean personalizadas en función de cada cliente, permitiendo así que los beneficios se incrementen [10].

3.3.2. Reglas de asociación

La reglas de asociación dentro del ámbito de la informática no son muy distintas, al menos en el concepto general, de la búsqueda de relaciones en cualquier ámbito. Las reglas de asociación se enmarcan dentro del aprendizaje automático o minería de datos y no es algo nuevo sino que llevan siendo usadas y estudiadas desde mucho tiempo atrás, datando una de las primeras referencias a estas, del año 1993 [11]. Su utilidad es la de obtener conocimiento relevante de grandes bases de datos y se representan según la forma **X->Y** donde **X**, es un conjunto de ítems que representa el antecedente e **Y** un ítem o conjunto de ítems consecuente, por ende, podemos concluir que los ítems **consecuentes** guardan una relación de co-ocurrencia con los ítems **antecedentes**. Esta relación puede ser obvia en algunos casos, pero en otros necesitará del uso de algoritmos de extracción de reglas de asociación que podrán desvelar relaciones no triviales y que puedan ser de mucho valor. Podremos presentar por tanto a las reglas de asociación, como un método de extracción de relaciones aparentemente ocultas entre ítems o elementos dentro de bases de datos transaccionales, *datawarehouses* u otros tipos de almacenes de datos de los que es interesante extraer información de ayuda en el proceso de toma de decisiones de las organizaciones.

Medidas

La forma clásica de medir la bondad o ajuste de las reglas de asociación a un determinado problema, vendrá dada por las medidas del **soporte**, la **confianza** y el **lift**, que podremos definir de la siguiente manera:

- Soporte: El soporte de un ítem se representa como *supp (X)*, y representa la fracción de las transacciones que contienen al ítem X entre el total de transacciones de la base de datos (D). El $supp(X \rightarrow Y)$ sería por consiguiente el total de transacciones que contiene tanto al ítem X como al ítem Y, y quedaría definido con la siguiente ecuación:

$$supp(X \rightarrow Y) = supp(X \cup Y) \tag{3.1}$$

- Confianza: Se representa como *conf (X→Y)*, y representa la fracción de transacciones en las que aparece el ítem Y, de entre aquellas transacciones donde aparece el ítem X. Su ecuación sería:

$$conf(X \rightarrow Y) = \frac{supp(X \rightarrow Y)}{supp(X)} \tag{3.2}$$

- Lift: El *lift*, es una medida útil para evaluar la independencia entre los ítems de una determinada regla de asociación. En una regla del tipo *lift (X→ Y)*, esta medida representa el grado en que X tiende a ser frecuente cuando A está presente en la regla, o viceversa. El lift, quedará definido matemáticamente de la siguiente manera:

$$lift(X \rightarrow Y) = \frac{conf(X \rightarrow Y)}{supp(Y)} \tag{3.3}$$

Pese a que estas medidas son las más comunes y extendidas, hay innumerables propuestas de medidas complementarias en la literatura, tales como la **convicción, factor de certeza, diferencia absoluta de confianza** entre otras muchas.

Obtención de reglas

Si nos centramos en la manera de obtener las reglas, estas pueden abordarse desde dos perspectivas, solución por fuerza bruta (prohibitivo) o desde un enfoque basado en dos etapas. La primera de estas etapas es la generación de itemsets frecuentes, a partir de los cuales, en la segunda etapa se obtienen las reglas de asociación, que tendrán, si todo ha ido correctamente, un valor de confianza aceptable o elevado. La primera etapa de obtención de itemsets frecuentes puede conllevar problemas de memoria ya que en una base de datos con muchos ítems o transacciones el número de estos será muy elevado, es por ello que surgen aproximaciones en el proceso de representación de itemsets frecuentes que nos permitirán obtener estos en bases de datos de gran tamaño. Estas aproximaciones son:

- Itemsets maximales: Son aquellos itemsets frecuentes para los que ninguno de los superconjuntos inmediatos al itemset en cuestión, son frecuentes. A partir de estos podremos recuperar todos los itemsets frecuentes de manera sencilla sin tener que mantenerlos todos en memoria.
- Itemsets cerrados: Son aquellos itemsets frecuentes para los que ninguno de los superconjuntos inmediatos al itemset en cuestión, tienen

un soporte igual. Con esta aproximación, tendremos soportes e itemsets frecuentes que podremos recuperar fácilmente, aunque al ser más numerosos que los maximales mantenerlos en memoria puede llegar a ser complicado.

Aplicaciones

Su uso ha sido extendido en campos como las telecomunicaciones, gestión de riesgos, control de inventarios [12] [13] o almacenes y recientemente en el minado de redes sociales representando en este ámbito una de las vertientes más estudiadas actualmente en campos de estudio como por ejemplo el análisis de sentimientos [14]. Dada la importancia de estas técnicas en nuestro trabajo las estudiaremos con detalle en el capítulo **??**, donde veremos su aplicación en diversos trabajos relacionados en menor o mayor medida con el nuestro.

3.4. Big Data

Como hemos visto en puntos anteriores, la 'explosión' y expansión de la era digital ha hecho que el volumen de datos de los que disponemos, así como de las fuentes que generan estos datos se hayan multiplicado exponencialmente. Erraríamos por tanto, si pensáramos que las técnicas tradicionales de carga, procesado y análisis de datos tradicionales pudieran ser aplicadas a estos grandes volúmenes de datos, por lo que ha sido necesario la implantación y creación de nuevas técnicas capaces de lidiar con estos grandes volúmenes de datos, y a esto es lo que conocemos como *Big Data Analytics*.

3.4.1. Historia

Pese a que es un término que llevamos pocos años escuchando, su acuñamiento data del año 1998, donde el libro *Predictive data mining: a practical guide.* [15], ya hacía referencia a los grandes volúmenes de datos y sus problemas relacionados, bajo el término de BigData, pero no fue hasta entrado el año 2000 cuando empezaron a aparecer los primeros artículos académicos, que podrían enmarcarse dentro del BigData. Pocos años después, con la apa-

rición y expansión de las redes sociales, estas empresas necesitaron nuevos paradigmas y algoritmos para procesar esta gran cantidad de información que venía de las mismas. Fue en este punto, y tras otros estudios como el llevado a cabo por Alex 'Sandy' Pentland en el MIT [16], cuando se comenzó a hablar de las **3 V's del Big Data** [17], tomando por tanto este nuevo concepto su forma actual y comenzando la expansión que le llevaría a ser hoy en día una de las 'tecnologías' más punteras.

3.4.2. Las V's del Big Data

En este punto, entraremos a hablar de las conocidas **V's del Big Data**, adjetivos que en su conjunción lo definen como tal y que en sus orígenes, fueron 3, aunque pronto se fueron complementando y extendiendo, hasta nuestros días donde el BigData quedaría caracterizado por 5 V's:

1. **Volumen**: La relación de esta palabra con el concepto Big Data es clara. Y es que el tamaño de los datos continúa aumentando, hasta volúmenes de los mismos nunca antes vistos.
2. **Variedad**: Los tipos de los datos son muy distintos y provienen de fuentes muy dispares.
3. **Velocidad**: Los datos son muy volubles y deben ser recogidos y analizados rápidamente, véase por ejemplo en el concepto de una aplicación de Big Data en bolsa, donde tan solo un segundo puede suponer perdidas o beneficios muy importantes.
4. **Variabilidad**: Los datos pueden cambiar de estructura o interpretación.
5. **Valor**: En última instancia, sin valor, no hay Big Data y es que estos datos una vez procesados deben aportar conocimiento y valor a la empresa y organización.

3.4.3. Aplicaciones

Las aplicaciones del BigData, dado su interés, están presentes en numerosas áreas, algunas de las cuales pueden ser las siguientes:

- Negocios y marketing: Análisis de comportamientos en el comprador, detección de comunidades.
- TIC: En este sector los beneficios son muy relevantes y evidentes, como por ejemplo reducir el tiempo de procesamiento de horas e incluso días a unos pocos segundos.
- Salud y ciencia: En este área el BigData ha supuesto una auténtica revolución. Disponer de nuevos algoritmos y formas de procesar datos más eficientes y potentes han supuesto la posibilidad de obtener el mapa genético de una persona en concreto a velocidades y costes antes nunca pensados, esto tiene grandes beneficios para la ciencia y la salud de esta persona que podrá incluso prevenir enfermedades futuras.

Todas estas aplicaciones, tienen como último fin mejorar los procesos de negocio y en última medida la vida diaria de las personas de a pie, lo que hace que aunque el concepto del Big Data esté aún en sus albores de lo que podrá ser en un futuro sus beneficios pueden notarse desde ya en el día a día de la sociedad.

3.5. Twitter

Twitter nace en Estados Unidos en el año 2006, partiendo de la idea de los antiguos mensajes de texto (SMS) limitó el número de caracteres en cada tuit a 140 favoreciendo que el intercambio de información fuera rápido, conciso y fluido, dando comienzo a una nueva vertiente en la web 2.0 que posteriormente se conocería como *microblogging*.

El crecimiento de la red social en los últimos años ha sido exponencial, hecho que no la ha alejado de tener serios problemas de rentabilidad, pero que confirman su éxito y aceptación por parte del gran público. A principios de 2010 el número de usuarios activos al mes de la misma se fijaba en torno a los 30 millones, número claramente superado en la actualidad donde se estima en torno a los 313 millones de usuarios activos mensuales (Dreamgrow Marketing, 2017).

Recientemente, la red social ha aumentado el número de palabras en cada tuit a 280, lo que en conjunción con nuevas medidas como la facilidad para in-

cluir videos o demás contenido multimedia y lo vistoso de estas publicaciones con un diseño intachable, constatan la salud de la red social.

3.5.1. Funcionamiento

El funcionamiento de la red social en sí trivial, en esta podemos acotar un rol muy sencillo, el de **seguidor** que serán aquellas personas que quieren seguir nuestras publicaciones y de las cuales podremos ser seguidores o no, es decir, puede no es de obligada existencia el carácter bidireccional en una relación de 'amistad' dentro de esta red social al igual que existe en otras redes sociales como por ejemplo Facebook.

Este tipos de relaciones, son muy interesantes y han sido estudiadas en la literatura como parte de la teoría de grafos y ha sido extendido a la minería de redes sociales, para la detección de comunidades o de personas influyentes dentro de la red social [43]. Dado que nuestro trabajo versa sobre la minería de opiniones, no es menester entrar en más detalle en las relaciones entre usuarios dentro de twitter (*retweets*, *follows*). Por otro lado, sí que es necesario dado nuestro problema, diseccionar las partes que componen un tuit para ver su estrecha relación con los trabajo de minería de opiniones y la importancia de estos datos en este ámbito de la minería de datos.

3.5.2. Anatomía de un tuit

En la figura 3.1, podemos ver el ejemplo de un tuit real. Este puede tener variantes como enlaces o imágenes, pero esencialmente es texto y algunos *hashtags* o etiquetas que sirven para acotar u opinar sobre temas en concreto.

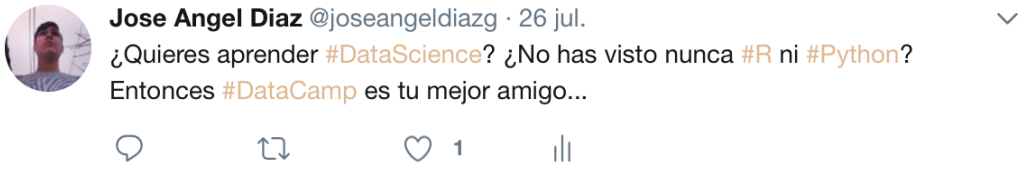

Figura 3.1: Ejemplo de un tuit.

Atendiendo al ejemplo anterior (figura 3.1), estaríamos hablando sobre *data science*, R, Pythton y el portal Data Camp. Cabe destacar, que aunque en este caso todos van precedidos del carácter # esto no tendría porque ser así, y alguna de estas palabras podría aparecer sin éste.

Tracemos ahora por tanto un pequeño paralelismo entre el tuit del ejemplo anterior y la definición dada por Liu de qué es una opinión que hemos podido ver en en la sección 3.2.

1. **Entidad**: En el caso de un tuit, la entidad sería sobre lo que se opina. En el caso del ejemplo anterior, sería #DataCamp.
2. **Emisor**: El paralelismo es obvio, el emisor es en el caso de un tuit la persona que lo tuitea en este caso, el usuario joseangeldiazg.
3. **Aspecto**: Recoge lo que se valora, y aunque puede parecer abstracto del anterior tuit podemos deducir que se valora la capacidad de un portal en internet para formar a las personas sobre conceptos tales como *data science*, R o Python.
4. **Orientación**: En el caso que nos ocupa, es positiva.
5. **Momento temporal**: Este elemento de la quintupla definida por Liu, al igual que el emisor siempre está presente dentro de un tuit y en este caso corresponde con el 26 de julio de 2017.

Es por tanto evidente, la relación entre un tuit y una opinión, poniendo al descubierto la importancia de este tipo de datos en el proceso y estudio de la minería de opiniones. Por otro lado, también es menester remarcar que no todos los tuits podrían enmarcarse dentro de la definición de opinión, como por ejemplo el que podemos ver en al figura 3.2 que simplemente es informativo.

3.5.3. Twitter API

Twitter abrió sus datos al mundo al hacer disponible una serie de APIs mediante las cuales se permite a terceros tanto la obtención de estos datos para su estudio como la implementación de software que trabaje sobre estos datos. Estas APIs necesitan del protocolo de seguridad y autenticación

Figura 3.2: Ejemplo de un tuit que no sería enmarcado como opinión.

OAuth, además ofrecen ciertas limitaciones a la hora de obtener los datos por lo que solo se permiten entre 150 y 300 solicitudes por hora y además hay una ventana temporal que cerrará el flujo de información cada 15 minutos. Las APIs disponibles son:

- **Search API**: Obtiene los tuits de hasta 7 días, es similar a lo que nos ofrecería la búsqueda básica de Twitter en la interfaz web al buscar por un término.
- **Streaming API**: Obtiene información en tiempo real.
- **REST API**: Obtiene los datos mediante HTTP, el formato puede venir en XML, HTML, o JSON, la limitación aquí viene definida por el número de resultados devueltos por página que no puede ser superior a 3200 tweets.

Como podemos observar, estas limitaciones pueden suponer un gran problema a la hora de obtener una gran cantidad de datos para que nuestro trabajo tenga un cierto rigor y peso, por ello, en los puntos siguientes ahondaremos en el proceso seguido y tecnologías usadas para la obtención y almacenamiento de un gran volumen de datos a pesar de estas restricciones.

Capítulo 4

Estado del arte

Capítulo 5

Arquitectura del sistema

Capítulo 6

Minería de datos

Capítulo 7

Conclusiones

Bibliografía

[1] Moosavi, S.A. and Jalali, M. Community detection in online social networks using actions of users. 2014 *Iranian Conference on Intelligent Systems, ICIS*.

[2] K. Kwon, Y. Jeon, C. Cho, J. Seo, In-Jeong Chung, H. Park: Sentiment trend analysis in social web environments. *BigComp 2017*, 261-268

[3] M. Pilar Salas-Zárate, J. Medina-Moreira, K. Lagos-Ortiz, H. Luna-Aveiga, M. Ángel Rodríguez-García, R. Valencia-García: Sentiment Analysis on Tweets about Diabetes: An Aspect-Level Approach. *Comp. Math. Methods in Medicine* 2017.

[4] Serrano-Cobos, Jorge. Big data y analítica web. Estudiar las corrientes y pescar en un océano de datos. *El profesional de la información*, 2014, vol. 23, n. 6, pp. 561-565.

[5] E. W. T. Ngai, Yong Hu, Y. H. Wong, Yijun Chen, and Xin Sun. 2011. The application of data mining techniques in financial fraud detection: A classification framework and an academic review of literature. Decis. Support Syst. 50, 3 (February 2011), 559-569.

[6] Pritam Gundecha, Huan Liu. Mining Social Media: A Brief Introduction. Arizona State University, Tempe, Arizona.

[7] B Liu, L Zhang . A survey of opinion mining and sentiment analysis. *Mining text data*, 2012. Springer.

[8] S. Noferesti, and M. Shamsfard. Resource Construction and Evaluation for Indirect Opinion Mining of Drug Reviews. *PLOS ONE*, 2015.

[9] Cambria E, Speer R, Havasi C, Hussain A. SenticNet: A publicly available semantic resource for opinion mining. *AAAI CSK*. 2010, 14-8.

[10] Baier D., Daniel I. Image Clustering for Marketing Purposes. In: Gaul W., Geyer-Schulz A., Schmidt-Thieme L., Kunze J. *Studies in Classification, Data Analysis, and Knowledge Organization*. Springer, Berlin, Heidelberg. 2012.

[11] Rakesh Agrawal, Tomasz Imieliski, and Arun Swami. Mining association rules between sets of items in large databases. *SIGMOD* Rec. 22, 1993, 207-216.

[12] P. Mandave, M. Mane, S. Patil. Data mining using Association rule based on APRIORI algorithm and improved approach with illustration. *International Journal of Latest Trends in Engineering and Technology (IJLTET)*, Vol. 3 Issue2 November 2013.

[13] Yong Yin, Ikou Kaku, Jiafu Tang, JianMing Zhu. Data Mining. Chapter 2, Association Rules Mining in Inventory Database (pp 9-23). Springer, 2011.

[14] R. Dehkharghani, H. Mercan, A. Javeed, Y. Saygin: Sentimental causal rule discovery from Twitter. *Expert Syst. Appl.* 41(10): 4950-4958 (2014).

[15] S. M. Weiss and N. Indurkhya. *Predictive data mining: a practical guide.* Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 1998.

[16] A. Petland. Reinventing society in the wake of big data. Edge.org, http://www.edge.org/conversation/reinventing-society-in-the-wake-of-big-data, 2012. Accedido el 1 de marzo de 2018.

[17] D. Laney. 3-D Data Management: Controlling Data Volume, Velocity and Variety. *META Group Research Note, February 6*, 2001.

[18] Han, J.W. and Kamber, M. (2001) Data Mining: Concepts and Techniques. Morgan Kaufmann Pulishers, Inc., San Francisco.

[19] Tan, P.N., Steinbach, M. and Kumar, V. (2006) Introduction to Data Mining. Pearson Education, Inc., London, 30-336.

[20] W. Seo, J. Yoon, H. Park, B. Coh, J. Lee, O. Kwon. Product opportunity identification based on internal capabilities using text mining and association rule mining. *Technological Forecasting & Social Change* 105 (2016) 94-104.

[21] M. Kaura, S. Kanga. Market Basket Analysis: Identify the changing trends of market data using association rule mining. International Conference on Computational Modeling and Security (CMS 2016). *Procedia Computer Science* 85 (2016) 78 - 85.

[22] K. Jayabal, Dr. P. Marikkannu. An Efficient Big Data processing for frequent itemset mining based on MapReduce Framework. *International Journal of Novel Research in Computer Science and Software Engineering* Vol. 3, Issue 1, pp: (130-134).

[23] Lin, Ming-Yen and Lee, Pei-Yu and Hsueh, Sue-Chen. Apriori-based Frequent Itemset Mining Algorithms on MapReduce. *ICUIMC* , 2012. pp(76:1-76:8).

[24] X. Zhou and Y. Huang. An improved parallel association rules algorithm based on MapReduce framework for big data. 11th International Conference on Fuzzy Systems and Knowledge Discovery (FSKD), Xiamen, 2014, pp. 284-288.

[25] Y. Chen, F. Li, J. Fan. Mining association rules in big data with NGEP. *Cluster Computin*, 2015, 18:2, 577-585.

[26] Dr. R Nedunchezhian and K Geethanandhini. Association Rule Mining on Big Data. International Journal of Engineering Research & Technology (IJERT). Volume 5 - Issue 05. (2015). 0

[27] M Adedoyin-Olowe, M Medhat Gaber, Frederic T. Stahl: A Survey of Data Mining Techniques for Social Media Analysis. *JDMDH* 2014.

[28] YZhou, N Sani, Chia-Kuei Lee, J Luo: Understanding Illicit Drug Use Behaviors by Mining Social Media. *CoRR* abs/1604.07096 (2016).

[29] L Cagliero and A Fiori. Analyzing Twitter User Behaviors and Topic Trends by Exploiting Dynamic Rules. Behavior Computing: Modeling, Analysis, Mining and Decision. Springer, 2012 pp. 267-287.

[30] L. Maria Aiello, G Petkos, Carlos J. Martín, D Corney, S Papadopoulos, R Skraba, A Göker, I Kompatsiaris, A Jaimes: Sensing Trending Topics in Twitter. *IEEE Trans.* Multimedia 15(6): 1268-1282 (2013).

[31] X Yu, S Miao, H Liu, Jenq-Neng Hwang, W Wan, J Lu: Association Rule Mining of Personal Hobbies in Social Networks. *Int. J. Web Service Res.* 14: 13-28 (2017).

[32] F Erlandsson, P Bródka, A Borg, H Johnson: Finding Influential Users in Social Media Using Association Rule Learning. Entropy 18: 164 (2016).

[33] A Pak, P Paroubek. Twitter as a Corpus for Sentiment Analysis and Opinion Mining. *Lrec.* 2010.

[34] Ana M. Popescu and O Etzioni. Extracting product features and opinions from reviews. *HLT '05 Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing* Pages 339-346. 2005.

[35] Hai Z., Chang K., Kim J. (2011) Implicit Feature Identification via Co-occurrence Association Rule Mining. In: Gelbukh A.F. (eds) *Computational Linguistics and Intelligent Text Processing. CICLing 2011.* Lecture Notes in Computer Science, vol 6608. Springer, Berlin, Heidelberg

[36] Yuan M., Ouyang Y., Xiong Z., Sheng H. (2013) Sentiment Classification of Web Review Using Association Rules. In: Ozok A.A., Zaphiris P. (eds) *Online Communities and Social Computing. OCSC 2013.* Lecture Notes in Computer Science, vol 8029. Springer, Berlin, Heidelberg

[37] Z Farzanyar, N Cercone: Efficient mining of frequent itemsets in social network data based on MapReduce framework. *ASONAM* 2013: 1183-1188.

[38] S. Gole and B. Tidke, Frequent itemset mining for Big Data in social media using ClustBigFIM algorithm. International Conference on Pervasive Computing (ICPC), Pune, 2015, pp. 1-6.

[39] S. Moens, E. Aksehirli, B. Goethals: Frequent Itemset Mining for Big Data. BigData Conference 2013: 111-118.

[40] J Yang and B Yecies. Open AccessMining Chinese social media UGC: a bigdata framework for analyzing Douban movie reviews, 2016, *Journal of Big Data*, vol 1.

[41] Abascal-Mena R., López-Ornelas E., Zepeda-Hernández J.S. User Generated Content: An Analysis of User Behavior by Mining Political Tweets. In: Ozok A.A., Zaphiris P. *Online Communities and Social Computing. OCSC* 2013. Lecture Notes in Computer Science, vol 8029. Springer, Berlin, Heidelberg

[42] Esuli, A., Sebastiani, F.: SENTIWORDNET: A Publicly Available Lexical Resource for Opinion Mining. *Proceedings of the 5th Conference on Language Resources and Evalua- tion*, LREC 2006, Genova, Italy, pp. 417-422 (2006)

[43] D. Ediger, K. Jiang, J. Riedy, D. A. Bader and C. Corley, "Massive Social Network Analysis: Mining Twitter for Social Good,"2010 39th International Conference on Parallel Processing, San Diego, CA, 2010, pp. 583-593.

[44] Web del proyecto MongoDB. https://www.mongodb.com. Accedido el 1 de marzo de 2018.

[45] Web del proyecto Tweepy. http://www.tweepy.org. Accedido el 1 de marzo de 2018.

[46] Web de Scrapy. https://scrapy.org. Accedido el 1 de marzo de 2018.

[47] Web de Scrapinghub. https://scrapinghub.com. Accedido el 1 de marzo de 2018.

[48] Matei Zaharia, Reynold S. Xin, Patrick Wendell, Tathagata Das, Michael Armbrust, Ankur Dave, Xiangrui Meng, Josh Rosen, Shivaram Venkataraman, Michael J. Franklin, Ali Ghodsi, Joseph Gonzalez, Scott Shenker, and Ion Stoica. 2016. Apache Spark: a unified engine for big data processing. Commun. ACM 59, 11 (October 2016), 56-65.

[49] I. Feinerer ,K. Hornik. (2017) Text Mining Package (Versión 0.7-3) [Software] Recuperado de https://cran.r-project.org/

[50] Jenny Rose Finkel, Trond Grenager, and Christopher Manning. 2005. Incorporating Non-local Information into Information Extraction Systems by Gibbs Sampling. *Proceedings of the 43nd Annual Meeting of the Association for Computational Linguistics* (ACL 2005), pp. 363-370.

[51] Manning, Christopher D., Mihai Surdeanu, John Bauer, Jenny Finkel, Steven J. Bethard, and David McClosky. 2014. The Stanford CoreNLP Natural Language Processing Toolkit In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pp. 55-60.

[52] Coupland, D. *Microserfs*. HarperCollins, Toronto, 1995.

[53] K. Hornik, C. Buchta, T. Hothorn, A. Karatzoglou, D. Meyer, A. Zeileis (2018) R/Weka Interface (Versión 3.9.2) [Software] Recuperado de https://cran.r-project.org/

[54] R. Agrawal and R. Srikant Fast algorithms for mining association rules in large databases. 1994. *Proceedings of the 20th International Conference on Very Large Data Bases*, VLDB, pp. 487-499.

[55] Han, J., Pei, H., Yin, Y.: Mining Frequent Patterns without Candidate Generation. 2000. *Proc. Conf. on the Management of Data* (SIGMOD 2000), Dallas, TX, pp. 1-12.