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

El mundo que nos rodea está en constante cambio y ha sido en las ultimas décadas cuanto este cambio ha sido notado y asimilado por la totalidad de los sectores económicos y sociales con una mayor fuerza. Este proceso ha sido influenciado sin duda alguna por la revolución de las *tecnologías de la información y la comunicación* y más recientemente de la inteligencia artificial.

Esta tecnología en conjunción con otras, ha propiciado el establecimiento de la conocida como *sociedad de la información*, denominación que se le da a una sociedad cambiante donde la manipulación de datos e información ha tomado un papel crucial en las actividades económicas, sociales y culturales. El tratamiento de estos datos para su posterior puesta en valor puede suponer una ardua labor, sobre todo cuando el volumen, naturaleza desestructurada de los mismos o necesidades de trabajo en tiempo real hacen que sea imposible su almacenamiento y procesado acorde a las técnicas habituales. Es aquí donde como una escisión o rama de la inteligencia artificial, nace el Big Data. El Big Data ha propiciado por tanto el nacimiento de aplicaciones basadas en el análisis de datos que antes eran simplemente imposibles de llevar a cabo por las limitaciones de la tecnología. Entre sus usos más habituales están aquellos sectores o ámbitos en los que el volumen de los datos son el principal escollo a salvar. Algunos ejemplos pueden ser el ámbito del internet de las cosas [Mourtzis et al.2016] donde se genera una gran cantidad de datos provenientes de sensores en apenas segundos, o las redes sociales [Fernandez-Basso et al.2019] donde la generación de contenido por parte de

los usuarios es tales dimensiones que las técnicas habituales de minería de datos no pueden abarcar.

Como hemos visto anteriormente, las redes sociales son una de las grandes factorías de datos actuales. La cantidad de datos generada actualmente por los usuarios de las mismas hacen que un procesado eficiente y útil de las mismas tenga que enmarcarse en soluciones de Big Data. Estas soluciones 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 de obtener valor de datos desestructurados y de gran volumen provenientes de redes sociales se denomina*social media mining*. Esta técnica está enmarcada, dependiendo de las necesidades de información, dentro de las técnicas de minería de textos, procesamiento del lenguaje natural o análisis de grafos y es una de las áreas de máximo apogeo actual entre los investigadores del ámbito de la inteligencia artificial.

En el presente proyecto, tratamos de aportar al ámbito de la minería de medios sociales, una metodología y un sistema final de análisis de tendencias u opiniones en un entorno de Big Data. El sistema será capaz de obtener de manera eficiente valor e información relevante de una gran cantidad de datos proveniente de plataformas de microblogging, como por ejemplo Twitter. En la siguiente sección veremos una breve motivación del proyecto para continuar con el enumerado de los objetivos principales que el proyecto cubre. El capitulo finaliza detallando la organización de la presente memoria.

1.1. Motivación

La totalidad de las actividades económicas y sociales del mundo actual se basan la posibilidad de acceder a golpe de click a ingentes bases de datos de información. Esto hace que los sistemas automáticos de procesado de datos usados para obtener, procesar y mostrar esta información tomen cada vez un papel más relevante en nuestro día a día. Como vimos en la introducción estas soluciones se enmarcan en multitud de sectores y herramientas como por ejemplo, las enfocadas al marketing [Serrano-Cobos2014] en pequeñas y grandes compañías, a la elaboración de modelos predictivos en ámbitos financieros o de seguros [Ngai et al.2011], y dada la reciente incursión de

las redes sociales en los paradigmas sociales, se ha propiciado la aparición de un gran número de soluciones enfocadas a la minería de redes sociales [Kwon et al.2017] [Salas-Zárate et al.2017].

El ámbito de la minería de medios sociales, se ha convertido recientemente en una las vertientes más estudiadas, aplicadas e investigadas dentro de la tradicional minería de datos. Esto es así, debido a que tanto el uso en alza que estas tienen, como el elevado nivel de implantación en la sociedad de las mismas, ha acrecentado la necesidad de sistemas y soluciones, capaces de extraer valor de estas cantidades ingentes de contenido. Es en esta necesidad de obtener valor de los datos desestructurados provenientes de redes sociales, donde surge el análisis de tendencias o minería de opiniones. La minería de opiniones, trata en última instancia de comprender o analizar comportamientos, actividades y opiniones, por ejemplo, de consumidores de cierto producto o usuarios de cierta red social. Si atendemos a su finalidad, esta radica en la obtención de conocimiento útil proveniente de las redes sociales y que pueda traducirse en ventajas competitivas en el proceso de toma de decisiones de una pequeña o gran compañía.

Dada la naturaleza, volátil y masiva de las redes sociales, las técnicas y paradigmas habituales de extracción de conocimiento se ven superadas. Es en esta encrucijada donde la minería de medios sociales se encuentra con el Big Data, siendo las soluciones enmarcadas dentro del mismo las únicas aplicables en innumerables ocasiones, sobre todo en aquellas en las que el volumen de datos o la necesidad de un procesado en tiempo real suponen un impedimento para los algoritmos y tecnologías de minería de datos tradicionales. La aplicación del Big Data a las redes sociales es bastante novedosa y es que no debemos perder de vista que las redes sociales mas famosas apenas tienen unos 13 años de vida, por lo que estaríamos hablando de una de las vías de investigación más novedosas del momento. Esto es así, tanto si ponemos el foco en la novedad que aportan las redes sociales, como si lo ponemos sobre la novedad de los aspectos puramente informáticos, siendo el Big Data 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.

Debido esta novedad y necesidad de soluciones de minería de opinión aplicadas a grandes cantidades de datos de manera eficiente donde surge este proyecto, cuya finalidad y objetivos veremos en la próxima sección.

1.2. Objetivos del proyecto

El presente proyecto fin de máster podría encuadrarse en un objetivo principal que a su vez quedaría definido por un conjunto de objetivos secundarios. El objetivo principal del proyecto sería por tanto, el estudio, desarrollo y validación de un sistema capaz de minar opiniones sobre plataformas de microblogging en un entorno de Big Data, permitiendo visualizar los resultados de una manera amigable y útil para el usuario final (investigador), aportando valor a las posibles preguntas de investigación que este pudiera formular. Este objetivo final, puede definirse como hemos visto antes en función de los siguientes objetivos secundarios:

- Estudio del estado del arte en el campo de la minería de opinión basada en Big Data en plataformas de microblogging .
- Obtención y salvado de un corpus de datasets de gran tamaño que permitan elaborar la experimentación y validación del sistema final.
- Desarrollo y aplicación de una metodología de preprocesado de datos eficaz para conjuntos de textos provenientes de plataformas de microblogging.
- Aplicación de técnicas de minería de datos descriptiva para obtener patrones interesantes en los datos.
- Pruebas y experimentación.
- Puesta en valor del sistema mediante técnicas de visualización dinámicas e interactivas basadas en web.
- Análisis de resultados y comparación de los mismos con posibles eventos políticos y sociales.

1.3. Organización de la memoria

Capítulo 2

Planificación del proyecto

Planificar un proyecto informático de manera correcta ser el factor crucial que determine el éxito, o en su defecto el fracaso del mismo. La necesidad de una correcta planificación se acentúa más en proyectos enmarcados dentro de ámbito del Big Data. Estos proyectos son verdaderas obras de ingeniería teniendo en cuenta la necesidad de equipos, recursos humanos, software, variables (tiempo de computo, memoria, costes) que pueden implicar. Es por ello, que en este capítulo haremos un pequeño resumen de la planificación del proyecto, aportando una visión general de los recursos implicados.

2.1. Gestión de recursos

En la primera sección de este capítulo, se hará un repaso por los principales recursos implicados pudiendo estos ser categorizados como personal, hardware y software, los cuales son a su vez los tres pilares clave de un proyecto de tecnologías de la información y la comunicación.

2.1.1. Personal

El personal a cargo del proyecto, consta principalmente del autor José Ángel Díaz García, encargado de desarrollar todas las partes del mismo mediante la supervisión de los tutores. Por otro lado se ha contado con cierta

asesoría y ayuda de miembros del equipo de investigación de bases de datos y sistemas de información inteligentes.

2.1.2. Hardware

Para elaboración de memorias, notas, artículos así como para los procesos de datos menos complejos se ha utilizado el sistema descrito en la tabla 2.1. Por otro lado, todo el proceso basado en Big Data se ha llevado a cabo en el cluster de procesado de datos del grupo de investigación, este cluster está formado por 4 maquinas cuyas especificaciones pueden verse en la tabla refcluster.

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.

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.

- **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**: Pasarela entre R y Spark que permite usar funciones de Spark de manera nativa en R.
- **Spark**: Entorno de computación en cluster usado para elaborar los procesos de Big Data más complejos.
- **Git**: Sistema utilizado para el control de versiones.

2.2. Planificación temporal

En este punto estudiaremos la planificación temporal seguida, así como los pequeños hitos que en cada una de las etapas se fueron consiguiendo y la duración de las mismas.

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	1mes	120

Tabla 2.3: Detalle de costes inventariables.

Si atendemos a los costes fungibles, aquellos cuyo inventariado es menos relevante como por ejemplo el material de oficina podríamos estimarlo en unos 100 euros.

Resumen de gastos

En función a los cálculos estimados de costes realizados en esta sección, en la tabla 2.4 podemos ver el total de los gastos que podrían derivarse del proyecto.

Gastos elegibles	Total
Personal	6000
Costes inventariables	201,25
Costes fungibles	100
TOTAL	6301,25

Tabla 2.4: Resumen final de costes.

Bibliografía

[Fernandez-Basso et al.2019] Fernandez-Basso, C., Francisco-Agra, A. J., Martin-Bautista, M. J., and Ruiz, M. D. (2019). Finding tendencies in streaming data using big data frequent itemset mining. *Knowledge-Based Systems*, 163:666–674.

[Kwon et al.2017] Kwon, K., Jeon, Y., Cho, C., Seo, J., Chung, I.-J., and Park, H. (2017). Sentiment trend analysis in social web environments. In *2017 IEEE International Conference on Big Data and Smart Computing (BigComp)*, pages 261–268. IEEE.

[Mourtzis et al.2016] Mourtzis, D., Vlachou, E., and Milas, N. (2016). Industrial big data as a result of iot adoption in manufacturing. *Procedia cirp*, 55:290–295.

[Ngai et al.2011] Ngai, E. W., Hu, Y., Wong, Y. H., Chen, Y., and Sun, X. (2011). The application of data mining techniques in financial fraud detection: A classification framework and an academic review of literature. *Decision support systems*, 50(3):559–569.

[Salas-Zárate et al.2017] Salas-Zárate, M. d. P., Medina-Moreira, J., Lagos-Ortiz, K., Luna-Aveiga, H., Rodriguez-Garcia, M. A., and Valencia-Garcia, R. (2017). Sentiment analysis on tweets about diabetes: an aspect-level approach. *Computational and mathematical methods in medicine*, 2017.

[Serrano-Cobos2014] Serrano-Cobos, J. (2014). Big data y analítica web. estudiar las corrientes y pescar en un océano de datos. *El profesional de la información*, 23(6):561–565.