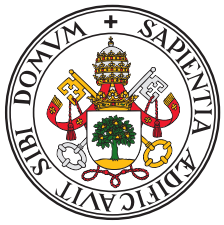


Universidad de Valladolid

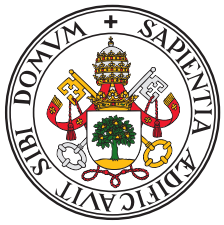
E.T.S Ingeniería Informática

Trabajo Fin de Grado

Grado en Ingeniería Informática,
mención en Computación

Algoritmos para Big Data

Autor:
Sergio García Prado



Universidad de Valladolid

E.T.S Ingeniería Informática

Trabajo Fin de Grado

Grado en Ingeniería Informática,
mención en Computación

Algoritmos para Big Data

Autor:
Sergio García Prado

Tutor:
Manuel Barrio Solórzano

Prefacio

Para entender el contenido de este documento así como la metodología seguida para su elaboración, se han de tener en cuenta diversos factores, entre los que se encuentran el contexto académico en que ha sido redactado, así como el tecnológico y social. Es por ello que a continuación se expone una breve descripción acerca de los mismo, para tratar de facilitar la compresión sobre el alcance de este texto.

Lo primero que se debe tener en cuenta es el contexto académico en que se ha llevado a cabo. Este documento se ha redactado para la asignatura de **Trabajo de Fin de Grado (mención en Computación)** para el *Grado de Ingeniería Informática*, impartido en la *E.T.S de Ingeniería Informática* de la *Universidad de Valladolid*. Dicha asignatura se caracteriza por ser necesaria la superación del resto de las asignaturas que componen los estudios del grado para su evaluacion. Su carga de trabajo es de **12 créditos ECTS**, cuyo equivalente temporal es de *300 horas* de trabajo del alumno, que se han llevado a cabo en un periodo de 4 meses.

La temática escogida para realizar dicho trabajo es **Algoritmos para Big Data**. El Big Data es la disciplina que se encarga de "todas las actividades relacionadas con los sistemas que manipulan grandes conjuntos de datos. Las dificultades más habituales vinculadas a la gestión de estas cantidades de datos se centran en la recolección y el almacenamiento, búsqueda, compartición, análisis, y visualización. La tendencia a manipular enormes cantidades de datos se debe a la necesidad en muchos casos de incluir dicha información para la creación de informes estadísticos y modelos predictivos utilizados en diversas materias."[Wik]

Uno de los puntos más importantes para entender la motivación por la cual se ha escogido dicha temática es el contexto tecnológico en que nos encontramos. Debido a la importante evolución que están sufriendo otras disciplinas dentro del mundo de la informática y las nuevas tecnologías, cada vez es más sencillo y económico recoger gran cantidad de información de cualquier proceso que se dé en la vida real. Esto se debe a una gran cantidad de factores, entre los que se destacan los siguientes:

- **Reducción de costes derivados de la recolección de información**: Debido a la constante evolución tecnológica cada vez es más barato disponer de mecanismos (tanto a nivel de hardware como de software), a partir de los cuales se puede recabar datos sobre un determinado suceso.

- **Mayor capacidad de cómputo y almacenamiento**: La recolección y manipulación de grandes cantidades de datos que se recogen a partir de sensores u otros métodos requieren por tanto del apoyo de altas capacidades de cómputo y almacenamiento. Las tendencias actuales se están apoyando en técnicas de virtualización que permiten gestionar sistemas de gran tamaño ubicados en distintas zonas geográficas como una unidad, lo cual proporciona grandes ventajas en cuanto a reducción de complejidad algorítmica a nivel de aplicación.

- **Mejora de las telecomunicaciones**: Uno de los factores que ha permitido una gran disminución de la problemática relacionada con la virtualización y su capacidad de respuesta ha sido el gran avance a nivel global que han sufrido las telecomunicaciones en los últimos años, permitiendo disminuir las barreras geográficas entre sistemas tecnológicos dispersos.

Gracias a este conjunto de mejoras se ha llegado al punto en que existe la oporturnidad de poder utilizar una gran cantidad de conocimiento, que individualmente o sin un apropiado procesamiento, carece de valor a nivel de información.

El tercer factor que es necesario tener en cuenta es la tendencia social actual, que cada vez más, está concienciada con el valor que tiene la información. Esto se ve reflejado en un amplio abanico de aspectos relacionados con el comportamiento de la población:

- **Monitorización de procesos laborales**: Muchas empresas están teniendo en cuenta la mejora de la productividad de sus empleados y máquinas. Por tanto, buscan nuevas técnicas que les permitan llevar a cabo dicha tarea. En los últimos años se ha dedicado mucho esfuerzo en implementar sistemas de monitorización que permitan obtener información para después procesarla y obtener resultados valiosos para dichas organizaciones.
- **Crecimiento exponencial de las plataformas de redes sociales**: La inherente naturaleza social del ser humano hace necesaria la expresión pública de sus sentimientos y acciones, lo cual, en el mundo de la tecnología se ha visto reflejado en un gran crecimiento de las plataformas de compartición de información así como de las de comunicación.
- **Iniciativas de datos abiertos por parte de las administraciones públicas**: Muchas insitituciones públicas están dedicando grandes esfuerzos en hacer visible la información que poseen, lo que conlleva una mejora social aumentando el grado de transparencia de las mismas, así como el nivel de conocimiento colectivo, que puede ser beneficioso tampo para ciudadanos como para empresas.

Como consecuencia de este cambio social, posiblemente propiciado por el avance tecnológico anteriormente citado, la población tiene un mayor grado de curiosidad por aspectos que antes no tenia la capacidad de entender, debido al nivel de complejidad derivado del tamaño de los conjuntos de muestra necesarios para obtener resultados fiables.

En este documento no se pretenden abordar temas relacionados con las técnicas utilizadas para recabar nuevos datos a partir de los ya existentes. A pesar de ello se realizará una breve introducción sobre dicho conjunto de estrategias, entre las que se encuentran: *Heurísticas*, *Regresión Lineal*, *Árboles de decisión*, *Máquinas de Vector Soporte (SVM)* o *Redes Neuronales Artificiales*.

Por contra, se pretende realizar un análisis acerca de los diferentes algoritmos necesarios para manejar dichas cantidades ingentes de información, en especial de su manipulación a nivel de operaciones básicas, como operaciones aritméticas, búsqueda o tratamiento de campos ausentes. Para ello, se tratará de acometer dicha problemática teniendo en cuenta estrategias de paralelización, que permitan aprovechar en mayor medida las capacidades de cómputo existentes en la actualidad.

Otro de los aspectos importantes en que se quiere orientar este trabajo es el factor dinámico necesario para entender la información, lo cual conlleva la búsqueda de nuevas estrategias algorítmicas de procesamiento en tiempo real. Por lo tanto, se pretende ilustrar un análisis acerca de las soluciones existentes en cada caso con respecto a la solución estática indicando las ventajas e inconvenientes de la versión dinámica según corresponda.

Índice general

Capítulo 1

Introducción

[TODO]

Capítulo 2

Algoritmos para Streaming

[TODO]

Capítulo 3

Estructuras de Datos de Resumen

3.1. Introducción

[TODO]

3.2. Tipos de Estructuras de Datos de Resumen

[TODO]

3.3. Sketching

[TODO]

3.4. Count-Min Sketch

[TODO]

3.5. Count Sketch

[TODO]

3.6. AMS Sketch

[TODO]

3.7. HyperLogLog

[TODO]

Capítulo 4

Algoritmos para Grafos

4.1. Introducción

[TODO]

4.2. Modelo en Semi-Streaming

[TODO]

Capítulo 5

Reducción de la Dimensionalidad

5.1. Introducción

[TODO]

5.2. Teorema de Johnson-Lindenstrauss

[TODO]

5.3. Busqueda de Vecinos más Cercanos

[TODO]

Capítulo 6

Técnicas de Minería de Datos

6.1. Introducción

[TODO]

6.2. Aprendizaje Supervisado y No Supervisado

[TODO]

6.3. Árboles de Decisión

[TODO]

6.4. Regresión Lineal

[TODO]

6.5. Redes Neuronales

[TODO]

6.6. Manifold Learning

[TODO]

Capítulo 7

Paralelización a Gran Escala

7.1. Introducción

[TODO]

7.2. Sistemas de Ficheros Distribuidos

[TODO]

7.3. Modelo de acceso a Memoria

[TODO]

7.4. Complejidad de la Comunicación

[TODO]

7.5. MapReduce

[TODO]

Apéndice A

Metodología de Trabajo

Apéndice B

¿Cómo ha sido generado este documento?

En este apéndice se describen tanto la estructura como las tecnologías utilizadas para redactar este documento. El estilo visual que se ha aplicado al documento se ha tratado de almoldar lo máximo posible a las especificaciones suministradas en la *guía docente* de la asignatura *Trabajo de Fin de Grado* [uva17].

Este documento ha sido redactado utilizando la herramienta de generación de documentos LaTeX[toog], en concreto se ha utilizado la distribución para sistemas *OS X* denominada *MacTeX* [tooh] desarrollada por la organización *TEX User Group*. Mediante esta estrategia todas las labores de compilación y generación de documentos *PDF* (tal y como se especifica en la guía docente) se realizan de manera local. Se ha preferido esta alternativa frente a otras como la utilización de plataformas online de redacción de documentos LaTeX como *ShareLateX* [tooj] u *Overleaf* [tooi] por razones de flexibilidad permitiendo trabajar en lugares en que la conexión a internet no esté disponible. Sin embargo, dichos servicios ofrecen son una buena alternativa para redactar documentos sin tener que preocuparse por todos aquellos aspectos referidos con la instalación de la distribución u otros aspectos como un editor de texto. Además garantizan un alto grado de confiabilidad respecto de pérdidas inesperadas.

Junto con la distribución LaTeX se han utilizado una gran cantidad de paquetes que extienden y simplifican el proceso de redactar documentos. Sin embargo, debido al tamaño de la lista de paquetes, esta será obviada en este apartado, pero puede ser consultada visualizando el correspondiente fichero `thestyle.sty` del documento.

Puesto que la alternativa escogida ha sido la de generar el documento mediante herramientas locales es necesario utilizar un editor de texto así como un visualizador de resultados. En este caso se ha utilizado *Atom* [tooa], un editor de texto de propósito general que destaca sobre el resto por ser desarrollado mediante licencia de software libre (*MIT License*) y estar mantenido por una amplia comunidad de desarrolladores además de una extensa cantidad de paquetes con los cuales se puede extender su funcionalidad. En este caso, para adaptar el comportamiento de *Atom* a las necesidades de escritura de texto con latex se han utilizados los siguientes paquetes: *latex* [tooc], *language-latex* [toob], *pdf-view* [tood] encargados de añadir la capacidad de compilar ficheros latex, añadir la sintaxis y permitir visualizar los resultados respectivamente.

Puesto que el *Trabajo de Fin de Grado* se refiere a algo que requiere de un periodo de tiempo de elaboración largo, que además sufrirá una gran cantidad de cambios, se ha creido conveniente la utilización de una herramienta de control de versiones que permita realizar un seguimiento de los cambios de manera organizada. Para ello se ha utilizado la tecnología *Git* [tooe] desarrollada originalmente por *Linus Torvalds*. En este caso en lugar de confiar en el entorno local u otro servidor propio se ha preferido utilizar la plataforma *GitHub* [toof], la cual ofrece un alto grado de confiabilidad respecto de posibles perdidas además de alojar un gran número de proyectos de software libre. A pesar de ofrecer licencias para estudiantes que permiten mantener el repositorio oculto al público, no se ha creido necesario en este caso, por lo cual se puede acceder al través de la siguiente url: `https://github.com/garciparedes/tfg-big-data-algorithms` [GP17]

```
.
+-- document
|   +-- bib
|   |   +-- article.bib
|   |   +-- ...
|   +-- img
|   |   +-- logo_uva.eps
|   |   +-- ...
|   +-- tex
|   |   +-- appendices
|   |   |   +-- how_it_was_build.tex
|   |   |   +-- ...
|   |   +-- chapters
|   |   |   +-- introduction.tex
|   |   |   +-- ...
|   |   +-- bibliography.tex
|   |   +-- ...
|   +-- document.tex
|   +-- ...
+-- notes
|   +-- readme.md
|   +-- ...
+-- summary
|   +-- summary.tex
|   +-- ...
+-- README.md
+-- ...
```

Figura B.1: Árbol de directorios del repositorio

Una vez descritas las distintas tecnologías y herramientas utilizadas para la elaboración de este trabajo, lo siguiente es hablar sobre la organización de ficheros. Todos los ficheros utilizados para este documento (obviando las referencias bibliográficas) han sido incluidos en el repositorio indicado anteriormente [GP17].

Para el documento, principal alojado en el directorio `/document/` se ha seguido una estructura modular, dividiendo los capítulos, apéndices y partes destacadas como portada, bibliografía o prefacio entre otros en distintos ficheros, lo cual permite un acceso sencillo a los mismos. Los apéndices y capítulos se han añadido en los subdirectorios separados. Para la labor de combinar el conjunto de ficheros en un único documento se ha utilizado el paquete *subfiles*. El fichero raiz a partir del cual se compila el documento es `document.tex`. La importación de los distintos paquetes así como la adaptación del estulo del documento a los requisitos impuestos se ha realizado en `thestyle.sty` mientras que el conjunto de variables necesarias como el nombre de los autores, del trabajo, etc. se han incluido en `thevars.sty`.

En cuanto al documento de resumen, en el cual se presenta una vista panorámica acerca de las distintas disciplinas de estudio relacionadas con el *Big Data* se ha preferido mantener un único fichero debido a la corta longitud del mismo. Este se encuentra en el directorio `/summary/`.

Por último se ha decidido añadir otro directorio denominado `/notes/` en el cual se han añadido distintas ideas de manera informal, así como enlaces a distintos cursos, árticulos y sitios web en que se ha basado la base bibliográfica del trabajo. En la figura B.1 se muestra la estructura del repositorio en forma de árbol.

Bibliografía

[GP17] Sergio García Prado. Trabajo de Fin de Grado: Algoritmos para Big Data, 2017. https://github.com/garciparedes/tfg-big-data-algorithms.

[tooa] Atom. https://atom.io.

[toob] Atom Package: Language LaTeX. https://atom.io/packages/language-latex.

[tooc] Atom Package: LaTeX. https://atom.io/packages/latex.

[tood] Atom Package: PDF View. https://atom.io/packages/pdf-view.

[tooe] Git. https://www.git-scm.com.

[toof] GitHub. https://github.com.

[toog] LaTeX. https://www.latex-project.org.

[tooh] MacTex. http://www.tug.org/mactex.

[tooi] Overleaf. https://www.overleaf.com.

[tooj] ShareLaTeX. https://www.sharelatex.com.

[uva17] Trabajo de Fin de Grado (Mención en Computación), 2016/17. https://www.inf.uva.es/wp-content/uploads/2016/06/G46978.pdf.

[Wik] Wikipedia. Big Data. https://es.wikipedia.org/wiki/Big_data.