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."[8]

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

El gran crecimiento tecnológico que se está llevando a cabo en la actualidad a todos los niveles está propiciando además un aumento exponencial en cuanto a la cantidad de información que se genera. La reducción de costes en cuanto a la instalación de sensores que permiten recoger información de muchos procesos productivos, así como la obtenición de metadatos a partir del uso de internet y las redes sociales por parte de los usuarios hace que el ritmo de crecimiento en cuanto a información generada por unidad de tiempo haya crecido a un gran ritmo.

Una de las razones que han facilitado dicha tendencia es la disminución de costes de almacenamiento de información a la vez que las capacidades de cómputo necesarias para procesar dicha información han aumentado. Sin embargo, debido al crecimiento exponencial en cuanto al tamaño del conjunto de datos, es necesario investigar nuevas técnicas y estrategias que permitan obtener respuestas satisfactorias basadas en la gran cantidad de información de la que se dispone en un tiempo razonable.

Tradicionalmente, la investigación en el campo de las *bases de datos* se ha centrado en obtener respuestas exactas a distintas consultas, tratando de hacerlo de la manera más eficiente posible, así como de tratar de reducir el espacio necesario para almacenar la información. *Acharya y otros* proponen en el artículo *Join synopses for approximate query answering* [1] el concepto de *Approximate Query Processing*. Dicha idea se expone en la subsección 3.1.1.

3.1.1. Approximate Query Processing

El *procesamiento aproximado de consultas*, (*Approximate Query Processing* o **AQP**) se presenta como una estrategia de consulta basada en conceptos y propiedades estadísticas que permiten una gran reducción de la complejidad computacional y espacial necesaria para la resolución de consultas a una base de datos. Por contra, dicha reducción a nivel de complejidad tiene como consecuencia la inserción de un determinado nivel de imprecisión en el resultado a la cual denominaremos tasa de error. Se pretende que dicha tasa de error pueda ser acotada en una desviación máxima determinada por ϵ y se cumpla con un índice de probabilidad δ. Al igual que en capítulos anteriores, en este caso también se presta especial importancia en la minimización del error relativo lo cual consigue que las soluciones mediante el *procesamiento aproximado de consultas* sean válidas tanto para consultas de tamaño reducido como de gran tamaño.

Durante el resto del capítulo se describen y analizan distintas estrategias que permiten llevar a cabo implementaciones basadas en *procesamiento aproximado de consultas* centrando especial atención en los *Sketches* por su similitud con el *Modelo en Streaming* descrito en el capítulo 2. En la sección 3.2 se realiza una decripción a partir de la cual se pretende aclarar las diferencias entre las distintas *estructuras de datos de resumen*. Posteriormente, en la sección 3.3 se

explican en detalle las cualidades de las estrategias basadas en *Sketching*. En las secciones 3.4, 3.5, 3.6 y 3.7 se habla de *Count-Min Sketch*, *Count Sketch*, *AMS Sketch* e *HyperLogLog* respectivamente.

3.2. Tipos de Estructuras de Datos de Resumen

Para el diseño de soluciones basadas en *procesamiento aproximado de consultas* en bases de datos existen distintas estrategias, las cuales presentan distintas ventajas e inconvenientes tal y como se pretende mostrar en esta sección. Dichas descripciones han sido extraidas del libro *Synopses for massive data* [4] redactado por *Cormode y otros*. En las secciones 3.2.1, 3.2.2, 3.2.3 y 3.2.4 se habla de *Sampling*, *Histogram*, *Wavelet* y *Sketches* respectivamente.

3.2.1. Sampling

El *Sampling* o *muestreo* es la estrategia más consolidada entre las que se presentan. Las razones se deben a su simplicidad conceptual así como su extendido uso en el mundo de la estadística. Uno de los primeros artículos en que se trata el muestreo aplicado a bases de datos es *Accurate estimation of the number of tuples satisfying a condition* [7] redactado por *Piatetsky-Shapiro* y *Connell*. La intuición en que se basa dicha estrategia es la selección de un subconjunto de elementos denominado *muestra* de entre el conjunto global al cual se denomina *población*. Una vez obtenida la *muestra* del conjunto de datos global cuyo tamaño es significativamente menor (lo cual reduce drásticamente el coste computacional), se realizan los cálculos que se pretendía realizar sobre toda la *población*, a partir de los cuales se obtiene un estimador del valor real que habría sido obtenido al realizarlos sobre el conjunto de datos global.

Para que las estrategias de sumarización de información obtengan resultados válidos o significativos respecto del conjunto de datos, es necesario que se escojan adecuadamente las instancias de la *muestra*, de manera que represente de manera fiel la información global. Para llevar a cabo dicha labor existen distintas estrategias, desde las más simples basadas en la selección aleatoria sin reemplazamiento como otras mucho más sofisticadas basadas en el mantenimiento de *muestras* estratificadas. Sea R la población y $|R|$ el tamaño de la misma. Denominaremos t_j al valor j-ésimo de la población y X_j al número de ocurrencias del mismo en la *muestra*. A continuación se describen distintas técnicas de muestreo:

- **Selección Aleatoria Sin Reemplazamiento**: Consiste en la estrategia más simple de generación de *muestras*. Se basa en la selección aleatoria de un valor entero r en el rango $[1, |R|]$ para después añadir el elemento localizado en la posición r de la *población* al subconjunto de *muestra*. Después repetir dicha secuencia durante n veces para generar una *muestra* de tamaño n. El estimador para la operación *SUMA* se muestra en la ecuación (3.1) además de la desviación de dicho estimador en la ecuación (3.2).

$$Y = \frac{|R|}{n} \sum_j X_j t_j \tag{3.1}$$

$$\sigma^2(Y) = \frac{|R|^2 \sigma^2(R)}{n} \tag{3.2}$$

- **Selección Aleatoria Con Reemplazamiento**: En este caso se supone que la selección de una instancia de la población tan solo se puede llevar a cabo una única vez como mucho, por lo tanto se cumple que $\forall X_j \in 0, 1$. La selección se lleva a cabo de la siguiente manera: se genera de manera aleatoria un valor entero r en el rango $[1, |R|]$ para después añadir el elemento localizado en la posición r de la *población* al subconjunto de *muestra* si este no ha sido añadido ya, sino volver a generar otro valor r. Después repetir dicha secuencia durante n veces para generar una *muestra* de tamaño n. Al igual que en la estrategia anterior, en este caso también se muestra el estimador para la operación *SUMA* en la ecuación (3.3). Nótese que el cálculo es el mismo que en el caso de

la estrategia sin reemplazamiento. Sin embargo, la varianza obtenida a partir de dicha estrategia es menor tal y como se muestra en la ecuación (3.4).

$$Y = \frac{|R|}{n} \sum_j X_j t_j \tag{3.3}$$

$$\sigma^2(Y) = \frac{|R|(|R| - n)\sigma^2(R)}{n} \tag{3.4}$$

- **Bernoulli y Poisson**: Mediante esta alternativa de muestreo se sigue una estrategia completamente distinta a las anteriores. En lugar de seleccionar la siguiente instancia aleatoriamente de entre todas las posibles, se decide generar $|R|$ valores aleatorios r_j independientes en el intervalo $[0, 1]$ de tal manera que si r_j es menor que un valor p_j fijado a priori, la instancia se añade al conjunto de *muestra*. Cuando se cumple que $\forall i, j \ p_i = p_j$ se dice que es un muestreo de *Bernoulli*, mientras que cuando no se cumple dicha condición se habla de muestreo de *Poisson*. El cálculo de la *SUMA* en este caso es muy diferente de los anteriores tal y como se muestra en la ecuación (3.5). La desviación de este estimador se muestra en la ecuación (3.6), que en general presenta peores resultados (mayor desviación) que las anteriores alternativas, sin embargo, esta alternativa posee la cualidad de aplicar distintos pesos a cada instancia de la población, lo que puede traducirse en que una selección adecuada de dichos valores p_j puede mejorar significativamente dichos resultados.

$$Y = \sum_{i \in muestra} \frac{t_i}{p_i} \tag{3.5}$$

$$\sigma^2(Y) = \sum_i (\frac{1}{p_i} - 1) t_i^2 \tag{3.6}$$

- **Muestreo Estratificado**: El muestreo estratificado trata de minimizar al máximo las diferencias entre la distribución del conjunto de datos de la *población* de la *muestra* que se pretende generar. Para ello existen distintas alternativas entre las que se encuentra una selección que actualiza los pesos p_j tras cada iteracción, lo que reduce la desviación de la *muestra*, sin embargo produce un elevado coste computacional en su generación. Por lo tanto se proponen otras estrategia más intuitiva basada en la partición del conjunto de datos de la *población* en subconjuntos disjuntos con varianza mínima entre las instancias que contienen a los cuales se denomina *estratos*. Posteriormente se selecciona mediante cualquiera de los métodos anteriores una *muestra* de cada *estrato*, lo cual reduce en gran medida la desviación típica global del estimador.

La estrategia de sumarización de información mediante *muestreo* tiene como ventajas la independencia de la complejidad con respecto a la dimensionalidad de los datos (algo que como se verá a continuación no sucede con otras alternativas) además de su simplicidad conceptual. También existen cotas de error para las consultas, para las cuales no ofrece restricciones en cuanto al tipo (debido a que se realizan sobre un subconjunto con la misma estructura que el global). El muestre es apropiado para conocer información general acerca del conjunto de datos que cada instancia del mismo posee. Además, presenta la cualidad de permitir la modificación en tiempo real, es decir, se pueden añadir o eliminar nuevas instancias a la muestra conforme se añaden o eliminan del conjunto de datos global.

Sin embargo, en entornos donde el ratio de addiciones/eliminaciones es muy elevado el coste del mantenimiento de la muestra puede hacerse impracticable . El *muestreo* es una buena alternativa para conjuntos de datos homogéneos en los cuales la presencia de valores atípicos es irrelevante. Tampoco obtiene buenos resultados en consultas relacionadas con el conteo de elementos distintos. En las siguientes secciones se describen alternativas que resuelven estas dificultades y limitaciones.

3.2.2. Histogram

[TODO introducción a los histogramas]

[TODO Descripción de esquemas de estimación]

- **Esquema Uniforme**:[TODO]
- **Esquema Basado en Splines**:[TODO]
- **Esquema Basado en Árboles**:[TODO]
- **Esquema Heterogéneo**:[TODO]
- **Esquema Probabilista**:[TODO]

[TODO Estrategias de particionamiento]

- **Particionamiento Heurístico**:[TODO]
- **Particionamiento con Garantías de Optimalidad**:[TODO]
- **Particionamiento Jerárquico**:[TODO]

[TODO Hablar de multidimensionalidad]

[TODO Ventajas y desventajas]

3.2.3. Wavelet

[TODO]

3.2.4. Sketch

[TODO]

3.3. Sketching

[TODO]

3.4. Count-Min Sketch

[TODO] *An improved data stream summary: the count-min sketch and its applications* [5]

3.5. Count Sketch

[TODO] *Finding frequent items in data streams* [3]

3.6. AMS Sketch

[TODO] *The space complexity of approximating the frequency moments* [2]

3.7. HyperLogLog

[TODO] *Hyperloglog: the analysis of a near-optimal cardinality estimation algorithm* [6]

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?

Bibliografía

[1] Acharya, S., Gibbons, P. B., Poosala, V., and Ramaswamy, S. Join synopses for approximate query answering. In *ACM SIGMOD Record* (1999), vol. 28, ACM, pp. 275–286.

[2] Alon, N., Matias, Y., and Szegedy, M. The space complexity of approximating the frequency moments. In *Proceedings of the twenty-eighth annual ACM symposium on Theory of computing* (1996), ACM, pp. 20–29.

[3] Charikar, M., Chen, K., and Farach-Colton, M. Finding frequent items in data streams. In *International Colloquium on Automata, Languages, and Programming* (2002), Springer, pp. 693–703.

[4] Cormode, G., Garofalakis, M., Haas, P. J., and Jermaine, C. Synopses for massive data: Samples, histograms, wavelets, sketches. *Foundations and Trends in Databases 4*, 1–3 (2012), 1–294.

[5] Cormode, G., and Muthukrishnan, S. An improved data stream summary: the count-min sketch and its applications. *Journal of Algorithms 55*, 1 (2005), 58–75.

[6] Flajolet, P., Fusy, É., Gandouet, O., and Meunier, F. Hyperloglog: the analysis of a near-optimal cardinality estimation algorithm. In *Analysis of Algorithms 2007 (AofA07)* (2007), pp. 127–146.

[7] Piatetsky-Shapiro, G., and Connell, C. Accurate estimation of the number of tuples satisfying a condition. *ACM Sigmod Record 14*, 2 (1984), 256–276.

[8] Wikipedia. Big Data.