*Jorge Andrés Galindo - 679155 Javier Aranda García- 679184*

*AlgoritmiA para problemas difÍciles*

Práctica 1

**INTRODUCCIÓN**

En esta práctica se ha desarrollado un sistema para dividir los productos de una cierta empresa en dos lugares distintos (Java). La complejidad está en que hay que reducir el número de productos separados en las dos empresas que han sido comprados alguna vez juntos, ya que esto puede dar lugar a cancelar la compra de los mismos si están divididos. Para resolver el problema, se ha identificado la división de los productos como un problema de grafos, representando cada producto como un vértice, y siendo las aristas la unión de dos productos si se han comprado alguna vez juntos.

Teniendo ya identificado el problema, se han implementados los algoritmos de “Karger” y “Karger-Stein” que encuentran el mínimo corte en un grafo. El corte mínimo consiste en minimizar el número de aristas que conectan dos vértices de los dos cortes disjuntos generados. Para conseguirlos, aparte de implementar los dos algoritmos nombrados, se han utilizado las estructuras de datos más convenientes para reducir al máximo el coste del algoritmo, generando también numerosas pruebas para comprobar su correcto funcionamiento.

**PRODUCTOS**

En primer lugar, antes de generar el grafo del que hay que calcular el corte mínimo, se debía tener una lista de productos. Esta lista se generó de manera iterativa con productos de valores aleatorios para los campos que se creyó que los definían (nombre, precio, unidades, marca, descuento…). Para acceder a la información del producto, se necesitaba una estructura de datos rápida capaz de dar con coste constante la información a partir del nombre. Dicha estructura no era otra que una tabla hash que identificara cada elemento por una clave única dentro de la tabla.

Producto: nombre, unidades, precio, descuento y marca

Teniendo ya la estructura decidida, el siguiente paso era definir cuál sería la clave de cada producto. Se pensó que el más adecuado sería el nombre, ya que en teoría no debería haber dos productos con el mismo nombre, pero eso dificultaba mucho las cosas al trabajar luego con el grafo y dar nombres a los vértices. Por ello, se decidió utilizar el nombre como clave en la tabla hash, pero utilizando además otra tabla hash auxiliar, independiente a la anterior, que relacionara el nombre del producto con una clave numérica (el identificador de lo que sería posteriormente el vértice).

|  |  |  |  |
| --- | --- | --- | --- |
| NombreProd1 | Producto1 | Clave1 | NombreProd1 |
| NombreProd2 | Producto2 | Clave2 | NombreProd2 |
| NombreProd3 | Producto3 | Clave3 | NombreProd3 |
| … | … | … | … |
| NombreProdN | ProductoN | ClaveN | NombreProdN |

**EMPAREJAR PRODUCTOS**

Con el listado de productos almacenados en la tabla hash, el siguiente paso era generar los emparejamientos entre los productos. Esta información se supone que se daría al algoritmo por parte de “Amazon” para iniciar el corte en dos partes, y consistía en una tabla de booleanos que indicara en la posición i-j si el producto i había sido comprado alguna vez con el producto j. La tabla tenía que ser consistente; es decir, si una posición i-j de la tabla tenía un cierto valor, la posición referenciada por j-i debía tener exactamente el mismo valor.

Así pues, la tabla era bidimensional, marcando sus límites el número total de productos a dividir. La posición de la tabla contenía el valor true si el producto había sido comprado con otro dado, y tomaba el valor false en caso contrario. En los ejemplos de prueba aleatorios que se realizaron, la tabla se rellenó de manera aleatoria recorriendo tan sólo la mitad, ya que la otra mitad era exactamente igual y se podía rellenar a la vez, reduciendo así el coste total de generar emparejamientos aleatorios. El valor a introducir en la tabla se obtenía con un número aleatorio, si era par se ponía a true.

**GENERACIÓN GRAFO**

Con los datos de los emparejamientos en la tabla booleana, se procedió a generar el grafo. Viendo el algoritmo de Karger, se apreció que se necesitaba a parte de la tabla de booleanos, una tabla de enteros que indicara el número de aristas que unía un cierto vértice con otro. La tabla inicial tan sólo contendría valores de 1 ó 0 según hubiera arista o no entro los vértices, pero al juntar dos ciertos vértices se daba el caso de que quedaban varias aristas entre dos de ellos, información que era necesario almacenar.

Para representar un grafo, las dos estructuras de datos más comunes eran matriz de adyacencia o lista de adyacencia. La primera de ellas correspondería a la tabla de emparejamientos nombrada anteriormente, valiendo true si un cierto producto (vértice) ha sido comprado con otro (existe arista entre ambos). La segunda es una lista dinámica con punteros a una lista de vértices con los que está conectado uno dado. Dichas estructuras eran muy eficientes para representar el grafo, pero nosotros no decidimos utilizar ninguna de ellas.

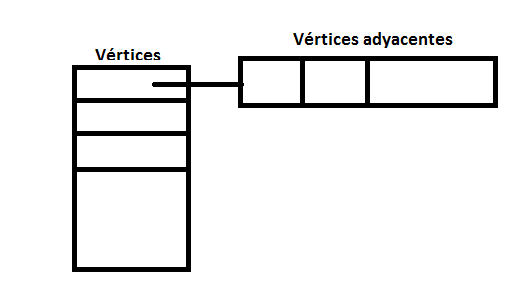
Si analizábamos el algoritmo a implementar, en cada iteración, se eliminaban dos vértices, creando uno nuevo con todas las aristas de los dos por separado. La estructura de matriz de adyacencia, además de tener un coste extra en espacio (se elimina un vértice en cada iteración sin reducir el coste en memoria), se necesitaba recorrerla entera actualizando el valor booleano de cada vértice con la correspondencia de la unión creada. Si mirábamos la lista de adyacencia, se reducía el problema del coste extra en espacio, pero nos encontrábamos ante el mismo problema de recorrer entera toda la estructura para saber si se encontraban los vértices modificados, y actualizar su información con la del nuevo vértice.

Viendo los problemas que generaban las estructuras de datos más comunes, se decidió crear una alternativa que reducía el coste de las operaciones comentadas anteriormente. Esta alternativa estaba compuesta por una tabla hash con todos los vértices que contenía el grafo. Para representar las aristas, se añadió una nueva tabla hash en cada posición de la anterior tabla nombrada. Esta segunda tabla contendría los vértices con los que el vértice de la tabla principal está conectado. La clave de cada índice de la tabla vendría dada por un número que representa el vértice, y cuya correspondencia con el producto se ha comentado en el apartado anterior. Para la segunda tabla hash de aristas, la clave era el vértice con el que estaba conectado seguido de “\_X”, siendo “X” el número de arista a ese vértice.

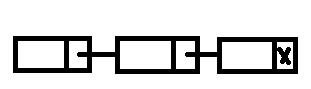
Con esta estructura, añadir un nuevo vértice con la información de los dos unidos tiene coste constante, igual que eliminar los otros dos. Juntar los vértices con los que tienen aristas los dos unidos tendría coste lineal en el número total de aristas de ambos vértices, algo que no mejora en nada sobre las dos estructuras anteriores. Donde está la verdadera mejora de esta estructura de datos es en la operación de actualizar el resto de vértices con la conexión del nuevo. Para realizarlo, la única manera era recorrer la tabla hash principal recorriendo todos los vértices, pero esta vez, no es necesario recorrer el resto de vértices para ver si alguno que era adyacente era alguno de los borrados.

En la matriz y lista de adyacencia, para actualizar la información en todos los vértices del grafo, era necesario recorrer la información del grafo entera; es decir, para cada vértice, mirar si tiene conexión con alguno de los dos borrados, y actualizarla. Para mirar esa conexión, partiendo de un vértice, se debían recorrer todos los demás para ver su conexión con el resto, y realizar el mismo procedimiento para todos. Por todo ello, esta operación tenía coste O(n2), siendo “n” el número de vértices.

Ahora, al estar esta información reflejada en una tabla hash, simplemente accediendo con coste constante a la posición en la que deberían estar los dos vértices borrados (si estuvieran), y modificándolos con la nueva información también con coste constante, se obtiene el grafo actualizado. Sabiendo el coste cuadrático de la operación en las anteriores estructuras, en la nueva que habíamos creado tan sólo tenía coste lineal en el número de vértices, ya que ahora no hace falta recorrerlos todos para ver si hay conexión con los borrados.



A parte de la tabla hash para los vértices adyacentes comentada, el nodo almacenado en la tabla principal tenía que tener también alguna estructura que guardara los vértices (o productos) que se habían unido para generar ese nuevo vértice. Como para este proceso no se necesitaba ningún tipo de eficiencia especial, ya que el coste de todos modos iba a ser lineal en la suma de los vértices unidos, se decidió utilizar una lista dinámica de los identificadores de los vértices ya comentados, haciéndola al menos de este modo eficiente en espacio.



En conclusión, el grafo es una estructura de datos de tipo tabla hash, en la que cada elemento contiene dos estructuras de datos adicionales. La primera de ellas otra tabla hash con los vértices adyacentes para facilitar y reducir el coste de las operaciones de actualización. La segunda una lista dinámica de los identificadores de los vértices que se han unido para generar dicho vértice del grafo. Con todo ello, ya se tiene el grafo final con los productos listo para aplicar el algoritmo de corte mínimo.

**ALGORITMO DE KARGER**

Como ya se ha expuesto en los anteriores apartados un poco para justificar la estructura de datos óptima para el problema, el algoritmo de Karger consiste en ir seleccionando sucesivamente dos vértices del grafo; es decir una arista, y colapsarlos o unirlos en uno sólo, incluyendo en éste las aristas que contenían los dos anteriores. Este proceso se repite un número de veces definido por el número de vértices del grafo hasta que tan sólo quedan dos vértices en total. Así pues, en esos dos vértices se han colapsado los demás que conformaban el grafo y han formado los dos conjuntos del corte realizado.

Este algoritmo no encuentra en todas las ocasiones el corte mínimo para el grafo, pero es una buena aproximación, aunque no acierte. Además, su probabilidad de acierto aumentaría considerablemente utilizando algunas técnicas como podría ser repetir el algoritmo un número alto de veces (aumentaría también el coste en tiempo), pero consiguiendo un acierto con alta probabilidad. La probabilidad de éxito del algoritmo se explica detalladamente y con cálculos matemáticos a continuación.

Todo grafo “G(V,E)” de n=|V| vértices tiene 2n-1 – 1 posibles cortes disjuntos, ya que se omiten dos cortes que harían alguno de los conjuntos finales vacíos (caso que no es posible), y si se intercambiarán los dos conjuntos también se obtendría el mismo resultado por lo que cada corte se está contando dos veces. El problema del corte mínimo es encontrar de entre todos estos posibles cortes disjuntos el mínimo de ellos (mínimo número de aristas entre vértices de los dos conjuntos generados).

Si definimos el peso del corte generado en el grafo matemáticamente se expresaría de la siguiente manera (el peso de todas nuestras aristas es 1):

Ahora, del total de cortes expresado anteriormente, al menos son cortes mínimos del grafo dado. Si denotamos C como las aristas de un coste específico de grado “k”, el algoritmo devuelve C si ninguna de las aristas elegidas aleatoriamente pertenece a C, lo que ocurre con una probabilidad de 1/k/|E|. El mínimo grado de G es al menos “k”, ya que obviamente el grado del grafo inicial es igual o superior al grado del mínimo corte encontrado, así que el número total de aristas es |E| = nk/2.

Dado los datos anteriores, ya se puede definir la probabilidad de que el algoritmo escoja una de las aristas que pertenecen a C, que viene dada por:

La probabilidad de que un cierto grafo con “n” vértices llegue a un corte mínimo C de grado “k” satisface una ecuación en recurrencias que viene dada por cada iteración del algoritmo juntando dos vértices: pn pn-1. Esta probabilidad puede ser expandida a la siguiente expresión:

A modo de información, si se realizase la mejora comentada de repetir varias veces el algoritmo, como podría ser veces, eligiendo de manera aleatoria independiente las aristas, y devolviendo el mínimo corte de todos, la probabilidad de no encontrar el corte mínimo sería:

**ALGORITMO DE KARGER-STEIN**

El algoritmo de Karger-Stein consiste en una mejora del algoritmo de Karger, que permite encontrar con una mayor probabilidad el corte mínimo de un grafo. El algoritmo se basa en que hay un 50% de probabilidad de unir dos nodos en el corte mínimo al contraer “n” nodos hasta tener n/ nodos.

Teniendo en cuenta lo anteriormente explicado, se pueden realizar dos llamadas al algoritmo reduciendo hasta n/ nodos ya que al ser la probabilidad de acierto del 50%, seguramente una de las dos llamadas obtendrá el resultado correcto. Por lo tanto, se elegirá el mejor grafo de ambos y se seguirá ejecutando el algoritmo sobre dicho grafo.

Una vez que el número de nodos sea lo suficientemente pequeño, se podrá continuar realizando el algoritmo sin necesidad de realizar varias llamadas, ya que la respuesta se podrá encontrar aplicando simplemente fuerza bruta (el algoritmo de contracción que se usaba en la versión de Karger). Así pues, simplemente consiste en llamadas sucesivas al algoritmo de Karger hasta obtener un cierto número de nodos, quedándose con el mejor resultado de ambas, y siguiendo el mismo procedimiento hasta quedar pocos nodos, momento en el que ya se aplica Karger directamente.

**GRAFO CON PESOS**

Para representar un grafo con pesos en este algoritmo, se realiza el mismo proceso que para una sola arista, sólo que ahora se generan varias aristas entre dos nodos. Para ello, en lugar de la tabla con valores booleanos que se empleaba anteriormente, se emplea directamente la tabla de enteros que indicaba el número de aristas entre dos vértices dados. En el grafo se generan tantas aristas entre dos vértices como el valor que haya en esa posición de la tabla.

Para representar este nuevo grafo se seguirá utilizando la misma estructura explicada anteriormente, indexando en la tabla hash principal los vértices de uno en uno del grafo, y en su correspondiente tabla hash de vértices adyacentes las claves de esos vértices. En este caso, igual que en el del grafo sin pesos, se indexan esos vértices con el código numérico indicado “\_X”, que permite identificar rápidamente a que vértice se dirige dicha arista. Por ejemplo, si entre los vértices 1 y 2 hay tres aristas, contendrá tres posiciones de la tabla hash ocupada por las claves “2\_0, 2\_1 y 2\_2”.

Con esto hecho, el funcionamiento del algoritmo será análogo al anteriormente explicado tanto para el algoritmo de Karger como para el de Karger-Stein. Al unir dos vértices, el vértice resultante recibirá tantas aristas como la suma de las que recibían sus dos mitades, con lo que aumentará la probabilidad de elegir aristas entre dos vértices conectados con muchas aristas. Esto es obvio ya que hay más probabilidad de elegir la unión de dos vértices cuanto más número de aristas hay que los une.

**PRUEBAS**

Una vez implementado todo lo requerido, se pasó a realizar las pruebas para saber si funcionaba bien todo lo realizado. Las pruebas principales consistieron en dos grafos, uno con pesos y otro sin pesos fijados a mano por los dos integrantes de la práctica, y ver que se obtenía la solución correcta en un gran número de casos. Además, a parte de estos dos grafos nombrados, se realizaron ejecuciones con grafos aleatorios creados por el programa, comprobando que los algoritmos funcionaban perfectamente.

**Grafo 1**

En este primer grafo, para obtener el mínimo corte de dos conjuntos de los vértices, se puede apreciar que en uno estarían los vértices {2,3,4,5} y en el otro {0,1}; sin embargo, el vértice 1 podría oscilar entre los dos conjuntos ya que el valor del número de aristas entre los dos conjuntos sería 1 en los dos casos. Así pues, teniendo el grafo construido, si insertaron a mano los valores de la tabla de booleanos, y se probaron con él el algoritmo de Karger y el de Karger-Stein.

Como resultado de ejecutar el algoritmo de Karger, se pudo ver que se obtenían en un gran número de casos cualquiera de los dos resultados óptimos, pero que también fallaba dando resultados cercanos, pero no el óptimo. Sin embargo, el algoritmo de Karger-Stein además de dar la solución óptima para el grafo, la daba en casi todos los casos, algo que se esperaba al ser una mejora del algoritmo de Karger.

**Grafo 2**

Este segundo grafo contenía los mismos vértices que el primero, pero además pesos en las aristas. En la imagen, se ha reflejado como lo va a ver internamente el programa, siendo el número de aristas entre dos vértices igual al peso de la arista inicial. Como se ve en la imagen, de igual modo que en el primer grafo, los dos conjuntos claramente diferenciados son {0,1} y {2,3,4,5}. Así pues, se introdujeron los datos en la tabla de número de vértices adyacentes a mano y se ejecutó el programa para dicho grafo (algoritmos de Karger y Karger-Stein).

Como resultado de la ejecución, se obtuvieron en ambos algoritmos los resultados esperados, y nuevamente, el algoritmo de Karger-Stein daba mejores resultados, ya que sacaba el resultado óptimo un mayor número de veces. Además, si no se daba el resultado óptimo, se obtenía uno bastante cercano al ideal, por lo que los algoritmos para los grafos con pesos funcionaban a la perfección.

Finalmente, se siguieron probando con varios grafos aleatorios con distinto número de vértices y emparejamientos aleatorios (no un número muy elevado para poder analizarlo) y se seguían obteniendo los resultados ideales en la mayoría de los casos. Además, se probaron los dos generadores aleatorios que proporciona el lenguaje con el que se había implementado la práctica (Java), y los resultados eran muy parecidos, no se apreciaba una mejora sustancial de uno a otro.

**CONCLUSIÓN**

En conclusión, se ha realizado un programa que, a partir de un grafo con pesos o sin pesos, devuelve un corte mínimo con una alta probabilidad. El programa implementa dos algoritmos como son el de Karger, y su mejora en Karger-Stein. Estos algoritmos, a pesar de que no encuentran en todas las ocasiones el corte mínimo del grafo sí que lo hacen con buena probabilidad.

Además, se han implementado analizando detalladamente los pasos que seguía el algoritmo para construir la estructura de datos que fuera más adecuada y aumentara todo lo posible su eficiencia. Estos algoritmos se han probado con numerosos ejemplos de distinto número de vértices, sabiendo de algunos de ellos el corte mínimo que se obtendría, para comprobar de este modo que funcionaban correctamente.

**DIVISIÓN DE TRABAJO**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | Métodos Grafos | Karger | Karger-Stein | Pruebas | Memoria |
| Jorge Andrés | 3 hora | 2 horas | 3 horas | 2 horas | 2 horas |
| Javier Aranda | 3 hora | 3 horas | 2 horas | 1 hora | 3 horas |

**REFERENCIAS**

<https://en.wikipedia.org/wiki/Karger's_algorithm>

<http://mathandshit.blogspot.com.es/2012/04/teorema-de-los-monos-infinitos.html>

<http://web.stanford.edu/class/archive/cs/cs161/cs161.1138/lectures/11/Small11.pdf>