CAIM, examen final

18 de gener de 2018. Temps: 2 hores 50 minuts

Exercise 1

(2 points) Suppose that we have a collection of documents where the term `Carles` appears in $5 \cdot 10^5$ documents, the term `Oriol` appears in $3 \cdot 10^5$ documents and the term `Agnes` appears in 10^4 documents. What would be the time cost of processing the queries

```
Carles and (Oriol or Agnes)
(Carles and Agnes) and Oriol
(Carles and Oriol) or Agnes
```

based on a worst case analysis? Assume that the `AND` and `OR` operations involve a merge-like algorithm of posting lists.

Answer of exercise 1

We define c_{inner} and c_{outer} as the worst case time cost of the inner and the outer operation, respectively. We define l_{inner} as the worst case length of the output of the inner operation. The total worst case time cost is $c = c_{inner} + c_{outer}$. We have

1. c[Carles and (Oriol or Agnes)] $= 11.2 \cdot 10^5$ with $c_{inner} = 3.1 \cdot 10^5$, $l_{inner} = 3.1 \cdot 10^5$ and $c_{outer} = 8.1 \cdot 10^5$.
2. c[(Carles and Agnes) and Oriol] $= 8.2 \cdot 10^5$ with $c_{inner} = 5.1 \cdot 10^5$, $l_{inner} = 10^4$ and $c_{outer} = 3.1 \cdot 10^5$.
3. c[(Carles and Oriol) or Agnes] $= 11.1 \cdot 10^5$ with $c_{inner} = 8 \cdot 10^5$, $l_{inner} = 3 \cdot 10^5$ and $c_{outer} = 3.1 \cdot 10^5$.

Exercise 2

(1.5 points) Suppose a posting list that consists of a sequence of n docid-frequency pairs, i.e.

$$x_1, y_1, ..., x_i, y_i, ...x_n, y_n$$

where x_i is the i-th docid and y_i is the frequency of occurrence of the term in document x_i. For instance, the sequence of integers

$$2, 6, 5, 1$$

indicates that the term appears six times in document 2 and once in document 5.

We have compressed a posting list following the format above and obtained the following string of bits

$$10001001100010010101111111000010 0110$$

Decode the bit string to obtain the original posting list assuming that

1. Frequencies have been coded using unary self-delimiting codes as a sequence of 1's ending by a 0.
2. Docids have been coded using gap compression and Elias γ codes (the unary self-delimiting code within the Elias γ code is a sequence of 0's ending by a 1).

Hint: the first element of the bit string is an Elias γ code representing the number 1.

Answer of exercise 2

Segmenting the sequence, one gets a list of pairs

$$(1, 0), (00100, 110), (00100, 10), (1, 0), (1, 1111110), (0001001, 10)$$

that encodes the list

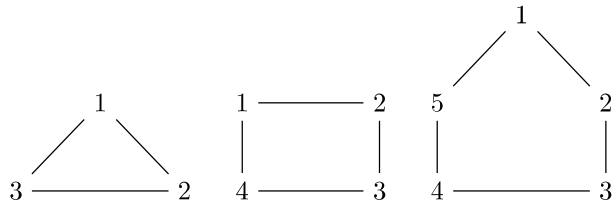
$$(1, 1), (4, 3), (4, 2), (1, 1), (1, 7), (9, 2)$$

Undoing gap compression, we finally obtain

$$(1, 1), (5, 3), (9, 2), (10, 1), (11, 7), (20, 2)$$

Exercise 3

(2 points) A cycle graph is a graph where all vertices are contained in a single cycle. Suppose that vertices are assigned unique integers from 1 to n as in the following cycle graphs



Then the edges

$$(1,2),(2,3),...,(i,i+1),...(n-1,n),(1,n)$$

define a cycle graph.

Calculate the betweenness centrality of the vertices of cycle graphs with $n = 3$, $n = 4$ and $n = 5$. Explain your answer.

Answer of exercise 3

We borrow the definition of betweenness centrality of a node i as

$$bc(i) = \sum_{j<k} \frac{g_{jk}(i)}{g_{jk}}.$$

In the summation above, notice that $g_{jk}(i) = 0$ implies

$$\frac{g_{jk}(i)}{g_{jk}} = 0$$

because $g_{jk} > 0$ for $j \neq k$. Notice also that

$$g_{ik}(i) = g_{ki}(i) = 0$$

We will focus on the calculation of $bc(1)$. The symmetry of the vertices as well as the calculations below allow one to see that $bc(1) = bc(2) = ... = bc(n)$.

When $n = 3$, one has $g_{jk}(1) = 0$ in all cases and then $bc(1) = 0$. When $n = 4$, one has $g_{jk}(1) = 0$ in all cases except for $j, k = 2, 4$. As $g_{2,4}(1) = 1$ and $g_{2,4} = 2$, bc$(1) = 1/2$. When $n = 5$, one has $g_{jk}(1) = 0$ in all cases except for $j, k = 2, 5$. As $g_{2,5}(1) = 1$ and $g_{2,4} = 1$, bc$(1) = 1$.

Exercise 4

(1.5 points) An edge of a weighted directed graph can be defined by a triple (u, v, w), where w is the weight of the arc from vertex u to vertex v. The in-strength of a vertex v is the total weight of the arcs ending at v while the out-strength of a vertex v is the total weight of the arcs beginning at v.

Given a list of the triples above in an arbitrary order, we want to produce the in-strength and out-strength of every vertex that has appeared in the list using the MapReduce framework. Provide pseudocode for map and reduce functions and, if appropriate, combine and partition functions. It is possible to use more than one map-reduce phase if required.

Answer of exercise 4

Every instance of made receives one triple of the list.

```
map(u, v, w)
   output(u, (w, 0))
   output(v, (0, w))

reduce(v, L) = combine(v, L)
  s is the sum of the 1st component of the elements of L
  t is the sum of the 2nd component of the elements of L
  output(v, (s, t))
```

Notice that the solution

```
map(u, v, w)
   output((u,"out"),  w)
   output((v,"in"),  w)

reduce(k, L) = combine(k, L)
  output(k, sum(L))
```

does not produce the in-strength and the out-strength of every vertex. In particular, it does not produce the in-degree of the vertices that have not appeared in the 2nd position of the triple. It does not produce either the out-degree of the vertices that have not appeared in the 1st position of the triple.

Notice also that calculating the in-degree and out-degree by separate map/reduce/combine functions is innefficient (as shown above the problem can be solved with a single map/reduce/combine). Submitting the whole graph (all triples) to a single instance of map is a bad solution for not exploiting parallel computation at least at the map stage.

Exercise 5

(1.5 points) Suppose a recommender system that has to estimate $r(a, s)$, i.e. the preference of user a for a certain item s based on the preferences of other users from a set U over the same product. Suppose that $r(a, s)$ is estimated with $pred(a, s)$, that is defined as a simple mean over U, namely

$$pred(a, s) = \frac{1}{|U|} \sum_{b \in U} r(b, s).$$

Analyze the limitations of the formula above and explain how it could be improved, proposing better formulae for estimating $pred(a, s)$. Suppose that U is given (cannot be changed) and that the scores of user a or users in U on other items are available. Explain the rationale behind your choices.

Answer of exercise 5

A problem of the formula above is that all user in U are given the same weight, namely $\frac{1}{|U|}$ regardless of their similarity with user a. Indeed, the formulae above is a particular case of a more general formula

$$pred(a, s) = \sum_{b \in U} w(a, b) r(b, s).$$

with $w(a, b) = \frac{1}{|U|}$. The original formula can be improved defining $w(a, s)$ as a normalized similarity, namely

$$w(a, b) = \frac{sim(a, b)}{\sum_{b \in U} sim(a, b)}.$$

After this improvement, a limitation that remains is that users in U are assumed to score in the same fashion (some user may tend to score high while other may tend to score low). Another limitation is that $pred(a, s)$ should take into account a's way of scoring. A simple solution consists of use the mean of the scores for reference, giving

$$pred(a, s) = \bar{r}_a + \sum_{b \in U} w(a, b)(r(b, s) - \bar{r}_b).$$

where $\bar{r}_b$ is the average of $r(b, s)$ over all items.

Exercise 6

(0.75 points) Discuss ethical issues involved in real applications of recommender system techniques.

Answer of exercise 6

Recommendations raise many ethical concerns. Some examples follow. An obvious concern is that they exploit private information (e.g., personal preferences). A deeper issue is that they solve the paradox of choice by reducing the set of possibilities when there are actually many or too many. Recommender systems can keep a user isolated in his bubble based on his culture, ideology or more specific personal preferences. Therefore, recommender systems are responsible for the balance, novelty,...of the recommendations.

Recommender system have to deal with a conflict of interest between at least four entities: the user, a society, a company and the invironment. In many commercial applications, recommender systems are biased toward corporate interests, usually aiming at increasing the number of purchases, manipulating the behavior of the user, who may end up buying products that he/she actually does not really need. The money spent could have been invested in the user health, his/her self-fullfillment,...or in reducing global warming or defending human rights.

Exercise 7

(0.75 points) Suppose that we wish to build a search engine for images. A query consists of an image and answers are images that resemble the query. For simplicity, suppose that images are restricted to grey scale bitmaps where gray scale is indicated by integer numbers. Thus an image is essentially a 2D matrix of integers. Discuss the limitations of the locality sensitive hashing methods seen in this course for retrieving images in a human-like fashion.

Answer of exercise 7

Each image would be represented as a long vector (a concatenation of rows of the 2D matrix) where each component would indicate the gray scale intensity of a given pixel of the image. Consider an image and a shifted, rotated or zoomed version of it. A serious limitation of the techniques that we have seen in the course is that they would consider the two images as different (if sufficiently shifted, rotated or zoomed) even if they are actually similar according to human perception. Similar problems may arise identical pictures that contrast in lighting.