Final exam for T.61-5060 and T.61-6010

Dec 17, 2013

This is an **open book** exam. It is allowed to use any textbook, printed material, or personal notes brought in the room.

There are **three** problems. Each problem receives the same number of points.

Problem 1

We are working in an application in which our data objects are represented as directed acyclic graphs (dags) over a set of items. Namely, $V = \{v_1, \ldots, v_n\}$ is a set of n items, and each data object is a dag $G = (V, E)$ over V. We want to define a distance function between such data objects. That is, we want to define a distance function $d(G_x, G_y)$ for any two dags $G_x = (V, E_x)$ and $G_y = (V, E_y)$. One attempt is the following:

1. Given two dags $G_x = (V, E_x)$ and $G_y = (V, E_y)$, put their edges together by taking the "overlay" graph $G_{x+y} = (V, E_x \cup E_y)$.

   (Note that G_{x+y} is directed by not necessarily acyclic)

2. Define $d(G_x, G_y)$ as the *minimum* number of edges required to *remove* from G_{x+y} so that it becomes acyclic.

Question 1.1. Provide an example of an application domain in which it is meaningful to represent data objects as dags over a set of items.

Question 1.2. Provide a justification for the distance function $d(G_x, G_y)$, as defined above. That is, try to explain what is the intuition behind the definition of $d(G_x, G_y)$.

Question 1.3. We have seen that a distance function should satisfy a number of properties. Which of these properties does the function $d(G_x, G_y)$ satisfy?

Question 1.4. Given your answer in Question 1.3, would you recommend adopting this distance function or not?

Problem 2

In class we discussed about representing documents as *sets of words*, i.e., each document is represented by the set of the words it contains. For example, the document $a\,b\,a\,c\,d\,a\,c$ is represented by the set $\{a, b, c, d\}$.

In many cases, in order to have a more accurate representation, we use *bags of words* (or *multisets of words*). According to this model, in addition to keeping the words in the document, we also record the *number of times* that each word appears. For example, the document $a\,b\,a\,c\,d\,a\,c$ is represented by the bag $\{(a, 3), (b, 1), (c, 2), (d, 1)\}$.

To deal with bags, we define some useful operations:

Bag size. The size of a bag A is defined as

$$||A|| = \sum_{(x,n)\in A} n.$$

Bag union. The bag union $\sqcup$ between two bags A and B is defined as

$$A \sqcup B = \{(x, \max\{n, m\}) \text{ where } (x, n) \in A \text{ and } (x, m) \in B\}.$$

Bag intersection. The bag intersection $\sqcap$ between two bags A and B is defined as

$$A \sqcap B = \{(x, \min\{n, m\}) \text{ where } (x, n) \in A \text{ and } (x, m) \in B\}.$$

For example, if $A = \{(a, 3), (b, 1), (c, 2), (d, 1)\}$ and $B = \{(a, 1), (b, 2), (d, 4), (e, 2)\}$, we have

$$||A|| = 7, \quad ||B|| = 9,$$

$$A \sqcup B = \{(a, 3), (b, 2), (c, 2), (d, 4), (e, 2)\},$$

and

$$A \sqcap B = \{(a, 1), (b, 1), (d, 1)\}.$$

We also extend the Jaccard coefficient between bags as

$$J(A, B) = \frac{||A \sqcap B||}{||A \sqcup B||}.$$

Question 2.1. Argue that the extension of the Jaccard coefficient to bags, as defined above, is meaningful and well-motivated.

Question 2.2. Provide a locality-sensitive hashing (LSH) scheme for the Jaccard coefficient to bags. In other words, design a family of hash functions $\mathcal{F}$ such that

$$\Pr[f(A) = f(B)] = J(A, B),$$

when f is drawn uniformly at random from $\mathcal{F}$. Discuss how exactly you will implement the locality-sensitive hashing scheme you designed.

Problem 3

The objective of this problem is to design an algorithm for counting distinct items in a data stream, which is *different* than the Flajolet-Martin algorithm.

Consider a data stream $x_1, x_2, \ldots$, potentially infinite, where each x_i is an item from a very large ground set $U = \{1, \ldots, N\}$.

Assume that we have access to a family of hash functions $\mathcal{F} : U \to [0, 1]$, such that any f from $\mathcal{F}$ maps each item in U to a *random number* in the interval $[0, 1]$.

Question 3.1. Assume that you observe the data stream $x_1, x_2, \ldots$, and you are able to keep in memory *only one* number. Describe how you can use your one-number-only memory space so that at any point you have an estimate of the distinct items that you have seen so far.

Question 3.2. What would you do if you could keep m numbers, instead of one?