Web Mining course, Academic year 2010-2011

Written exam

Mauro Brunato

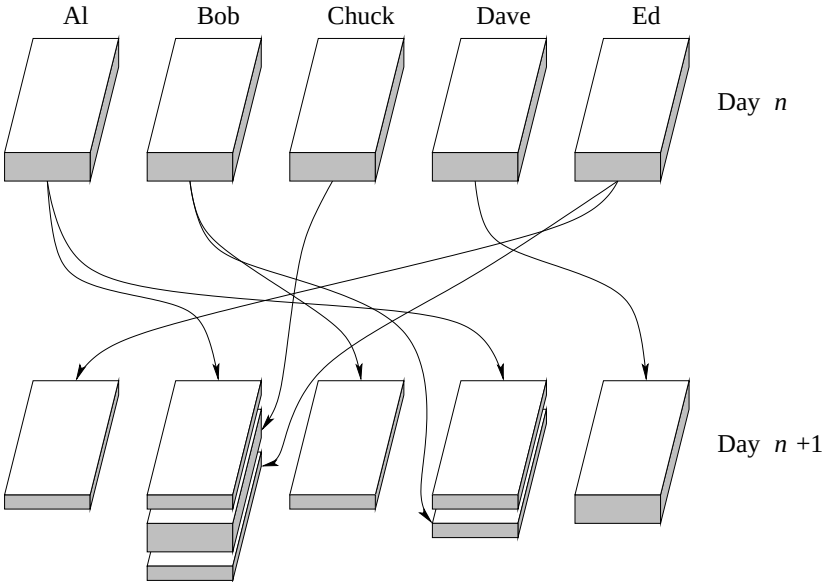
Tuesday, June 14, 2011

Exercise 1

Five friends (Al, Bob, Chuck, Dave and Ed), like play a trading cards game every morning. In order to make the game more interesting, they decide to exchange their decks of cards every day before playing, according to the following rules:

- Al splits his deck in two equal parts and gives the two halves to Bob and to Dave;
- Bob splits his deck in two equal parts and gives the two halves to Chuck and to Dave;
- Chuck gives his deck to Bob;
- Dave gives his deck to Ed;
- Ed splits his deck in two equal parts and gives the two halves to Bob and to Al.

Here is an example of how the decks are dealt every morning:



Let N be the overall number of cards owned by the five friends. Assume that at every step all decks can be precisely split in half.

1.1) Show that, no matter how the cards are initially distributed, after a (possibly large) number of days the number of cards owned by each of the five friends stabilizes (i.e., does not change anymore).

1.2) Determine the number of cards in each player's deck when it stabilizes, expressed as a fraction of N. Check the solution for $N = 380$.

1.3) Suppose now that Bob changes his own policy and gives his whole deck to Chuck every morning. Prove that the number of cards owned by each friend does not necessarily stabilize. What is the only stable combination, given as the fraction of N owned by each friend? Check the solution for $N = 380$.

Exercise 2

Let D be any set, and $s : D \times D \to \mathbb{R}$ a similarity measure between any two elements of D. Discuss some possible ways to generalize s to a similarity measure between any pair of subsets of D (possibly excluding the empty set).

Exercise 3

Google Analytics logs the time of permanence of each user on each visited web page on a simple text file whose lines are in the following format:

page_ID user_ISO_country_code time_of_permanence

Consider, for example, the following lines:

```
13287 IT 32
36453 UK 12
32789 IT 45
435 FR 10
2348 UK 30
```

The first one says that the webpage having ID 13287 has been visited for 32 seconds by a user coming from Italy, etc.
The log file is very large and it is stored in a distributed file system (or, possibly, there are many log files).
3.1) Provide a basic MapReduce implementation (i.e., mapper, combiner, reducer) that computes the average permanence time per country code. For example, the outcome of the above log file should be the following:

```
IT 38.5
UK 21
FR 10
```

meaning that users from Italy spent 38.5 seconds on average ($\frac{32+45}{2}$) on each visited page.
3.2) Provide an example in which using a combiner function results in a clear advantage for the framework.
3.3) Suppose that ten thousand machines are available. However, the ISO country codes are only 248. Propose a solution for using all machines in the Reduce phase.

Use any language or pseudolanguage you like, as long as the algorithm is clear.