COS 435, Spring 2011 - Problem Set 6

Due at 1:30pm, Wednesday, April 13, 2011

Collaboration and Reference Policy

You may discuss the general methods of solving the problems with other students in the class. However, each student must work out the details and write up his or her own solution to each problem independently.

Some problems have been used in previous offerings of COS 435. You are NOT allowed to use any solutions posted for previous offerings of COS 435 or any solutions produced by anyone else for the assigned problems. You may use other reference materials; you must give citations to all reference materials that you use.

Lateness Policy

A late penalty will be applied, unless there are extraordinary circumstances and/or prior arrangements:

- No penalty if in Prof. LaPaugh's office or inbox by 5pm Wednesday (4/13/11).
- Penalized 10% of the earned score if submitted by 11:59 pm Wed. (4/13/11).
- Penalized 25% of the earned score if submitted by 5pm Friday (4/15/11).
- Penalized 50% if submitted later than 5pm Friday (4/15/11).

Problem 1: Clustering -- iterative improvement for divisive partitioning

Slide #21 of the slides for *general clustering algorithms* posted under March 30 presents an iterative improvement algorithm for divisive partitioning. This problem addresses recalculating the total relative cut cost (slides #17 and #18) incrementally for use with that algorithm.

Let U denote the set of objects to be clustered. Assume that for any objects v and w, sim(v,w)=sim(w,v) (we have been assuming this in class). Also assume that for any object v, sim(v,v)=0. Let C_p be an arbitrary cluster containing object x, C_q be an arbitrary cluster that does not contain x. (The set notation $C_p - \{x\}$ denotes C_p with x removed, and $C_q \cup \{x\}$ denotes C_q with x added.)

The following relationship holds for incremental changes to the intracost of a cluster when removing or adding an object x.

$$\text{intracost}(C_p) - \text{intracost}(C_p - \{x\}) = \sum_{v_i \text{ in } C_p - \{x\}} \text{sim}(v_i, x)$$

$$= \sum_{v_i \text{ in } C_p} \text{sim}(v_i, x) \qquad \text{since sim}(x,x) = 0$$

From this relationship we derive the incremental cost changes for intracost:

$$\text{intracost}(C_p - \{x\}) = \text{intracost}(C_p) - \sum_{v_i \text{ in } C_p} \text{sim}(v_i, x)$$

$$\text{intracost}(C_q \cup \{x\}) = \text{intracost}(C_q) + \sum_{v_i \text{ in } C_q} \text{sim}(v_i, x)$$

Your task is to derive incremental cost equations for cutcost:

Part a: Give an equation for

$$\text{cutcost}(C_p) - \text{cutcost}(C_p - \{x\})$$

when x is an object in C_p. Your equation should be in terms of similarities between x and other objects.

Hint: the quantity

$$\sum_{v_i \text{ in } U} \text{sim}(v_i, x) \quad \text{where U is the set of all objects}$$

is useful because it is a function of x independent of the clustering and can be precomputed before the clustering construction is begun.

Part b: Using your equation of Part a, derive equations for

i. $\text{cutcost}(C_p - \{x\})$ as an incremental change to $\text{cutcost}(C_p)$;
ii. $\text{cutcost}(C_q \cup \{x\})$ as an incremental change to $\text{cutcost}(C_q)$.

Problem 2: Latent Semantic Indexing

The computational cost of comparing a query to all documents by calculating $C^T\mathbf{q}$ is M*N multiplications and (M-1)N additions, assuming C is stored as an array of M×N elements and **q** is stored as a vector of M elements. As in class, **q** denotes the vector representation of the query and C denotes the matrix whose columns are the vector representations of the documents. There are M terms and N documents.

In contrast, the computational cost of doing a comparison of a query to all documents after Latent Semantic Indexing has been used to compute the rank-k approximation is

i. Step 1: compute $\boldsymbol{q}_k = (\Sigma'_k)^{-1} (U'_k)^T \boldsymbol{q}$. This uses k*M multiplications and k*(M-1) additions.

plus

ii. Step 2: compare $\boldsymbol{q}_k$ to all documents by computing $V'_k(\Sigma'_k)^2 \boldsymbol{q}_k$. This uses N*k multiplications and N*(k-1) additions.

Note that matrices U'_k, Σ'_k, V'_k, $(V'_k(\Sigma'_k)^2)$, and $((\Sigma'_k)^{-1}(U'_k)^T)$ are all computed in a preprocessing step before any queries are evaluted, and this cost is not included in the cost of processing a query.

This analysis suggests a significant computational savings using Latent Semantic Indexing when k is small. However the matrix C and query vector **q** are sparse, particularly **q**. In contrast, the matrices $(V'_k(\Sigma'_k)^2)$, and $((\Sigma'_k)^{-1}(U'_k)^T)$) and transformed vector $\boldsymbol{q}_k$ are generally not sparse. Accounting for this may change the picture.

Part a: Redo the calculation of the cost of comparing a query to all documents by calculating $C^T\mathbf{q}$ versus the cost of comparing a query to all documents using steps 1 and 2 above if

- Query **q** consists of *t* non-zero entries (terms) and is stored in sparse form.
- Each row of C^T is stored in sparse form.
- Documents contain, on average, α*M terms.

Count each access of an element not used in a computation as well as counting multiplications and additions. Your calculation should be in terms of parameters k, *t*, α.

Part b: Holding *t* fixed, for what values of k and α does the rank-k approximation save computational cost?

Problem 3: Detecting near-duplicate documents

Part a: Let D denote a document that is 500 words long and contains each of the words "philanthrepist" , "pendantic" and "androgenous" exactly once each, with "philanthrepist" occurring in word position 100, "pendantic" in position 205, and "androgenous" in position 320. Each of these words is misspelled. Let D_{cor} be the document with these spelling errors corrected ("philanthropist" , "pedantic" and "androgynous"). What is the value of the resemblance $r(D, D_{cor})$ for a 5-shingling of each document if, for each document, 25% of all possible shingles are repeated shingles?

Part b: Let E denote a document that is 500 words long and contains each of the words "philanthrepist" , "pendantic" and "androgenous" exactly once each but as the phrase "pendantic androgenous philanthrepist" starting at word position 200. Let E_{cor} be the document with the spelling errors in this phrase corrected ("pedantic androgynous philanthropist"). What is the value of the resemblance $r(E, E_{cor})$ for a 5-shingling of each document if, for each document, 25% of all possible shingles are repeated shingles?

Part c: For what threshold would one of the pairs *(D, D_{cor})* and *(E, E_{cor})* be considered near-duplicates and the other not? Which is which? In your opinion, is this a desirable outcome?