CSCI 6370 IR and Web Search
ASSIGNMENT 5
Due is 07/06/2020 23:59

Ulvi Bajarani
Student ID 20539914
E-mail: ulvi.bajarani01@utrgv.edu

Questions and Answers:

Problem 1. [35 points] Assume that we use cosine similarity as the similarity measure. In the hierarchical agglomerative clustering (HAC), we need to define a good way to measure the similarity of two clusters. One usual way is to use the group average similarity between documents in two clusters. Formally, for two cluster C_i and C_j, let $C = C_i \cup C_j$, $n = |C|$ we define

$$sim(C_i, C_j) = \frac{1}{n \cdot (n-1)} \sum_{x,y \,\in\, C,\; x \neq y} s\,(x, y)$$

Where $s\,(x, y)$ is the cosine similarity between x and y.

Given a list of clusters $C_1, C_2, C_3, ..., C_m$, assume that their pairwise similarities are saved in a two dimensional array of size m^2. Given three clusters C_i, C_j and C_k, show that there is a way to compute $sim\,(C_i \cup C_j,\ C_k)$ in constant time. Note that we ignore the dimensionality in time complexity.

Answer 1. It is possible to do in the constant time by:

$$sim\,(C_i \cup C_j,\ C_k) = \frac{(s\,(c_i \cup c_j) + s\,(c_k)) \bullet (s\,(c_i \cup c_j) + s\,(c_k)) - (|\,(c_i \cup c_j)\,| + |c_k|)}{(|\,(c_i \cup c_j)\,| \;\; + |c_k|)\,(|\,(c_i \cup c_j)\,| + |c_k| - 1)}$$

where $s\,(c_j)$ is the sum vector of the cluster c_j equal to

$$s\,(c_j) = \sum_{x \in c_j} x$$

Problem 2. [30 points] For a list of m documents in d-dimensional vector space, each iteration of the k-means clustering has a time complexity of $O(kdm)$, which the hierarchical agglomerative clustering (HAC) has a time complexity of $O(dm^2)$. We know that the overall performance of the k-means clustering depends on the choices of initial k centroids. However, there is no such an issue for HAC. Describe a method to use HAC to help the k-means clustering, but the method shall maintain the same time complexity for the k-means clustering.

Answer 2. The algorithm above assumes that all m clusters are the examples. If we take only $\sqrt{m}$ clusters as a sample, the complexity will be $O(\sqrt{m^2}) = O(m)$. This called as the Buckshot Algorithm. After taking $\sqrt{m}$ examples, the group-average HAC algorithm should be conducted on the sample. The results are used as initial seeds for K-means.

Problem 3. [35 points] Assume you are working for amazon.com. You need to find some way to recommend products for an online shopper without asking for relevance feedback from any user. Provide one solution.

Answer 3. One possible solution is using the content-based recommending system. It might be implemented by a Bayesian text-categorization algorithm, which will evaluate the probability in the categories.