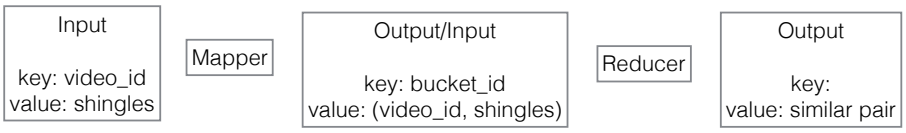
Data Mining: Learning from Large Data Sets - Fall Semester 2015

luchen@student.ethz.ch
zhhan@student.ethz.ch
myhshirley@student.ethz.ch

October 12, 2015

Approximate near-duplicate search using Locality Sensitive Hashing

Procedure



Mapper

1. Pick a number of bands b and a number of rows r, then $n = br$ is the number of minhash functions. Set $N = 20001$ as the number of shingles, and $Prime = 150000001$ as the divisor for the modulo operation in linear hash functions.
2. Generate random numbers as the parameters for n minhash functions and b linear hash functions.
3. For each video, use minhash functions to generate its signature. Divide the the signature into b bands with r rows in each band. Hash each band to a bucket, and output <bucket_id, (video_id, shingles)>.

Reducer

1. For each pair of two videos in one bucket, compute Jaccard similarity of their shingles, and print the video_ids if the similarity is no less than 0.9.

Choice of r and b The probability that two videos become a candidate pair is $P(s; b, r) = 1-(1-s^r)^b$, where s is the Jaccard similarity of the two videos. $P(s; b, r)$ is steepest at $\frac{1}{b}^{\frac{1}{r}}$, so we shall choose b and r to approach the threshold 0.9 as close as possible, but not to exceed 0.9 to keep the false negative low. Fix $br = 1000$ or 1024, Figure 1 reveals that $b = 50, r = 20$ is an optimal choice as $\frac{1}{b}^{\frac{1}{r}} = 0.82$ is only a

little bit smaller than the threshold 0.9. Figure 2 plots $P(s; 50, 20) = 1-(1-s^{20})^{50}$ against the similarity s. $P(0.9; 50, 20) = 9985$ while $P(0.8; 50, 20) = 0.4400$. This guarantees a low false negative, while the false positive can be eliminated by the pairwise comparison. If we fix $br = 1024$, then $b = r = 32$ seems to be an optimal choice leading to $\frac{1}{b}^{\frac{1}{r}} = 0.897$. In this case, however, $P(0.9; 32, 32) = 0.67$ would result in a high false negative.

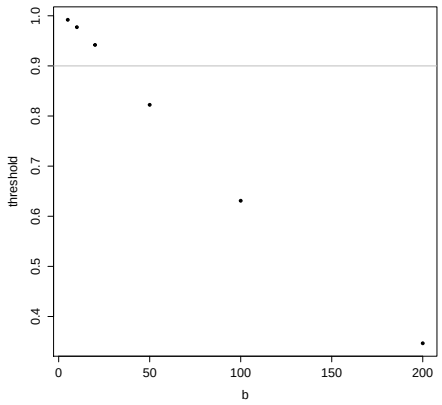


Figure 1: Threshold $(\frac{1}{b}^{\frac{1}{r}})$ against value of b.

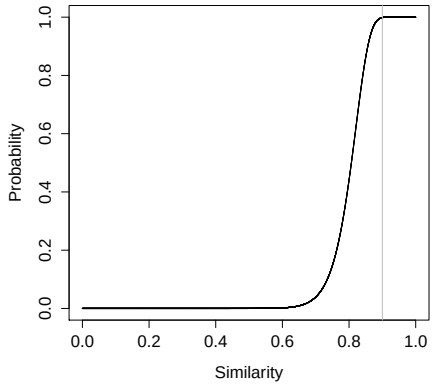


Figure 2: Probability of the two videos becoming a candidate pair against similarity s.

Contribution Each of us worked out one solution, and we handed in the one with the best result with recision = recal = 1. We also tried to use signatures instead of shingles to compute the similarity in the reducer step. It would reduce the running time of reducer step while increase the false negative and false positive.