Topic: Implementation of SimHash

Yuchen Hong, A20283450

Junzhe Zheng, A20254389

Abstract

Most hash functions are used to separate and obscure data, so that similar data hashes to very different keys. We use hash functions to detected similarities between data.

As storage capacities become larger it is increasingly difficult to organize and manage growing file systems. Consolidating or removing multiple versions of a file becomes desirable. However, deduplication technologies do not extend well to the case where files could also be useful for classification purposes and as an aid to search. A standard technique in similarity detection is to map features of a _le into some high- dimensional space, and then use distance within this space as a measure of similarity. Unfortunately, this typically involves computing the distance between all pairs of _les, which leads to O(n^2) similarity detection. Our goal was to create a similarity hash function, which maps similar data to a very similar hash key.