Bansilal Ramnath Agarwal Charitable Trust's
VISHWKARMA INSTITUTE OF INFORMATION TECHNOLOGY, PUNE
DEPARTMENT OF COMPUTER ENGINEERING

PROJECT SYNOPSIS

1 Group Id

STUDENT: -

1) Kedar Nikhil -	221059	(17U127)
2) Kesharwani Sourabh -	221060	(17U355)
3) Khandelwal Adarsh -	221061	(17U102)
4) Khandelwal Onkar -	221062	(17U018)

2 Project Title

Implementation of Locality-Sensitive Hashing over txt documents

3 Technical Keywords (As per ACM Keywords)

LSH
DATA CLUSTERING
K NEAREST NEIGHOUR

4 Problem Statement

C++ implementation of LSH over txt documents, using Jaccard Similarity.

5 Abstract

Locality Sensitive Hashing (LSH) to get sub-linear dependence on the data-size for high-dimensional dataPreprocessing :Hash the data-point using several LSH functions so that probability of collision is higher for closer objects

In this project we are using Locality-Sensitive Hashing technique over text documents . For this we are also using Jaccard Similarity . Using LSH we design algorithms for fast search of similar keywords in the text documents. Basically Lsh increases the frequency of collision in hashing so that we can put similar keyword in the same bucket . This technique is used on massive datasets for fast searching operation. This project work like K Nearest Neighbors Algorithms.

6 Goals and Objectives

As LSH algorithm is use for fast searching so our goal is fast searching of similar keywords.The main goal is replication of K Nearest Neighbors algorithms.

7 Mathematics associated with project:

An *LSH family* F is defined for a metric space M=(m,d) , a threshold R>0 and approximation factor C>1. This family F is a family of functions h: M->S which map elements from the metric space to a bucket s {S. The LSH family satisfies the following conditions for any two points p, q{M , using a function h{F which is chosen uniformly at random:

1. if d(p,q)<=R , then h(p)=h(q) (i.e.,*p* and *q* collide) with probability at least P1
2. if d(p,q) >=R, then h(p)=h(q) with probability at most P2.

A family is interesting when P1>P2. Such a family F is called (R,cR,P1,P2) *-sensitive*.

Alternatively it is defined with respect to a universe of items *U* that have a similarity function $ = U*U ->[0,1]. An LSH scheme is a family of hash functions *H* coupled with a probability distribution *D* over the functions such that a function h{H chosen according to *D* satisfies the property .

References

Koga, Hisashi, Tetsuo Ishibashi, and Toshinori Watanabe (2007), "Fast agglomerative hierarchical clustering algorithm using Locality-Sensitive Hashing", Knowledge and Information Systems, ***12*** *(1): 25–53, doi:10.1007/s10115-006-0027-5.*

1 *Rajaraman, A.; Ullman, J. (2010). "Mining of Massive Datasets, Ch. 3".*