minihash

2017年3月23日 星期四　　上午10:44

一、简介

是LSH技术的一种。可以用于快速评估两个集合的相似度。

从LSH的角度来看，相似度函数用的是Jaccard index（Jaccard similarity coef

值最小的那个值相等的概率" 估计Jaccard index。

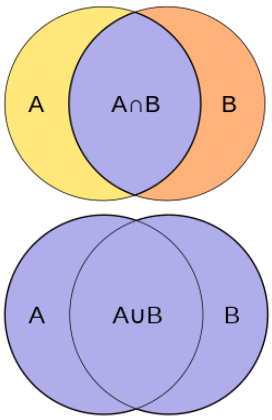
也称作“min-wise independent permutations locality sensitive hashing s

二、原理

2.1 Jaccard index（Jaccard 相似度）

评估两个集合A、B的相似度，可以用两个集合"交集占并集"的比例。

$$J(A,B) = \frac{|A \cap B|}{|A \cup B|}.$$



2.2 MiniHash 原理

如何快速估计Jaccard相似度。结合LSH思想，应该提出一种hash函数，has

A和B集合中的各个元素，通过hash函数映射之后，所有hash value的最小值

如果hash映射是均匀的，并且hash映射是不冲突的。A和B集合中，最小的

同）的概率和Jaccard 相似度是成正比的。也就是：P[hmin(A) = hmin(

所以，LSH的hash函数是hmin(A)，用P[hmin(A) = hmin(B)]的概率快速

fficient ），然后用"两个集合中，各个元素的hash

scheme”

sh value相同的key具有相同的相似度。

值定义为hmin(A)、hmin(B)

的hash value相同（也就是hash value最小的元素相

B)] = J(A,B)。

值估计J(A,B)

所以，LSH的hash函数是hmin(A)，用P[hmin(A) = hmin(B)]的概率快速

2.3 MiniHash 方案

由于hmin(A) = hmin(B)只有一个case：值要么是0，要么是1。

我们想用概率估计真实值，得用大量的case去估计，才具有统计意义，我

2.3.1 多个hash函数

2.3.2 一个hash函数，取最小的K的进行比较

三、应用

可以应用于信息检索任务中，判断两个文档是否雷同。

这里两个文档是集合A、B，文档中的term是集合中的元素。

http://baike.baidu.com/link?url=_2HnyhfLFFInXdDAiDhYcINPimZrG6Oo3rUD5qUHUGIdBf
https://en.wikipedia.org/wiki/MinHash
https://en.wikipedia.org/wiki/Jaccard_index

的计为(A,B)。

我们有两种方案。

YRxaVQfWg_M9uekYD3-jtKESscHVqrHn7USfj6AK