1. 使用哪些数据来计算Simhash

计算Simhash需要对一段文字逐字符的计算。那么用于计算Simhash的字符串的长短直接决定计算Simhash所需话费的时间。

如果选择title作为计算simhash的种子，那么计算耗时将较低。但是会发生如下错误

Problems:

Title: 10 things you need to know today: November 16, 2012

Title: 10 things you need to know today: November 17, 2012

Title: 10 things you need to know today: November 18, 2012

We can see that these RSS Items are different RSS Items, but the hamming distances between each other are very short.

所以我们选择Title + Brief Description 作为计算Simhash的种子。

1. 结果分析：

在去重算法中，相似RSS Item海明距离的阈值设置非常重要。如果阈值太低，重复RSS将会被漏过。如果太高，本不重复的RSS 会被认为是重复的RSS Item

下面是一些字符串海明距离的实例：

|  |  |  |
| --- | --- | --- |
| String1 | String2 | Hamming Distance |
| List of National League MVP award winners | List of National League MVP award winners | 6 |
| Golf-McIlroy set to miss cut in Hong Kong after putting problems | McIlroy set to miss cut in Hong Kong after putting problems | 3 |
| Battery maker A123 files for bankruptcy protection | Alien Life May Require Rare 'Just-Right' Asteroid Belts | 18 |
| Explosion on bus in Kenya capital kills at least 5 | Riot in Kenya after bus bombing kills 7 in capital | 14 |
| Charles Johnson criticizes unnamed Panthers teammates on Twitter after Bucs loss | An ‘Encore’ Life Beckons … on the Far Side of Midlife | 19 |

We can see that when hamming distance above 14, 结果会非常糟糕。

所以hamming distance 距离设置成8-12 比较好。

1. RSS 数据的抓取量

每一个RSS源一次发布的RSS新闻条目的数量是不一致的，大概从10-25条，平均在18条。每天大概更新3次。如果设置了100个RSS源，一天RSS的抓取量就是：100\*18\*3=4800条。实际统计数据请看下图：

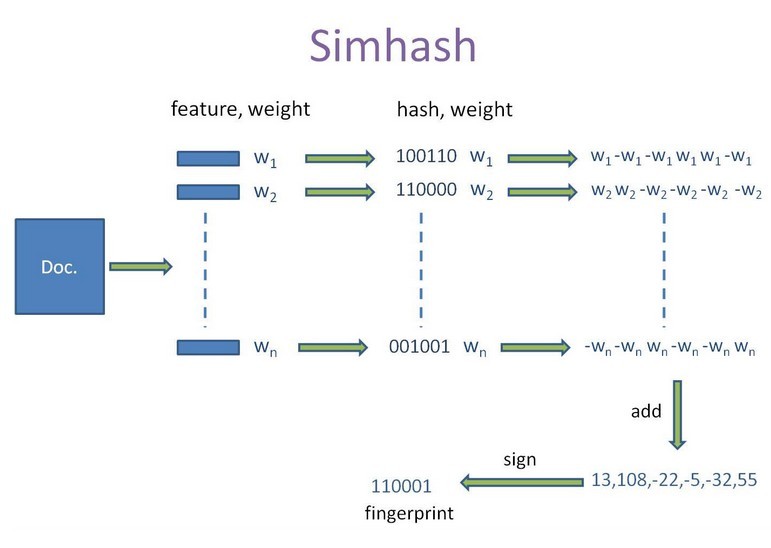
1. Simhash原理的介绍：

这个网址不错，简单简要：你翻译翻译就能用了。

不过你要答辩，估计算法的基本原理还是要搞清楚

<http://baike.baidu.com/view/9251886.htm>

下面这个图也不错



1. 所用到的library
2. RSSLib4j 主要用于对RSS抓取和解析。

sourceforge.net/projects/**rsslib4j**/

1. Simhash： 主要用于计算simhash值

是从**commoncrawl**.org项目中摘取出来的。

**Commoncrawl是一个开源的网络爬虫项目**

**3 Cassandra： 主要用于对数据库的操作。**

**对数据库Cassandra的操作是通过Hector API 完成的**

http://hector-client.github.com/hector/build/html/index.html#