2023数据库设计与实践 实验报告

项目名：利用HITS算法分析流行短语和网站

姓名：陈家豪 学号：19307130210

1. **项目简介**

本项目的主要目的是利用HITS算法分析一段时间内网络中最流行的短语以及最流行的网站。

项目的数据集来自于斯坦福大学的Memetracker项目，该项目通过分析来自100万个网络资源(从大众媒体到个人博客)每天的新闻报道和博客文章，跟踪范围内出现最频繁的短语。这使得我们可以分析每天有哪些不同的事件在竞争新闻和博客头条。项目调查的时间为2008年美国总统大选期间，项目统计了新闻和博客中出现的前50条热词的频率。

1. **算法简介**

本项目采用的算法是HITS（Hyperlink-Induced Topic Search）算法，该算法在1999年由Jon Kleinberg 提出，原论文为《Authoritative Sources in a Hyperlinked Environment》。该方法最初用于分析网络中的权威信息来源。

**2.1 算法原理**

在HITS算法中，网页被分为两种：hub页面和authority页面。hub页面是指包含了很多指向authority页面的链接的网页，比如国内的一些门户网站，比如博客、新闻网站等；authority页面则指那些包含有实质性内容的网页。HITS算法的最初目的是：当用户查询时，返回给用户高质量的authority页面。

HITS算法基于下面两个假设：1）一个高质量的authority页面会被很多高质量的hub页面所指向。2）一个高质量的hub页面会指向很多高质量的authority页面。

而一个页面的“质量”，由这个页面的hub值和authority值确定。其确定方法为：

页面hub值等于所有它指向的页面的authority值之和。

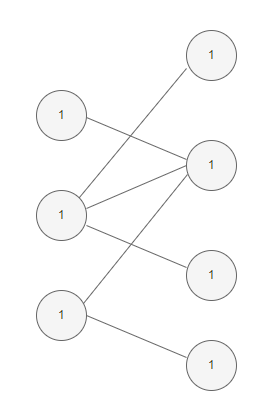
页面authority值等于所有指向它的页面的hub值之和。

我们在开始时将所有页面的hub,authority值置为1，并通过以上两个步骤不断迭代页面的hub,authority值，则最终会得到收敛的值。

**2.2 算法流程演示**

首先我们需要将数据集转换成一个二分图。hub节点集和authority节点集分别为二分图的两个节点集，如果一个hub节点引用了某authority节点，则在它们之间建立一条边。我们用矩阵来表示边的关系。

假设有如下节点：



则矩阵表示为：

hub向量为：， 标准化后为， authority向量为：，标准化

第一轮迭代后，

在HITS算法中，网页被分为两种：hub页面和authority页面。hub页面是指包含了很多指向authority页面的链接的网页，比如国内的一些门户网站，比如博客、新闻网站等；authority页面则指那些包含有实质性内容的网页。HITS算法的最初目的是：当用户查询时，返回给用户高质量的authority页面。】