**HITS算法简介**

**一、英文名称**

Hyperlink-Induced Topic Search

**二、中文名称**

超文本敏感标题搜索

**三、背景**

HITS 算法是由康奈尔大学的Jon Kleinberg 博士于1997 年首先提出的,为IBM 公司阿尔马登研究中心的名为"CLEVER"的研究项目中的一部分。

作为几乎是与PageRank同一时期被提出的算法，HITS同样以更精确的搜索为目的，并到今天仍然是一个优秀的算法。在HITS算法中，每个页面被赋予两个属性：hub属性和authority属性。同时，网页被分为两种：hub页面和authority页面。hub，中心的意思，所以hub页面指那些包含了很多指向authority页面的链接的网页，比如国内的一些门户网站；authority页面则指那些包含有实质性内容的网页。HITS算法的目的是：当用户查询时，返回给用户高质量的authority页面。

**四、思想**

按照HITS算法，用户输入关键词后，算法对返回的匹配页面计算两种值，一种是枢纽值，另一种是权威值,这两种值是互相依存、互相影响的。所谓枢纽值，指的是页面上所有导出链接指向页面的权威值之和。权威值是指所有导入[链接](https://baike.so.com/doc/854036-903001.html" \t "_blank)所在的页面中枢纽之和。通常HITS算法的基本思想是：

好的Hub型网页指向好的Authority型网页。

好的Authority型网页是由好的Hub型网页所指向的网页。

**五、流程**

HITS算法是利用HubPAuthority的搜索方法,

具体算法如下:

将查询q提交给基于关键字查询的[检索](https://baike.so.com/doc/5668894-5881558.html" \t "_blank)系统，从返回结果页面的集合中取前n个网页(如n=200)，作为根集合(root set)，记为S，则S满足:

（1）S中的网页数量较少

（2）S中的网页是与查询q相关的网页

（3）S中的网页包含较多的权威(Authority)网页

通过向S 中加入被S 引用的网页和引用S 的网页,将S 扩展成一个更大的集合T. 以T 中的Hub 网页为顶点集V1 ,以权威网页为顶点集V2 。

V1 中的网页到V2 中的网页的超链接为边集E ,形成一个二分有向图. 对V1 中的任一个顶点v ,用h ( v) 表示网页v 的Hub 值,且h ( v)收敛;对V2 中的顶点u ,用a ( u) 表示网页的Authority 值。

开始时h ( v) = a ( u) = 1 ,对u 执行I 操作,修改它的a ( u) ,对v执行O操作,修改它的h ( v) ,然后规范化a ( u),h ( v) ,如此不断的重复计算下面的I操作和O操作,直到a ( u),h(v)收敛 。

其中I操作:a ( u) = Σh ( v) ;O 操作: h ( v) = Σa ( u) 。每次迭代对a ( u) 、h ( v) 进行规范化处理: a ( u) = a ( u)/Σ[ a ( q) ]2 ; h ( v) = h ( v)/Σ[ h ( q) ]2 。

**HITS算法伪代码**

G= set of pages

for eachpage in G do

p.auth = 1 // 权威值

p.hub = 1 //枢纽值

HubsAndAuthorities(G)

for step from 1 to k do// run the algorithm for k steps

norm = 0

for eachpage in G do// update all authority values first

p.auth = 0

for eachpage in p.incomingNeighbors do//计算平方的总和

p.auth +=q.hub

norm += square(p.auth) // 计算auth平方的和

norm = sqrt(norm)

for eachpage in G do // update the auth scores

p.auth =p.auth / norm // 归一

norm = 0

for eachpage in p.outgoingNeighbors do

p.hub +=r.auth

norm += square(p.hub) // 计算中心值平方的和

norm = sqrt(norm)

for eachpagepinGdo

p.hub =p.hub / norm

**六、应用**

HITS算法是Web结构挖掘中最具有权威性和使用最广泛的算法。HITS算法是利用Web的链接结构进行挖掘典型算法，其核心思想是建立在页面链接关系的基础上，对链接结构的改进算法。HITS算法通过两个评价权值--内容权威度(Authority)和链接权威度(Hub)来对网页质量进行评估。其基本思想是利用页面之间的引用链来挖掘隐含在其中的有用信息(如权威性)，具有计算简单且效率高的特点。HITS算法认为对每一个网页应该将其内容权威度和链接权威度分开来考虑，在对网页内容权威度做出评价的基础上再对页面的链接权威度进行评价，然后给出该页面的综合评价。内容权威度与网页自身直接提供内容信息的质量相关，被越多网页所引用的网页，其内容权威度越高;链接权威度与网页提供的超链接页面的质量相关，引用越多高质量页面的网页，其链接权威度越高。

**七、缺点**

1.计算效率低

这里说的“效率低”是针对其实时计算的特点而提出的。HITS算法是在用户提出搜索请求之后才开始运行的，然而计算出结果又需要多次迭代计算，所以就这点上来说HITS算法效率仍然较低。

2.主题漂移

在算法原理部分介绍了HITS算法是如何生成初始集合。从根集合我们通过链接添加网页的方法进行扩展，但这也很可能添加进与搜索主题无关的网页。若是这部分网页中又恰恰有着一些高质量的authority页面，则很有可能返回给用户，降低用户的搜索体验。

3. 易被作弊者操纵结果

试想我们弄一个页面指向很多高质量的authority页面，那么这个页面就成为了一个高质量的hub页面。然后再弄个链接指向自己的网页，按照HITS算法，将大大提升自己的网页的权威值。