

文章编号:1006-2475(2007)07-0023-03

Web 结构挖掘及 HITS 算法分析

黄英铭

(揭阳职业技术学院计算机中心,广东 揭阳 522000)

摘要:在介绍 Web 结构的基础上,研究了基于 Web 超链接的 HITS 算法,分析该算法存在的若干问题并提出了两种改进的算法。

关键词:Web 结构挖掘;超链接;HITS 算法

中图分类号:TP301.6 **文献标识码:**A

Web Structure Mining and Analysis of HITS Algorithms

HUANG Ying-ming

(Computer Center, Jieyang Professional & Technical College, Jieyang 522000, China)

Abstract: On the basis of the introduction of the Web construction, this paper studies the HITS algorithms based on Web-hyperlinks, analyzes some problems existing in this algorithm and proposes two improvements in algorithms.

Key words: Web construction mining; hyperlinks; HITS algorithms

0 引言

据统计,Internet 包含了 100 亿以上的静态网页和 5500 亿以上的动态网页,并且这个数字还在不断增长。据调查,99% 的信息对 99% 的人群是无用的;大约 85% 的 Internet 用户通过搜索引擎查找特定的 Web 信息。对于搜索引擎的评价标准一般是查全率和查准率。由于技术上的原因,目前基于关键字的搜索引擎存在一些问题,例如返回的文档数量过于庞大,其中许多文档与目标话题的相关性并不大;或者与话题相关的网页并不包含相应的关键字等等。问题的根源是:传统的搜索引擎主要采用内容分析(Content analysis),其排序算法主要基于经典信息检索技术中的相关性度量而无质量度量,无法满足用户对于检索信息相关性和准确性的要求。为了从大量半结构化数据中发现可用的模式,基于网页超链接的 Web 结构挖掘技术应运而生。

1 Web 结构挖掘

Internet 由大量 Web 站点组成,每个 Web 站点又

包含众多 Web 页面。Web 页面除了包含文本外,还包含了从一个页面指向另一个页面的超链接。超链接包含了大量潜在的注释,它有助于推动出权威性概念。Web 结构挖掘即挖掘 Web 潜在的超链接结构模式,它有助于用户找到相关主题的权威站点,并且可以概括指向与众多权威站点相关主题的站点。当一个 Web 页面作者建立指向另一页面的指针时,可以认为该作者对另一页面的认可,且两个网页的主题具有相关性。把一个页面的来自不同的作者的注解收集起来,就可以用来反映该页面的重要性,并可以很容易地用于权威 Web 页面的发现。因此,大量的 Web 超链接信息提供了丰富的关于 Web 内容相关性、质量和结构等方面的信息。

通过挖掘 Web 结构信息,可以揭示许多蕴藏在 Web 内容中的有用信息。如 URL 可以反映页面的类型及在存储位置和内容方面的层次关系;页内链接主要是用于对包含大量内容的 Web 页起到页面导航的作用,通过分析其结构,可以得到其结构特征,并可用于寻找相关的页面集合;Web 页之间的超链接结构说明了 Web 页的权威性,如指向一个文档的超链接

收稿日期:2007-04-30

作者简介:黄英铭(1969-),男,广东揭阳人,揭阳职业技术学院计算机中心讲师,研究方向:数据安全,数据库技术。

体现了该文档被引用的情况。如果大量的链接都指向某一 Web 页,就可以认为它是一个权威 Web 页。

当前,有许多研究机构对 Web 上超文本系统的链接结构进行了大量的研究,分析的主要方法是将 Web 映射成有向图或无向图的形式,根据一定的启发规则,用图论的方法对其进行分析。该方法的研究成果已经在信息检索领域得到了良好的运用。目前已有许多有关 Web 结构分析的算法,其中最有名的是 Brin 和 Page 提出的 PageRank 算法,以及 Kleinberg 提出的 HITS 算法。PageRank 的理论基础是:忽略 Web 页上的文本和其它内容,只考虑页面的超链接,把 Web 看成是一个巨大的有向图 $G(V, E)$, 结点 $v \in V$ 代表一个 Web 页面,有向边 $(p, q) \in E$ 代表从结点 p 指向结点 q 的超链接,结点 p 的出度是指从页面 p 出发的超链接(outlink)的总数,而入度是指所有指向结点 p 的超链接(inlink)的总数。

PageRank 算法效率高且计算简单,但由于忽略 Web 的内容,存在一定的缺陷。因此, Kleinberg 提出了 HITS 算法来评定网页内容的重要性,并以此作为核心技术解决了搜索引擎 Clever 的检索结果相关度的排序问题。

2 HITS 算法

2.1 HITS 算法基本思想

HITS (Hypertext Induced Topic Search) 算法是 Kleinberg 于 1999 年提出的关于超链接的检索算法。该算法通过对网络中超链接的分析,利用页面的被引用次数及其链接数目来决定不同网页的权威性。HITS 涉及两个重要的概念:

Authority: 表示一个权威网页被其它网页所引用的数量,即该权威网页的入度值。若某网页被引用的数量越大,则该网页的入度值越大, Authority 越大;

Hub: 表示一个 Web 页面指向其它网页的数量,即该 Web 页的出度值,它提供了指向权威页面的链接集合。若某网页的出度值越大,则该网页的 Hub 值越大。Hub 起到了隐含说明某话题权威页面的作用。

Hub 页面本身可能并不突出,但它却提供了指向就某个公共话题而言最为突出的站点链接,如一门课程主页上推荐的参考文献站点。一般情况下,好的 Hub 页是指向许多具有较高 Authority 值的页面;反过来,好的 Authority 页是由许多具有较高的 Hub 值所指向的页面。这种 Authority 和 Hub 的相互作用可用于权威网页的挖掘和高质量 Web 结构和资源的自动发现,这就是 HITS 算法的基本思想。Hub 和 Au-

thority 的关系可以用图 1 来表示。

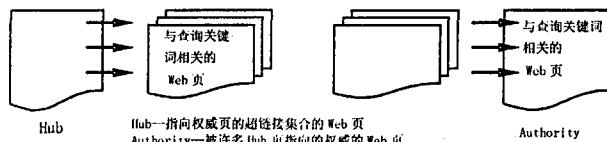


图 1 Hub 和 Authority 的关系

2.2 HITS 算法步骤

首先, HITS 由查询词 Q 得到一初始查询结果集,如基于搜索引擎得到 200 个 Web 页面。这些页面构成了根集 (Root Set), 记为 R_Q 。由于这些页面中的许多页面是假定与搜索内容相关的,因此它们中应包含指向最权威页面的指针。所以, R_Q 可以进一步扩展为基本集 (Base Set), 记为 S_Q 。包含了所有由 R_Q 中的页所指向的页,以及所有指向根集页的页。可以为 S_Q 设定一个上限,如 1000 个 Web 页,用于指明扩展的一个尺度。

接着,开始权重传播阶段,这是一个递归过程,用于决定 Hub 与权威权重的值。具体操作如下:

(1) 为基本集中的每个页面赋予一个非负的权威权重 a_p 和非负的 Hub 权重 h_p , 并将所有的 a 和 h 值初始化为同一个常数,如 $a_p = 1, h_p = 1$ 。

(2) Hub 与 Authority 的权重可按如下公式进行迭代计算:

$$a_p = \sum_{(q \text{ 满足 } q \rightarrow p)} h_q \quad (1)$$

$$h_p = \sum_{(q \text{ 满足 } q \rightarrow p)} a_q \quad (2)$$

式(1)反映了若一个页面由许多好的 Hub 所指,则其权威权重会相应增加(即增加为所有指向它的页面的现有 Hub 权重之和)。

式(2)反映了若一个页面指向许多好的权威页,则 Hub 权重也会相应增加(即权重增加为该页面链接的所有页面的权威权重之和)。

(3) 每次迭代后使用下面公式进行规范化处理,保证不变性:

$$\sum_p (a_p)^2 = 1 \quad (3)$$

$$\sum_p (h_p)^2 = 1 \quad (4)$$

(4) 当 a 和 h 值没有收敛时,转向(2)

(5) 实验证明,经过大约 10 ~ 15 次迭代计算, a 和 h 值将趋于稳定,迭代结束。此时可设置阈值 T , 将所有 a 和 h 大于 T 的网页挑选出来,排序输出查询结果。

实践证明,该算法对于许多查询具有良好的查准率和查全率。

3 HITS 算法存在的主要问题

虽然 HITS 算法取得了很大的成功,但也存在着一些问题,主要有:

(1) 容易发生主题偏移。由于 HITS 算法局限于 Web 页面之间的链接关系,忽略了页面的内容,在应用过程中表现出不稳定性,有时会出现主题偏移问题。一般地,一个好的 Hub 应指向许多好的 Authority 页面,但是,如果一个 Hub 包含有多个主题,就会有許多链接指向不相关的 Web 页面,这些不相关页面的 Authority 权重虽然较低,但是 HITS 算法将它们 Authority 权重之和作为该 Hub 的权重,该权重就可能很大,而实际上只有很少链接指向有价值的 Authority 页面。

(2) 容易产生不合理结果。

作为一种基于页面粒度的算法, HITS 往往意味着给不同的文档或 Web 站点作者规定不平等的影响权重,导致筛选出来的结果不尽客观合理。

(3) 无链接的影响。

通常情况下,一个页面上的链接并不都与主题有关,它包括站点内的导航链接,广告链接等,这些链接对权威没有贡献,若不从权重传播分析中去除,将会极大影响 HITS 算法的效果。但有些链接的过滤比较复杂,技术要求较高,不容易过滤。

(4) 无关页面的影响。

无关页面的引入有两个途径:一是基于相似度的搜索引擎返回的根集中就包含无关页面;二是根据链接关系生成基本集时引入的。由于 HITS 算法只是简单地根据链接关系确定权重,缺乏对页面有效性的判断,容易造成无关页面获得较大的 Hub 权重和 Authority 权重,从而导致输出的 Hub 页面和 Authority 页面与查询主题无关。

4 HITS 算法的改进

针对 HITS 存在的诸多问题,许多研究者提出了许多改进算法,并取得了一些成果,例如通过引入页面文本的语义信息解决主题漂移问题;通过对超链接引入相关权值来修正 Authority 权值和 Hub 权值;将大的 Hub 分裂成较小的单元等等。从实验结果来看,这些改进算法确实能获得比较满意的结果。下面介绍几个行之有效的改进方案。

4.1 Hub 权重算法的改进

当一个 Hub 页面链接到许多无关 Web 页时,它

的 Hub 权重一般较高,导致输出的 Hub 页面偏离主题。这是由于 HITS 算法在计算 Hub 权重时,采用 Hub 所指向的所有 Authority 权重之和所致。为此,可将公式 (2) 修改为:

$$h_p = \left(\sum_{(q \text{ 满足 } q \rightarrow p)} a_q \right) / n \quad (5)$$

其中, n 表示 q 的总数,即该 Hub 指向的 Web 页数。该式将 Hub 权重的计算由原 Authority 权重之和改为 Authority 之和的平均值,这样可以避免由于 Hub 指向较多无关页面而造成 Hub 权重的异常升高,影响查准率和查全率。

4.2 超链权值计算优化

对超链接上下文的分析主要集中在链接地址和链接描述上。通过对链接文本的语义分析,容易去除站点内的导航链接。设有一链接 L , 赋予其权重为 L_q 。如果 L 为导航链接,则令 $L_q = 0$, 否则,对 L 的链接文本 T (即 HTML 标记 $\langle a \rangle T \langle a \rangle$ 间的文本) 进行分析。文本 T 通常是该链接所指向 Web 页面的主题概括,对 T 进行分析比分析它所指向的整个页面显得简单而精确。根据搜索主题和链接文本 T 的匹配程度可以确定链接权重 L , 而匹配程度的计算要用到较复杂的技术,如语义分析,文本挖掘技术等,这样当然也增加了算法的复杂度。

设查询关键词 Q 在链接文本 T 中出现的次数为 N_Q , 则权重 L 可计算如下:

$$L_q = N_Q + 1 \quad (6)$$

这样, (5) 式可以改写为:

$$h_p = \left(\sum_{(q \text{ 满足 } q \rightarrow p)} a_q L_q \right) / n \quad (7)$$

(1) 式也可以改写为:

$$a_p = \sum_{(q \text{ 满足 } q \rightarrow p)} h_q L_q \quad (8)$$

通过对链接地址和链接文本的分析,产生链接权重 L_q 作为 Hub 和 Authority 权重的系数,可以有效地解决无链接的影响。

Clever 系统利用超链文字及其周围文字计算超链的权值,将页面文本的语义信息引入 THIS 算法,收到了一定的效果。在某些情况下,仅仅依靠超链文本 T 进行词频分析还是不够的,如果同时对超链文本周围的文字的词频信息进行分析能够提高算法的精度。然而,由于网络信息的异常复杂化,有时超链周围的文字无法代表链宿页面的内容,甚至与链宿相差径庭。此时,应通过适当的系数控制其对超链接权值的影响。设 S_q 为查询关键字在超链接周围文字中出现的次数,系数 α 用于控制周围文字的语义信息在超链接权值中的比例,则式 (6) 可以进一步修改为: (下转第 37 页)

务质量的阈值

```

if ( simbas > f1 && simfunc > f2 ) {
  If ( reS.QoS = null )
    matchList.append ( adS ) ;
  else if ( simQoS > f3 )
    matchList.append ( adS ) ;
  }
}

matchListResult = sort ( matchList ) ;
//按服务综合相似度降序排列
return matchListResult ;
}

```

在服务调用时,如果最佳服务暂时不能参加此次服务处理操作,则系统可以选择次优的服务,这免去了重新查找服务的麻烦。

5 结束语

本文通过采用 OWL-S 作为服务描述语言,为 Web 服务添加丰富的语义描述,使服务提供者、服务

请求者以及服务处理程序都能充分理解服务,并以此构建基于语义的 Web 服务发现模型,在模型中将服务请求进行语义解析,生成服务请求文档,在服务注册库中通过进行服务文本匹配、功能匹配、服务质量匹配,得到候选服务集,借助线形评价标准对所选服务进行语义评估,从中选择最优的服务;最后使用服务查准率和查全率作为度量 Web 服务发现性能指标,对 Web 服务发现机制进行分析。

参考文献:

- [1] 满君丰,杨伟丰,朱艳辉,等. 语义 Web 服务的相似性方法研究[J]. 计算机应用与软件,2006,23(10):15~17.
- [2] 沈玮韡,蔡鸿明,姜丽红. 一种基于语义 Web 服务的服务自动发现的实现[J]. 计算机工程,2006,32(18):211~213.
- [3] 仲梅,宋顺林. 一种语义 Web 服务的多层次匹配方法[J]. 计算机应用,2007,27(1):199~201,204.

(上接第 25 页)

$$L_Q = N_Q + 1 + \alpha * S_Q \quad (9)$$

将式(9)代入式(7)和式(8),可以获得更好的查询效果。

5 结束语

Web 是一个海量信息库,蕴含着巨量的信息资源,同时包含了丰富的动态和静态超链接信息,为数据挖掘提供了丰富资源。传统的搜索引擎往往基于文本检索技术,而对于一些文字描述较少的页面无法有效检索,通过对超链接结构的分析可以克服这一点。

HITS 算法是 Web 结构挖掘中重要的算法之一,针对该算法存在的一些问题,许多学者提出了各种改进算法。HITS 的改进算法有许多种,并且还在不断研究发展中。通过改进的 HITS 算法,可以获得较高的查询精确度,当然,也可能增加了算法的复杂度。如何改进 HITS 算法,使其具有较高查询准率和查全率,同时又能降低算法的复杂度,这应是 HITS 算法研究的方向。

参考文献:

- [1] Jiawei Han, Micheline Kamber. Data Mining: Concepts and Techniques[M]. Morgan Kaufmann Publishers, Inc., 2001.
- [2] Henzinger M. Hyperlink analysis for the Web[J]. IEEE Internet Computing, 2001, 5(1):45-50.
- [3] Linoff G S, Brry M J A. Mining the Web: Transforming Customer Data into Customer Value[M]. Publishing House of Electronics Industry, 2002.
- [4] Sepandar D Kanmvar, Taher H Haveliwala, Christopher D Manning, et al. Exploiting the Block Structure of the Web for Computing Pagerank[R]. Stanford: Stanford University, 2003.
- [5] 王艳华,张纪. Web 结构挖掘及其算法[J]. 计算机工程, 2005(21).
- [6] 杨炳儒,李岩,等. Web 结构挖掘[J]. 计算机工程, 2003, 29(20):28~30.
- [7] 韩亚洪,许卓明,等. Web 信息检索中主题精选算法的研究与改进[J]. 计算机工程与应用, 2004, 40(17):174~178.
- [8] 钟敏娟,林亚平,等. 基于超链接和标记文本的信息检索算法[J]. 小型微型计算机系统, 2004, 25(7):1344~1347.