一种基于相似度分析的主题提取和发现算法*

王晓宇[1,3+], 熊 方[1], 凌 波[1], 周傲英[1,2]

[1](复旦大学 计算机科学与工程系,上海 200433)

[2](复旦大学 智能信息处理开放实验室,上海 200433)

[3](同济大学 汽车电子研究所,上海 200092)

A Similarity-Based Algorithm for Topic Exploration and Distillation

WANG Xiao-Yu[1,3+], XIONG Fang[1], LING Bo[1], ZHOU Ao-Ying[1,2]

[1](Department of Computer Science and Engineering, Fudan University, Shanghai 200433, China)

[2](The Laboratory for Intelligent Information Processing, Fudan University, Shanghai 200433, China)

[3](Institute of Vehicle Electronics, Tongji University, Shanghai 200092, China)

+ Corresponding author: E-mail: xiaoyu_w@yahoo.com; xiaoyuwang@fudan.edu.cn

http://www.fudan.edu.cn

Received 2002-06-05; Accepted 2002-08-14

Wang XY, Xiong F, Ling B, Zhou AY. A similarity-based algorithm for topic exploration and distillation. *Journal of Software*, 2003,14(9):1578~1585.

http://www.jos.org.cn/1000-9825/14/1578.htm

Abstract: In this paper, the authors attempt to revisit the behaviour of HITS from a different point of view. Namely, a similarity-based analysis model is proposed to observe the distillation procedure. By defining a generalized similarity, an algorithm is presented, which can improve the quality of distillation using only hyperlinks. A topic exploration function is also integrated into the algorithm framework, which enables end-users to search less popular topics when multi-topics are involved in queries. The experimental results reveal two benefits from the new algorithm: the improvement of distillation quality without utilizing any content information of pages, and an additional ability to explore the topics emerging in the query results.

Key words: topic distillation; topic exploration; linkage analysis; Web searching

摘 要: 试图从另一个角度来考察主题提取算法 HITS,即提出一种基于相似度的链接分析模型来观察主题提取的过程.通过给出一种一般化的相似度定义,提出了一种仅使用链接分析来改善主题提取的质量的主题提取算法.同时,还将主题发现的功能也结合到了算法的框架中.通过该功能,用户可以搜索到次流行的主题.实验结果显示了这一新算法的两个优点:不必使用内容分析即能改善主题提取的质量以及能够进一步发现在查询结果中显现出来的不同主题.

关键词: 主题提取;主题发现;链接分析;Web 搜索

* Supported by the National Natural Science Foundation of China under Grant No.60003016 (国家自然科学基金); the National Grand Fundamental Research 973 Program of China under Grant No.G1998030404 (国家重点基础研究发展规划(973))

第一作者简介: 王晓宇(1975－),男,安徽濉溪人,博士,讲师,主要研究领域为数据库技术,互联网环境下的数据检索.

中图法分类号: TP311 文献标识码: A

在万维网上,对于给定的用户查询,搜索引擎典型的做法是返回大量与查询关键词相匹配的文档;但用户所愿意浏览的只是其中的极小一部分.许多研究者试图针对特定的查询,确定文档的相对的权威值[1~5],依据这种权威值搜索引擎返回给用户那些权威值最高的文档.这种找出高质量页面的过程称为主题提取(topic distillation)[1].但是,在互联网环境中,提交给搜索引擎的查询常常是不明确的,有时候包含了若干个可能的主题.在许多情况下,用户可能更有兴趣找出与给定查询相关的几个主题,但是主题提取的目标是找出那些最流行的主题的权威网页,也就是说,它排除了与查询关键词相关的其他可能主题.因此,有必要在主题提取的过程中同时找出所有可能的查询主题,这种功能我们称为主题发现(topic exploration).

著名的 HITS 算法是一种有效的基于链接分析的主题提取方法,它所依赖的是对超链接环境下链接结构的分析[3].但在过去几年中,由几位研究者所继续的实验却显示出 HITS 算法提取质量的恶化[1,4~6].因此,他们试图对 HITS 算法进行改进以避免主题漂移(topic drift)问题.所采取的方法是对文档进行内容分析.

许多人也许认为,内容与链接相结合的混合型方法是对纯链接分析方法研究的终结,但实际的情形却更像是一种相持不下的军备竞赛.在混合型方法中,需要获取并存储整个页面,还要对内容文本进行处理,这些都增加了算法额外的负担.而且,当查询中蕴含若干主题时,这些算法都放弃了那些次流行的主题,我们把这种缺陷称为主题遗失(topic missing).

对于主题提取与主题发现,文档之间的链接信息是一个丰富的资源,应该被我们更好地利用.我们对 HITS 算法进行了透彻的分析,试图换一种角度来重新看待它的行为.也就是说,我们将提出一个基于相似度的分析模型来观察其提取的过程.我们在技术上的主要贡献是一种新的基于相似度的主题提取算法,它仅使用链接的信息而能改善提取的质量;主题发现的功能也结合到了这一算法的框架中,它使得在查询蕴含多个主题的时候,用户能够搜索到那些次流行的主题.

1 相关工作

许多研究者都提出了使用链接结构进行主题提取的方案[1,3~5,7].而其中最为著名的则莫过于 HITS 算法[3].一些研究者对于这一基本算法进行了扩展.Chakrabarti 等人针对特定主题的查询,基于超链接周围文本的相似度,对链接赋予了权重[4].Bharat 和 Henzinger 作了几个重要的扩充:首先,他们根据文档与查询主题的相似度,对文档赋予了权重;其次,他们仅考虑来自不同域名的文档之间的链接,并且把来自同一个域名的全部链接所产生的影响平均到某一特定文档[1].Chakrabarti 提出了一种统一的细粒度模型:在网上所有的页面都由它们的标记树来表示,这样在权威性页面(authority)和中心性页面(hub)之间的相互影响就包含了 DOM 树中的块[5].

这些扩展的方法都致力于解决在 HITS 算法中遇到的主题漂移问题[3],尽管所采用的技术各有不同[1,4,5],他们都使用了文档的内容.我们的主要目标也和这些扩展方法相同,但仅使用链接结构的信息来解决主题漂移问题,因此避免了获取页面以及内容分析的额外负担.

2 预备知识与分析模型

2.1 对HITS算法的回顾

在 HITS 算法中,对每个文档都要计算两个值:权威值(authority)与中心值(hub)[3].开始时,由用户发出查询,HITS 算法使用一个基于文本的搜索引擎,得到许多被返回的页面,构成根集合(root set)R.把根集合中的页面所指向的页面都包括进来,再把指向根集合中的页面的页面也包括进来,这样就扩充成了基础集合(base set)T.处于同一个域的各个页面之间的链接仅仅起导航的作用,应不作考虑.

假定所得到的图是 $G=(V,E)$.在 V 中的每个页面 p 都有一对非负的权重值$\langle a_p,h_p\rangle$,其中 a_p 表示权威值,h_p 表示中心值.所有的权威值和中心值都初始化成 1.设指向页面 p 的页面为 q,a_p 的值则更新为所有 h_q 的和:$a_p=\sum_{q|q\to p}h_q$;与此严格对应的是,如果把页面 p 所指向的页面称为 q,则 h_p 的值更新为所有 a_q 之

和:$h_p=\sum_{q|q\to p}a_q$. 这两步将被重复多次,最后按照得到的权威值和中心值对页面进行排序.如果图被表示成为相邻矩阵 A 的形式(即当且仅当在图中 i 指向 j 时,矩阵 A 中的第(i,j)项为 1,否则为 0),则以上的操作可以被写成矩阵运算的形式:$a\leftarrow A^Th\leftarrow A^TA\ a=(A^TA)a$ 和 $h\leftarrow Aa\leftarrow AA^Th=(AA^T)h$,并且与规范化的过程相交替,以保持$|a|=|h|=1$.因此,经过多次迭代所得到的向量 a 也就是对 A^TA 进行幂迭代所得到的结果,而且,这在矩阵计算[8]中有一个标准的结论.

2.2 分析模型

这个模型的基础思想就是,那些有协同引用(指向同一页面)或有联结(被同一页面所指)的页面很有可能在语义上是相关的.

现在我们对基础集合 T 中的那些页面进行编号:{1,2,…,n},并且把每个页面都表示成为两个 n 维向量:v^{out} 和 v^{in}.给定 T 中的页面 p,向量 v_p^{out} 中的第 i 个分量为 1 当且仅当页面 p 指向页面 i,否则为 0.同样地,向量 v_p^{in} 的第 i 个分量为 1 当且仅当页面 p 被页面 i 所指,否则为 0.这里假设没有页面 p 指向自身的链接,即,向量 v_i 的第 i 个分量为 0.内积被用来对页面对之间的语义关系进行度量:$Similarity^{\text{in}}(p,q)=v_p^{\text{in}}\cdot v_q^{\text{in}}$,$Similarity^{\text{out}}(p,q)=v_p^{\text{out}}\cdot v_p^{\text{out}}$.

很明显,假设 p 和 q 是不同的页面,$Similarity^{\text{out}}(p,q)$表示那些为 p 和 q 共同所指的页面的数量,而 $Similarity^{\text{in}}(p,q)$表示的是那些既指向 p 又指向 q 的页面的数量.这里,相似度的意思类似于在协同引用分析中的协同引用强度(co-citation strength).如果 p 与 q 相同,则两个相似度的值分别表示该页面的链出数和链入数.实际上,以 $Similarity(i,j)$作为第(i,j)项的相似矩阵 S,恰恰就是在 HITS 算法中幂迭代的基础,其准确的表达如下:

定理 2.1. 存在相似矩阵 S^{in} 和 S^{out},其项分别为 $S^{\text{in}}(i,j)=Similarity^{\text{in}}(i,j)$和 $S^{\text{out}}(i,j)=Similarity^{\text{out}}(i,j)$,有 $S^{\text{in}}=A^TA$,$S^{\text{out}}=AA^T$.这里,A 是在 HITS 算法中对于给定查询的图的相邻矩阵.

接下来,由 S^{in} 和 S^{out} 的 k 次幂矩阵$(S^{\text{in}})^k$和$(S^{\text{out}})^k$,k 阶相似度 $Similarity_k^{\text{in}}(i,j)$和 $Similarity_k^{\text{out}}(i,j)$也可随之被定义,有 $Similarity_k^{\text{in}}(i,j)=(S^{\text{in}})^k(i,j)$以及 $Similarity_k^{\text{out}}(i,j)=(S^{\text{out}})^k(i,j)$.当 k 为 1 时,$Similarity_k^{\text{in}}(i,j)$和 $Similarity_k^{\text{out}}(i,j)$即是最初的相似度定义:$Similarity^{\text{in}}(i,j)$和 $Similarity^{\text{out}}(i,j)$,这在后面分别记为 $Similarity_1^{\text{in}}(i,j)$和 $Similarity_1^{\text{out}}(i,j)$.

在 HITS 算法中经过第 k 次迭代,权威性向量 a_k 和中心性向量 h_k 分别是在$(A^TA)^ku$ 方向和$(AA^T)^ku$ 方向的单位向量,其中,u 表示向量{1,1,…,1}.根据定理 2.1,我们可以得到 $a_k(i)=\sum_{j=1}^{n}Similarrity_k^{\text{in}}(i,j)$ 和 $h_k(i)=\sum_{j=1}^{n}Similarrity_k^{\text{out}}(i,j)$.

我们现在可以来定义无向带权图 G_s^{in}.以基础集合 T 中的节点作为图的顶点集;对两个顶点 i 和 j,如果 $Similarity_1^{\text{in}}(i,j)$不为 0,就在它们之间画一条边,赋予权重值 $Similarity_1^{\text{in}}(i,j)$.当 $i=j$ 且 $Similarity_1^{\text{in}}(i,j)$不为 0 时,就在该节点上画一条自环.在图 G_s^{in} 中,若节点 i 与节点 j 之间存在 u 条不同的长度为 k 的路径,把每条路径上的边的权重之和记为 σ_k^{ij},则由图论中邻接矩阵的幂定理可知,$Similarity_k^{\text{in}}(i,j)=\sum_{t=1}^{u}\sigma_k^{ij}(t)$.图 G_s^{out} 和相似度 $Similarity_k^{\text{out}}(i,j)$也可以用类似的方法来定义.总结一下上述的讨论,可以形成如下的引理和定理:

引理 2.1. 如果在图 G_s^{in} 中,节点 i 与 j 之间存在 u 条长度为 k 的路径,则这两个页面之间的 k 阶相似度 $Similarity_k^{\text{in}}(i,j)$等于 $\sum_{t=1}^{u}\sigma_k^{ij}(t)$.如果在图 G_s^{out} 中,节点 i 与 j 之间存在 v 条长度为 k 的路径,则这两个页面之间的 k 阶相似度 $Similarity_k^{\text{out}}(i,j)$等于 $\sum_{t=1}^{v}\sigma_k^{ij}(t)$.

定理 2.2. 给定一个包含 n 个页面的基础集合,在 HITS 算法中经过 k 次迭代所得到的权威性向量 a_k 的第 i 个分量,等于 $\sum_{j=1}^{n}\sum_{t=1}^{u}\sigma_k^{ij}(t)$,其中 u 表示在图 G_s^{in} 中节点 i 和 j 之间长度为 k 的路径的数目.在 HITS 算法中经过 k 次迭代所得到的中心性向量 h_k 的第 i 个分量,等于 $\sum_{j=1}^{n}\sum_{t=1}^{v}\sigma_k^{ij}(t)$,其中 v 表示在图 G_s^{out} 中节点 i 和 j 之间长度为 k 的路径的数目.

3 问题的陈述

在带权图中,两节点之间长度为 k 的路径的 σ_k 值可以被看成是,在基础集合中的两页面间的以 k 为半径的相似度的传递.在很多与特定查询相关的图中,主题极易漂移,使得这种相似度的传递很容易被扭曲,而由引理 2.1 我们可以知道,这实际上就是对高阶相似度的扭曲.根据定理 2.2 我们不难了解,当很多存在于节点对之间的高阶相似度被扭曲的时候,往往就会发生主题漂移的问题.

由定理 2.2 可以看出,在 HITS 算法中所得到的权威性页面和中心性页面,实际上就是由 n 个节点所构成的图 G_s 中,所有节点间的 n 阶两两相似度所决定.这种节点对之间的相似度的传递,最终导致了高阶相似度的扭曲,进而错误地得到对网页权威性的排序.定理 2.2 还告诉我们,HITS 算法是致力于找出那些有最高 n 阶相似度的页面,其比较的范围为整个带权图.但是,通过实验我们发现,如果一个查询蕴含多个主题,那么在图 G_s 中会有多个主连通分支.在这种情况下,HITS 算法的这种全局的比较方法很难发现所有可能与查询有关的主题.

4 使用关联规则的广义相似度定义

为了缓解高阶相似度的扭曲,对于基数大于 2 的集合,我们提出了一种广义的协同引用相似度的定义.如果涉及的对象超过了两个节点,那么找出协同引用的问题可以被看做是,在关联规则挖掘中的频繁项集(frequent itemset)的生成.这里,由所有页面组成的集合对应于项集,而对基础集合中每一页面的引用对应于事务,使用关联规则发现算法而得到的频繁项集则对应于那些有共同引用或被共同引用的页面的集合.频繁项集所捕捉的是那些个数大于 2 的项之间的关系.

要度量在频繁项集中的页面间的相互关系,一个可行的办法是使用每个频繁项集的支持率(support),它体现了这些页面的协同引用的强度.另外一个办法是,以频繁项集相关的关联规则的置信度(confidence)为参数,定义一个函数.在集合规模为 2 时,两种方法的效果差不多.实际上,如果两个项 A 和 B 在所有事务中出现的次数相同,那在支持率和规则的置信度之间有直接的联系.但是,如果集合的规模超过 2,使用支持率的方法所能提供的信息要少得多,因为一般来讲,集合规模越大,支持率就越低.而且,对于那些较大的频繁项集,使用其相关的关联规则的置信度可以反映项之间的相互关系,而使用支持率则不会.

若以数学公式来表达,给定某个频繁项集 I,如果其基本关联规则的置信度是 $\{\mu_1,\mu_2,\ldots,\mu_k\}$,那么其关联规则度量则为

$$\varepsilon(I)=\left(\sum_{t=1}^{k}\mu_t\right)\Big/k\ .$$

页面中心值或权威值的计算,是以一个无向带权图为基础的,其中的边表示节点对之间的一阶相似度.但是,任意基数的集合中的相似度度量要求,将相似度图中的边推广到超边,也就是说,一条边可能涉及到超过两个的顶点.尽管把中心值或权威值的计算推广成这样一个相似度的超图是比较困难的,我们可以先采取一种较简易的混合型的方式来实现我们的算法,它是在标准的两两节点间相似度度量的基础上,通过把所有包含此节点对的页面集合的关联规则度量相加,来显现任意基数集合的相似度度量的特性.形式化的表示如下:

$$\zeta_1(i,j)=\sum_{\{I|i,j\in I\}}\varepsilon(I)\ .$$

这样的度量方式很容易被用来计算页面的中心值或权威值.这里,一个节点对于它自身的一阶相似度我们并未重新定义.

5 算 法

5.1 主题提取中的基本迭代算法

我们所提出的一阶相似度是用来构建无向带权图 G_s^{in} 和 G_s^{out} 的,它们是计算页面权威值和中心值的基础.而构建图 G_s 的相似矩阵 S 则是通过下面的两个步骤:

(1) 对于矩阵 S 中的项 $S(i,j)$,如果 i 等于 j,则有 $S(i,j)=Similarity_1(i,j)$;

(2) 对于矩阵 S 中的项 $S(i,j)$,如果 i 不等于 j,则有 $S(i,j)=\zeta_1(i,j)$.

在算法中,我们将使用迭代操作,一般地,我们将迭代步数设为 200.通过这样的迭代过程,我们可以得到最具代表性的返回页面.

5.2 控制主题漂移的参数

这种基于广义相似度定义的迭代操作,在很多情况下可以有效地改善主题漂移的问题.但有些查询在某些相关主题上极易发生漂移.受那些将内容与链接分析相结合方法的启发,我们可以对上节所提出的一阶相似度

定义作进一步的扩展,以便更有效地控制主题漂移的问题.大多数既使用内容又使用链接信息的混合型方法都基于这样的假设,在根集合中的页面是不太容易偏离查询主题的,这也是我们控制主题漂移问题的基准.但那些方法是以根集合中的网页为基准的,通过内容分析在图中删去与根集合文本相似度较小的邻近节点,而我们则试图在一阶相似度的定义中包含此基准信息.经过扩展的一阶相似度定义如下:

给定全部由关联规则挖掘算法所发现的频繁项集,我们以 I_{root} 表示那些包含了根集合中页面的频繁项集,而以 $I_{neighbor}$ 表示那些不包含根集合中页面的频繁项集,则一阶相似度可扩展定义为

$$\zeta_1(i,j)=\sum_{\{I_{root}|i,j\in I_{root}\}}\varepsilon(I_{root})+\delta\cdot\sum_{\{I_{neighbor}|i,j\in I_{neighbor}\}}\varepsilon(I_{neighbor}).$$

其中,$i\neq j$ 并且 $0\leq\delta\leq1$.δ 称为漂移控制因子,它反映了在多大程度上考虑那些未涉及根集合的邻近页面间关系.如果 δ=0,在算法中就完全不考虑那些不涉及根集合中页面的关系.如果 δ=1,就对所有的页面间关系作相等的考虑.理论上来说,δ 的值越小,对涉及到根集合之外的页面的关系就考虑得越少.通过实验我们发现,对于大多数相关主题范围较广的查询,在 δ=0 的时候就可以返回那些具有代表性的页面了.如果给定查询的相关主题范围较窄,我们可以将 δ 赋一个较大的值,这样就能考虑到更多的关系了.在大多数情况下,这种算法被证明是能够有效改善主题漂移问题的.

5.3 主题发现算法

在许多情况下,我们需要发现与给定查询相关的若干个主题.但是,正如我们在第 2.2 节中所讨论的,当在与特定查询相关的图中出现了几个主题时,HITS 算法容易遗失那些次流行的主题.而使用我们的可视化工具,如果给定的查询与多个主题相关,那么在无向带权图 G_s 中可以观察到多个主连通分支.这里,我们假设图 G_s 中的主连通分支对应于给定查询的潜在的相关主题,那么,首要的是发现图中这些主要的主题社区,而不是直接对图 G_s 进行迭代操作.任何的连通分支搜索算法(connected component search,简称 CCS)都可以运用于我们的算法框架中,其输入为图 G_s 的相似矩阵 S,其输出为所有的连通分支的集合$\{CC_i\}$.

算法 1. 主题发现算法.

```
Exploration(S,τ)
S: the matrix of the undirected weighted graph G_s
τ: a natural number
begin procedure Exploration
{CC_i}=CCS(S)
j=0
for each element CC_i in {CC_i}
        if |CC_i|>τ
        // τ must be set no larger than max (|CC_i|).
        // |CC_i| is the number of nodes in the connected component CC_i.
Construct matrix T_j with the nodes in set CC_i following their relationships in matrix S
        j++
        endif
endfor
return {T_j}
end procedure Exploration
```

其中,参数 τ 是一个门限值,如果连通分支中所包含的节点数小于 τ 就把该连同分支丢掉.这种操作基于如下假设:在图 G_s 中,包含的节点数较少的连通分支,所携带的与查询主题有关的信息也会较少.用户可以根据其查询要求来设置 τ 的值,如果需要较详细的搜索,τ 的值可以设得低一点,如果只是要关于该查询的最流行的信息,则门限值就可以设得高一点.

6 实验分析

我们实现了 TED 算法和 HITS 算法.实验中基础集合的生成是遵循 Kleinberg 的论文[3]中所提出的过程.

6.1 总体评估

为了对 TED 算法和 HITS 算法的排序进行比较,我们在实验中设计了 18 个查询.表 1 中列出了这些查询以及用它们所生成的基础集合的大小.

Table 1 Queries with the size of base set in the experiments

表 1 实验中的查询以及它们的基础集合大小

No.	Query	Size of base set	No.	Query	Size of base set
1	Jaguar Jaguars	2 921	10	Cancer	4 111
2	Affirmative action	4 367	11	Java	5 793
3	Movie award	4 121	12	Virus	4 156
4	Table tennis	3 648	13	Abduction	2 874
5	Human rights	3 912	14	HIV	3 896
6	Latex	4 520	15	Computational geometry	1 299
7	Data mining	1 634	16	Gulf war	2 698
8	Computation complex	1 686	17	Gun control	2 337
9	Abortion	3 244	18	Bin laden	2 874

6.2 总体评估

我们通过这 18 个查询对不同算法的整体主题提取质量进行了评估.图 1 中显示了对于 TED 算法和 HITS 算法每个查询的漂移程度(drift extent).在实验中,漂移控制因子 δ 的值设为 1.漂移程度被分为 3 个等级,“大漂移”(large drift)表示有超过一半的结果偏离了主题,“小漂移”(small drift)表示不到一半的结果偏离了主题,而“无漂移”(no drift)表示所有的结果都是与查询主题相关的.

从整体的评估结果来看,在δ=1 的情况下,TED 算法就能极大地改善在 HITS 算法中所遇到的主题漂移问题.但是,有些查询主题实在极易漂移,像 Human Rights,Cancer 等,这时候的效果就不是很理想.为了更有效地控制漂移问题,我们使用不同的δ值在 TED 算法上又重新进行了实验,如图 2 中的评估结果所示,通过在 TED 算法中调节漂移控制因子δ的值,大多数情况下我们可以成功地解决主题漂移的问题.

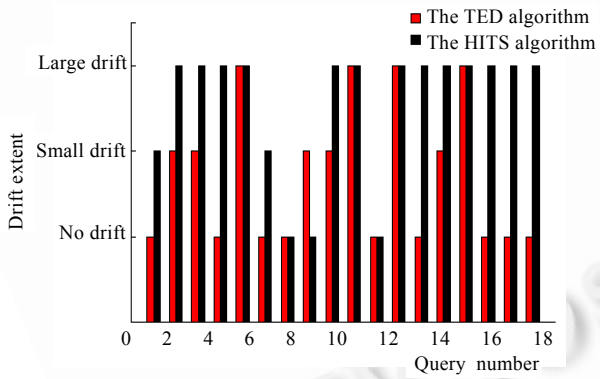

Fig.1 Drift extents of HITS and TED for each of the topics. The drift control parameter δ is equal to 1 in TED

图 1 查询的漂移程度.在 TED 算法中,漂移控制因子δ的值设为 1

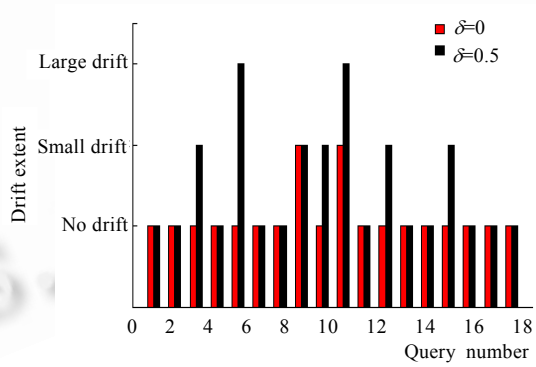

Fig.2 Drift extent of the TED algorithm on two different values of parameter δ

图 2 在 TED 算法中使用不同的δ值,所得到的每一查询的漂移程度

但是,以上对算法的评估仅仅依据于返回结果是否偏离主题,而并没有考虑算法返回结果的质量(即权威性).因此,我们设置了一个人工实验来评估 TED 算法所返回的页面质量,所选择的比较基准是 YAHOO[9].对于同样查询的返回结果,实验的参与者被要求通过使用这些返回的查询结果作为了解相关主题的一组“起点”.我们让参与者用定量的度量来评估这些返回的结果是否能有效地了解和查询相关的主题.度量被分为 10 个等级,从 1~10(10 代表非常有助于了解该查询主题).图 3 展示了 TED 算法得分和 YAHOO 得分的一个比率.y 轴的 1 刻度表示两算法没有差异,高于 1 表示 TED 算法比 YAHOO 优越,低于 1 表示 TED 算法没有 YAHOO 优越.

从平均得分比率来看,TED 算法已经相当接近 YAHOO.在不同的查询上,得分比率的差异很大,比如对诸如 Bin Laden,Human Rights 和 Affirmative Action 这样的查询,TED 的算法表现较差,这是因为这些主题在网上具有

非常丰富的资源,YAHOO 的人工分类具有更好的效果.而对于诸如 Data Mining,Computation Complex 和 Computational Geometry 这样的主题,TED 算法具有较好的表现,这是因为这些主题在网上的表现相对较少,TED 算法在这类主题上就表现出明显的优势.

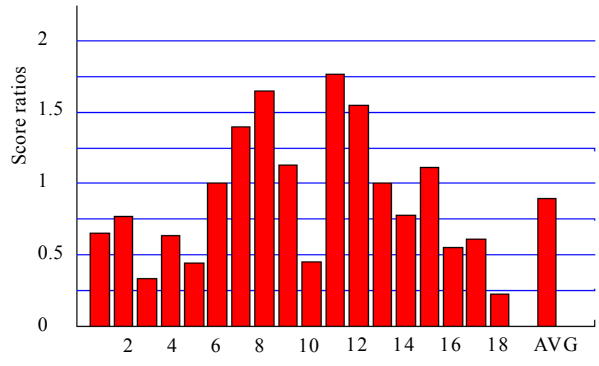

Fig.3 Ratio of quality TED scores to Yahoo scores for each of the queries. The last bar shows the average ratio of scores

图 3 TED 算法和 Yahoo 在每个查询上的得分比率,最后一条代表平均得分比率

6.3 主题发现

在这一节中,我们将详细地列举一些实验例子来说明 TED 算法中的主题发现功能.其中,引用到了 Kleinberg 论文[3]中的实验结果,其目的是要揭示在 HITS 算法中使用非主特征向量遇到的主题遗失问题.

例 1:查询主题 Jaguar 是一个很有用的查询,包括文献[3]在内的许多论文都使用了这一查询.该例的实验结果摘自文献[3].那些与主题相关的有代表性的页面是通过手工从一些非主特征向量中提取的,其中权威集合涉及了 Atari Jaguar Product,NFL Football Team from Jacksonville 和 the Jaguar Automobile 这些主题.要选择表现了这多个主题的非主特征向量,是一项手工工作,它依赖于与此查询有关的图的拓扑结构.它们分别是查询 Jaguar*和 Jaguar Jaguars.主题 Jaguar Products 是从查询 Jaguar*中得出的,而主题 Jaguar Mobile 和 Jaguar Football 是从查询 Jaguar Jaguars 中得到的.

实际上,通过我们自己的实验,我们发现在与查询 Jaguar*和 Jaguar Jaguars 相关的图中,都包含了在例 1 中发现的 3 个主题,只是 HITS 算法没有能在同一个查询子图中发现这 3 个主题.为了与我们的算法进行比较,我们用 HITS 算法实现了查询 Jaguar Jaguars,实验结果在例 2 中给出.

例 2:在我们的实验中,查询 Jaguar Jaguars 的根集合包含了 200 张页面,在扩展后得到的基础集合中有 2 921 张页面.在这次测试中,实验结果的第 2、第 3 特征向量并不含负值项.带权图 G_s^{in} 的结构则可见于图 4,A,B 和 C 分别是它的 3 个主连通分支,而 D 则是图中其他连通分支,它们所包含的节点较少,或者包含的是孤立点.在关于查询 Jaguar Jaguars 的图 G_s^{in} 中,A,B 和 C 分别包含了 290,209 和 104 个节点,而 D 则包含了 1 767 个孤立点和 21 个较小的连通分支,它们所包含的节点数在 2~38 之间不等.

值得注意的是,前述的 3 个主题在图 4 中显现为 A,B 和 C.但是,在例 1 中遇到的现象又重现了,从例 2 中只能得到其中的两个主题,即 NFL Football Team from Jacksonville 和 the Jaguar Automobile.当然,使用更多的特征向量可以得到进一步的主题信息,但是要通过手工把有价值的信息从一大堆非主特征向量中抽取出来,也是一件琐碎的事情,况且,如果矩阵的维数很高,要计算更多的非主特征向量也是不太现实的.

例 3:由 TED 算法所构建的拓扑结构图 G_s^{in} 如图 5 所示.令人惊讶的是,它包含了 4 个主连通区域,细致地观察后发现这 4 个主连通域对应表现了 4 个主题,除了上面提到的 3 个主题以外,还有 Jaguar as Mammal.基于图 G_s^{in}.

观察例 3 中的实验结果,我们发现与该查询相关的 4 个潜在主题都出现了,而且它们每一个都返回了高质量的页面.在例 3 中没有返回无关页面,与例 2 相比,表明了 TED 算法相较于 HITS 算法的优越性,即对主题漂移问题的控制及主题发现的能力.

实验的结果揭示了 HITS 算法与 TED 算法的不同之处,尤其是在计算那些最具权威性的页面时,HITS 算法只重视那些最流行的主题,而 TED 算法不仅会注意到那些次流行的页面,而且还能按不同的主题将结果分类返回.

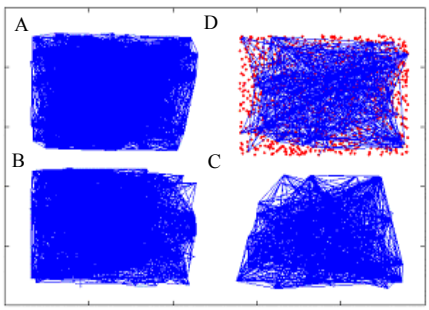

Fig.4 The topology of graph G_s^{in} of query "Jaguar Jaguars"

图 4 查询 Jaguar Jaguars 的图 G_s^{in} 的拓扑结构

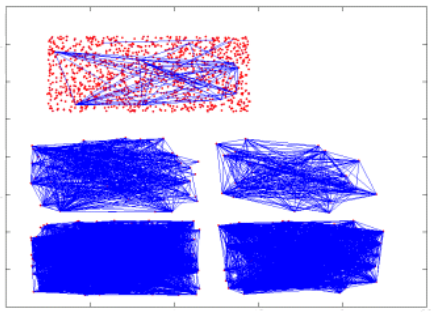

Fig.5 Topology of graph G^{in} constructed by the TED algorithm

图 5 由 TED 算法所构建的拓扑结构图 G_s^{in}

7 结论与将来的工作

本文从基于相似度的分析模型的角度重新审视了 HITS 算法.为了控制主题漂移的问题,我们在主题提取的过程中,将关联规则挖掘与常规的链接分析结合起来.这样,我们的方法即使没有用到任何文本信息,也在主题提取的质量上取得了很大的改善.而且,我们还在算法框架中加入了主题发现的功能,它使得用户可以搜索到与查询有关的次流行的主题.

对于未来的工作,我们还有另外两个想法:(1) 要从任意基数相似度的超图中得出那些具有代表性的页面,会是很有趣的工作;(2) 如果相似度的定义得到进一步的完善,我们也许可以一步得到那些具有代表性的页面,而无须若干次的迭代过程.

References:

[1] Bharat K, Henzinger M. Improved algorithms for topic distillation in a hyperlinked environment. In: Voorhees E, Kirsch S, eds. Proceedings of the 21st ACMSIGIR International Conference on Research and Development in Information Retrieval. Melbourne: ACM Press, 1998. 104~111.

[2] Brin S, Page L. The anatomy of a large-scale hypertextual Web search engine. In: Thistlewaite P, *et al.* eds. Proceedings of the 7th ACM-WWW International Conference. Brisbane: ACM Press, 1998. 107~117.

[3] Kleinberg J. Authoritative sources in a hyperlinked environment. In: Tarjan RE, Baecker T, eds. Proceedings of the 9th ACM-SIAM Symposium on Discrete Algorithms. New Orleans: ACM Press, 1997. 668~677.

[4] Chakrabarti S, Dom B, Gibson D, Kleinberg J, Raghavan P, Rajagopalan S. Automatic resource compilation by analyzing hyperlink structure and associated text. In: Thistlewaite P, *et al.* eds. Proceedings of the 7th ACM-WWW International Conference. Brisbane: ACM Press, 1998. 65~74.

[5] Chakrabarti S. Integrating the document object model with hyperlinks for Enhanced topic distillation and information extraction. In: Vincent Y S, *et al.* eds. Proceedings of the 10th ACM-WWW International Conference. Hong Kong: ACM Press, 2001. 211~220.

[6] Borodin A, Roberts G, Rosenthal J, Tsaparas P. Finding authorities and hubs from link structures on the World Wide Web. In: Vincent Y S, *et al.* eds. Proceedings of the 10th ACM-WWW International Conference. Hong Kong: ACM Press, 2001. 415~429.

[7] Davison B, Gerasoulis A, Kleisouris K, Lu Y, Seo H, Wang W, Wu B. DiscoWeb: Applying link analysis to web search (extended abstract). In: Vezza A, Maloney M, Cailliau R, eds. Proceedings of the 8th ACM-WWW International Conference. Toronto: ACM Press, 1999. 148~149.

[8] Golub GH, Van Loan CF. Matrix Computations. London: Johns Hopkins University Press, 1989. 40~45.

[9] http://www.yahoo.com. 2001.