

搜索引擎的主题提取算法研究*

郑利荣

(广东医学情报研究所, 广州 510180)

摘要: 以一个自行开发的搜索引擎系统为背景研究主题提取算法, 通过对几种经典主题提取算法的分析、融合, 提出了一个新的主题提取算法。用该搜索引擎证明了新提出算法比经典的 HITS 算法在性能上有很大的提高。

关键词: 主题提取; 搜索引擎; 链接分析

0 引言

在 WWW 中, 如何给用户返回主题最相关的网页一直是搜索引擎研究的热点。互联网链接结构往往隐含着大量与主题相关的信息, 因此, 近年来出现了许多基于分析链接结构来判断某网页是否符合查询主题的算法^[1,2,3,4]。HITS 算法^[1]是其中一个经典算法, 它在描述网页的主题相关度时引入了权威网页(Authority)和中心页面(Hub)的概念。该算法反映了权威网页和中心页面的相互加强关系。但 HITS 算法存在主题漂移(Topic Drift)问题, 文献[5,6,7,8]通过结合文本分析和链接分析来改善这个问题。PageRank 算法^[2,9]是另一个著名的链接分析算法, 它基于随机行走(Random Walks)模型, 在互联网大环境下计算网页重要度 PageRank 的值, 该值不包含主题相关信息, 它反映的是网页的流行程度。SALSA^[6]结合 PageRank 算法和 HITS 算法, 提出了一种新的链接分析随机模型方法, 该方法比 HITS 算法更有效率。主题信息提取算法中一个关键的问题是对基础集(Base Set)T 的选取, 文献[5]通过对文本进行相似度分析来优化基础集, 较好地解决了主题漂移问题。

1 相关工作

1.1 HITS 算法

HITS 算法考虑的是权威网页和中心网页之间的加强关系。每个网页都会有一个权威值和中心值与之对应, 如果某个网页有许多中心值高的网页指向它, 则它将具有高的权威值; 同样, 如果某个网页指向了

许多高权威值的网页, 那么它将具有较高的中心值。算法描述如下:

将查询 q 提交给搜索引擎, 从搜索引擎返回的网页中提取前 n 个网页作为根集, 用 S 表示。

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 , 称为基础集。

令 E 为边集, 元素 (v, u) 表示网页 v 链向网页 u , 对于 T 中的每个页面 u , 通过下式计算它的权威值 $a(u)$:

I 操作:

$$a(u) = \sum_{v:(v,u) \in E} h(v) \quad (1)$$

即页面 u 的权威值等于所有指向它的页面 v 的中心值之和, 该公式称为 I 操作。

同理, 对于 T 中的每个页面 v , 通过下式计算它的中心值 $h(v)$:

O 操作:

$$h(v) = \sum_{u:(v,u) \in E} a(u) \quad (2)$$

即页面 v 的中心值等于该页面所指向的所有页面 u 的权威值之和, 该公式称为 O 操作。

重复 I 操作和 O 操作直到 $a(u)$ 和 $h(v)$ 收敛, 每次迭代后需要对 $a(u)$, $h(v)$ 进行规范化处理。

也可用矩阵形式来描述算法。

1.2 PageRank 算法

PageRank 算法的基本思想是如果网页 u 有一个指向网页 v 的链接, 那么网页 u 隐式地传递一些“重要度”给网页 v 。算法描述如下: $B(u)$ 是指网页 u 的网

* 基金项目: 国家自然科学基金(No.60573097)、广东省自然科学基金(No.06104916)

收稿日期: 2008-05-21 修稿日期: 2008-07-05

作者简介: 郑利荣(1971-)女, 硕士, 副研究员, 研究方向为数据挖掘、信息检索等

页集合, $N(v)$ 是网页 v 指向外的链接数, c 是一个用于规范化的因子, 则网页 u 的 PageRank 值 $R(u)$ 计算如下:

$$R(u) = c \sum_{v \in B(u)} R(v) / N(v)$$

也可以以矩阵形式来描述此算法, 设 A 为一个方阵, 行与列对应网页集的网页。如果网页 u 有指向网页 v 的链接, 则 $A_{u,v} = 1/N_u$, 否则 $A_{u,v} = 0$ 。设 R 是对应网页集的一个向量, 有 $R = cAR$, 因此 R 是 A 的以 c 为特征根的特征向量。只需求出最大特征根对应的特征向量, 就是网页集对应的最终 PageRank 值, 可以用迭代的方法计算。

这里存在一个小问题: 设 2 个相互指向的网页 a, b , 它们不指向其他任何网页, 另外有某个网页 c , 指向 a 和 b 中的其中一个, 那么在迭代计算中, a 和 b 的 PageRank 值由于无法分布出去而不断地累计。

为了解决这个问题, 引入一个对应网页集的向量 $E(u)$, 它对应 PageRank 的初始值, 上式改为:

$$R'(u) = c \sum_{v \in B(u)} R(v) / N(v) + cE(u)$$

对应的矩阵形式为 $R' = c(AR' + E)$ 。

2 改进后的算法

2.1 根集的获取和扩展

HITS 算法最大的弱点是处理不好主题飘移问题; 如果在基础集 T 中有一些与查询主题无关的相互紧密链接的网页, 则算法的结果可能就是这些网页, 因为这些网页相互加强得到了很高的权威值和中心值, 从而偏离了原来的查询主题。而用 HITS 进行窄主题查询时, 则可能产生主题泛化问题; 扩展以后引入了比原来主题范畴更广的新主题, 新主题也可能与原始查询无关。泛化的原因是因为网页中包含不同主题的向外链接, 而且新主题的链接可能具有更大的权重。

根集是算法的起点, 如果根集质量不高, 则扩展后的基础集会增加很多无关的网页, 很容易产生主题漂移、主题泛化等问题, 计算量也会增多。因此本算法首先对根集的获取和扩展进行优化。

我们采用和文献[5]类似的方法: 对搜索引擎返回的网页进行相似度处理, 消除镜像网页和相似的网页, 获取根集时按包含查询关键词排序, 使得根集 S 中大多数是与查询关键词密切相关的网页。

更进一步, 我们提取根集 S 中每个网页的前

1000 个关键词, 取交集作为查询主题 Q , 网页 D_j 和主题 Q 的相似度按如下余弦相似度公式^[10]计算:

$$\text{similarity}(Q, D_j) = \frac{\sum_{i=1}^t (w_{q_i} \times w_{d_j})}{\sqrt{\sum_{i=1}^t (w_{q_i})^2 \times \sum_{i=1}^t (w_{d_j})^2}}$$

其中:

$$w_{q_i} = \text{freq}_{q_i} \times \text{IDF}_i$$

$$w_{d_j} = \text{freq}_{d_j} \times \text{IDF}_i$$

freq_{q_i} = 项 i 在查询 q 中的出现次数;

freq_{d_j} = 项 i 在网页 D_j 中的出现次数;

IDF_i 称为逆向文件频率^[9]。

S 扩展到 T 后, 计算每个网页的主题相似度, 根据不同的阈值进行筛选, 可以选择所有网页相似度的中值、根集网页相似度的中值、最大网页相似度的分数等作为阈值。根据不同阈值进行处理, 删除不满足条件的网页。

2.2 对 I 操作和 O 操作的改进

某一主机 a 上的很多网页可能指向另外一台主机 b 上的同一网页, 这就使得 b 上该网页的权威值过大; 同理, 某一主机 a 上某个网页可能指向另外一台主机 b 上很多网页, 这就使得 a 上该网页的中心值过大。而 HITS 是假定某一网页的权威值是由不同的组织或者个人决定的, 因此上述情况影响了对网页权威值和中心值的计算。

通过重新定义 I 操作和 O 操作解决该问题^[9]: 假定主机 A 上有 k 个网页指向主机 B 上的某个网页 d , 则 A 上的 k 个网页对 B 的权威值的贡献值总共为 1, 每个网页贡献 $1/k$, 而不是 HITS 中的每个网页贡献 1, 总共贡献 k 。类似的, 假定主机 A 上某个网页 t 指向主机 B 上的 m 个网页, 则 B 上 m 个网页对 t 的中心值的总贡献值为 1, 每个网页贡献 $1/m$ 。相应的 I 操作和 O 操作如下:

I 操作:

$$a(u) = \sum_{v:(v,u) \in E} h(v) \times \text{au_wt}(v, u)$$

O 操作:

$$h(v) = \sum_{u:(v,u) \in E} a(u) \times \text{hub_wt}(v, u)$$

2.3 基本的迭代算法

本算法采用 SALSA 算法中的迭代思想^[4], 精简了 HITS 中相互加强迭代的过程, 计算量远小于 HITS, 描述如下:

根据基础集 T 构造二分无向图 $G'=(V_h, V_a, E)$, 其中:

$$V_h = \{S_h | S_h \in T \text{ and out-degree}(s) > 0\} \text{ (G' 的中心网页集)}$$

$$V_a = \{S_a | S_a \in T \text{ and in-degree}(s) > 0\} \text{ (G' 的权威网页集)}$$

$$E = \{(s_h, r_a) | s \rightarrow r \text{ in } T\}$$

这就定义了 2 条马尔可夫链: 权威链和中心链。接着定义 2 条马尔可夫链的变化矩阵, 也是随机矩阵, 分别是中心矩阵 H 和权威矩阵 A :

$$H_{i,j} = \sum_{k \in F(i) \cap F(j)} \frac{1}{F(i)} \times \frac{1}{B(k)}$$

$$A_{i,j} = \sum_{k \in B(i) \cap B(j)} \frac{1}{B(i)} \times \frac{1}{F(k)}$$

求出矩阵 H 和 A 的主特征向量, 就得到对应的马尔可夫链的静态分布, H 和 A 中值大的项对应的网页就是所要找的重要网页。

2.4 算法描述

改进后的算法描述如下:

(1) 按以下步骤构造根集

① 将查询关键词 q 提交给搜索引擎, 从搜索引擎返回的结果中取前 n 个网页作为根集, 用 S 表示;

② 清除无关的链接, 如同一个站点内部的导航链接等。

(2) 扩展根集 S 为基础集 T

① 向根集 S 中加入被 S 引用的网页和引用 S 的网页, 将其扩展为基础集 T , 并除去孤立节点;

② 通过相似度计算在 T 中去除主题相似度值小于阈值的网页, 得到最终的基础集 T 。

(3) 迭代计算

① 根据基础集 T 构造无向图;

② 生成 2 条马尔可夫链的变化矩阵: Hub 矩阵 H 和 Authority 矩阵 A ;

③ 求出矩阵 H, A 的主特征向量;

④ 提取 A 中值大的对应的网页。

(4) 优化索引

3 实验

在一个自行开发的搜索引擎系统中测试算法, 使用采集的网页对算法进行检索对比, 证明算法比 HITS 算法在性能上有较大提高。

3.1 测试数据

使用搜索引擎中的网络爬虫程序抓取网页, 收集了上百个网站的近 30 万网页, 这些网页经过分析处理后按照表 1 的格式保存在数据库中。

表 1 数据库中保存网页信息的主要字段

docid	网页的唯一 id
title	网页的标题
content	网页的正文, 去掉了 html 信息的纯文本
url	网页的 url
subject	网页的主题
mimetype	网页的类型 (Word 文档、PDF 文档等)

首先为数据库中的网页信息按 HITS 算法建立索引, 保存在原始的索引文件夹中。通过这些索引构建一个搜索器, 将该索引映射到内存中, 对提交的查询关键词提供快速的检索。

然后对数据库中的网页信息按改进后的算法计算优化索引, 保存在新的索引文件夹中。

这样我们就获得了可比较的两种索引。

3.2 测试结果

(1) 生成基础集的质量方面: 针对两种不同的算法, 分别在计算中将生成的基础集记录下来并进行比较。

对同样的 10 个关键词, 生成基础集, 测试统计数据, 如表 2 所示。

表 2 改进算法和 HITS 算法生成基础集质量的比较

关键词	改进算法的基础集主题符合?	HITS 算法的基础集主题符合?
Java	是	是
中山大学	是	是
电脑	否	否
计算机	是	是
电信	是	否
科学研究	是	否
中国	是	是
论文	否	否
搜索	是	是
优化	是	是

以上测试数据显示改进算法产生主题漂移的概率下降了 20%。

(2) 索引质量的测试比较:

在主题吻合方面, 用搜索器分别加载经典算法和改进算法生成的索引, 提交 10 个关键词给搜索器, 该搜索器从索引中查询出排名前 10 位的网页。然后再判断这些返回的结果是否符合检索词的主题。两种搜索的结果比较如表 3。

由表可知, 改进算法比 HITS 算法得到的检索结果更主题相关。

为了证明该结论的一般性, 我们分 5 次分别提交了 10 个、20 个、50 个、80 个、100 个关键词进行检索, 得出如图 1 的测试结果:

表 3 改进算法和 HITS 算法索引质量的比较 1

检索词	改进算法	经典算法
java	100%	100%
中山大学	100%	100%
计算机 电脑	100%	90%
计算机	100%	100%
电信	90%	80%
科学研究	90%	90%
中国	100%	100%
论文	80%	70%
搜索	80%	70%
优化	80%	50%

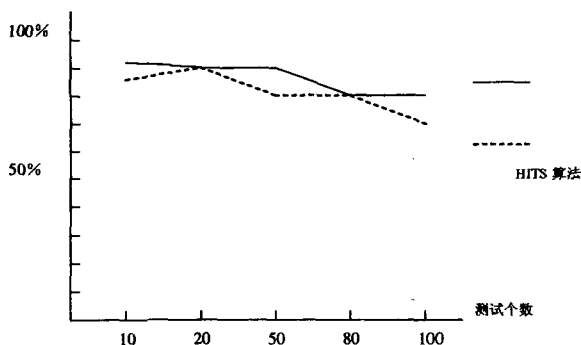


图 1 改进算法和 HITS 算法索引质量的比较 2

由上图也可看出,改进算法比 HITS 算法在返回查询结果时更加主题相关。

4 结语

本文以一个自行开发的搜索引擎系统为背景研究主题提取算法,同时提出了一个融合几种经典算法优点的主题提取算法,最后用该搜索引擎证明了新提出算法比 HITS 算法在性能上有很大提高。

虽然用相似度计算优化基础集的方法较好地解决了主题飘移问题,但时间开销较大,寻找更好的代替方法是我们下一步的目标。

参考文献

[1]J. Kleinberg. Authoritative Sources in a Hyperlinked Environment. *Journal of the ACM*, 46(5):604-632, 1999
 [2]Lawrence Page, Sergey Brin, Rajeev Motwani, and Terry Winograd. The PageRank Citation Ranking: Bringing Order to the Web. Technical Report, Stanford Digital Library

Technologies Project, 1998. Paper SIDL-WP-1999-0120 (version of 11/11/1999)
 [3]David Cohn and Huan Chang. Learning to Probabilistically Identify Authoritative Documents. In *Proceedings of the 17th International Conference on Machine Learning*, pages 167-174, Stanford, CA, 2000
 [4]R. Lempel and S. Moran. The Stochastic Approach for Link-Structure Analysis (SALSA) and the TKC Effect. In *9th International World Wide Web Conference*, Amsterdam, Netherlands, May 2000
 [5]Krishna Bharat and Monika R. Henzinger. Improved Algorithms for Topic Distillation in Hyperlinked Environments. In *21st International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR)*, pages 104-111, 1998
 [6]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
 [7]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
 [8]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
 [9]Sergey Brin and Lawrence Page. The Anatomy of a Large-Scale Hypertextual Web Search Engine. *Computer Networks and ISDN Systems*, 30(1-7):107-117, 1998
 [10]Salton, G and Buckley, C. Term-Weighting Approaches in Automatic Text Retrieval. *Information Processing and Management*, 24(5), 513-523,1988

(下转第 25 页)

参考文献

- [1]刘秀丽,彭复员. 基于小波变换的加权特征脸识别算法. 计算机应用研究,2007(10)
- [2]罗来平,官辉力,刘先林. 基于决策树算法的遥感图像分类研究与实现. 计算机应用研究,2007-01
- [3]王林泉,邱伟峰. 国际图像系统(I2S)上手写汉字识别研究. 计算机工程,1994,20
- [4]苗琦龙,栾新. 基于遗传算法和BP网络的文字识别方法. 计算机应用,2005(12)
- [5]四维科技. Visual C++/MatLab 图像处理与识别实用案例精选. 北京:人民邮电出版社,2004
- [6]阮秋琦. 数字图像处理学. 北京:电子工业出版社,2001

Research on Algorithm for Handwritten Character Recognition

GAO Liang , WU Jian-guo , LU Ying , CHA Shou-li

(Ministry of Education's Key Laboratory of Intelligence Computing and signal Processing, Anhui University, Hefei 230039)

Abstract: Introduces the basic principles of pattern recognition, the recognition to handwritten characters and word, expatiates the process of pattern recognition, the format of image file, and describes the pattern recognition theory based on bitmap technology in detail, discusses the optimization of the classifier's feature space design, the criteria of classifier, the basic methods and the knowledge of discriminant functions, gives a principle on using template matching method to recognize image.

Keywords: Pattern Recognition; handwritten Character; Template Matching; Discriminant Function

(上接第7页)

Research on Topic Distillation Algorithm for Search Engine System

ZHENG Li-rong

(Guangdong Institute of Medical Information, Guangzhou 510180)

Abstract: Takes an own designed search engine system as background to study topic distillation algorithm, through analyzing and combining with several classic topic distillation algorithms, proposes a novel topic distillation algorithm. Experiments show that the algorithm is more effective than HITS algorithm.

Keywords: Topic Distillation; Search Engine; Linkage Analysis