基于超链接引导和链接图分析的主题搜索引擎

唐 苏,刘 循
(四川大学 计算机学院,四川 成都 610064)

摘 要:主题搜索引擎是专为查询某一学科或主题信息而出现的查询工具。针对目前各种主题搜索引擎在主题搜索上的优缺点,提出将基于文字内容启发的超链接引导技术与基于Web链接图的PageRank算法相结合的IPageRank - IND算法,以提高链接相关度判断的准确性和主题资源搜索的覆盖率,并将网页按照VSM算法进行内容相关度判断和自动分类,从而提高检索效率。最后构建一个搜索引擎进行实验,通过比较该算法与其他几种算法的实验结果,能够看到IPageRank-IND算法的优势是明显的。

关键词:主题搜索引擎;超链接分析;PageRank算法;自动分类
中图分类号:TP31 **文献标识码**:A **文章编号**:1673-629X(2011)02-0155-04

Research on Focused Search Engine Based on Hyperlink Induced and Web Structure

TANG Su, LIU Xun
(Dept. of Computer Science, Sichuan University, Chengdu 610064, China)

Abstract: Focused search engine is a tool designed to query information on a particular subject or theme information. Considering the advantages and disadvantages of current focused search engine technologies, put forward the IPageRank - IND algorithm that combining the hyperlink - induced technology based on text-inspired with the PageRank algorithm based on web structure analysis to improve the accuracy of relativity judgment and the coverage of focused resources research, and classifying the web page by sub-topic in order to retrieve efficiently. Then, experiment with a search engine to build, to compare the algorithm with several other algorithms, see the advantage of IPageRank-IND algorithm is obvious.

Key words: focused crawler; hyperlink analysis; PageRank algorithm; automatic classification

1 概 述

主题搜索引擎[1]的目的是查找信息,对某一特定主题或主题进行查询的工具。鉴于主题搜索引擎的搜索只局限于一个特定的主题或专门领域。在搜索过程中,是不需要遍历整个网站的,只要选择含有要访问的主题网页,因此,以哪种爬行策略接入网络,使其抓取尽可能多的网页,尽量少抓取无关网页,并确保网页的质量,是主题搜索引擎设计的关键问题之一。

"目前常用的主题搜索爬行策略主要有2类:基于文字内容的启发策略和基于Web超链接图评价策略"[2]。基于文字内容的启发策略起源于文本检索中对文本相似度的评价,以J. Cho、Herseovici等人的研究成果Best first search[3]及Shark[4]为代表,其原理是利用了Web网页文本内容、URL字符串、锚文字等文字内容信息来判断相关性。然而,这些方法忽略了链接结构信息,使得预测值的准确性较差。以PageRank[5]和HITS[6]为代表的Web链接结构为基础的搜索策略,通过分析网络页面之间的相互作用关系来表示网页的重要性,以此来确定链接的访问顺序。虽然这种方法考虑了链接结构与网页之间的引用关系,但忽视了页面与主题相关的关联性。"HITS算法由于hub页面的多主题性而使得主题存在漂移现象"[7];"PageRank计算独立于用户查询,没有考虑用户查询的具体要求,从而不能够应用于特定主题获取信息,算法过分强调网页的链入链接而贬低链出链接、忽视专业站点以及偏重旧网页等"[8]。

另外,当用户使用搜索引擎查找资料经常会面对着成千上万条的检索结果,这样就很容易忽略掉他们所要查找的信息。现有搜索引擎的主要缺陷是没有对检索结果分类和按人们查询习惯来进行再组织,检索结果自动分类能很好地解决这个问题。

收稿日期:2010-06-09;**修回日期**:2010-09-13
基金项目:国家自然科学基金(60773169)
作者简介:唐 苏(1984-),男,四川南充人,硕士研究生,研究方向为智能信息处理;刘 循,博士,副教授,研究方向为图像处理、模式识别及智能信息处理。

考虑到链接 URL 的真实价值并不能通过单一的评价方法进行有效预测，文中提出了将基于内容评价的搜索策略和基于 Web 链接结构的搜索策略相结合，并使用向量空间算法[9]（VSM）将网页进行筛选并自动分类，这样就能利用基于内容和主题相似性评价，来提高搜索的相关性，同时又以链接结构为基础来提高主题资源搜索的检出率。

2 主题搜索引擎模型

主题搜索引擎与普通搜索引擎的结构非常相似，但主题搜索引擎通过配备一个主题模型来进行主题相关资源的优先检索，实现了对网页中出现的超链接进行链接相关度分析，保证尽可能全面准确地检索到与主题相关的网络信息。并对网页内容进行内容相关度分析并进行自动分类，提高检索的准确率和效率。其体系结构如图 1 所示。

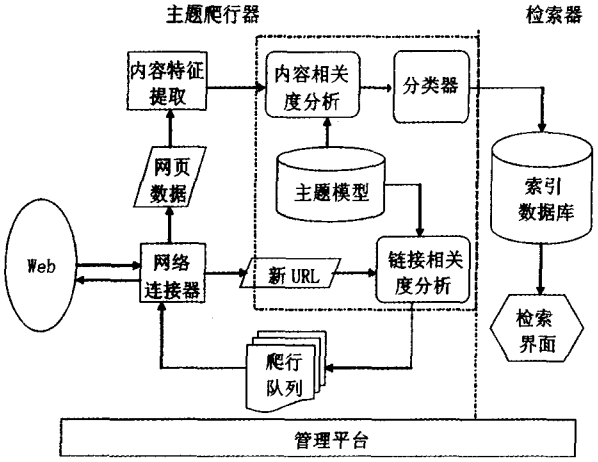

图 1 主题搜索引擎体系结构

主题模型主要包括链接相关度分析、内容相关度分析和分类器三个部分：

链接相关度分析是指系统测算从网页中提取的超链接信息，得出每个 URL 所对应页面与指定主题的相关度，将符合主题相关度要求的 URL 加入到爬行队列中并对其进行爬行优先度排序，以确保高优先级的相关网页被检索到。

内容相关度分析是指通过系统分析经过内容特征提取后的网页数据，判定网页内容与指定主题相关度的关联程度，过滤掉不相关的网页，保留相关度达到阈值的网页。

分类器将采用自动分类算法，分类器将按照被考察对象的内部或者外部特征，将相近、相似或者相同特征的对象聚合在一起的过程。

文中分别对主题模型的三个部分进行研究，提出一种结合了内容评价的搜索策略和 Web 链接结构的搜索策略，并具有自动分类功能的主题模型。

3 链接相关度分析

链接相关度分析是将超链接信息进行测算，得出每个 URL 所指页面与指定主题的相关度。目前主要利用超链接中存在的描述文本（Anchor Text[10]）信息来预测链接的相关度，但是由于链接描述文本（Anchor Text）通常包括很短的文本，单单利用这些很少的文本不能准确推测其与主题相关度。文中通过对主题样本网页集进行分析，将所有链接中的链接描述文本（Anchor Text）进行分词后得到的引导词集合，计算出每个引导词对主题的平均指示度，从而得到链接的主题相关度。

3.1 主题指示度算法

目前，在信息处理方向上，文本的表示主要采用向量空间模型（VSM），文本被表示为向量空间中的一个向量（$W1, W2, W3, \cdots, Wn$），其中将文本分词后得到的特征项作为向量的维数来表示文本，用词频来表示特征项对应的向量分量。文中采用基于特征向量的主题表示，即用主题相关的网页集合进行特征提取得到的主题特征向量来表示主题，然后根据下面的 IND 算法计算链接和主题的相关度。

定义 1：样本网页链接图 $G = (V, E)$ 是有向图，G 由非空的网页节点集合 $V = \{v_1, v_2, \cdots, v_n\}$ 和链接集合 $E = \{l_1, l_2, \cdots, l_m\}$ 组成，其中每个链接 $l_k (l_k \in E)$ 是由两个网页结点和链接引导词集合 A 组成的有序对来表示：$l_k = < v_i, v_j, A > (l_k \in E, v_i, v_j \in V)$，其中链接引导词集合 A 是由链接描述文字（Anchor Text）分词后得到的一系列单词。例如有序对 $< v_i, v_j, A > (l_k \in E, v_i, v_j \in V)$，表示一条从网页 v_i 指向网页 v_j 的链接，其链接引导词集合为 A。

链接引导词集合 A 中每个引导词 w 对主题的平均指示度 IND(w) 可由公式（1）计算：

$$\mathrm{IND}(w) = \frac{\sum_{<v_i, v_j, A> \in E \wedge w \in A} \mathrm{Sim}(v_t, v_j)}{| \{ < v_i, v_j, A > | < v_i, v_j, A > \in E \wedge w \in A \} |} \quad (1)$$

这里 v_t 表示主题的特征向量，$\mathrm{Sim}(v_t, v_j)$ 表示网页 v_j 与主题的相关度，利用向量空间模型来计算网页向量的相似度。

在向量空间模型中，网页的特征向量 v 和主题特征向量 t 之间的内容相关度 $\mathrm{Sim}(t, v)$ 常用向量之间夹角的余弦值表示，如公式（2）所示：

$$\mathrm{Sim}(t, v) = \cos\theta = \frac{\sum_{k=1}^{n} W_{tk} \times W_{vk}}{\sqrt{(\sum_{k=1}^{n} W_{tk}^2)(\sum_{k=1}^{n} W_{vk}^2)}} \quad (2)$$

其中,W_{tk}、W_{vk}分别表示特征向量t和v的第k个特征项的权值,$1 \leqslant k \leqslant N$。

公式IND(w)是计算链接引导词集合A中每个引导词w对主题的平均指示度,而对于每一条链接$L = < v_i, v_j, A >$与主题的相关度,可以通过计算集合A中每个引导词w的平均指示度之和来得到,因此,链接L的相关度IND(L)可按照公式(3)计算:

$$\mathrm{IND}(L) = \sum_{w \in A} \frac{\sum_{<v_i, v_j, A> \in E \wedge w \in A} \mathrm{Sim}(v_t, v_j)}{| \{ < v_i, v_j, A > | < v_i, v_j, A > \in E \wedge w \in A \} |} \tag{3}$$

3.2 基于IPageRank-IND算法的链接相关度分析

链接相关度分析是获取主题信息的核心技术,它直接关系到主题信息采集的效率和质量。文中针对传统的PageRank算法有利于发现权威网页,不利于发现主题资源的不足,以及针对链接描述文本(Anchor Text)信息量偏少,对其挖掘不足的缺点,提出了IPageRank-IND算法。该算法将链接引导与链接图分析相结合,可显著提高主题性判定的准确度。

这里将PageRank与IND两种算法做一定修改,在链接关系的基础上,加上一定的语义信息权重[11],使得所产生的重要页面是关于某一个特定主题的,这样就形成了IPageRank - IND算法,如公式(4)所示:

$$\mathrm{IPR}(A) = (1 - d) + d \sum_{i=1}^{n} (\mathrm{IPR}(T_i) \frac{\mathrm{IND}(L_i)}{\sum_{k=1}^{k_i} \mathrm{IND}(L_{ik})}) \tag{4}$$

其中,A为给定的一个网页,假设指向的网页有$T_1, T_2, \cdots, T_n$。L_i分别是网页T_i指向A的链接,k_i分别是网页T_i中所含的导出链接数,L_{ik}是网页T_i中的导出链接。IPR(A)为A的IPageRank值,d为衰减因子(也设成0.85)。

可以看出,当有许多PageRank值高的网页指向一个网页,那么这个网页的PageRank就会比较高,但IPR值不一定很高,除非它们中大部分与主题相关,这个页面的IPR值才会很高。

3.3 主题爬行策略

通用搜索引擎一般采取宽度优先(Breadth-First Search[12])的爬行策略,来使得资源覆盖率尽可能高,但这样只能获得较低的主题搜索率。主题搜索引擎采用主题优先策略,即按照IPageRank-IND算法计算出爬行队列中待爬链接的主题相关度,对主题相关性高的页面进行优先访问,保证爬行器始终访问主题相关度较高的网页。

4 内容相关度分析及自动分类

为了进一步提高页面采集的准确率,需要对已采集的页面进行主题相关度分析,也就是页面过滤。通过对评价结果较低的页面(小于设定的阈值)进行过滤,来提高所采集主题页面的准确率。最后将主题相关页面自动分类[13],把相关子类型的文档分为一类,方便用户查找,提高检索效率。

4.1 内容相关度分析

采取的方法就是基于向量空间模型的相关度算法,具体流程如下:

训练:

根据主题训练样本,提取主题特征向量v(t)。

筛选:

给定一个网页文档d:

计算其文档特征向量与主题特征向量的相似值Sim(d,v(t));

如果相似值Sim(d,v(t))小于设定阈值q,说明网页d与主题相关度低,为非主题文档,应将其排除。

4.2 自动分类

自动分类能够抽取文档中包含的重要概念自动进行分类,将相似的文档归类到同一个主题,帮助用户迅速地判断返回的结果是否符合自己的检索要求。如果采用了自动分类技术,就可将不同的内容分到不同的类目中去,从而节省用户的判断时间,提高检索效率。

本主题模型采取的分类方法是基于向量空间模型的分类算法,此时每个类型用一个中心向量(Centroid)代表。分类时通过检查待分类文档和这些中心向量的相似度,把它分到最相似的中心向量所代表的类中。具体流程如下:

令$C = \{c_i\}_{i=1}^{m}$代表预定义的类别集合,用于分类的向量空间法如下:

训练:

对每个类c_i,计算该类中所有样本文档向量的算术平均值作为该类的中心向量$v(c_i)$。

分类:

给定一个待分类文档d_q:

根据上公式计算m个相似值$\mathrm{Sim}(d_q, v(c_i)), i = 1, \cdots, m$。

返回$\hat{c}(dq) = \arg\max \mathrm{Sim}(d_q, v(c_i)), c_i \in C$

在训练过程中,可以顺序扫描一遍训练集所有文档,将每个向量累加到它对应的类中心向量上,然后用每个类中心向量除以该类的文档数,以求得算术平均。设整个训练集的文档数为N,则训练阶段的时间复杂度为$O(N)$。在分类阶段,对于每一个待分类文档,由于要计算m个相似度的值,所以时间复杂度为$O(m)$。

5 实验结果与分析

基于 IPageRank-IND 算法，构建了一个主题搜索引擎“Ergate”进行实验。

以财经作为主题，挑选财经主题网站 10 个，其中共含超过 5000 个页面作为主题训练样本，将其特征提取后得到主题特征向量。将主题样本按照股票、基金、外汇、期货、债券 5 个子类型进行分类，构成类型训练样本，将其特征提取后得到各子类型的中心向量。然后对相同的初始 URL 集合，分别用几种算法对数据进行采集。最后的实验结果如表 1 所示。

表 1 实验结果

	主题准确率	资源发现率
宽度优先	39%	40%
PageRank	29%	30%
IPageRank-RW	68%	86%
IPageRank-IND	75%	88%

从表 1 可以看出，IPageRank-IND 算法在主题准确率和资源发现率比宽度优先算法、PageRank 算法都有较大幅度提升；与 IPageRank-RW[14] 算法相比，也有明显的提高。对于自动分类算法，其准确率可达到 80% 左右，效果也比较令人满意。

6 结束语

随着人们对个性化信息服务需要的日益增长，主题搜索引擎的发展将成为搜索引擎发展的主要趋势之一。对主题爬虫爬行策略的研究，以及对主题搜索引擎的应用和发展具有重要意义。文中针对原始的 PageRank 算法[15] 有利于发现权威网页，但不利于发现主题资源的特点，以及基于链接引导算法表现出了很高的主题采集准确率，但却在资源发现率方面表现得不尽人意的特点，综合各自的优点，提出将基于文字内容启发的超链接引导技术与基于 Web 链接图的 PageRank 算法相结合，来提高链接相关度判断的准确性和主题资源搜索的覆盖率，并按照 VSM 算法将网页进行内容相关度判断，并进行自动分类，提高检索效率。实验表明基于 IPageRank-IND 算法的主题搜索引擎优势是明显的。

参考文献：

[1] Zhao Zhongmeng, He Shi li, Yuan Wei. Research on the Algorithm of Specialty Web Page Indices Collected in Topic-Specific Search Engines[J]. Microelectronics & Computer, 2005, 1:6-9.

[2] 刘金红，陆余良. 主题网络爬虫研究综述[J]. 计算机应用研究，2007，24(10)：26-29.

[3] Cho J, Garcia-Molina H, Page L. Efficient crawling through URL ordering [C]//In: Proc. of the Seventh World-Wide Web Conference. New York: ACM, 1998: 161-172.

[4] Hersovici M, Jacovi M, Yoelle S. The shark-search algorithm-An application: tailored web site mapping[C]//In: Proc. of the 7th International Conference on World Wide Web 7. New York: ACM, 1998: 317-326.

[5] Brin S, Page L. The anatomy of a large-scale hypertexual web search engine[C]//In: Proc. of the 7th World Wide Web Conference. Stanford: Stanford University, 1998.

[6] Kleinberg J. Authoritative sources in a hyperlinked environment[C]//In: Proc. of the 9th ACM-SIAM Symposium on Discrete Algorithms. New York: ACM, 1998: 668-677.

[7] 宋聚平，王永成，尹中航，等. 对网页 PageRank 算法的改进[J]. 上海交通大学学报，2003，37(3)：397-400.

[8] 宋建康，张礼平. Web 结构挖掘算法探讨[J]. 华东理工大学学报，2003，29(5)：537-540.

[9] Salton G, Wong A, Yang C S. A Vector Space Model for Automatic Indexing [J]. Communication of the ACM, 1975, 18(11): 613-620.

[10] Eiron N, Mccurley K S. Analysis of Anchor Text for Web Search[C]//Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. Toronto, Canada: [s. n.], 2003.

[11] 白光祖，吕俊生. 基于 WebSPHINX 的主题搜索引擎原理研究与结构设计[J]. 现代图书情报技术，2007，1(11)：58-62.

[12] Cormen T H, Leiserson C E, Rivest R L, et al. Introduction to Algorithms[M]. 2nd ed. Cambridge: The MIT Press, 2001: 532-545.

[13] Soong Mu-Hee, Lim Soo-Yen, Kang Dong-Jin, et al. Automatic Classification of Web Pages Based on the Concept of Domain Ontology [C]//Proc. of the 12th Asia-Pacific Software Engineering Conference. Washington: IEEE Computer Society, 2005: 645-651.

[14] 李盛韬. 基于主题的 Web 信息采集技术研究[D]. 北京：中国科学院研究生院，2002

[15] 常 庆，周明全，耿国华. 基于 PageRank 和 HITS 的 Web 搜索[J]. 计算机技术与发展，2008，18(7)：77-79.