基于内容评价与超链分析的主题爬虫策略

陈志雄,朱向庆
(嘉应学院电子信息工程学院,广东 梅州 514015)

【摘 要】 分析当前主题爬虫策略的优点和不足,提出基于内容评价与超链分析的主题爬虫策略。实验结果表明,基于该策略的主题爬虫准确率和召回率均优于基于内容评价策略的主题爬虫。

【关键词】 主题爬虫;内容评价;超链分析
【中图分类号】 TP311 【文献标识码】 A 【文章编号】 1003-2673(2011)03-66-02

1 前言

互联网技术的发展使得人们越来越依赖网络来获取信息。通用搜索引擎如 Baidu,Google 为人们获取信息提供了便捷的工具。然而返回结果过多、主题针对性不强,使得通用搜索引擎很难为用户提供准确、专业、细致的信息搜索服务。为适应用户专业化、个性化的信息获取要求,各种主题搜索引擎应运而生。主题爬虫作为主题搜索引擎的核心和基础,能够从浩瀚的网络中抓取用户感兴趣的主题网页,获得越来越多的关注,成为研究的热点之一。

目前常用的主题爬虫策略主要有:(1) 基于内容评价的爬行策略[1][2]。(2)基于网页链接结构的爬行策略[3,4]。基于内容评价的爬行策略是在通用网络爬虫的基础上计算新抓取网页与主题的相关度,判断网页是否与主题相关,以达到抓取主题网页的目的。该方法的优点是理论基础好、便以计算,但是缺点是忽略了链接信息的作用,预测链接价值的能力较差。基于网页链接结构的爬行策略通过分析网页之间的链接关系计算网页的重要性,并按网页的重要程度对链接进行排序,优先下载重要的网页。这种方法可以抓取到重要的网页,但是它忽略了网页与主题的相关性,容易造成“主题偏移”,准确率较低。另外,基于网络结构的爬行策略计算网页链接结构的成本偏高,效率较低。

以上两种主题爬虫策略均忽略了 URL 链接字符串本身对预测链接价值的作用。本文提出一种新的主题爬虫策略,在内容评价的基础上,综合了超链分析的一些特性,对待爬行链接进行优先级调整,具有较好的主题爬行效果。

2 主题爬虫策略

2.1 问题陈述

常用的基于内容评价的主题爬虫策略通过计算新抓取网页与主题的相关度,判断网页是否与主题相关。对于相关网页,提取网页包含的 URL 链接,赋予它们相应的优先级。对于不相关网页则抛弃不做处理。待爬行链接按照优先级的大小顺序保存在待爬行链接队列中,Crawler 优先从待爬行队列中读取优先级较高的链接抓取网页。常用的优先级计算方法是网页的主题相关度值与网页包含链接数量的比值。这种方法忽略了网页中大量的主题无关链接对链接优先级计算的影响。无关链接的数量越多,单个链接的优先级越小。

考察发现,通常同一网站上的网页具有相同的排版风格。包含大量相同的导航、广告等与主题无关的链接。同时,为方便用户浏览同一主题的网页,通常同一主题的网页也会互相链接在一起。显然,每次都无差别的抓取、分析相同链接指向的网页是低效的、不科学的,不仅占用了大量的网络带宽,还占用了大量的硬件资源。因此,避免重复抓取具有相同链接的网页是非常必要的,不管这些网页是主题相关的还是主题不相关的。

考察还发现,通常同一网站的 URL 链接都有统一命名规则。因此分析已下载相关和不相关网页的 URL 字符串的规律,对于发现、预测链接的主题相关性是非常有用的。计算待爬行网页的 URL 链接与已下载网页的 URL 链接字符串的相关性,预测待爬行链接的价值,并对它们的优先级调整,显然更有利于抓取到与主题相关的网页。

2.2 解决方案

为解决上述存在问题,本文所设计系统在传统的基于内容评价的主题爬虫基础上增加了 URL 分析器和两个 URL 队列。

URL 分析器实现三个功能,一是筛选待爬行的 URL 链接。二是计算待爬行 URL 的优先级。三是调整待爬行 URL 的优先级。

系统设置了三个 URL 队列,除了原有的队列用于保存待下载的 URL 链接外,新增的两个队列用于保存已下载的相关网页和不相关网页的 URL 链接。三个队列分别命名为,待下载 URL 队列(队列 1);相关 URL 队列(队列 2);不相关 URL 队列(队列 3)。

URL 分析器和三个 URL 队列共同作用,实现 URL 分析器的三个功能。

2.2.1 筛选待爬行 URL 链接

对于主题爬虫抓取的新网页,首先通过计算与主题的相关度来判断是否与主题相关。相关度的计算通常采用下列公式[1]:

$$sim(q,p)=\frac{\sum_{k\in q\cap p} W_{kq}\times W_{kp}}{\sqrt{(\sum_{k\in p} W_{kp}^2)\times(\sum_{k\in q} W_{kq}^2)}} \quad (1)$$

其中 q 表示主题,p 表示网页文档,Wkp 表示词条 k 在文档 p 中的权值,词条的权值一般采用 tf-idf 方法计算,Wkq 表示主题的特征词权值。计算所得的结果为网页与主题的相关度。

【作者简介】陈志雄(1980-),男,广东梅县人,实验师,硕士,研究方向:信息检索,数据挖掘。
【基金项目】梅州市科学技术局、嘉应学院联合自然科学研究项目(08KJ08)资助

设定相关度阈值，若新网页与主题的相关度小于则判定该网页为不相关网页，抛弃该网页。否则，判定该网页为相关网页，提取该网页中包含的所有 URL 链接。然后逐一判断这些链接是否已经存在于三个 URL 队列中。若是，则剔除该链接，其余链接作为待爬行链接保存在队列 1 中。这种机制避免了重复抓取具有相同链接的网页，节约了网络带宽和硬件资源。

2.2.2 计算待爬行 URL 的优先级

基于内容评价的主题爬虫策略，首先使用公式(1)计算新网页与主题的相关度，判断网页是否与主题相关。若网页与主题不相关，则抛弃该网页。若网页与主题相关，则提取网页中包含的所有 URL 链接，赋予它们相同的优先级。优先级的计算通常采用下列公式：

$$\text{priority}(U_{pi}) = \frac{\text{sim}(q, p)}{N_p} \quad (2)$$

其中，其中 q 表示主题，p 表示网页文档，sim(q,p)表示网页 p 与主题 q 的相关度，Np 表示网页 p 包含的 URL 链接数量，Upi 为网页 p 的第 i 个链接，计算结果为链接 Upi 的优先级值。

这种方法可能导致大量与主题无关的链接被赋予与主题相关链接相同的优先级，不利于主题爬虫抓取主题相关网页。因为主题爬虫不能区分哪些链接是主题相关的，哪些链接是不相关的。同时无关链接数量的增加，还将导致主题相关链接的优先级变小。为解决上述存在的问题，很自然地想到统计链接数量时只要剔除网页中的无关链接就可以解决问题，已下载的主题无关链接保存在队列 3 中。需要讨论的是统计网页链接数量时，是否应该包含已下载的相关网页的 URL 链接数量。

如果不包含已下载相关网页的 URL 链接，那么 Np 可能会进一步变小，相当于已下载的相关网页链接的优先级被平均分配给其余待下载 URL 链接。随着主题爬虫抓取的相关网页数量增加，后加入队列 1 的链接优先级会越来越大，最先加入队列 1 中的链接可能会被饿死。显然这种方法对于已经在队列 1 中的链接是不公平的。因此在统计 Np 时，应包含已下载相关网页的 URL 链接。

2.2.3 调整待爬行 URL 的优先级

随着主题爬虫的持续工作，队列 2 和队列 3 积累了相当数量的已下载的相关网页和不相关网页的 URL 链接。分析这些链接，对于寻找规律用于预测 URL 链接的价值是非常有意义的。如搜狐网上一个关于天气的新闻网页的链接为：

(1)http://news.sohu.com/20110225/n279530112.shtml

该网页包含的部分 URL 链接为：

(2)http://news.sohu.com/20110224/n279505996.shtml

(3)http://news.sohu.com/20110124/n279044809.shtml

(4)http://news.sohu.com/20110224/n279500720.shtml

(5)http://star.news.sohu.com/

(6)http://t.sohu.com/p/m/484111946

显然，链接(2) (3) (4)与链接(1)有相同的命名规则，比链接(5) (6)更有可能指向与链接(1)相关的网页。通过大量的考察，发现事实也是如此。因此链接(2) (3) (4)应该赋予比链接 (5) (6)更高的优先级，以便主题爬虫优先抓取这些链接指向的网页。当然链接(2) (3) (4)并不一定都指向与链接(1)相关的网页，只是与链接(1)同为“新闻报道”，所以具有相同的 URL 命名风格而已。实际上链接(4)指向的网页与链接(1)并不相关。但是至少可以预测链接 (5)(6)的主题相关性要低于链接(2)(3) (4)。

实施时，URL 链接被看作一个有序的字符串序列。对队列 2、队列 3 中的所有 URL 链接，以斜线‘/’为终止符，从左至右抽取频繁出现的最大字符串(简称，字符串)，作为判断 URL 链接是否指向相关或不相关网页的依据。

例如，假设链接(1) (2) (3)都在队列 2 中，都是同一主题的 URL 链接。设字符串频繁度阈值为 f=3，则以斜线‘/’为终止符抽取出的最大字符串为“http://news.sohu.com/”。

定义从队列 2 中抽取的字符串集为相关字符串集 S。与集合 S 中任一字符串匹配的待爬行链接为相关链接，其优先级调整为初始优先级乘以系数 a=1.5。

相应的，定义从队列 3 中抽取的字符串集 M 与队列 2 中抽取的字符串集 S 的差 U 为不相关字符串集，U=M-S。与集合 U 中任一字符串匹配的待爬行链接为不相关链接。其优先级调整为初始优先级乘以系数 b=0.5。

2.3 系统设计

本文设计系统在传统基于内容评价主题爬虫的基础上增加了 URL 分析器和两个 URL 队列，图 2-1 中阴影部分为增加的部分。整个系统包含待爬行 URL 队列；相关 URL 队列；不相关 URL 队列；Crawler 模块； 文本分类器；URL 提取模块和 URL 分析器等。系统设计如图 1 所示。

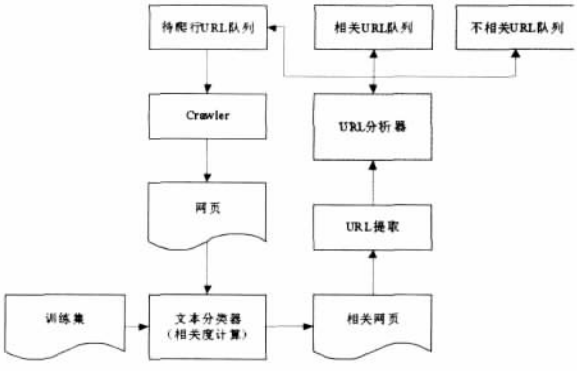

图 1 主题爬虫系统设计

系统设置了三个 URL 队列用于保存待下载网页的 URL 链接，已下载的相关网页和不相关网页的 URL 链接。分别命名为，待下载 URL 队列(队列 1)；相关 URL 队列(队列 2)；不相关 URL 队列(队列 3)。初始状态下，种子 URL 链接加载在队列 1 中，其余两个队列为空。

Crawler 爬虫模块从待爬行队列 1 中读取一条优先级最高的 URL 链接，把该链接指向的网页下载到本地。

文本分类器对下载到本地的网页进行分类。分类器通过计算新网页与训练集的相关度，判断该网页是否与主题相关。若该网页被判定为不相关网页， 把指向该网页链接从队列 1 移到队列 3 中。若该网页被判定为相关网页，则提取该网页的 URL 链接，保存该网页，并把指向该网页链接从队列 1 移到队列 2 中。

URL 提取模块用来提取相关网页的 URL 链接。

URL 分析器分析从网页中提取的 URL 链接，筛选出待爬

(下转第 75 页)

响的，经过反复调试对比，在得到的实验结果中，最佳的识别率达到 98%。

4 结论与展望

本文用最小距离法对 0-9 的数字图片进行了识别，从试验结果看识别效果差别比较大，这主要和一些参数的选取有关，比如特征值矩阵的分块，当选取合适的 9 块空间可以使识别率达到 98%。最小距离法的特点就是算法简单，针对小样本问题有一定的优势，但不能保证有良好的识别率。针对此问题，我们可以在此法的基础上和其他方法智能算法结合起来来进行识别。

参考文献

[1]杨淑莹.模式识别与智能计算—Matlab 技术实现[M].北京:电子工业出版社,2008.

[2]宋曰聪.手写体数字识别系统中一种新的特征提取方案[J].计算机科学,2007,(9).

[3]王耀南,李树涛,毛建旭.计算机图像处理与识别技术[M].北京:高等教育出版社,2001,(6).

[4]边肇祺.模式识别[M].北京:清华大学出版社,2000,(1).

(上接第 67 页)

行链接，并且计算、调整这些链接的优先级。

3 实验结果与分析

实验所用平台为 PC (AMD Athlon (tm) 64 X2 Dual Core 4200+ 2.21GHz，内存 1GB)，操作系统是 Windows SP2。算法使用 JAVA 语言编写(JVM1.5)，使用了开源软件包 HTML Parser2.0。

常用的评价爬虫系统的指标有两个：准确率(Precision)和召回率(Recall)。定义如下[5]：

$$Precision = \frac{\text{采集的目标页面数}}{\text{总爬行页面数}} \quad (3)$$

$$Recall = \frac{\text{采集的目标页面数}}{\text{总目标页面数}} \quad (4)$$

实验所用的训练集来自搜狐、新浪、腾讯、网易等门户网站。种子链接从训练集网页中提取。主题相关度阈值设置为，队列 2 和队列 3 的 URL 字符串频繁度阈值设置为。主题爬虫分别对天气、购物、娱乐、财经、体育等五个主题进行测试。实验所得数据如表 1 所示。

表 1 两种主题爬虫策略性能比较

主题	准确率(%)		召回率(%)	
	T&L	T	T&L	T
天气	62.7	44.1	72.3	53.3
购物	61.4	45.6	73.2	52.1
娱乐	65.1	50.1	77.6	54.9
财经	60.6	43.7	71.8	47.7
体育	64.0	47.5	75.5	55.3
Average	62.8	46.2	74.1	52.7

注：T&L 表示基于内容评价与超链分析的主题爬虫策略；T 表示基于内容评价的主题爬虫策略。

表 3-1 列举了两种主题爬虫策略分别在 5 个主题上进行爬行的准确率和召回率。从表中不难看出基于内容评价与超链分析的主题爬行策略的性能全面优于基于内容评价的主题爬虫策略。

4 结束语

本文设计和实现了一个基于内容评价与超链分析策略的主题爬虫。系统通过计算网页与主题的相关度对网页进行分类。对已下载网页的链接进行分析，找出有用规律用于预测新链接的价值，并根据预测结果调整新链接的优先级。主题爬虫优先抓取优先级高的链接指向的网页，达到抓取主题网页的目的。实验证明，改进后的系统性能显著提高。

参考文献

[1]林海霞,原福永,陈金森,刘俊峰.一种改进的主题网络蜘蛛搜索算法[J].计算机工程与应用,2007,43(10):174-176.

[2]李卫疆,赵铁军,朴星海.一种新的面向主题的爬行算法[J].计算机应用研究,2009,26(5):1663-1666.

[3]张翔,周明全,李智杰,黄丽丽.基于 PageRank 与 Bagging 的主题爬虫研究[J].计算机工程与设计, 2010,31(14):3309-3312.

[4]蒋宗礼,徐学可,李帅.一种基于超链接引导的主题搜索的主题敏感爬行方法[J].计算机应用,2008,28(4):942-944,950.

[5]Menczer F, Pant G, Srinivasan P, et al. Evaluating topic-driven web crawlers[C]. Proc 24th Annual Int'l ACM SIGIR Conference on Research and Development in Information Retrieval, 2001: 241-249.