计算机系统应用 ISSN 1003-3254, CODEN CSAOBN
Computer Systems & Applications,2020,29(3):253−260 [doi: 10.15888/j.cnki.csa.007290]

E-mail: csa@iscas.ac.cn
http://www.c-s-a.org.cn
Tel: +86-10-62661041

基于动态隧道技术的主题爬行策略①

姜　琨, 朱　磊, 王一川

(西安理工大学 计算机科学与工程学院, 西安 710048)
通讯作者: 姜　琨, E-mail: jk_365@126.com

摘　要: 互联网网页所形成的主题孤岛严重影响了搜索引擎系统的主题爬虫性能, 通过人工增加大量的初始种子链接来发现新主题的方法无法保证主题网页的全面性. 在分析传统基于内容分析、基于链接分析和基于语境图的主题爬行策略的基础上, 提出了一种基于动态隧道技术的主题爬虫爬行策略. 该策略结合页面主题相关度计算和URL 链接相关度预测的方法确定主题孤岛之间的网页页面主题相关性, 并构建层次化的主题判断模型来解决主题孤岛之间的弱链接问题. 同时, 该策略能有效防止主题爬虫因采集过多的主题无关页面而导致的主题漂移现象, 从而可以实现在保持主题语义信息的爬行方向上的动态隧道控制. 实验过程利用主题网页层次结构检测页面主题相关性并抽取“体育”主题关键词, 然后以此对采集的主题网页进行索引查询测试. 结果表明, 基于动态隧道技术的爬行策略能够较好的解决主题孤岛问题, 明显提升了“体育”主题搜索引擎的准确率和召回率.

关键词: 网络爬虫; 主题孤岛; 动态隧道; 爬行策略

引用格式: 姜琨,朱磊,王一川.基于动态隧道技术的主题爬行策略.计算机系统应用,2020,29(3):253–260. http://www.c-s-a.org.cn/1003-3254/7290.html

Dynamic Tunneling Heuristic for Focused Crawling

JIANG Kun, ZHU Lei, WANG Yi-Chuan

(Faculty of Computer Science and Engineering, Xi’an University of Technology, Xi’an 710048, China)

Abstract: Topic island on Internet Web pages has seriously affected the performance of focused crawlers. The metric of setting more initial links to find new topics cannot guarantee the comprehensiveness of Web pages. On the basis of analyzing typical crawling strategies and taking into account the hierarchy of topic relevant, we propose a crawling strategy using dynamic tunneling. The crawling strategy uses the tunneling technology based on the topic of Web pages to discover new topics, and constructs a hierarchical topic model to solve the problem of weak link between two topic islands. Meanwhile, the strategy can effectively prevent topic drift caused by collecting too many topic-independent pages, thus dynamic controls the tunneling depth in the crawling direction with the semantic information of the topic maintained. Experimental results show that the proposed method can better address the topic island issue, thereby enhancing the recall of focused search engines.

Key words: focused crawler; topic island; crawling schema; dynamic tunneling

互联网网页的聚集特性表明主题页面容易聚集出现, 因主题相关或相近而链接在一起的互联网网页被称为主题岛或者主题团. 主题爬虫依据主题团的聚集特性对网页进行采集. 然而, 并非所有的主题相关网页都是链接在一起的, 它们之间可能要跨过几个主题不相关页面的链接. 许多主题岛被这些主题无关的页面

① 基金项目: 国家自然科学基金 (61602374)
Foundation item: National Natural Science Foundation of China (61602374)
收稿时间: 2019-07-19; 修改时间: 2019-08-22; 采用时间: 2019-08-27; csa 在线出版时间: 2020-02-28

链接, 使主题岛之间被分隔, 这种现象被称为主题孤岛. 如图 1 所示, 这些无关页面的链接分布在互联网上待采集的主题团之间, 形成连接主题孤岛的一个隧道, 这就是 Web 页面的隧道特性[1,2]. 实际的 Web 中存在大量这样的主题孤岛, 如果主题爬虫系统只通过父页面来预测子页面的相关度, 只提取主题相关页面中的超链接作为种子链接, 那么就会丢失大量的主题孤岛. 因为如果子页面是主题无关的, 爬虫就可能不会访问该页面中的超链接, 这些链接可能穿过数次链接而连着另一个主题孤岛.

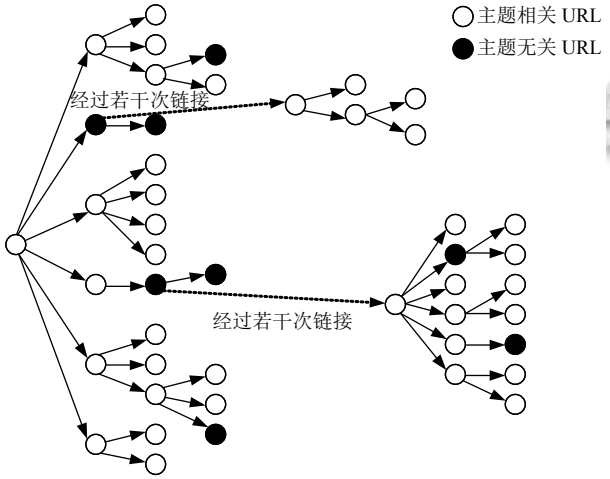

图 1 主题孤岛问题示意图

如果为了提高爬虫采集页面的准确度而提高爬行策略中的相关度阈值, 则会过滤掉大量的隧道, 这样就访问不到隧道另一端可能存在的主题孤岛, 导致爬虫的对主题网页的召回率较低. 为了提高爬虫一次爬行过程所采集主题相关页面的数量, 往往会降低爬行策略中判断主题相关与否的相关度阈值. 如果这个隧道很长, 那么降低相关度阈值去访问这些主题孤岛, 又会采集隧道路径上的大量主题无关页面, 但是有一定概率发现新的主题相关页面[2].

本文在分析现有主题爬虫爬行策略特点的基础上, 针对现有主题爬行策略不能很好解决主题孤岛问题, 提出一种能在爬虫爬行过程中对不相关页面中提取的 URL 链接对应页面的主题相关度进行预测, 并动态调整隧道长度的主题爬行策略模型, 从而可以挖掘不相关网页信息及发现隐藏的主题团相关链接.

1 主题爬行策略研究现状

普通网页爬虫一般采用广度或者深度优先的爬行策略, 而主题爬虫采用的爬行策略按照判断网页相关度的不同分为: 基于内容分析的爬行策略、基于链接分析的爬行策略和基于语境图的爬行策略等[3].

1.1 基于内容分析的爬行策略

基于内容分析的主题爬行策略主要是利用页面或者链接的内容特征对页面与主题的相关程度进行打分评价, 进而对待采集网页和爬行方向进行优化选择[4]. 基于内容分析的主题爬行策略主要有: 最佳优先搜索 (Best First Search, BFS)、Fish Search、Shark Search 等 3 种策略[5].

BFS 策略的基本思想是利用主题团特征, 通过分析当前已经获取的页面, 使用一定的打分策略来预测与其连接的页面的主题相关度, 然后使用最好优先的原则每次优先选择主题相关度最高的页面作为下一个处理的对象. 与主题关系比较密切的页面, 它所包含链接的优先级就高, 这样就确定了等待处理的链接队列中链接的前后顺序. 该策略每次添加到爬虫种子优先级队列的链接的优先级分数是相同的.

在 Fish Search 策略中, 当通过某一链接发现主题相关页面时, 沿着这个方向的爬行深度增加, 且后代链接的爬行深度保持不变. 如果没有发现主题相关页面, 这个链接的爬行深度不变, 但是后代链接的爬行深度递减. 如果沿着某个方向经过多次采集仍然没有找到主题相关页面, 那么它的爬行深度会逐渐降低直至为零. Fish Search 策略在主题不相关方向上的采集具有一定的动态特性, 但是其主题相关性的判断仅仅是一种二值分类判断, 不能评价相关程度的高低.

Shark Search 策略是对 Fish Search 策略中主题相关度打分策略的改进, 其页面与主题之间的相关程度是一个介于 0 到 1 之间的连续值, 这一改进的优点是可以获得一个 URL 与主题的相关程度. 然而, 因为 Shark Search 和 Fish Search 策略在主题不相关页面方向上采用了降低爬行深度的数据采集, 而且对主题不相关页面采取了和之前页面相同的分析方法, 因而导致其提升召回率的代价是牺牲了爬虫的准确率.

1.2 基于链接分析的爬行策略

基于链接分析的爬行策略主要是依据网页之间的引用关系和页面已知重要度分数来判断网页之间的重要程度. 基于链接分析的爬行策略主要是基于以下两个条件: (1) 如果在网页 A 中包含网页 B 的链接, 则表明网页 A 对网页 B 重要性的推荐; (2) 此时, 如果在网

页 B 中也同时包含网页 A 的链接, 则网页 A 和网页 B 一般有共同的主题. 比较有代表性的基于链接分析的爬行策略如: 基于 PageRank 的链接分析方法等.

PageRank (PR)[6]用于搜索引擎中对查询结果进行排序, 近年来也被用于预测主题爬虫的链接优先级. PR 链接分析方法对网页重要性的打分评价主要依据 3 个方面: (1) 内链越多的网页越重要, 即其他网页对该网页的推荐较多; (2) 内链的网页重要度越高, 被这些高质量网页的链接指向的网页也越重要; (3) 外链数越少的网页相对越重要, 即一般重要网页中的链接都是其子链接. 然而, 为了降低动态计算每个待爬取队列里 URL 链接的 PR 值的代价, 实际获得 PR 值都是非精确的.

基于内容分析的爬行策略和基于链接分析的爬行策略都属于立即回报型爬行策略, 这类爬行算法通过分析当前的页面内容或者链接信息, 目的是要通过这样的分析来及时指导紧接着的爬行方向. 这类主题爬行策略虽然在主题页面附近的时候能够表现出较好的性能, 但是对那些有潜在主题相关性的链接不够关注甚至过早丢弃, 所以在距离主题页面较远的地方就有可能会出现“主题漂移”的现象, 也难以有效解决主题孤岛问题[5].

1.3 基于语境图的爬行策略

为了解决有效主题孤岛问题, 研究人员提出语境图爬行策略 (context graph)[7]. 这种策略的训练过程首先要给系统提供一组种子主题页面, 然后利用 Google 反向链接服务寻找到所有拥有指向种子页面链接的页面作为第一层页面, 而所有拥有指向第一层页面链接的页面被称作第二层页面, 依次类推, 层数由用户控制. 图 2 展示了一个深度为 2 的语境图. 当每一个种子页面都建立好一个语境图后, 将不同的语境图的相应各层进行合并, 形成一个合并语境图 (merged context graph). 然后为合并语境图的每一层训练一个贝叶斯分类器. 在爬行过程中, 分类器被用来确定所要爬行的页面应该属于哪一层, 从而有效识别主题相关度较低网页的所属的层数并计算爬行优先分数.

基于语境图的爬行策略避免了立即回报型爬行策略只关注能带来立即效益链接的缺点. 然而基于语境图的爬行策略需要为其建立语境图模型, 因此这种方法无疑加重了主题搜索引擎的复杂度. 本文在考虑到基于语境图的爬行策略的在线复杂度和其采用的利用 Google 反向链接服务的局限性, 受基于语境图的爬行策略中采用的主题层次思想降低隧道长度的启发, 提出一种新的主题爬行策略, 其可以通过预测 URL 链接相关度方法分析隐含的主题层次结构, 并动态维护各个主题层次的隧道长度, 并使爬行过程具有较低在线复杂度和更好可操作性.

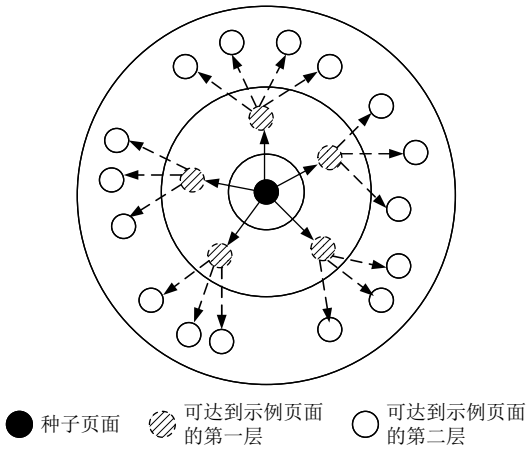

图 2 一个两层的语境图模型

2 基于动态隧道技术的爬行策略

基于语境图的爬行策略为我们提供了一个发现隐藏主题相关链接的很好的思路, 对于爬虫发现的主题不相关链接也不能轻易抛弃, 而是要看它是否属于主题相关链接的前驱链接. 如果一个爬虫的目标是获取与“体育”主题相关的网页内容, 那么一些体育高校的主页可能是很有价值的, 虽然这些页面本身并不一定直接与“体育”的主题有关系, 但是这些主页可能会链接到某些和“体育”相关的新闻页面, 在这些新闻的页面中则对应的着“体育”主题相关的页面. 如: “北京体育大学”主页中包含“媒体北体”页面, 然后进一步链接到“新华社”等多个“体育”主题相关新闻网站. 在这种情况下, 体育高校的主页、学校的新闻页面或者论坛主页等与“体育”主题相关或相近的页面和“体育”主题目标页面之间就形成了一种既有联系又有区别的层次结构, 而在这种层次结构中就隐含了能够找到目标主题页面的爬行路径. 互联网网页的主题相关层次示意图如图 3.

基于语境图的爬行策略认为主题爬虫在互联网上查找某个特定主题的信息时, 如果发现某一网页的主题和给定主题存在某种预定义的相关性时, 就可以认

为沿着这些在层次结构中的相近页面必定能找到更多的主题页面. 在爬虫实现中, 通过建立主题相关词典模型和主题相近词典模型, 对主题不相关链接进行进一步语义挖掘, 就可能发现更多的主题团, 从而在一定程度上解决主题孤岛问题. 本节采用 URL 链接相关度预测的方法进行定量的语义挖掘, 得出在两个词袋 (bag-of-words) 模型下不相关网页的相似度, 结合动态隧道模型来确定爬虫在不相关网页上的预测深度.

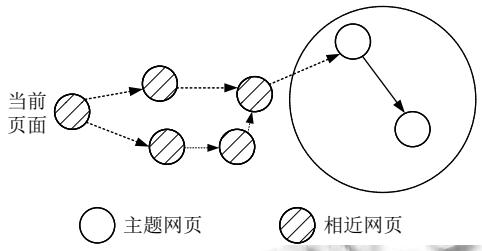

图 3 互联网网页主题相关层次图

2.1 动态隧道长度

Bergmark 等提出了隧道技术来描述和解决主题孤岛问题[8]. 使用隧道技术的主题爬虫在碰到主题不相关的网页时会继续在该链接方向上向前探索 k 步. 这样主题爬虫可以从一个主题团游走到另外一个主题团, 其中可能经过多层主题相关度较低的页面. 如果在两个主题团之间的距离不大的前提下, 就可能发现互联网中所有与预定义主题相关的网页. Bergmark 在对 500 000 个网页的分析表明主题孤岛现象的普遍性以及大多数属于主题孤岛的距离在 1 和 12 之间, 平均距离是 5. 然而, 采用隧道技术的主题爬虫在遇到主题相关度较低的页面时会扩大探索范围. 也就是说, 爬虫以种子集为圆心, 以 k 为半径的圆周范围中探索其它主题团, 随着半径 k 的增大, 发现其它主题团的概率也在增大, 但是需要处理的主题无关网页也以显著增加. 实际上, 当 k 无限增大时, 主题爬虫对每个预测不相关的网页都要进行采集, 这样的主题爬虫就成为了通用爬虫. 因此可以说, 这种方法是放松了对主题爬虫的定义来提高召回率, 从而极大地降低了爬虫的效率[9].

尽管可以采用人工动态调整主题相关阈值的办法来改变一个链接的主题相关情况, 但也只有链接相关和不相关两种情况. 因而该技术在检测页面不相关时, 对该方向上的链接爬行深度的设定完全没有考虑到在该方向上爬行每层页面的动态情况. 因此, Bergmark 等提出的隧道技术属于在主题相关度较低链接方向上的静态探索技术, 即在主题不相关时仍然搜索 k 步, 而不去关注这 k 步的搜索中获取的链接的反馈信息. 而 Fish Search 策略的动态隧道思想表现在如果出现链接主题不相关, 则减少该方向上下一个链接的隧道长度. 如果遇到潜在的 URL 链接相关度较高, 但是页面主题相关性不够高的情况, 那么原来的方法难以将这一信息及时反馈到主题爬虫. 这一问题很大限制了主题爬虫发现主题孤岛的能力.

互联网网页的主题相关层次表明, 对于主题不相关网页还需要进一步分析其是否属于主题语境图中的某一层. 如果该链接属于语境图层次结构中的某一层时, 沿着这个链接方向的爬行深度增加, 并且后代链接的爬行深度保持不变. 如果通过该链接不属于主题层次结构的任何一层, 则这个链接本身的爬行深度不变, 但是后代链接的爬行深度才需要递减. 因此, 采用动态控制主题不相关方向上的搜索深度, 可以发现潜在的优质主题 URL 链接, 从而增加发现主题孤岛的可能性.

2.2 主题爬行模型

结合主题相关层次和隧道长度的分析, 本文提出的解决主题孤岛问题的爬行策略的主要思想为: 爬虫在遇到主题相关页面时, 将该页面中的所有 URL 链接和其优先值 pv (pv=主题相关度) 送到爬虫的优先级队列, 相关度越高的页面其 URL 外链优先级越高, 在优先级队列中也应当被优先采集; 此时候选 URL 链接非常多, 爬虫不可能出现优先队列空的现象; 此时采用广度优先的方式对相同页面的相同优先级的页面进行采集.

主题爬虫通过式 (1) 计算得到主题不相关的页面时 (pv=0), 并不是停止获取其页面中的 URL 外链, 而是继续在所获取的 URL 外链上向前探索 k 步路径. 对于路径上的每一层页面, 若在此路径上通过下一节所阐述的 URL 链接相关度预测方法发现潜在主题相关页面 (pv=0), 爬虫在这个链接方向上的爬行深度保持不变, 否则 k 值递减. 此时采用深度优先的方法判断获得的页面是否仍然是不相关页面, 直到达到某一个主题相关页面为止 (pv=主题相关度). 策略流程如图 4 所示, 工作流程为: 对采集到的某网页去噪之后得到正文内容, 之后调用主题词库进行相关度计算. 如果与主题相关, 则将当前爬行深度设为∞, 表示按照原有方式进行采集. 如果与主题不相关, 检查爬行深度值 k. 如果 $k = 0$, 表示在此链接方向上已经无需再采集, 并停止采集. 如果$k = \infty$, 表示 k 值未被设置过, 并设置$k = k_{\text{depth}}$,

递减 k 值之后交由后续模块处理. 如果$0 \le k < \infty$, 调用下节所述的 URL 链接相关度预测方法进行页面相关度计算, 在这种情况下, 要 URL 和主题内容相关, 则使当前深度不变, 并交后续模块处理; 要是不相关, 则使爬行深度递减, 并交后续模块处理.

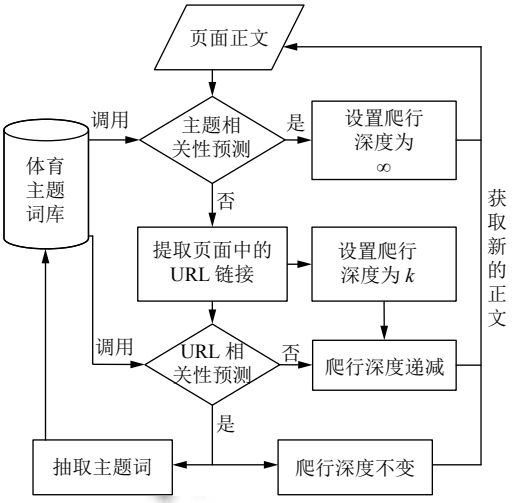

图 4 动态隧道技术策略流程图

这种策略的优点在于, 在不相关页面方向上设定的爬行深度是动态变化的, 它把不相关页面方向上的信息反馈到对隧道爬行深度 k 的动态控制, 因此被称为动态隧道技术 (Dynamic Tunneling Heuristic, DTH). 因此, 该策略减少了在此方向上的搜索, 这样可以有效的降低了主题爬虫的在主题无关方向上的爬行范围; 而对于通过 URL 链接相关度预测后发现可能有潜在主题相关的链接, 该策略加大了在此方向上的爬行深度, 这样能进一步发现隐藏的主题团. 因此, 该策略利用 URL 链接相关度预测和动态隧道控制技术对潜在的主题团进行搜索.

主题爬虫如果仅仅采用页面主题相关度计算方法, 则随着爬虫不断的爬取新的主题页面, 新的主题关键词会不断加入主题词库并获得新的权重, 从而出现“主题漂移”现象. 这主要是因为主题页面的缺失导致的, 此时虽然出现大量主题无关页面, 但是主题爬虫却无法发现新的主题团, 因此会制约主题爬虫的准确率. 本文方法对主题无关页面进行 URL 链接相关度分析, 能够提升主题爬虫发现新主题页面的准确率. 如果将页面主题相关度计算和 URL 链接主题相关度计算结合, 则会明显影响主题爬虫在主题团内部爬取主题页面时的性能.

2.3 URL 链接相关度预测

主题爬虫系统的主题相关度判断方法: 爬虫系统需要维护一个主题词库, 其中包括了由大量主题相关的关键词组成的主题向量和每个主题词出现在网页中的个数 IDF. 主题词典的关键词来源是预先给定的网页页面, 包括爬虫系统初始化时给定 URL 链接种子对应的页面和主题词库更新过程中添加的该领域比较有代表性的网页.

主题爬虫系统运行过程中对于主题页面的选择规则如下: 含有“default”、“index”等信息的 URL 链接可以初步作为主题页面; 不能作为主题页面的规则为: 入链小于一定阈值的页面; 锚文本过长的页面; 锚文本中包含“下一页”、“更多”等信息的页面; URL 过长的页面等. 对于利用上述规则选择的多个主题页面, 再通过 TextRank 策略进行主题向量抽取, 形成主题词库. 主题向量 T 是由基于 TextRank 的关键词抽取方法提取的关键词及其权重$w_{i,r}$组成. TextRank 是一种非监督式的主题抽取策略, 不依赖于其他语料, 直接从文本中抽取主题关键词; 适用于对于少量网页文本的主题关键词进行分析. 主题词典可以在爬虫未启动时进行更新维护, 输入发现的新的网页正文进行重新计算.

本文在对网页P_j进行正文提取后首先采用向量空间模型 (VSM) 来计算网页内容与主题的相关度, 即利用基于 TextRank 的抽取得到的主题向量和给定网页特征向量计算当前页面的主题相关度, 计算公式如下:

$$Sim\left(P_j,T\right)=\frac{\sum_{i=1}^{t} w_{i,j}\times w_{i,r}}{\sqrt{\sum_{i=1}^{t} w_{i,j}^2\times\sum_{i=1}^{t} w_{i,r}^2}} \tag{1}$$

其中, $w_{i,j}$表示特征向量在给定网页文本中的权重值, $w_{i,r}$表示特征向量 i 在主题向量中的权值, T 代表主题向量, $Sim\left(P_j,T\right)$表示文本P_j与给定主题向量的相关度. 计算文本权重值$w_{i,j}$的策略是 TF-IDF, 即:

$$w_{i,j}=tf_{i,j}\times\log\left(\frac{N}{df_i}\right) \tag{2}$$

其中, $tf_{i,j}$表示关键词t_i在给定网页正文P_j中出现的次数, df_i则表明当前关键词t_i在已经采集的网页中出现次数, N 为已经采集的网页数量.

在 2.2 节的主题爬行模型中, 如果当前网页通过

式 (1) 计算得出是主题不相关的, 则进一步对该网页中的 URL 链接的主题相关度进行预测. 其中需要考虑 URL 链接的锚文本本身、URL 链接的上下文环境以及 URL 链接字符串的主题相关度等 3 个因素. 因此, 待采集 URL 链接 p 对应页面的主题相关度计算如下:

$$Relevance(p) = RArchor(p)\cdot w_1 + RContext(p)\cdot w_2 + RUrl(p)\cdot w_3 \quad (3)$$

其中, $w_1 + w_2 + w_3 = 1$.

$RArchor(p)$指的是 URL 链接 p 对应的锚文本的主题相关度. 链接的锚文本一般都明显包含了出链网页的信息, 因此有助于预测对应网页的主题相关度. $RContext(p)$指的是 p 对应的 URL 链接在当前网页中的附近文本信息的主题相关度. 一个链接附近的信息也能在一定程度上说明该出链网页的主题. 以上两部分基于主题向量和式 (1) 来计算主题相关度. $RUrl(p)$是链接 p 字符串信息的主题相关度. 这是因为域名往往包含有该网页的主题相关信息. 比如对某页面提取的 URL 链接: https://sports.ifeng.com/, 其中包括字符串 sports, 据此就可以推断出该网页主要描述的是“体育”类主题信息. 本文在判断未知链接字符串相关性时, 采用了分析主题爬虫采集的主题页面 URL 字符串, 以及人工收集主题页面 URL 链接常用字符串的方法, 但是最终通过人工确定所采用的 URL 字符串集合. 因为主题页面 URL 链接的主题字符串范围比较小, 通过上述方法基本能够保证 URL 链接主题字符串的全面性.

3 实验测试与结果分析

3.1 实验设置

本实验在 Windows 10 下采用 Java 语言实现了一个多线程的主题爬虫原型系统, 采用了本文提出的基于动态隧道技术的 DTH 主题爬行策略, 对比策略为基于内容分析的爬行策略 BFS 和基于链接分析的 PR 爬行策略. 其中, BFS 策略实现过程中主要通过优先级队列实现对 URL 链接的准确率, 即优先级高 (主题相关度高) 的 URL 链接的外链的优先级也要高. PR 值策略因为仅仅采用链接重要度来确定优先级, 因此实现过程中加入了主题相关度的检测来避免“主题漂移”问题; 此外, PR 值的计算范围是当前待爬取队列的 URL 链接. 对于当前每个爬取到的页面, 分析其包含的所有 URL 是否存在于待爬取队列中, 如果存在则增加该 URL 的 PR 值; 如果不存在则赋予其 PR 初值. DTH 策略实现过程的特点主要是对不相关网页的爬行路径上“隧道”的深入挖掘和处理.

该爬虫系统的输入是特定领域的主题词库 (采用结巴分词所带的 TextRank 模块进行主题关键词抽取获得[10]) 和一组种子 URL 链接, 输出是主题相关的结果页面集合. 实验从互联网网站 (如: 新浪、搜狐、凤凰、体育高校等) 中采用上述爬行策略下载“体育”相关网页. 通过正文提取、中文分词 (采用“结巴”分词器进行分词[10])、除去停用词等预处理步骤后构建索引数据. 实验测试中, 分别统计采集页面的数量在 500、1000、1500、…、4000 时的情况. 通过构建不同网页数量的倒排索引数据, 采用“体育”主题查询词集合在搜索引擎中进行主题关键词的查询. 实验评价指标是查询的准确率和召回率主题页面数量 R, 定义 $Precision = M/N$, 其中, M 是搜索到的体育主题相关文档数, N 是搜索到的全部文档数; R 是系统采集的全部主题相关的文档数, 表明爬虫系统采集到主题网页的能力.

3.2 结果分析

互联网中各个话题相关的主题团为吸引用户浏览不能独立存在, 必然是通过一定的链接相互联系, 而主题团之间的隧道长度究竟是多长. 图 5 给出主题爬虫原型系统在 URL 链接主题不相关条件下 (共 100 000 次) 采用动态隧道技术 DTH 找到新的主题团的爬行深度 k 的分布. 实验中不相关方向上的初始化最大爬行深度k_{depth}的值可以调整 (初始设为 12), 实际隧道长度为为找到主题相关页面时在该方向上的爬行深度. 可以看出, 对于找到主题相关页面的情况, 隧道长度平均值为 4. 考虑到主题爬虫系统采集 URL 链接的随机性, 可以得出大部分主题相关节点之间的最短距离不超过 6 (six degrees of separation). 因此, 可以初步推测这一现象可能符合 Web 中主题团是小世界网络 (small world) 的假设.

主题搜索引擎在采用不同爬行策略时的准确率随采集页面的变化趋势如图 6 所示. 实验结果可以发现 BFS 策略的准确率高于 PR 策略. 这是因为基于链接分析的 PR 爬行策略只是依据页面的 PR 值来确定待爬行链接的优先级, 而忽视了页面内容的主题相关情况, 随着爬取深度的增加就容易出现主题漂移的情况, 从而导致爬行策略的准确率较低. 基于内容分析的 BFS 爬行策略可以比较有效的对页面主题相关的程度进行预测, 但是这种方法会忽略页面之间的链接结构信息,

对主题团内部链接的重要性区分不够, 制约了在给定 URL 链接采集数量时策略的准确率. 本文所提处的 DTH 策略能够通过“隧道”到达新的主题团, 进而发现更多的主题相关网页, 所以其准确率较前两种策略要高, 尤其是随着下载网页个数的增多这一优势更加明显.

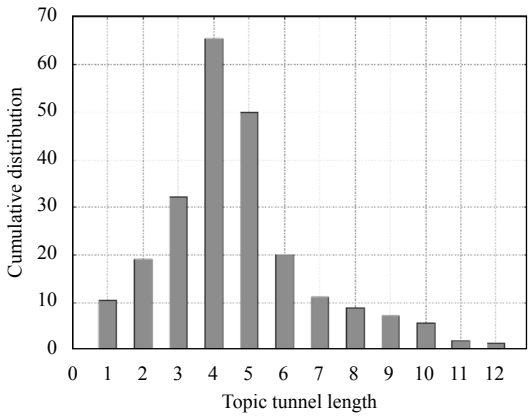

图 5 主题团之间的隧道长度 k 的分布

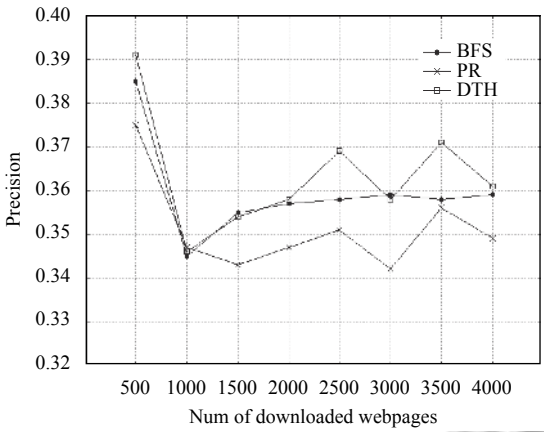

图 6 主题查询的准确率的变化趋势

主题搜索引擎在采用不同爬行策略时的返回页面数随总采集页面的变化趋势如图 7 所示. 实验结果可以发现, BFS 策略返回页面的数量略低于 PR 策略, 这是因为 PR 策略发生主题漂移, 却有可能有利于发现新的链接. DTH 策略在 2000 以下时的返回页面数量和前两者差不多, 但是在 2000 到 3000 时就出现返回页面数量的上升速度急剧下降, 这可能是因为初始化 URL 链接形成的主题团中的链接已经采集完毕, 而主题团的大小据统计一般在 1500 到 2000 之间. 在此之后 DTH 策略在 3000 到 4000 时上升的速度又能恢复正常, 这可能因为在 3000 到 3500 之间时本文提出的爬行策略找到了新的主题团, 前两种策略却在 3000 到 4000 之间没有变化.

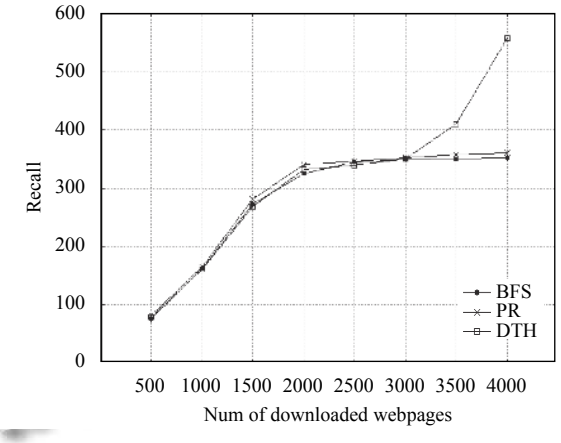

图 7 主题页面数量的变化趋势

综上, 本文设计的主题搜索引擎原型系统的准确率不低于采用 BFS 或者 PR 爬行策略的主题搜索引擎, 而召回率和采用 BFS 或者 PR 爬行策略的主题搜索引擎相比有了很大的提升. 实验进一步表明, 本文提出的基于动态隧道技术的爬行策略对改进主题搜索引擎的性能是有效的.

4 结论

针对现有主题爬行策略存在的主题孤岛问题, 提出了一种基于动态隧道技术的主题爬虫爬行策略. 该策略利用 URL 链接相关度预测方法动态调整不相关链接方向上的爬行深度, 使得爬虫能够进一步发现较多隐藏的主题相关链接. 同时, 该策略能有效防止主题爬虫因采集过多的主题无关页面而导致的主题漂移现象, 从而可以实现在保持主题语义信息的爬行方向上的动态隧道控制. 面向互联网网页的爬虫采集实验结果表明, 基于动态隧道技术的主题爬行策略提升了主题搜索引擎的准确率和召回率, 能够比较好的解决现有主题爬虫存在的主题孤岛问题.

参考文献

1 Santos A, Pasini B, Freire J. A first study on temporal dynamics of topics on the web. Proceedings of the 25th International Conference Companion on World Wide Web. Montréal, QB, Canada. 2016. 849–854. [doi: 10.1145/2872518.2889291]

2 彭涛, 孟宇, 左万利, 等. 主题爬行中的隧道穿越技术. 计算

机研究与发展, 2010, 47(4): 628–637.
3 Liakos P, Ntoulas A, Labrinidis A, *et al*. Focused crawling for the hidden web. World Wide Web, 2016, 19(4): 605–631. [doi: 10.1007/s11280-015-0349-x]
4 Du YJ, Liu WJ, Lv XJ, *et al*. An improved focused crawler based on semantic similarity vector space model. Applied Soft Computing, 2015, 36: 392–407. [doi: 10.1016/j.asoc.2015.07.026]
5 龚勇. 搜索引擎中网络爬虫的研究[硕士学位论文]. 武汉: 武汉理工大学, 2010.
6 张乃洲, 李石君, 余伟, 等. 使用联合链接相似度评估爬取 Web 资源. 计算机学报, 2010, 33(12): 2267–2280.
7 Diligenti M, Coetzee FM, Lawrence S, *et al*. Focused crawling using context graphs. Proceedings of the 26th International Conference on Very Large Data Bases. San Francisco, CA, USA. 2000. 527–534.
8 Bergmark D, Lagoze C, Sbityakov A. Focused crawls, tunneling, and digital libraries. Proceedings of the 6th European Conference on Research and Advanced Technology for Digital Libraries. Berlin, Germany. 2002. [doi: 10.1007/3-540-45747-X_7]
9 张涛. 基于 Nutch 的垂直搜索引擎研究与实现[硕士学位论文]. 天津: 南开大学, 2009.
10 Huaba. 结巴分词 (Java 版). https://github.com/huaban/jieba-analysis. [2019-05-12].