针对爬虫的域名链接过滤算法

文阳 陈文宇 袁野 朱建

[摘 要] 认为传统的基于主题的链接过滤算法虽然在某一领域的主题爬虫中使用广泛,但该方法只关心抓取的网页与主题之间的相关性,忽略了网站自身链接的结构特点。提出基于域名的链接过滤算法,该方法对基于网页链接中域名的结构特点进行比较,同时以基于主题的链接过滤算法作为辅助,判断出无用的垃圾链接。与单一基于主题的链接过滤算法相比较,基于域名的链接过滤算法的判断方式更为全面,链接过滤效率更高,从而能有效地提高网络爬虫的抓取效率和情报检索的效率。最后,通过仿真实验证明该算法的有效性。

[关键词] 网络爬虫 链接过滤 域名过滤 主题过滤
[分类号] G350.7
DOI:10.13266/j.issn.0252-3116.2014.20.019

1 引 言

通用搜索引擎主要包含网络爬虫系统和信息检索系统两部分:网络爬虫主要用于根据链接库来抓取互联网上的网页数据并建立相应的索引数据库,而信息检索系统主要用于在索引数据库中进行相关搜索,这两部分是搜索引擎必不可少的部分。其中链接过滤算法主要用于在网络爬虫中过滤掉无关的垃圾链接,减少爬虫的资源消耗,提高爬虫的抓取效率[1-2]。

目前,许多网络爬虫所使用的链接过滤算法都是基于主题的链接过滤算法,虽然该类方法在某一领域的主题爬虫中使用广泛,但该方法只关心抓取的网页与主题之间的相关性,忽略了自身网站链接的结构特点[3-4],本文提出了针对爬虫的域名链接过滤算法。该算法根据网页链接中域名的结构特点,并将基于主题的链接过滤算法作为辅助,判断出不属于该网站或主题不相关的页面。与单一基于主题的链接过滤算法相比较[4-6],基于域名的链接过滤算法的判断方式更为全面,链接过滤效率更高,从而能有效地提高网络爬虫的抓取速度,缩短索引数据库的刷新周期。

基于域名的链接过滤算法主要分为两个阶段:第一阶段是域名过滤阶段,该阶段为必选阶段。即将抓取链接的核心域名与关键域名数据库中的域名集合进行匹配,如果匹配成功,则直接保留页面,否则直接丢弃或交给第二阶段进行再次过滤(开启第二阶段过滤);第二阶段是主题过滤阶段,为可选阶段。即将页面内容与主题关键词集合进行相关度评分的计算,当网页的相关度评分高于某个事先制定好的阈值时,该页面将被保留,否则直接丢弃。

最后,在仿真实验中,将基于域名的链接过滤算法与原始网络爬虫进行对比,最终证明该算法的有效性。

2 基于域名的链接过滤算法

2.1 算法核心思想

2.1.1 域名过滤阶段 网页链接一般主要由 HTTP 协议名、网站域名和网页所在位置三部分组成,网站域名中独一无二的核心域名是过滤算法的关键依据。域名过滤阶段的核心思想就是根据网页链接中的核心域名是否在事先创建好的关键域名数据库中,来判断该网页链接是否属于某一指定范围内的网页。

2.1.2 主题过滤阶段 主题过滤阶段作为域名过滤阶段的补充,主要是为了保留一部分不属于指定域名范围但网页内容与指定主题十分相似的页面,该阶段可选择性开启。

通常情况下,每一个网页的自身内容都有一个主题,只有小部分广告链接或诈骗链接没有相关主题,所以可以通过判断一个网页的主题是否与事先制定的主题相关,来判断该网页是否应该被过滤。

主题过滤阶段每个网页的主题相关度评分计算方

[作者简介] 文阳,电子科技大学图书馆馆员,硕士;陈文宇,电子科技大学计算机学院教授,博士,通讯作者,E-mail:cwy@uestc.edu.cn;袁野,电子科技大学计算机学院硕士研究生;朱建,电子科技大学计算机学院硕士研究生。
收稿日期:2014-07-21 修回日期:2014-09-01 本文起止页码:125-130 本文责任编辑:杜杏叶

式如公式(1)所示:

$$Score(D) = (\sum_{q \in keys} D.ct(q) * ct.weight + D.tl(q) * tl.weight) * (1 + d * (num(keys, D) - 1)) \quad (1)$$

其中,*Score* 表示文档 *D* 的主题评分;*keys* 表示事先创建的主题关键词集合;*q* 表示一个主题关键词;*D.ct*(*q*)表示关键词 *q* 在文档 *D* 的正文段中是否出现,出现值为 1,否则为 0;*ct. weight* 表示关键词出现在正文段所获得的评分;*D. tl*(*q*)表示关键词 *q* 在文档 *D* 的标题段中是否出现,出现值为 1,否则为 0;*tl. weight* 表示关键词出现在标题段所获得的评分;*d* 表示一个增益系数,当有多个不同的关键词出现在文档中的时候,它适当地提高评分;*num*(*keys*,*D*)表示在主题关键词集合 keys 中,有多少个不同的关键词出现在文档 D 中。公式(1)的核心思想是在文档中出现的主题关键词越多,该文档的主题相关度评分越高。

2.1.3 算法架构 基于域名的链接过滤算法的组织架构主要分为核心数据库和链接过滤分析组件两部分。核心数据库主要包含起始链接数据库、常用域名数据库、关键域名数据库和主题数据库,而链接过滤分析组件主要包含域名过滤分析器和主题过滤分析器。

基于域名的链接过滤算法的框架如图 1 所示:

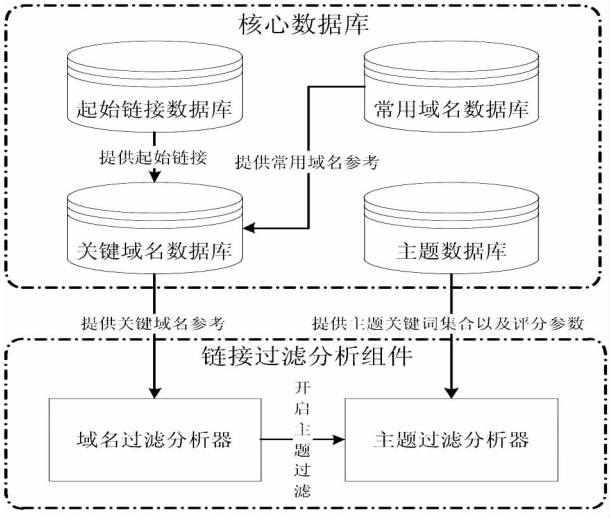

图 1 基于域名的链接过滤算法的架构

其中,关键域名数据库、常用域名数据库、主题数据库、域名过滤分析器和主题过滤分析器是基于域名的链接过滤算法架构的核心组件,共同构建起基于域名的链接过滤算法的框架结构。

2.1.4 算法工作流程 基于域名的链接过滤算法主要分为初始化数据库和链接过滤两个部分:初始化数据库部分是对核心数据库中的关键域名数据库和主题数据库进行初始化设置,为链接过滤部分准备好参考数据源;链接过滤部分是对抓取的链接或页面进行过滤分析,丢弃掉无关的垃圾链接或广告链接。

图 2 展示了该算法的流程:

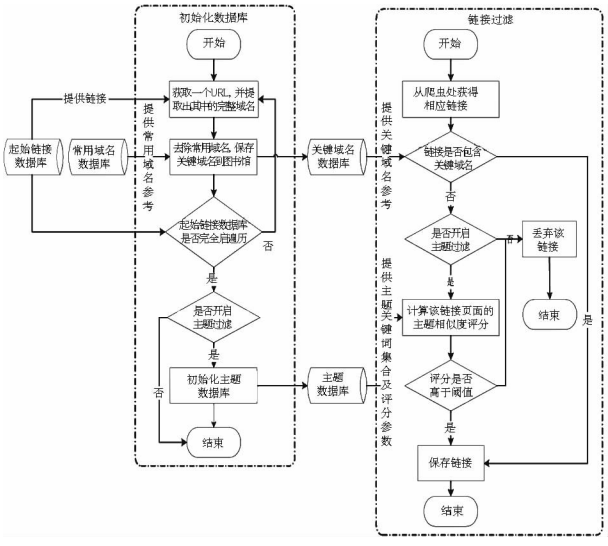

图 2 基于域名的链接过滤算法的流程

(1)初始化数据库流程:

- 从起始链接数据库中获得一个未被访问过的抓取链接,并去除掉该链接中与域名过滤无关的 *HTTP* 协议名和网页所在位置,只保留相应的完整域名。如获得的链接地址为 *http://www. sina. com. cn/*,只提取出其中的完整域名 *www. sina. com. cn*。
- 根据常用域名数据库中所提供的常用域名,去除完整域名中所包含的常用子域名,同时将剩下的子域名存储到关键域名数据库中。如完整域名为 *www. sina. com. cn*,而常用域名数据库中包含 *www*、*com* 和 *cn* 这些子域名,就将 *sina* 这个子域名作为关键域名存储到关键域名数据库中。
- 判断起始链接数据库中是否有未被分析过的网页链接:如果有,回到第(1)步,继续分析起始链接数据库中剩下的抓取链接;如果没有,表示关键域名数据库构建完成。
- 判断是否启用了主题过滤:如果没有启用主题过滤,则初始化数据库流程结束;否则根据配置文件对主题数据库进行创建和初始化设置,其中主要包含主题关键词集合、评分参数和过滤阈值的设定。设置关键词在正文段和标题段出现的评分分别是 1 和 3,增益系数为 0.2,过滤阈值为 8。

(2)链接过滤流程:

- 从网络爬虫处获得需要被过滤分析的网页链接,该链接通常是由网络爬虫从已抓取的页面内容中

解析出来的。

• 根据已经创建的关键域名数据库所提供的关键域名参考，对网页链接中是否包含关键域名进行判断：如果该网页链接包含关键域名数据库中的任一关键域名，就保存该链接等待网络爬虫下一层抓取。如网页链接 *http://mil. news. sina. com. cn/*2014 - 01 - 16/0754760321. *html* 包含关键域名 *sina*，所以该链接不会被丢弃；否则进入下一步过滤判断。

• 判断是否开启主题过滤：如果没有开启主题过滤，就丢弃该网页链接；否则进行主题过滤分析。

• 根据主题数据库所提供的主题关键词集合与评分参数计算该链接所指页面的主题相关度评分，计算方法见公式(1)。

• 将链接所指页面的主题相关度评分与事先制定好的过滤阈值进行比较：如果该链接的主题相关度评分高于过滤阈值，则保留该链接等待网络爬虫的下一层抓取；否则丢弃该网页链接。

综上所述，初始化数据库部分是该算法参考数据源的构建阶段，而链接过滤分析部分才是该算法的主体，这两部分共同组成了基于域名的链接过滤算法。

2.1.5 *关键数据结构* *Trie* 树(字典树)是一种常用于统计大量字符串的树形结构，该数据结构利用字符串的公共前缀来进行字符串的查找操作，从而最大限度地避免无谓的字符串比较，以致查询速度非常快，查询效率甚至比哈希表还高。所以核心数据库中的常用域名数据库和关键域名数据库使用 *Trie* 树作为自身的数据组织结构，以提高域名匹配的速度。

Trie 树也用于搜索引擎的文本词频统计，它主要包含 3 个基本特性：①树的根节点不包含任何字符，而除根节点以外的每一个节点都只包含一个字符。②从树的根节点到某一节点可以组成一条路径，将该路径上所经过的字符连接起来，就是该节点所对应的字符串。③树中每个节点的所有子节点包含的字符都各不相同。图 3 为展示一个 *Trie* 树的示例：

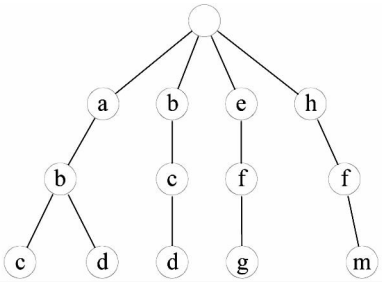

图 3 Trie 树示例

该 *Trie* 树中存储着 *a*、*b*、*ab*、*abd*、*bcd*、*hf*、*hfm* 等各不相同的单词。同时，该 *Trie* 树的根节点不包含任何字符，并且除根节点以外的每一个节点都包含一个字符。

当查询单词 *efg* 时，只需在该 *Trie* 树中比较 3 次就可知该单词存储在 *Trie* 树中，即查找单词成功时需要比较的次数就是单词的长度，远远快于一个一个单词进行比较的方法。而且，当 *Trie* 树用于储存含有大量相同前缀的单词集合时，其时间效率和空间效率更高。

综上所述，*Trie* 树在字符串的查找方面速度优势明显，唯一的缺点就是空间占用率较大，但考虑到常用域名数据库和关键域名数据库中存储的域名有限，所以最终决定使用 *Trie* 树作为常用域名数据库和关键域名数据库的数据存储结构。

3 仿真实验

本节基于域名的链接算法的衡量参数，展示仿真实验的结果，并对实验数据进行讨论。

3.1 算法衡量参数

基于域名的链接过滤算法的主要作用是过滤掉链接集合中的广告链接和垃圾链接。该算法对搜索引擎的好处有两点：一是该算法能有效地减少网络爬虫需要抓取的链接数量，避免无用的资源消耗，以此提高网络爬虫的抓取速度；二是经过该算法过滤后的链接所指网页质量更高且数量更少，有效地提高了全文索引数据库的文档质量，也减少了全文索引数据库所占用的存储空间。所以，要衡量基于域名的链接过滤算法的好坏，应该从网络爬虫的执行速度和抓取的链接数量这两方面进行考虑。

本节提出时间链接积(TLP)这个概念，用来评判基于域名的链接过滤算法的高效性，计算方法如公式(2)所示：

$$TLP = Time * LinkNum \quad (2)$$

其中，*Time* 表示网络爬虫执行所需要的时间，以秒为单位；*LinkNum* 表示网络爬虫所抓取的链接数量。

在相同的抓取条件下，如果网络爬虫的 TLP 值越小，则网络爬虫的抓取效率越高。

3.2 实验与讨论

基于域名的链接过滤算法的仿真实验主要分为两个部分：仿真实验一对指定的网站分别进行两次抓取，一次使用原始网络爬虫(即 Nutch 自带的网络爬虫)[7-8]，另一次使用实现了基于域名的链接过滤算法的网络爬虫，但只开启域名过滤阶段，最后分别记录这两次抓取的实验数据并进行对比；仿真实验二对仿真

实验一中的网站进行再一次抓取,这次也使用实现了基于域名的链接过滤算法的网络爬虫,但开启了域名过滤和主题过滤这两个阶段,最后记录相应的实验数据并与仿真实验一中的实验数据进行对比。

3.2.1 *仿真实验一* 在网络爬虫开始抓取前,需要对一部分组件进行初始化。原始网络爬虫只需要初始化起始链接数据库,而使用域名过滤的网络爬虫需要初始化的组件包含起始链接数据库和常用域名数据库。所以本小节将首先介绍起始链接数据库和常用域名数据库的初始状态,具体内容如下:

(1)起始链接数据库:http://www.163.com/(网易门户)、http://www.souhu.com/(搜狐门户)、http://www.uestc.edu.cn/(电子科技大学)、http://www.ctbu.edu.cn/(重庆工商大学)。

(2)常用域名数据库:www、com、cn、edu、org、gov、net。

同时,根据起始链接数据库和常用域名数据库的初始状态,可以推理出关键域名数据库将包含 163、souhu、uestc 和 ctbu 这几个关键域名。

将起始链接数据库和常用域名数据库的初始状态设置好后,便开启网络爬虫进行抓取,所得的实验数据如表 1、表 2 所示:

表 1 使用原始网络爬虫的实验数据

实验数据 / 指定网站	抓取层数	原始网络爬虫		
		抓取链接总数(个)	爬虫执行时间(秒)	TLP(秒·个)
网易门户	4	12 155	406.078s	4 935 878.09
搜狐门户	4	6 086 284	30 060.578s	182 957 214 912.152
电子科技大学	4	619 957	5 966.187s	3 698 779 393.959
重庆工商大学	4	551 444	8 477.375s	4 674 797 579.5

表 2 使用域名过滤的网络爬虫的实验数据

实验数据 / 指定网站	抓取层数	使用域名过滤的网络爬虫		
		抓取链接总数(个)	爬虫执行时间(秒)	TLP(秒·个)
网易门户	4	11 259	378.938s	4 266 462.942
搜狐门户	4	5 662 259	25 556.313s	144 706 463 291.067
电子科技大学	4	235 550	3 044.109s	717 039 874.95
重庆工商大学	4	374 389	6 631.046s	2 482 590 680.894

比较表 1 和表 2 中抓取的链接总数这一列,可知当抓取相同的网站时,使用域名过滤的网络爬虫能够有效地过滤掉无关的链接,从而降低抓取的链接总数,而且抓取的链接数量越多,过滤的垃圾链接也越多。被排除掉的链接中垃圾链接所占的比例,网易门户为 7.37%,搜狐门户为 6.97%,电子科技大学为 62%,重庆工商大学为 32.11%。

图 4 是各个网站被抓取的链接总数的对比:

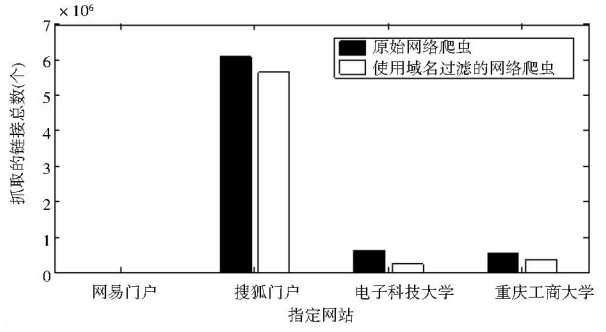

图 4 各个网站被抓取的链接总数的对比(仿真实验一)

由于网易门户被抓取的链接总数与其他 3 个网站被抓取的链接总数相比差距较大,所以图 4 无法很好地展示其对比结果。而网易门户被抓取的链接总数这么少的原因可能是网易门户所链接的子网站拒绝网络爬虫的抓取,故网络爬虫只抓取了一小部分网页。根据图 4,可知抓取的网站不同,过滤掉的无关链接所占比重也不同,如搜狐门户被过滤掉的链接只占总链接的一小部分,而电子科技大学被过滤掉的链接超过了总链接的一半。总体来说,抓取的链接数量越多,过滤掉的无关链接也越多。

然后,比较表 1 和表 2 中爬虫执行时间这一列,可知当抓取相同的网站时,使用域名过滤的网络爬虫的执行速度快于原始网络爬虫。图 5 为爬虫执行时间的对比:

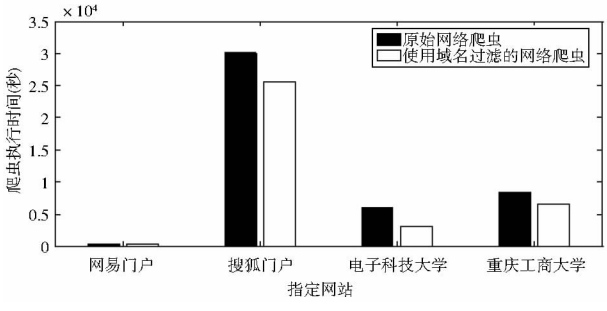

图 5 爬虫执行时间对比(仿真实验一)

由于网络爬虫在抓取网易门户时,其执行时间与抓取其他 3 个网站时相比少很多,所以图 5 无法很好地展示其比较结果。根据图 5,可知使用域名过滤的网络爬虫和原始网络爬虫抓取相同的网站时,前者消耗的时间更少,而且两者的执行时间与抓取的链接总数有一定的关系,即抓取的链接总数越少,速度越快。

比较表 1 和表 2 中 TLP 这一列,可知当抓取相同的网站时,使用域名过滤的网络爬虫的 TLP 比原始网络爬虫的 TLP 小,也就是说使用域名过滤的网络爬虫

速度更快,过滤的无关链接更多。图 6 为 TLP 的对比:

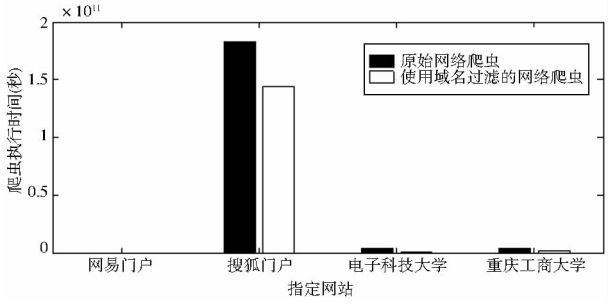

图 6 TLP 对比(仿真实验一)

根据图 6,可知使用域名过滤阶段的网络爬虫的抓取效率更高。

综上所述,在抓取的链接总数、爬虫执行时间和 TLP 3 个方面,使用域名过滤的网络爬虫相比于原始网络爬虫都更为优秀,所以仿真实验一证明了域名过滤阶段的有效性和高效性。

3.2.2 *仿真实验二* 仿真实验二中所使用的网络爬虫开启了域名过滤和主题过滤两个阶段,即使用的是完整的基于域名的链接过滤算法,所以除了设置仿真实验一中的组件,还需要设置主题数据库。其中起始链接数据库和常用域名数据库与仿真实验一的初始设置相同,而主题数据库的初始参数如下:

(1)主题关键词集合:“研究生”、“论文”、“毕业”、“体育”、“篮球”。

(2)ct. weight 和 tl. weight:0.1 和 0.3。

(3)d:0.2。

(4)过滤阈值:0.5。

将需要初始化的组件设置完毕后,便开启网络爬虫进行抓取,最后所得的实验数据如表 3 所示:

表 3 使用域名过滤和主题过滤的网络爬虫的实验数据

实验数据 / 指定网站	抓取层数	使用域名过滤和主题过滤的网络爬虫		
		抓取链接总数(个)	爬虫执行时间(秒)	TLP(秒·个)
网易门户	4	11 265	380.112s	4 281 961.68
搜狐门户	4	5 662 411	25 566.999s	144 827 480 484.589
电子科技大学	4	235 889	3 059.873s	721 790 382.097
重庆工商大学	4	374 617	6 645.084s	2 489 361 432.828

将表 3 中的实验数据与仿真实验一中的实验数据进行对比,可知表 3 中的各项数据与表 2 中的各项数据差距不大,只是数值略高,这是因为开启了主题过滤阶段后,会保留一小部分主题相关度符合标准但域名不符合标准的链接,这样会增加网络爬虫的消耗,所以就会导致出现上述情况。

为了更好地理解实验数据的比较情况,以图 7、图 8、图 9 分别展示所抓取的链接总数的对比情况、爬虫执行时间的对比情况和 TPL 的对比情况:

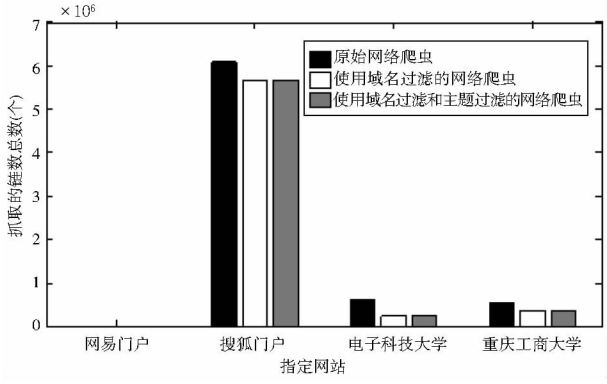

图 7 抓取的链接总数的对比(仿真实验二)

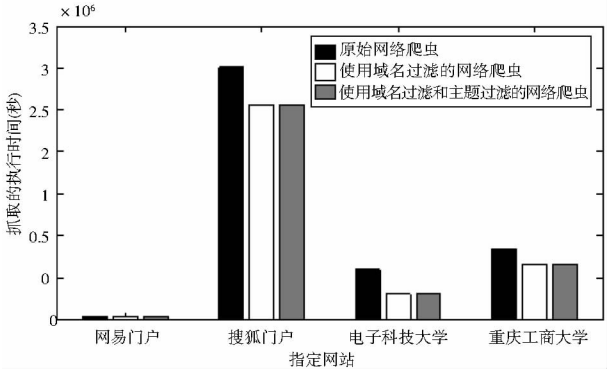

图 8 爬虫执行时间对比(仿真实验二)

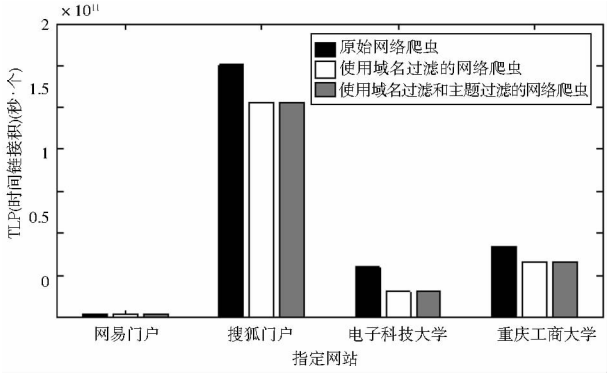

图 9 TLP 对比(仿真实验二)

从图 7、图 8 和图 9 可以看出,使用域名过滤和主题过滤的网络爬虫与原始网络爬虫相比,各方面都较为优秀,而与只使用域名过滤的网络爬虫相比,能够增加少量的消耗以有效地保留主题相关性高的页面。总而言之,仿真实验二证明了主题过滤阶段的有效性,仿真实验一和仿真实验二共同证明了基于域名的链接过滤算法的有效性和高效性。

4 总 结

随着搜索引擎的不断发展,网络爬虫中的链接过滤算法也越来越重要。本文提出了一种基于域名的链

接过滤算法,该算法对基于网页链接中域名的结构特点进行比较,同时以基于主题的链接过滤算法作为辅助,使其判断方式更为全面,链接过滤效率更高,从而能有效地提高网络爬虫的抓取效率。从仿真实验的结果可知,基于域名的链接过滤算法不但能够有效地过滤无关的垃圾链接,还能有效地提高网页的抓取速度。

在未来的工作中,主要有以下 3 点需要改进:①如何通过对模板网页进行分析自动设置常用域名数据库和主题数据库的配置文件;②如何使主题过滤阶段的评分公式更加公平,使计算出来的评分能更准确地反映网页的相关程度;③如何对以 IP 地址形式出现的垃圾链接进行判断。

如果解决了上述问题,本文所提出的算法会更加高效,而其通用性也会得到极大的提高。

参考文献:

[1] 张云秋,安文秀,冯佳. 探索式信息搜索行为研究[J]. 图书情报工作,2012,56(14):67 - 72.

[2] A. Emtage, P. Deutsch. Archie: An electronic directory service for the Internet[C]//Proceedings of the Winter 2010 Usenix Conference. California:USENIX, 2010:93 - 110

[3] Alberti B, Anklesaria F, Lindner P, et al. The Internet Gopher protocol: A distributed document search and retrieval protocol[J]. The Journal of Universal Computer Science, 1991, 24(2): 235 - 246.

[4] Pant G, Srinivasan P. Learning to crawl: Comparing classification schemes[J]. ACM Transactions on Information Systems (TOIS), 2005, 23(4): 430 - 462.

[5] Knoblock C A, Arens Y. An architecture for information retrieval agents[C]//Working Notes of the AAAI Spring Symposium on Software Agents. New York:SIGIR, 2010:49 - 56.

[6] Abiteboul S, Preda M, Cobena G. Adaptive on - line page importance computation[C]//Proceedings of the 12th International Conference on World Wide Web. Budapest:Springer, 2012:280 - 290.

[7] Cutting D, Pedersen J. Optimization for dynamic inverted index maintenance[C]//Proceedings of the 13th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. New York:SIGIR, 2011:405 - 411.

[8] Salton G, Buckley C. Term-weighting approaches in automatic text retrieval[J]. Information Processing & Management, 2009, 24(5): 513 - 523.

Link Filtering Algorithm of Domain Name in View of the Crawler

Wen Yang[1] Chen Wenyu[2] Yuan Ye[2] Zhu Jian[2]

[1]Library of University of Electronic Science & Technology of China, Chengdu 611731

[2]School of Computer Science and Engineering, University of Electronic Science & Technology of China, Chengdu 611731

[**Abstract**] Traditional link filtering algorithm based on topic even though the topic in the field of a crawler is widely used, but this method only cares about fetching the correlation between subject and the website, and ignoring the website links to the structure characteristics of itself. The connection filtering algorithm is proposed based on domain name, and this method is based on the structure characteristics of the domain name in the web link. Link filtering algorithm will be based on the theme at the same time as the auxiliary, judge the useless garbage links. Compared with the single link filtering algorithm based on theme, link filtering algorithm based on domain name is a more comprehensive judgment way. Besides, link filter is more effective, which can effectively improve the efficiency of the web crawler capture, and improve the efficiency of information retrieval. Finally, through the simulation experiment proves the validity of the algorithm.

[**Keywords**] Web crawler connection filtering domain filtering theme filtering