

文章编号:1001-9081(2004)12Z-0174-03

## 链接分析对主题爬虫的改进

汪涛<sup>1,2</sup>, 樊孝忠<sup>1</sup>

(1. 北京理工大学 计算机科学与工程系, 北京 100081; 2. 炮兵学院 三系, 安徽 合肥 230031)

(wander@bit.edu.cn)

**摘要:**在分析总结两种主题爬虫设计的基础之上,研究了用链接分析改进主题爬虫的方法。通过实验,比较引入链接分析前后的结果,论证了其设计可行性与可操作性,为实现定向信息采集奠定了良好的基础。

**关键词:**链接分析; Web 结构挖掘; 主题爬虫; 相关度; 信息采集

**中图分类号:** TP393 **文献标识码:** A

## 1 主题爬虫的作用

爬虫是搜索引擎的核心部件,搜索引擎利用它从 Web 中采集网页,完成重要的第一步工作。传统搜索引擎的页面采集面向整个 Web,普通爬虫能够顺利完成工作,但 Web 信息急剧膨胀使搜索引擎专用化成为发展趋势,定向采集信息成为搜索引擎一个重要研究方向,主题爬虫也应运而生,它根据事先确立的主题,在受限领域内进行定向页面采集。

我实验室承担了某公司资助展开研究的领域信息专家评估系统项目,该系统从网上获取受限领域内的信息,针对用户要求进行相关信息的提取和评估,并针对具体情况给出适当的建议。显然,系统首先要完成的工作就是信息采集,如果采用普通爬虫将会得到大量的无关网页,使信息评估的工作量加大,可信度降低,而采用主题爬虫是较优的解决方案,因此主题爬虫的设计成为一个相对独立的部分从系统中划分出来。

## 2 两种成功的设计方案

## 2.1 主题爬虫的系统组成

主题爬虫最初的设计思想是考虑对页面的过滤,不像普通爬虫对所有页面的链接进行处理,而是先对页面与受限领域的主题相关度进行分析,只有当其主题相关度符合要求时才处理该页面中的链接,因为如果该页面和本领域比较相关,它所包含的链接和领域相关的几率也较大,这样提高了爬行精度,虽然会遗漏少数页面,但综合效果是令人满意的。

因此,主题相关度的分析是主题爬虫设计的关键,最简单的可以基于关键词进行分析,更深入的可以上升到语义和概念层次。第一阶段基于关键词的主题相关度分析的主题爬虫设计取得了较好的效果,主要思路是:首先在领域专家的参与下,确定一组带有权重的能够代表受限领域的关键词,用它表示确定的主题;然后对页面进行关键词提取,采用向量空间模型算法计算网页的主题相关度决定页面的取舍。第二阶段,上升到语义和概念层次进行主题相关度的分析,在同等耗时

量级条件下,获得了更高的精度。

不论是基于关键词还是基于概念,主题爬虫基本的系统组成一样,以普通爬虫为基础,对其进行功能上的扩充,在对网页的处理过程中增加如下模块:主题确立模块、优化初始种子模块、主题相关度分析模块和排序模块,不同之处在于主题

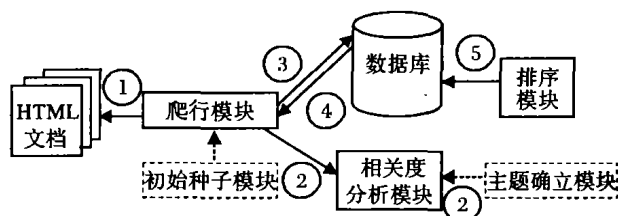


图1 系统组成

- ① 爬虫模块取回网页;
- ② 调用相关度分析模块,对网页进行相关度分析;
- ③ 爬行模块根据分析的不同结果进行相应的处理;
- ④ 爬行模块从数据库取出等待处理的 URL 继续工作,循环到①,直至没有新的 URL;
- ⑤ 对网页的重要程度进行排序。

## 2.2 基于关键词的方案

在基于关键词的方案中,把关键词的个数  $n$  作为向量空间的维数,每个关键词的权重  $w_i$  作为每一维分量的大小,则主题用向量表示为:

$$\alpha = (a_1, a_2, \dots, a_n), i = 1, 2, \dots, n, a_i = w_i$$

对页面进行分析,统计关键词出现频率,求出频率之比,以出现频率最高的关键词为基准,其频率用  $x_i = 1$  表示;通过频率比,求出其他关键词的频率  $x_i$ ,则该页面对应向量的每一维分量为  $x_i w_i$ ,页面主题用向量表示为:

$$\beta = (x_1 w_1, x_2 w_2, \dots, x_n w_n), i = 1, 2, \dots, n$$

用两个向量夹角的余弦表示页面的主题相关度:

$$\cos \langle \alpha, \beta \rangle = (\alpha, \beta) / |\alpha| |\beta| = \frac{x_1 w_1^2 + x_2 w_2^2 + \dots + x_n w_n^2}{\sqrt{w_1^2 + w_2^2 + \dots + w_n^2} \sqrt{x_1^2 w_1^2 + x_2^2 w_2^2 + \dots + x_n^2 w_n^2}}$$

收稿日期:2004-03-18;修订日期:2004-05-27

作者简介:汪涛(1977-),男,湖北钟祥人,博士研究生,主要研究方向:计算机网络、Web 信息处理; 樊孝忠(1948-),男,河南人,教授,博士生导师,主要研究方向:自然语言处理、数字化网络教学。

指定一个阈值  $r$ , 当  $\cos\langle\alpha, \beta\rangle \geq r$  时就可以认为该页面和主题是比较相关的,  $r$  的取值根据经验和实际要求确定。如果想获得较多的页面, 可以把  $r$  设小一点; 要获得较少的页面, 可以把  $r$  设得大一点。

### 2.3 基于概念的方案

在基于概念的方案中, 由于“知网”(英文名称为 HowNet) 能够为汉语语义处理提供一个比较全面的语义知识库, 可以此作为基于概念的主题相关度分析的概念网络。所有的关键词  $t$  形成关键词集合  $T = (t_1, t_2, \dots, t_n)$ , 设关键词  $t_i$  有  $a_i$  ( $a_i \geq 1$ ) 个概念义, 记  $C(t_i) = (c_i^1, c_i^2, \dots, c_i^{a_i})$ , 从关键词集合  $T$  转化扩充得到关键词概念集合  $C_T$ , 其元素为每个关键词的每个概念义, 令  $tc_i^j = (t_i, c_i^j)$ , 其中  $j = 1, 2, \dots, a_i$ ,  $tc_i^j$  表示第  $i$  个关键词的第  $j$  个概念义, 则有  $C_T = \{(t_1, c_1^1), (t_1, c_1^2), \dots, (t_1, c_1^{a_1}), \dots, (t_n, c_n^1), \dots, (t_n, c_n^{a_n})\} = (tc_1^1, tc_1^2, \dots, tc_1^{a_1}, \dots, tc_n^1, \dots, tc_n^{a_n})$ 。考虑一词多义, 删除其他完全相同的概念, 将集合  $C_T$  变为概念集合  $C$ , 设由  $m$  个概念组成, 记为  $C = (c_1, c_2, \dots, c_m)$ 。把概念的个数  $m$  为向量的维数, 每个概念的权重  $w_i$  (该概念对于主题贡献度) 作为每一维分量的大小, 得到主题的新的向量表示  $\alpha$ 。

然后, 与基于关键词的方案对应, 对页面进行概念提取, 在概念集合  $C$  上页面  $d$  的特征向量为  $V_c(d) = (c_1, w_{c_1(d)}; \dots; c_i, w_{c_i(d)}; \dots; c_m, w_{c_m(d)})$ , 其中  $w_{c_i(d)}$  为概念  $c_i$  在页面  $d$  中的权重<sup>[9]</sup>, 最终得到新的页面主题向量表示  $\beta$ 。

最后, 计算  $\cos\langle\alpha, \beta\rangle$ , 并将它和指定的阈值  $r$  进行比较, 来判断页面的主题相关度, 决定页面的取舍。

## 3 链接分析对设计方案的改进

基于关键词和基于概念的方案均能取得较好的爬行精度, 二者都是基于向量空间模型算法, 论证了用向量空间模型算法进行主题相关度计算的可行性和有效性。但从另一个角度来看, 想进一步提高爬行效率, 单纯使用向量空间模型算法显然不会取得太好的效果, 必须辅助采用其他方法。

上述两个方案的设计思想是考虑对页面的过滤, 不像普通爬虫对所有页面的链接进行处理。如果更进一步, 能避免对页面的所有链接进行处理, 理论上爬虫效率可以再次得到提高。从爬虫工作的基本原理可知, 爬虫工作的源动力就是网页间的链接, 链接又是有规律可寻, 如果能够充分分析利用链接信息, 应该可用于改进主题爬虫的设计。

### 3.1 链接分析研究现状

链接分析属于 Web 结构挖掘研究的范畴, Web 结构挖掘主要是从 Web 组织结构和链接关系中推导信息和知识。根据科学引文分析理论, 文档之间的互联数据中蕴涵着丰富有用的信息, 在通常的搜索引擎中由于考虑到结构的复杂性, 仅将 Web 看作是一个平面文档的集合, 忽略其结构信息。挖掘页面的结构和 Web 结构, 可以用来指导对页面进行分类和聚类, 找到权威页面、中心页面, 从而提高检索的性能; 同时还可以用来指导网页采集工作, 提高采集效率。

### 3.2 用链接分析 Web 站点结构

Web 上的信息看似杂乱无章, 其组织仍然有一定的结构, 这既包括由 URL 目录层次反映出来的物理结构, 也包括页面和页面间链接构成的逻辑结构。可以通过分析 Web 站点信息结构, 初步判断信息的类别, 作为预测采集的结构基础。

#### 1) 物理结构

一个完整的 URL 包括协议和路径两个部分。在 Web 站点中使用的大多数是 HTTP 协议, 其 URL 语法形式如下:

http://<host>:<port>/<path>?<searchpart>

其中 <host> 表示站点主机名(域名或 IP 地址); <port> 表示端口号; <path> 表示页面的路径; <searchpart> 表示 CGI 接口 GET 方法的参数表达式。对于一个站点来说, 能够用来表示站点结构的只有 <path> 部分。页面的路径和 Web 站点的文件系统是对应的, 也是一种分层的树型结构, 每个层之间通过“/”分开。

一个设计比较规范的站点, 内容的组织都是人工按照栏目目进行的, 每个栏目包括某个主题的内容。一个站点设置若干栏目, 内容较多的栏目则设置子栏目, 每个栏目或者子栏目的文件分目录进行存放。这样, 站点物理结构相同的页面就属于同一栏目, 它们有相同或相似的主题, 可以利用站点的物理结构来进行信息的采集。

#### 2) 逻辑结构

站点物理结构反映的是页面的存储方式, Web 页面内容之间的主要联系是通过页面间的链接来进行的。如图 2 所示, 页面之间的链接主要包括以下 5 种类型:

downward——下行链, 目标页面是当前页面的下级页面;

upward——上行链, 目标页面是当前页面的上级页面;

horizontal——水平链, 目标页面和当前页面处于同一目录;

crosswise——交叉链, 目标页面和当前页面不在同一路径上;

outward——外向链, 目标页面和当前页面不在同一站点。

通常情况下, 下行链的目标页面是对当前页面的详细描述, 上行链的目标页面是对当前页面的概括, 水平链的目标页面和当前页面属于同一领域内容, 交叉链和外向链主要表示和锚点信息指向内容相关。

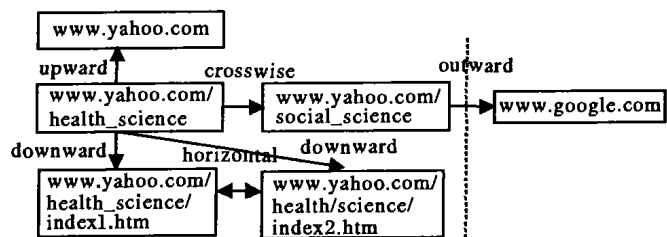


图 2 链接类型

### 3.3 用链接分析对主题爬虫进行改进

首先, 分析一下目前两个方案存在的问题, 无论是基于关键词的方案还是基于概念的方案, 在进行完页面的主题相关度分析后, 当其主题相关度符合要求时将处理该页面中的所有链接, 显然是不合适的, 因为即使页面本身和主题很相关,

但其中的链接指向的页面也可能有许多偏离了主题,如果不加区分地取回所有页面并分析主题相关度,不仅没有必要,反而会增加系统的工作量,影响爬虫的总体工作效率。

如果爬虫在取回页面之前能够对页面上的链接进行一次过滤,剔除那些明显偏离主题的链接,在后续工作中就不必分析这些链接指向的页面,可节省大量时间,因为相比之下,主题相关度分析耗时较多。

上述链接分析理论可以用于链接过滤,通过链接分析可以把页面中的链接分成五种不同的类型,对于不同类型链接采取不同的处理。水平链的两页面属于同一内容,下行链的目标页面是对当前页面的详细描述,对于水平链和下行链的目标页面应该取回;上行链的目标页面是对当前页面的概括,从统计的角度看,并非所有的目标页面和主题相关,对于上行链的处理以一定的概率去取页面,取中间值 0.5 较合适;交叉链和外向链的目标页面主要取决于锚点信息,应该对锚点进行分析,可以取链接周围的文本来确定目标页面是否和主题相关,考虑链接  $p \rightarrow q$ ,  $p$  中有若干链接标记,文本  $1 < a \text{ href} = "q" >$  锚文本  $</a >$  文本 2,统计在文本 1、锚文本、文本 2 出现指定主题的关键词次数,从而判断目标页面是否和主题相关来决定是否取回目标页面,文本 1 和文本 2 的长度经过试验设为 50 字节。这时只需要停留在关键词层次,不必上升到概念层次,关键词已经能够保证足够的精度和效率。

通过对不同类型链接采取不同的处理,可使主题爬虫的性能得到提高:链接分析过滤了大量无链接指向的网页(也会遗漏极少有关网页),降低了主题爬虫的负载;对于相关概率较低的交叉链和外向链,采取先简单统计锚点周围关键词出现次数的方法决定是否对目标页面采取进一步处理,减少了爬虫的工作时间。

修改后的算法如下:

1) 主题的关键词或概念表示。对选定的主题,从关键词或者概念的层次用向量将其形式化。

2) 页面的关键词或概念提取。从关键词或概念层次分析页面,得到它的向量表示。

3) 相关度计算。计算上述两个向量夹角的余弦。

4) 相关度分析。根据预先设定的阈值进行分析,决定页面取舍。

5) 对于相关度符合要求的页面进行链接分析,对页面中的链接进行过滤。

## 4 实验与结论

按照上述思想,分别对基于关键词的主题爬虫和基于概念的主题爬虫进行修改,并通过实验比较修改前后的数据。

软硬件环境不变:系统用 Jbuilder9 开发,运行在配置为 CPU:P4 2.4C,内存:1G 的机器上。为了和未引入链接分析的主题爬虫进行比较,实验仍然围绕显示器这个主题,因为目前该领域信息评估系统主要是做和监视器相关的信息评估。

实验参数:搜索深度 = 2(设得较小,防止搜索规模过大),线程数 = 200(要求在网络环境较好的情况下),起始种子 = 10(经过人工选择的较好的种子),阈值  $r = 0.1$ 。

表 1 基于关键词的主题爬虫的数据

类别	基于关键词		基于概念	
	链接分析前	链接分析后	链接分析前	链接分析后
提取文档/个	2568	1899	2032	1301
提取失败/个	436	103	125	27
拒绝文档/个	1	0	0	0
发现文档/个	3005	2002	2157	1328
收集数据/Byte	93 135 421	62 562 782	70 524 356	44 158 432
爬行时间/s	351	241	537	423

比较引入链接分析前后数据,可以看出主要有两个方面的变化:

1) 提取文档数减少了,原因是链接分析剔除了大量无关网页,仔细比较发现,极少数相关度较大的页面也被剔除,这些链接主要是上行链、交叉链和外向链,和开始的分析吻合。

2) 爬虫时间减少了,原因是剔除大量无关网页后,降低了爬虫工作负载。

总体上看,引入链接分析后,没有影响爬虫精度,但是爬虫速度却提高了,说明引入链接分析对主题爬虫的改进方案是可行的。

通过对主题爬虫的改进研究,得到如下结论:

1) 主题爬虫整体性能的提高必须从多方面入手,单纯使用一种方法只能单方面在一定限度内取得效果,例如单纯使用向量空间模型算法就存在其局限性。

2) 链接分析和向量空间模型算法的综合运用,同时实现了页面过滤和链接过滤,既可以保证主题爬虫的爬行精度,又可以保证较高的爬行速度,改进后的主题爬虫能够以更高的效率工作。

3) 通过基于关键词的主题爬虫、基于概念的主题爬虫和链接分析对主题爬虫的改进等一系列研究,为进一步深入研究奠定了基础,其研究方法思路也有很大的理论指导意义。

### 参考文献:

- [1] HEATON J. 网络机器人 Java 编程指南[M]. 童兆丰,李纯,刘润杰,译. 北京:电子工业出版社,2002.
- [2] 厉亮,等. 主题搜索引擎的探讨[A]. 李晓明,李星. 搜索引擎与 Web 挖掘进展[C]. 北京:高等教育出版社,2003. 34-40.
- [3] 李盛韬,等. 主题 Web 信息采集的研究与设计[A]. 孙茂松,陈群秀. 语言计算与基于内容的文本处理[C]. 北京:清华大学出版社,2003. 488-494.
- [4] 李宏乔,等. 基于关键词与概念相结合的混合信息检索模型[A]. 李晓明,李星. 搜索引擎与 Web 挖掘进展[C]. 北京:高等教育出版社,2003. 41-45.
- [5] 李振星,等. 专用搜索引擎中信息采集的预测与过滤方法[A]. 李晓明,李星. 搜索引擎与 Web 挖掘进展[C]. 北京:高等教育出版社,2003. 107-115.
- [6] 曹军. Google 的 PageRank 技术剖析[J]. 情报杂志,2002,(10): 15-18.
- [7] PAGE L, BRIN S, MOTWANI R, et al. The PageRank citation ranking: bringing order to the web[EB/OL]. <http://www-db.stanford.edu/~backrub/pageranksub.ps>, 2003-03-25.
- [8] 汪涛,樊孝忠. 主题爬虫的设计与实现[J]. 计算机应用,2004,24(6Z).
- [9] 汪涛,等. 基于概念的主题爬虫设计[J]. 北京理工大学学报,2004,24(10).