基于元搜索技术的主题新闻门户系统①

Topic-Specific News Portal System Based on Meta Search

白　鹤[1,2]　王劲林[1]　赵志强[1]　(1.中国科学院声 学研究所 国家网络新媒体工程技术研究中心　北京 100190; 2.中国科学院 研究生院 北京 100049)

摘　要： 主题新闻门户提供个性化的新闻信息集成服务，是现在企业级信息化应用的一个热点。实现了一个无需人工干预的主题新闻系统。它基于元搜索技术获得主题相关的新闻索引，之后构造分装器和采用启发式算法准确提取双层页面中的结构化信息；同时设计了结合 URL 和标题的新闻重复检测策略，提高抓取质量。原型系统运行以来得到了较优异的实验效果。

关键词： 新闻门户 元搜索 Web 信息提取

1　引言

专业搜索引擎中的新闻搜索覆盖了大量新闻源站点，它的出现是对门户站点的有效补充，可以为用户搜索与聚合主题相关的不同媒体多角度的新闻报道；另外也为主题新闻门户的自动实现和快速部署提供了一种思路和技术基础。

主题新闻门户技术的基本功能是信息集成和个性化服务，它根据企业级或个人应用的需求，通过 Web 挖掘技术及时地从 Web 新闻源中获取所关心的的新闻，在本地存储并呈现。这是企业信息化应用的一个热点。早期它的实现使用人工采集；然后出现了基于神经网络的主题追踪[1]的方法，需要有样本学习和页面相关度判断的开销，并且查准率一般。我们基于新闻搜索引擎的元搜索技术，提出启发式的新闻正文提取算法，实现了一个无需人工干预的主题新闻门户系统。

2　相关研究

2.1 元搜索

元搜索引擎[2](meta search engine)，是一种调用其它独立搜索引擎的引擎，它对多个独立搜索引擎进行整合、调用、控制和优化利用。当用户查询一个关键词时，它把用户的查询请求转换成其他搜索引擎能够接受的命令格式，并行地访问多个传统的搜索引擎来查询这个关键词，然后将返回的结果进行合并、重新排序等处理后，作为自己的结果返回给用户。严格地讲，元搜索引擎只是一个搜索代理程序，算不上一个真正独立的搜索引擎。从检索机制的角度看，元搜索引擎可算是一种分布式信息检索系统，有其检索覆盖面广、系统复杂度低等优点。

2.2 Web 内容挖掘

Web 内容挖掘[3](Web Data Mining)，属于 Web 数据挖掘的范畴，主要处理 Web 文本内容进行知识发现，它包括了页面相关性分析、页面分割和页面信息提取等任务。

在新闻门户技术中它的应用主要是页面信息提取，现在主要两种方法：利用信号处理中快速傅里叶变换(FFT)的提取方法[4]和基于分装器(wrapper)的方法。前者结合统计学原理实现，不需要先验知识，自动地抽取正文，对中文“正文式”文本比较有效，但计算复杂度较高；后者基于定义好的规则，快速准确地匹配目标信息，针对每个目标页面类型需要实现一个匹配模板。本系统需要对商用新闻搜索引擎返回的索引页面，解析以获得新闻页面的链接、新闻标题和

① 基金项目:国家高技术研究发展计划(863)(2008AA01A307)
收稿时间:2009-03-04

摘要等结构化信息，使用 wrapper 方法可以满足精确度要求。

3　架构和关键技术

本文实现的主题新闻门户系统的架构如图 1 所示，包含元搜索器、爬取器和索引器三大主要模块。系统的后台定义了主题需求，会按照一定时间间隔动态调用元搜索器请求新闻搜索引擎接口；分装器负责解析新闻搜索返回的列表，得到新闻链接的结构化信息；爬取器从 URL 数据库提取链接获取新闻页面，并用索引器解析得到正文；把正文和结构化信息一起存入关系数据库。搜索引擎接口是人工设置的新闻抓取源站点的入口。如表 1 所示。

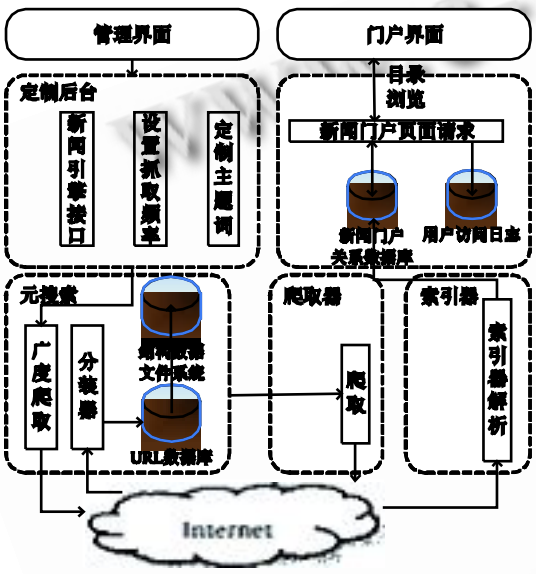

图 1　主题新闻门户系统体系架构图

表 1　权威新闻搜索接口

引擎	接口
Google	http://news.google.cn/news?q=##QUERY##&start=##START##
Baidu	http://news.baidu.com/ns?tn=news&word=##QUERY##&pn=##START##

“##QUERY##”，“##START##”是预定义字符，分别代表查询主题词和单页偏移量，系统运行中替换之进行动态赋值。

系统中需要解决三个关键问题：分装器的构造、新闻页面正文提取算法的设计和正文的重复性判断。

3.1 分装器

本模块负责对新闻搜索返回的目录页面进行解析，提取新闻链接的结构化信息：链接、新闻站点、摘要和标题。图 2 上是 Google 新闻搜索中对关键字“旅游”查询返回的索引页面一部分，包含了两个结构相似的实体模块。图 2 下是从实体模块提取结构化信息后生成的 XML 数据。

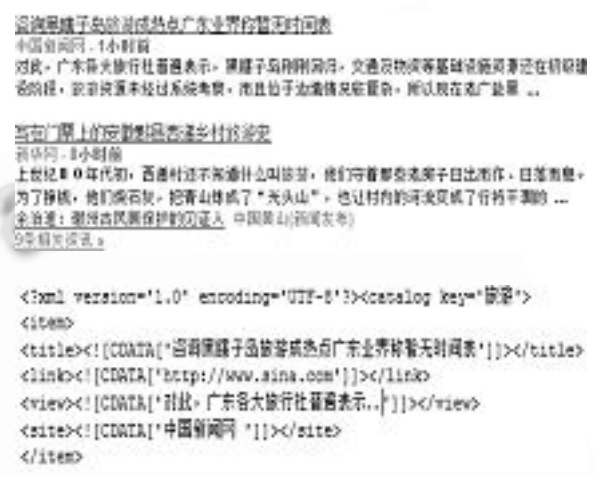

图 2　目录页的结构化结果

源站点的新闻目录页面由机器代码生成，是规范的半结构化页面，所以设计提取结构化信息的分匹配模板成为可能。本系统实现的分装器采用正则表达式作为匹配规则的表示语言。它的处理流程如图 3 所示：

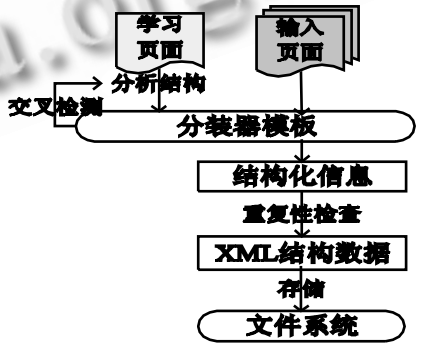

图 3　分装器处理流程

所示流程中，生成模板的过程要兼容特殊情况，比如模块内会出现的附图，模块尾部出现的添加信息；交叉检测是输入不同页面对模板进行测试，以保证其健壮性。本分装器把一个新闻实体模块归纳为一条规则，相对于每一个结构元素提取出规则，可以提高匹

配程序的运行效率

3.2 启发式新闻正文提取算法

索引器模块的输入包括两部分：通过网络请求获得的新闻正文页面和分装器解析后得到的结构化信息。本模块主要解决准确地提取新闻正文的问题。中文源新闻站点数量众多且更新频繁，如果用设计分装器的方法来提取正文，则开发和维护的成本过高，所以必须实现一个智能提取新闻正文的算法，这也是本文的重要创新点。

经过对大量新闻正文页面和获得信息的分析，得到了如下几条启发式经验：

1) 新闻页面中的正文包含在页面的块标签中，即 、<table>或标签内，这三类标签是算法解析的目标；

2) 结构化信息中的摘要文本一般位于新闻正文的前部，一般直接截取自正文文本，附带的 HTML 编码字符格式与文本相匹配；

3) 不匹配情况：当摘要首句中出现“讯”“报道”等词组后，新闻搜索引擎会习惯性改变其后空格和括号的格式；

根据以上启发性规则，设计的正文提取算法有如下思路：清理摘要信息格式，得到前 n(30<n<60)个字符，计为 p，作为正文首部可能的匹配字段；截取匹配到 p 位置之前的新闻文本，设置一个游标在其中执行后向查找，检索第一个未闭合的块标签位置 fpos；在 fpos 后的文本内查找对应前标签的闭合位置 rpos；fpos 与 rpos 之间的文本段作为此块标签的对应文本；综合长度等信息对比每种块标签抽取文本，判决最大可能的正文段。其流程见如下代码：

```
/*
INPUT: string, abstract
OUTPUT: main text
*/
1 pattern:="clean(abstract)->substr(0,n)
2 pos:=string->index_last(pattern)
3 fstr:="string->substr(0,pos)
4 labels:={div, table, span}
5 for i:=1 to labels.length do
6       fposs:=fstr->match_all(labels[i])
7       l:=fposs.count()
8       n:=0
9       while l-- do
10            if string[fposs[l]-1]=="/"
11                then n++
12                else if n==0 do
13                    fpos:=fposs[l]
14                    break
15                then
16                    n--
17      end while
18      str:=string->substr(fpos,-1)
19      rpos:="str->index_rear_label(labels[i])
20      p[i]=[fpos,rpos]
21 end for
22 main_text:=""
23 for i:=1 to labels.length do
24     if(main_text.length<(p[i].rpos-p[i].fpos))
25         main_text=string.substr(p[i].fpos,p[i].rpos)
26 end for
```

图 4　启发式正文提取算法伪码

算法充分利用了摘要和正文的特征，以及 HTML 结构化语言的一些标记，在代码中只处理了位置和游标信息，时间和空间复杂度比较小。文献[5]同样是利用摘要信息对正文进行提取，它采用了基于 DOM 树的做法，需要对整个文本进行处理，然后再匹配整段摘要，这样一方面处理较繁琐，占用了较多内存空间，另外没有考虑搜索元站点已经得到的摘要信息，导致不能精确匹配。

3.3 新闻重复判断机制

新闻抓取系统中对新闻的重复性判断是一个重要的问题，影响到新闻门户系统的质量和效率。系统在分装器得到结构化信息后判决新闻对象的重复性，重复就抛弃，否则就存储以备抓取器调用。系统需要处理两类重复新闻：同源新闻和转载新闻。

同源新闻的新闻对象来自不同元引擎，或者是同一元引擎不同时段的返回结果，它们的新闻本体集合{站点，标题，摘要，链接}相同，系统以 URL 作为同源新闻重复性的判断条件。对于 URL 这样的字符串查询，基于 Trie 多叉树结构算法效率要优于二叉树和 Hash 表查找算法，其算法时间复杂度跟组成树的节点无关，只与检索对象字符串中字符个数有关。基于新闻搜索引擎对结果按照时间排序的特点，判断的样本集合是在前 24 个小时内的抓取新闻形成的 Trie 树数据。

不同新闻源站点经常转载同一新闻，转载新闻的站点和链接不同，但新闻标题和正文最大程度相似。分装器解析后的结构化信息中包括标题和摘要，系统基于提高效率的角度考虑，只从标题出发判断是否转载新闻，同样可以达到较高的准确率。具体方法是：对新闻标题按照词库进行分词，在已抓取到的新闻标题倒查索引(Inverted Index)表中进行匹配，如果按词组在标题中的位置顺序找到全部匹配项，就可判定相同新闻已存在。

4　仿真和讨论

新闻门户系统基于元搜索技术，从高质量的新闻搜索引擎获得对应主题的不断更新的新闻源。系统内

管理员可以设置多个主题，在测试系统中定制了关键词“旅游”、“汽车”和“经济”，快速部署了对应主题的新闻门户，具有良好的主题的兼容性；从新闻引擎返回页面解析得到的新闻目标链接，起到了对抓取对象站点的负载进行离散化的目的，可以避免因为抓取频率太高会被目标新闻站点屏蔽，增强了系统的鲁棒性；系统中对解析成功的正文结构进行缓存，一定时间内如果重复访问同样站点，不经过正文提取算法，直接调用存储的匹配结构进行处理，提高了系统的效率。

我们在 1G 内存、P4.2.8G 的 CPU、安装 FreeBSD 系统的 PC 机上，使用 PHP 和 C 语言实现了主题新闻门户的原形系统。运行前在后台设置了 Google 新闻搜索和 Baidu 新闻搜索两个接口，分别订制了“旅游”、“汽车”和“经济”三个主题，设置抓取时间间隔为 1 个小时，元引擎的广度搜索设置为 5 页，对一个小时来说，5 页既能保证主题相关新闻的质量和数量，另外又避免了重复采集新闻目标链接。然后各自主题分别运行 24 个小时，得到如表 2 所示的数据。

表 2 原型系统运行数据

名称	旅游	汽车	经济
重复新闻	2391	2014	1765
目标链接	1932	1805	2163
解析页面	1590	1579	1900
解析成功	1401	1396	1697

表 2 中的“搜索引擎索引页”记录了请求元搜索返回的总页数，每页 20 条新闻，对应每个主题是 4800 条新闻；分装器从索引页解析出的正确的结构化结果集包含了两类数据：“重复新闻”和“目标链接”，分别记录了检测出来的重复数据和经过重复性验证的结构化新闻数据；分装器的提取正确率是两部分数据数目之和与总条目的比值，将三个领域的比值经过算术平均，其值为 83.8%，说明了模板规则的兼容性还有改进的地方。重复信息条目较多，大部分是同源新闻重复抓取，除了两个引擎返回一样新闻情况外，主要在于在搜索引擎更新较慢的时段，比如凌晨，页数设置过多只会增加冗余，可以采用分时段设置抓取广度的策略。

目标链接压入爬取器的 URL 队列，爬取器成功获取的页面成为了解析和正文提取的对象，一部分不能被解析器识别，处理过程中止；其余可以经过解析器处理的页面数据被统计为解析页面数目。对新闻页面解析算法来进行度量：查准率(precise)用解析成功数目与解析页面数目的比值来计算，

$$Precise\_rate = \frac{harvest\_pages}{parser\_pages}$$

三个领域的比值经过算术平均后的正确率为 88.2%；查全率(recall)用解析成功数目与目标链接数目的比值来计算，

$$Recall\_rate = \frac{harvest\_pages}{target\_pages}$$

三个领域的比值经过算术平均后的正确率为 76.1%。可见启发式算法可以达到较高的正文提取准确度，但识别率还有待提高。

5 结论

本系统基于元搜索技术，从权威度高的新闻搜索引擎抓取不断更新的新闻页面，构造分装器对得到的页面进行解析，提取新闻对象的结构化信息，并以此作为输入的先验知识实现提取新闻正文的启发式算法；同时设计双层的新闻重复性判断机制，避免了同源或者转载的重复新闻，提高了抓取新闻的质量。原型系统经测试，其查准率和效率达到较高水平，超过了文献[5]中的抽取方法，以此系统为基础实现了旅游资讯模块在“E 游天下”网站正式上线运行。

参考文献

1 李宝利,俞士汶.话题识别与跟踪研究.计算机工程与应用, 2003,39(17):7－10.
2 门凤超,濮德敏,王东菊.论元搜索引擎的实现技术与发展趋势.现代情报, 2008,7:61－63.
3 马保国,侯存军,王文丰.Web 数据挖掘技术及应用.计算机与数字工程, 2006,34(6):20－23.
4 李蕾,王劲林,白鹤,胡晶晶.基于 FFT 的网页正文提取算法研究与实现.计算机工程与应用, 2007,30(12):35－38.
5 刘敏,何渝.基于元搜索引擎技术的新闻对象抽取方法研究.北京工商大学学报, 2008,26(3):66－69.