互联网新闻搜索设计①

Design on Internet News Search Engine

于静波　余敦一　陈秋月　胡文学　（北京印刷学院 教学技术与网络中心 北京 102600）

摘　要：随着互联网的普及和信息技术的飞速发展，互联网新闻成了人们获取信息的重要渠道。为了方便广大网民看新闻，设计了一个互联网新闻搜索引擎。构建了一个新闻搜索引擎系统的总体架构。设计了一个轮询监控快速收录的新闻网页采集模块；另外把传统的文本分类算法、文本聚类算法应用到新闻网页智能分析处理中，设计了一个新闻网页的智能分析处理模块。

关键词：新闻采集 网页分类 相关新闻 新闻搜索 网页内容抽取

看新闻是许多网民上网的主要目的，新闻搜索也就成了查看新闻的重要工具。目前的新闻搜索主要提供新闻搜索、聚合浏览、订阅等功能。相对于传统的搜索引擎技术，新闻搜索有许多独特之处：新闻网页的采集对时效性要求很高；实现新闻的浏览功能需要应用各种信息智能处理技术[1]。

实现新闻的搜索和订阅功能必须解决网页的抓取、新闻索引查询两个主要功能。新闻的网页抓取对实时性要求比较高。新闻检索对时效性要求比较高。新闻的聚合浏览，首先也要解决网页的抓取问题，同时还要对新闻网页进行分析处理。新闻网页的分析处理主要有新闻网页的内容抽取、新闻网页的自动分类、相关新闻的计算、相同新闻的识别等。本文将详细介绍一个完整新闻搜索的总体架构、新闻搜索各个模块的设计。

1　新闻搜索的系统架构

一个完整的新闻搜索需要包括新闻采集、新闻网页分析处理、新闻网页的检索三大部分[2]。新闻网页的采集，具体架构图见图1。

新闻网页抓取主要解决的问题是把互联网上各个新闻网站发布的新闻采集到本地的新闻数据库。新闻采集主要解决抓取的实效性问题、新闻网页的识别问题。

新闻网页的检索相对于传统的全文检索系统，需要解决的问题检索的实效性问题，也就是一篇新闻网页从互联网中采集到本地网页库中后要立刻能检索到，具体算法在本文的第三部分会提到。

新闻网页的分析处理部分主要解决的是：为了生成新闻的聚合浏览网页需要对新闻网页按类别进行分类，另外还需要对新闻进行相关新闻计算，相同新闻去重等。

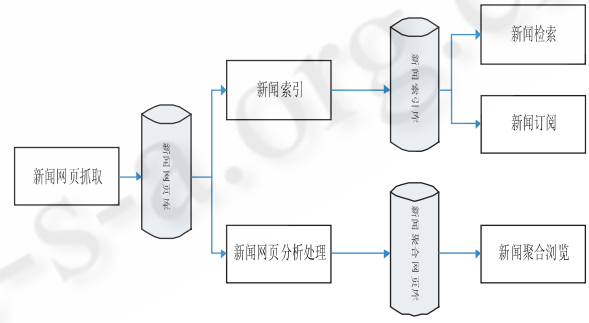

图1　新闻搜索系统架构图

2　新闻搜索各项模块设计

2.1　新闻网页采集

建立一个新闻搜索系统的第一步，就是从网络上抓取新闻页面，然后提出正文/时间/图片，并且按照某种方式保存起来，提供给其它模块建立索引，进行检索和浏览。新闻网页抓取模块的主要任务是把新闻站点上新出现的新闻网页及时抓取到本地。一个比较合理的新闻采集系统主要由两部分组成抓取模块和调控

① 基金项目：北京印刷学院青年基金科技研究项目（09190107054）

模块。

2.1.1 新闻抓取模块

抓取的过程并不只是通过 HTTP 协议在网络上抓取网页这样简单，我们需要处理各种复杂的网页重定向的情况。

通常抓取模块采用多线程的方式进行抓取，在初始化后，在指定端口监听。每接收一个连接请求，就创建一个线程，然后在该连接上，接受网页抓取请求，返回得到的页面。整个流程如图 2 所示：

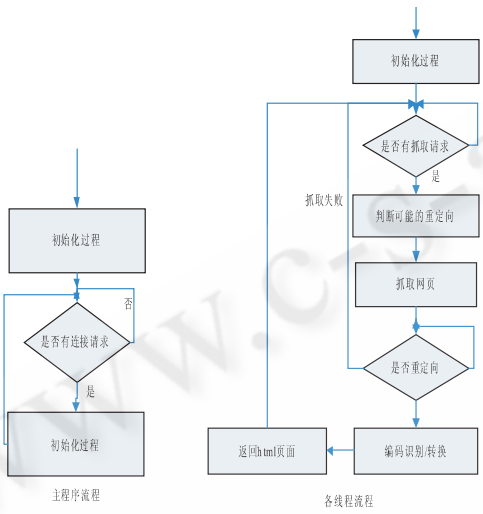

图 2　网页抓取流程图

2.1.2 新闻抓取控制模块

抓取的过程中我们需要控制抓取的速度，是否是新闻网页，哪些网页已经抓取过了，等等。控制模块就是完成抓取调度、新的链接抽取、新闻网页的识别等功能。

控制模块对各个站点进行调度，确定对各个站点抓取的时间，对站点上新出现的新闻网页进行抓取。一般新闻站点的调度周期可以通过该新闻网站在某一段时间内的更新频率进行动态调整。为了快速地收录新出现新闻网页，本文采用轮询监控的方式对各新闻网站进行监控。

此外新闻网站上除了新闻内容网页、新闻索引网页之外，还有各种各样的评论页面、论坛页面等其它对于新闻搜索无用的页面，控制模块需要判断识别出新闻内容网页和新闻索引网页，并将其它页面抛弃。

控制模块还有一个功能就是要从更新的网页中抽取新的新闻链接，放入到待抓取队列中。这个过程主要是对 html 文件进行解析的过程，抽取其中可能是新闻网页的 URL。控制模块具体流程见图 3。

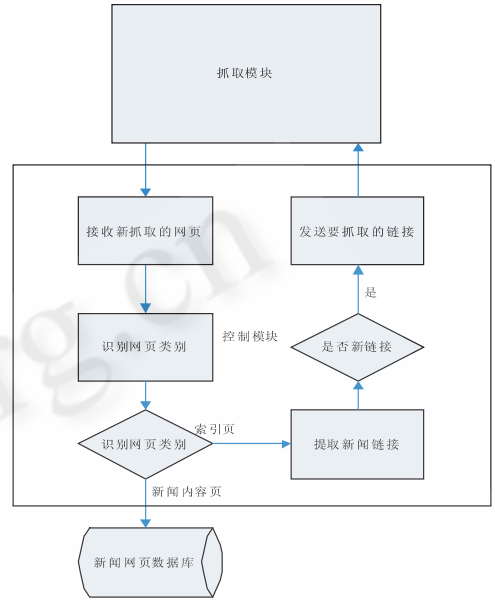

图 3　控制模块流程图

2.2 新闻网页分析处理模块

新闻网页被抓取回来之后，为了给用户提供搜索和浏览服务，需要对新闻网页进行处理。新闻网页处理主要需要解决新闻网页的内容抽取、新闻网页的分类、相关新闻网页的计算、相同新闻网页的去重等问题[3]。

2.2.1 新闻网页内容提取

新闻网页内容提取一般采用模板提取的方法。如果采集的新闻源比较少，主要是一些大的新闻站点的话可以人工配置模板。人工配置模板的优点是准确。但是缺点也很明显：人工配置模板人力消耗大，网页结构变化的话提取内容就会失败。目前应用比较多的是通过学习的方法自动获取新闻网页的模板。

2.2.2 新闻网页分类

新闻网页分类方法比较成熟，算法也比较多。常见的网页分类算法有 KNN、SVM、Naive Bayes 等，本文采用 KNN 算法[4]进行分类。

KNN 分类算法是一种传统的基于统计的模式识别方法。算法思想很简单：对于一篇待分类文档 $\vec{x}$，系统在训练集中找到 k 个最相近的邻居，使用这 k 个邻居的类别为该文档的候选类别。该文档与 k 个邻居之间的相似度为候选类别的权重，然后使用预先得到的最

优截尾阈值,就可以得到该文档的最终分类列表。kNN 算法的定义如公式(1)所示。

$$y(\vec{x},c_j)=\sum_{d_i\subset knn} sim(\vec{x},\vec{d_i})y(\vec{d_i},c_j)-b_j \quad (1)$$

其中:$\vec{x}$ 为一篇待分类网页向量表示;$\vec{d}$ 为训练集中的一篇实例网页向量表示;c_i 为一类别;$y(\vec{x}c_i)\in\{0,1\}$(当 $\vec{d}$ 属于 c_i 时取 1;当 $\vec{d}$ 不属于 c_i 时取 0);b_i 为预先计算得到的 c_i 的最优截尾阈值;

$sim(\vec{x}\,\vec{d}_i)$ 为待分类网页与网页实例之间的相似度,由公式(2)计算得到:

$$Cos(\vec{x},\vec{d})=\frac{\vec{x}\bullet\vec{d}}{|\vec{x}||\vec{d}|} \quad (2)$$

KNN 算法本身简单有效,它是一种 lazy – learning 算法,分类器不需要使用训练集进行训练,训练时间复杂度为 0。KNN 分类的计算复杂度和训练集中的文档数目成正比,也就是说,如果训练集中文档总数为 n,那么 KNN 的分类时间复杂度为 O(n)。

在实验中 k 的取值根据总的训练样本数量确定。具有相同类别的实例的相关度相加作为待分类网页的类别相关度。最后,我们把这些实例的类别为候选类别。

2.2.3 相关新闻计算

相关新闻的计算最常用的是聚类的方法,通过聚类把相关的新闻网页聚在一起。相关新闻的计算一般包括四个步骤:

(1)新闻网页表示:包括特征抽取和特征选择。特征选择是选择那些最具有区分性的特征,也就是最能把不同类别区分开来的特征,而不是大多数对象都具有的特征。

(2)相似度计算:主要根据网页表示的距离函数来定义,本文采用向量夹角余弦值两表示两个特征向量的距离。

(3)聚类:根据网页表示和相似度计算的结果,按照一定的规则将聚类网页分成不同的类。

(4)给出聚类的标识:在最后形成的每一类中抽取一定具有代表性的特征,作为该类的标识。

2.2.4 相同新闻去重

要判断两个新闻网页是否相同,首先要抽取网页的特征,然后通过比较两个网页特征是否一致,来判断网页是否相同。

新闻网页特征提取方法有很多种:最长句子、特征向量法等。

本文采用特征向量法是指对新闻正文进行分词提取主题特征词,通过对网页特征词向量的比较来判断两个网页是否相同。

2.3 新闻检索

全文检索方面的技术已经很成熟了,一般是采用倒排索引的方法。网上这种开源的检索系统很多。本文采用目前较流行的基于 Java 的 Lucene [5] 全文检索系统。

Lucene 是一个基于 Java 的全文索引工具包,它可以方便的嵌入到各种应用中实现针对应用的全文索引和检索功能。Lucene 的 API 接口设计的比较通用,输入输出结构都很像数据库的表、记录、字段,所以很多传统的应用的文件、数据库等都可以比较方便的映射到 Lucene 的存储结构和接口中。总体上看:可以先把 Lucene 当成一个支持全文索引的数据库系统。

3 结语

本文设计了一个互联网新闻搜索引擎。采用轮询监控的算法实现了新闻网页的快速采集。传统的文本分类、聚类等算法被应用到新闻网页的浏览中。本文能为互联网应用的开发人员提供一些有益的参考。

参考文献

1 刘悦,许洪波,程学旗. 互联网挖掘和搜索的研究进展. 见:曹右琦,孙茂松. 中文信息处理前沿进展——中国中文信息学会二十五周年学术会议. 北京:清华大学出版社,2006. 18 – 33.

2 李晓明, 闫宏飞,王继民. 搜索引擎(原理技术与系统). 北京:科学出版社,2005. 107 – 118.

3 徐宝文,张卫丰. 搜索引擎与信息获取技术. 北京:清华大学出版社,2003. 92 – 111.

4 龚笔宏,冯是聪. 针对中文网页评测 KNN 与 NB 分类算法. 见:李晓明,李星. 搜索引擎与 Web 挖掘进展. 北京:高等教育出版社,2003. 73 – 79.

5 郎小伟,王申康. 基于 Lucene 的全文检索系统研究与开发. 计算机工程,2006,32(4):94 – 96.