

商品评论聚焦爬虫算法设计与实现

方美玉^{1,2}, 郑小林², 陈德人², 华 艺¹, 施 艳¹

(1. 浙江外国语学院 科学技术学院, 杭州 310012; 2. 浙江大学 计算机科学与技术学院, 杭州 310027)

摘要:从商品评论抽取算法出发,使用开放应用编程接口调用和链接跟踪的方法,分别设计了基于通用爬虫原理和开放应用编程接口(OpenAPI)的商品评论聚焦爬虫算法。在此基础上实现了淘宝网和京东网商品评论收集程序。最后将两者与通用爬虫算法进行比较,证实了二者的程序设计具有针对性强、数据采集实时性好、易嵌入开发等优点,为实时评论数据采集技术的研究提供了新思路。

关键词:计算机软件;商品评论;开放应用编程接口;聚焦爬虫;爬虫算法

中图分类号:TP312 **文献标志码:**A **文章编号:**1671-5497(2012)Sup.1-0377-05

Design and implementation of focused crawler algorithms of product reviews

FANG Mei-yu^{1,2}, ZHENG Xiao-lin², CHEN De-ren², HUA Yi¹, SHI Yan¹

(1. Institute of Science and Technology, Zhejiang International Studies University, Hangzhou 310012, China;
2. College of Computer Science and Technology, Zhejiang University, Hangzhou 310027, China)

Abstract: From the product reviews extracting algorithm, focused crawler algorithms of product reviews were designed based on the general crawler principle and the OpenAPI respectively, using OpenAPI call and link tracking method. Then the comments' collection procedure of taobao.com and 360buy.com were realized. Finally both algorithms are compared with the general crawler algorithm, and some merits who are targeted program, timeliness data collection and easily embedded development were demonstrated. The results provide a good idea for the research on real-time reviews data collecting technology.

Key words: computer software; product reviews; open application programming interface; focused crawler; crawler algorithm

对评论语言进行分析是研究产品信任度和商家声誉的重要方法之一,其前提就是要收集到实时的评论数据。为了提高商品评论数据采集程序的针对性、实时性和嵌入性,有必要专门针对商品评论采集算法进行研究。

数据采集程序也称之为网络爬虫。主要有通用网络爬虫(百度,Google等)、增量式网络爬虫(目前流行的搜索引擎)、基于Agent的网络爬虫

和迁移的网络爬虫^[1]。这些研究大多针对搜索引擎领域,由于商品评论语言自身的特点和后续处理的需要^[2],以前的研究具有借鉴价值但不能完全适用。目前声誉研究的实验数据大部分采用实验随机产生并处理后的数据^[3]、仿真实验通用数据^[4]和通用网络爬虫算法采集的数据^[5]。其中,前两种不是实际环境数据,通用网络爬虫算法采集的数据具有真实性和实际应用性。但是通用爬

收稿日期:2012-04-17.

基金项目:国家自然科学基金项目(61003254);国家科技支撑计划项目(2008BAH24B03);浙江省自然科学基金项目(Y1080130, Y1101304).

作者简介:方美玉(1971-),女,副教授,博士.研究方向:电子商务. E-mail:hwdfmy@yahoo.com.cn

虫算法实时性不能满足要求^[6]。聚焦爬虫是一种主题网络爬虫,是近年来研究的热点^[5-6],采集数据围绕主题内容进行,具有很强的针对性。开放应用编程接口(Open application programming interface, OpenAPI)是网站服务商将自己的网站服务(包括评论数据收集服务)封装成的一系列 API,具有很强的针对性和易于嵌入第三程序的特点,在此基础上开发的评论收集算法,能够提高商品评论采集程序的针对性、实时性和嵌入性,相对于通用爬虫算法有明显的优势。

基于此,本文提出了基于通用爬虫厚理和 OpenAPI 的商品聚焦爬虫算法。在此基础上实现了淘宝网和京东网商品评论收集程序,为实时评价数据采集技术的研究提供了新思路。

1 相关概念

定义 1 聚焦爬虫是一个自动下载网页的程序,它根据既定的抓取目标,有选择的访问万维网上的网页与相关的链接获取所需要的信息^[4]。

聚焦爬虫需要解决的 3 个问题:①对抓取目标的描述或定义;②对网页或数据的分析和过滤;③对统一资源定位符(Uniform resource locator, URL)的搜索策略。

定义 2 基于通用爬虫算法的商品评论聚焦爬虫是以抓取网络商品评论为目的,在一定范围内对网页与评论相关的链接模式作分析,根据正则规则进行匹配并下载匹配段网页片段,对其他无关链接和页面内容进行过滤的 URL 搜索和评论收集策略。

定义 3 开放 API 也称开放平台,是服务型网站(包括电子商务平台)常见的一种应用,网站的服务商将自己的网站服务封装成一系列 API 开放出去,供第三方开发者使用,这种行为就叫做开放网站的 API,所以开放的 API 就被称作 OpenAPI 或开放 API。

定义 4 基于开放 API 的商品评论聚焦爬虫是以抓取网络商品评论为目的,给定 API 说明中指定的参数,调用开放 API 中的评论搜集函数,在一定范围内进行搜集并返回 XML 或 JSON 等开放格式的评价数据的评论收集策略。

2 基于通用爬虫原理的商品评论聚焦爬虫算法

2.1 通用爬虫算法与聚焦爬虫算法的区别

通用爬虫算法和聚焦爬虫算法的流程图如图 1 所示。通用爬虫从一个或若干初始网页的 URL 开始,获得初始网页上的 URL,在抓取网页的过程中,不断从当前页面上抽取新的 URL 放入队列,直到满足系统的一定停止条件,如图 1(a)所示。聚焦爬虫的工作流程较为复杂,需要根据一定的网页分析算法过滤与主题无关的链接,保留有用的链接并将其放入等待抓取的 URL 队列。然后,它将根据一定的搜索策略从队列中选择下一步要抓取的网页 URL,并重复上述过程,直到达到系统的某一条件时停止,如图 1(b)所示。另外,所有被爬虫抓取的网页将会被系统存贮,进行一定的分析、过滤,并

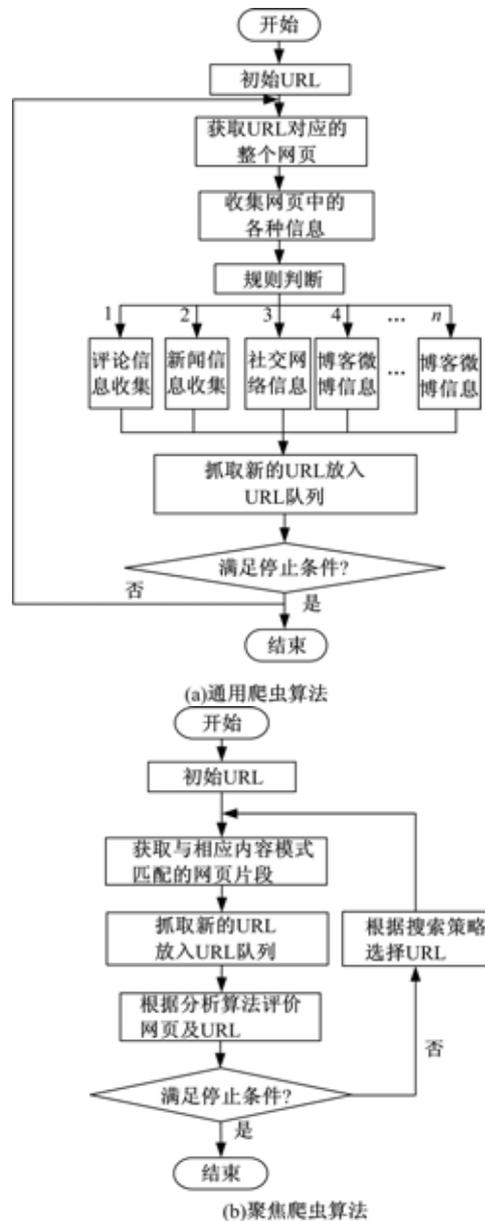


图 1 算法流程图

Fig. 1 Flow chart of algorithm

建立索引,以便之后的查询和检索。对于聚焦爬虫来说,这一过程所得到的分析结果还可能对以后的抓取过程给出反馈和指导。

2.2 算法设计

基于商品评论的聚焦爬虫算法流程类似于图 1(b),图中的初始 URL 匹配结果几乎是包含整个网页代码,经过进一步聚焦分析出评价网页 URL 对应的代码子模块后,再根据搜索策略从子模块识别出评价网页的 URL。其算法针对目标是评价网页的评价者、评价时间、评价优点、评价缺点等信息。以京东网为例,首先给定产品首页 URL,如:体育器械类页面,查看其源代码文件,

分析得出,其商品评论页面链接为“http://club.360buy.com/review”开头,后面跟若干位数字,以“-1-1.html”结尾的字符串构成,因此采用 d*正则表达式来匹配 http://club.360buy.com/review\d*-1-1.html,“*”代表任意个任意字符;“?”代表任意一个字符。然后从对应的页面查找对应的商品名页面代码对应的正则表达式模式:“商品名称(. *?)”。同理确定评价内容、优点、不足点等部分各自的正则表达式模式。从而获取与相应内容模式匹配的网页片段,过滤掉不需要的内容,获取与评价信息相关的内容写入数据库。表 1 为 2012 年 3 月实验收集的部分数据。

表 1 商品评论聚焦爬虫算法实验收集的数据样本

Table 1 Data samples collected from the experiment of focus crawler algorithm of product reviews

商品名	采集时间	用户名	优点	缺点
捷安特折叠自行车	2012-03-12 14:53:32	gyj9821	总体不错,满意	车身有点重
捷安特折叠自行车	2012-03-12 14:53:32	7357850947	便捷轻快,自行车不错,儿子喜欢	暂时还没发现缺点哦!
捷安特折叠自行车	2012-03-12 14:53:32	yljssorry	轻巧方便,适合短途上班或游玩。	刹车不好调,坐垫太小,坐着不舒服,车子有些刮花。
捷安特折叠自行车	2012-03-12 14:53:32	13810128986	正品行货,有发票,送货很快	暂时还没发现缺点哦!

3 基于 OpenAPI 的商品评论聚焦爬虫算法

3.1 淘宝开放平台(TOP)

TOP(Taobao open platform)是一个以综合性、商业性为特点的开放平台^[7]。TOP 正式发布的 API 涵盖了用户、商品、产品、类目、交易、评价、物流等不同专业领域的开放接口。使用这些预先定义好的函数,可以更加方便、快捷地调用这些接口来完成一些有关处理工作。

3.2 淘宝 API 中商品评论信息获取接口剖析

获取淘宝相应数据的 API 集中在用户 API、商品 API、类目 API、评价 API、店铺 API 等。评价 API 用来专门收集商品评论,其参数和返回值情况参见文献^[5]。使用 taobao.traderates.search 获取评价信息接口,是 TraderatesSearchRequist 类(请求类)与 TraderatesSearchResponse 类(响应类)组合完成的,请求类继承 TaobaoRequest 接口,而响应类继承了 TaobaoResponse 类。TraderatesSearchRequist 类中定义了接口中的应用参数,商品 ID 号、评价页号、每页显示评价总数、卖家昵称等,并对这些参数设计了 get/set 方法,最后将参数以 map 的形式存放。响应类中定义了 trade Rates 类型的参数,trade Rates 类型由 Java Bean 的形式写于 Trade Rate 类中,其中定义了评价列表的属性以

及 get/set 方法。

3.3 算法设计

商品评论聚焦爬虫算法的基本流程首先要符合淘宝 API 的调用流程^[7]。其次根据淘宝 API 文档调用规则,设计和实现接口进行调用,爬虫部分的主体代码都封装在 API 中。用淘宝 API (taobao.traderates.search) 来设计算法,其算法流程如图 2 所示。

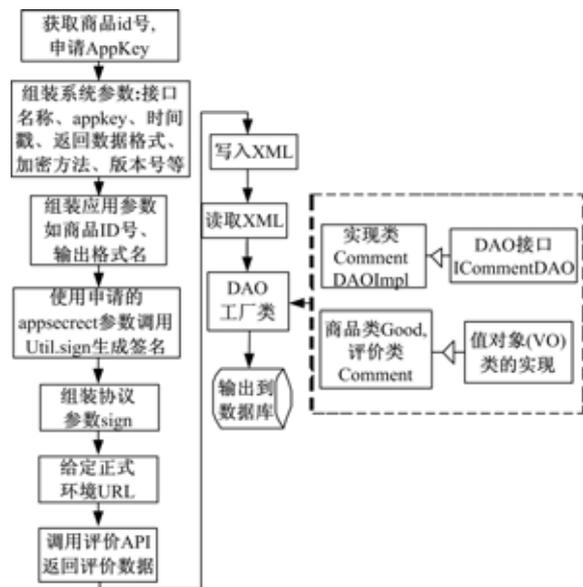


图 2 基于淘宝 API 商品评论聚焦爬虫算法的设计流程

Fig. 2 Flow chart of focus crawler by Taobao API

图 2 中数据访问对象(Data access object, DAO)商品接口 IGoodDAO 和评价信息接口 ICommentDAO 中定义了所有的用户操作,如添加记录、修改、删除、查找等,CommentDAOImpl 实现了 DAO 接口 ICommentDAO,并实现了 DAO 接口中所有抽象方法,可在 DAO 实现类中通过数据库连接类操作数据库。图 2 中值对象 VO(Value object)类是一个包含数据库表对应属

性和字段的类,包括商品类 Good 和评价类 Comment。类中提供了 setter 和 getter 方法来设置和获得该类中的属性值,属于 JavaBean 的一种。为了增强了代码的移植性,设计了 DAO 工厂类,通过该 DAO 工厂类的一个静态方法来获得实现类的实例。实现了一个淘宝商品评论收集程序。实验收集的 4 条样本数据如表 2 所示。

表 2 2012 年 5 月实验, ID=13979410881 的淘宝商品实验部分收集的样本

Table 2 One part of data samples collected from the experiment about Taobao product of ID=13979410881 on May 2012

淘宝账号名	评价时间	评价信息
蓝色微笑 0688	2012-05-2 09:16	帮朋友买的,很满意!赞!
胡洋洋 19924	2012-05-01 21:31	质量很不错,比图片上的还要好看哦,好喜欢,还会下次光临,希望亲再多晒点好看的衣服哈,我的同学们都很喜欢啊
马梅洁	2012-05-01 20:34	裙子很合身,很喜欢
丫丫 111965	2012-04-30 11:30	东西收到了,质量很不错,就是胸前有点小问题,不过不影响整体效果。担心长度太短,我身高 170,可是穿上刚刚好。谢谢卖家的小礼物。

4 对比分析

目前应用广泛的通用爬虫与本文所设计的两种算法比较,各有优劣,通用爬虫以网络神采为例(由郑州神采软件技术有限公司开发)。其适应性强,下载内容范围广,但使用前要设置很多参数,基于商品评论聚焦爬虫专门针对商品评论设计,虽然在算法流程上有两个新程序模块,但其针对

性内容少,因此其效率比通用爬虫快些,在程序编写复杂度方面,通用爬虫考虑的因素比较全面,因此其程序复杂度比聚焦商品评论算法要复杂很多,而基于 API 的聚焦爬虫算法只需要考虑接口调用准备工作,程序编写复杂度上大大降低。详细比较结果如表 3 所示。表中每项指标在括号里注明了各项结果表现的依据。

表 3 三种算法比较

Table 3 Comparison of three algorithms

指标	通用网络爬虫算法	基于通用爬虫原理的商品评论	基于 OpenAPI 的商品评论
	(网络神采)	聚焦爬虫算法	聚焦爬虫算法
通用性	最强(根据制定采集规则,可以采集任何通过浏览器看得到的东西)	较强(通过简单修改可以移植到其他网站)	不能通用(不同网站发布只属于自身的 API)
灵活性	最高(对于目的明确的评论收集来说不必要这种高灵活性)	一般(网页结构变化,程序和算法匹配模块需要做相应改变)	不够灵活(只适用商家 API 对应的自身网站)
采集规则	难度很高	不需要设定(程序员直接考虑在信息采集程序中)	不需要设定(图 3 只需要输入商品 ID 号,也可以自动搜索 ID 直接顺序扫描所有 ID 的评价信息)
针对性	一般(全面考虑各种信息采集)	高(只关注评论信息采集)	高(选择评价 API 直接有效)
速度(实时性)	一般(约 200 条信息/min)	快(约 240 条信息/min)	快(约 240 条信息/min)
复杂度	复杂(需要全面考虑多方面、多形式的信息收集情况,如图 1(a)所示)	一般(需要规则比较和形成 URL 队列)	低(不需要考虑信息收集的内部细节)
程序更新	慢(通用性强决定的)	快(网页更新速度决定)	慢(API 规则会维持一定稳定性)

5 结束语

设计了基于通用爬虫和 OpenAPI 的商品评论聚焦爬虫算法。在此基础上,实现了淘宝网和京东网的商品评价收集程序。基于通用爬虫的商品评论聚焦爬虫有一定通用性,高效性,实时性,

因此可以考虑作为实时商品评论采集算法,但程序更新速度频繁,给后期维护增加了难度;基于 OpenAPI 的商品评论聚焦爬虫针对性强、实时性高、复杂度低,程序更新速度慢,因此将是实时采集商品评论数据首选算法,为文献[3]的数据采集实时性奠定了基础。

参考文献:

- [1] Wang Bo, Wang Hou-feng. Bootstrapping both product properties and opinion words from chinese reviews with cross-training[C]//IEEE/WIC/ACM International Conference on Web Intelligence, Beijing,2007: 259-262.
- [2] 白鸽,左万利,赵乾坤,等.使用机器学习对汉语评论进行情感分类[J].吉林大学学报:理学版,2009,47(6):1260-1263.
Bai Ge, Zuo Wan-li, Zhao Qian-kun, et al. Sentiment classification for chinese reviews using machine learning[J]. Journal of Jilin University(Science Edition),2009,47(6):1260-1263.
- [3] Fang Mei-yu, Zheng Xiao-lin, Chen De-ren. A reputation evaluation approach based on fuzzy relation [J]. International Journal of Computational Intelligence Systems, 2011, 4(5), 759-767.
- [4] Miller R C, Bharat K. SPHINX: a framework for creating personal, site-specific Web crawlers [J]. Computer Networks and ISDN Systems,1998,30(1-7):119-130.
- [5] Arun Manicka Raja M, Winstler S G, Swamynathan S. Review analyzer: analyzing consumer product reviews from review collections[C]//2012 International Conference on Recent Advances in Computing and Software Systems(RACSS),2012: 287-292.
- [6] 张红云. 基于页面分析的主题网络爬虫的研究[D]. 武汉:武汉理工大学计算机学院,2010.
Zhang Hong-yun. The research of thematic reptile's based on analysis of network page[D]. College of Computer, Wuhan University of Technology,2010.
- [7] Taobao. com: API 调用原理 [EB/OL]. [2012-04-28]. <http://open.taobao.com/doc/detail.htm?id=55#s2>. 2012.