

面向个性化推荐的教育新闻爬取及展示系统

刘 灿, 任剑宇, 李 伟, 张强强

(重庆科技学院, 重庆 401331)

摘 要: 为了满足人们能快速、准确地获取个人最关心的教育类新闻的实际需求, 本文提出了一种面向个性化推荐的教育新闻爬取及展示系统。本系统采用主题爬虫技术, 在对爬取策略及爬取内容进行文本解析的基础上, 获取教育新闻数据。结合用户需求特征, 利用协同过滤的推荐策略, 生成满足不同用户需求的个性化页面, 并以词云和列表的形式展示给用户。该系统为人们获取最具时效性的教育新闻信息提供了一种可行的途径。

关键词: 个性化推荐; 教育新闻; 主题爬虫; 词云

中图分类号: TP319 **文献标识码:** A

The Personalized Recommendation-Oriented Education News Crawling and Displaying System

LIU Can, REN Jianyu, LI Wei, ZHANG Qiangqiang

(Chongqing University of Science and Technology, Chongqing 401331, China)

Abstract: In order to meet the actual needs of people who can quickly and accurately obtain the education news that individuals are most concerned about, we propose a customized crawling and displaying system for education news. Based on the text analysis mechanism of crawling strategy and crawling content, this system obtains the educational news data, combines the characteristics of user needs, and uses the recommendation strategy of collaborative filtering to generate a personalized page that meets the needs of different users, and presents the page to users in the form of word clouds and lists. The system provides a viable way for people to obtain the most up-to-date educational news information.

Keywords: personalized recommendation; education news; topical crawler; word cloud

1 引言(Introduction)

随着国家及个人对各类教育重视程度的不断提升, 关注教育热点信息的社会群体越来越多。然而, 由于教育新闻种类繁多, 既有中小学、大学等课堂教育相关的新闻, 又有培训、继续教育等非学历教育相关的信息, 还有国家及地方政府出台的各类教育制度等, 人们很难快速、准确地获取到最关心的教育新闻。如何从浩如烟海的互联网大数据环境中快速、准确地获取最权威的教育新闻热点信息成为一个亟待解决的问题。此外, 从个性化推荐的角度分析, 多数新闻站点对所有用户提供的都是同样的资源, 缺乏针对不同用户的个性化推荐功能。因此, 研发一种面向个性化推荐的教育新闻爬取及展示系统已成为一个重要研究课题, 这对于更好地服务民众具有重要的现实意义。

2 相关研究(Related research)

2.1 主题爬虫技术

网络爬虫(Web Crawler), 又称为网页蜘蛛, 是一种按照一定的规则, 自动地抓取互联网信息的程序或者脚本^[1]。网络爬虫的结构大致可以分为页面爬行模块、页面分析模块、链接过滤模块、页面数据库、URL队列、初始URL集合等几部分。为了提高工作效率, 网络爬虫通常会采取深度优先策略、广度优先策略等爬行策略。由于通用的网络爬虫获取的结果主题性不够突出, 结果不够深入和专业化, 在个性化推荐领域通常采用主题爬虫技术。主题爬虫(Topical Crawler), 也称为聚焦爬虫(Focused Crawler), 是一种下载特定主题网页的程序^[2]。和通用网络爬虫相比, 主题爬虫只需要爬行与主题相关的页面, 极大地节省了硬件和网络资源,

保存的页面也由于数量少而更新快,还可以很好地满足一些特定人群对特定领域信息的需求。

近年来,大量的学者对主题爬虫技术进行了相关的研究和应用,并取得了突出的成果。例如,黄仁等提出一种基于主题相关概念和网页页的主题爬虫,在垂直搜索引擎和舆情分析应用方面具有较好的搜索效果^[3]。吴岳廷等采用扩展主题特征库引入基于爬虫的网页过滤算法,提高了网页抓取精度,具有较高可用性^[4]。陈一峰等利用一种基于遗传算法的主题爬虫改进策略,在不影响查全率的情况下,与主题相关且重要的网页数量可提升5%以上^[5]。此外,李璐等提出一种基于SVM分类的主题爬虫技术,能够较好地区分主题相关网页和不相关网页,提高了主题相关网页的查准率和查全率^[6]。

2.2 个性化推荐技术

个性化推荐技术通过建立用户与信息之间的二元关系,利用已有的选择过程或相似性关系挖掘每个用户潜在感兴趣的对象,进而进行个性化推荐^[7]。传统的推荐方法主要包括:基于内容(content-based)的推荐系统,协同过滤(collaborative filtering)推荐系统和混合(hybrid)推荐系统^[8,9]。近年来,深度学习在图像处理、自然语言理解和语音识别等领域取得了突破性进展^[10],基于深度学习的推荐系统为推荐系统的研究带来了新的机遇^[11]。

个性化推荐系统已广泛应用于书籍推荐、互联网信息推荐、信息过滤等领域。例如,王永固等人研究了一个基于协同过滤技术的学习资源个性化推荐的理论模型,有效提高了学习资源个性化推荐的精度和效率^[12]。朱夏等人提出了一种云计算环境下基于协同过滤的个性化推荐机制RAC,构建基于分布式存储系统的两个阶段平分索引,在此基础上提出候选邻居的协同过滤算法,有良好的推荐精度和推荐效率^[13]。吕学强等提出了一种结合影评内容相似度和长短期兴趣模型来计算电影相似度的推荐方法^[14]。Wang等研究了融入图像内容信息来提升兴趣点推荐的问题^[15]。Huang等提出了一种神经概率模型用于基于情境的引用推荐^[16]。将深度学习技术融入推荐系统中,研究如何整合海量的多源异构数据,构建更加贴合用户偏好需求的用户模型,以提高推荐系统的性能和用户满意度,成为基于深度学习的推荐系统的主要任务^[17]。

3 系统设计(System design)

3.1 系统架构

本文设计了一个面向个性化推荐的教育新闻爬取及展示系统。系统通过爬取给定待爬取列表中的相关数据,及时地把爬取到的教育新闻信息进行解析,并把所有的信息进行分类存储;然后,根据用户喜好,利用个性化推荐方法,以词云或列表形式把用户最关心的教育新闻展示给用户,系统总体架构如图1所示。

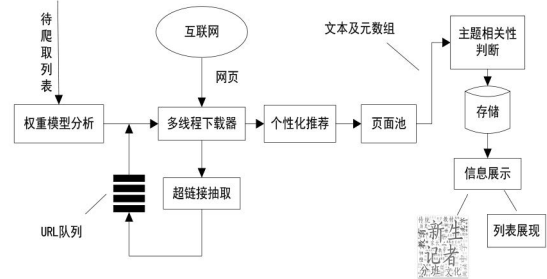


图1 系统总体架构图

Fig.1 The overall system framework

在信息爬取方面,系统采用Scrapy框架和轮循策略,爬虫系统在多线程方式下工作,采用生产者/消费者模型。使用权重值的爬取策略,对于影响力较大的教育新闻网站,通过给定较高的初始权重而优先爬取其数据,从而有利于获取最权威、最真实的信息。

在个性化推荐方面,使用一种基于协同过滤的推荐算法。首先进行新闻用户的跟踪,推测出用户感兴趣的主题;再通过分析用户的特征数据;然后根据用户的特征信息对其进行新闻推荐,从而实现教育新闻的个性化推荐。

在信息展示方面,根据信息来源站点的权威性和用户需求,对相关新闻以权重比进行排序,并以词云和列表的方式展示给用户。

3.2 信息爬取

本系统中,信息爬取部分采用主题爬虫技术,根据功能需求进行模块划分,采用调用组件的方法运行,便于根据功能需求进行系统修改,且维护简单,扩展性能好。主题爬虫模块结构如图2所示。

主题爬虫按照权重值大小对从给定的待爬取列表进行分组,使用生产者/消费者模型,以多线程的方式依次爬取每个权重组内的站点。生产者把抽取到的链接加入URL队列,消费者从URL队列中取出链接进行爬取,并把爬取的网页放入网页池。消费者从网页池中提取出网页进行解析及主题相关度判定,并把最终数据存储存储在数据库中,供展示系统调用。

在爬取过程中,首先判断该URL列表中的链接是否爬取过。如果已经爬取过,则转向下一个链接的处理。对于一个新链接,考虑该页面是否包含Robots协议。若存在Robots协议,则采用适当的反爬技术。由于网络或网站本身的原因,导致网页未下载成功,则放弃下载,并且转向下一个链接的处理。

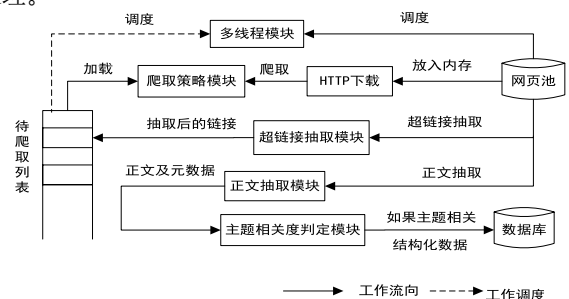


图2 主题爬虫模块结构图

Fig.2 Topical crawler module structure

3.3 个性化推荐

通过跟踪用户的使用轨迹,应用基于余弦相似度的协同过滤推荐方法分别对新闻用户和教育新闻的特征量建模,从而实现根据用户的需求推送用户感兴趣的信息。个性化推荐模块图如图3所示。

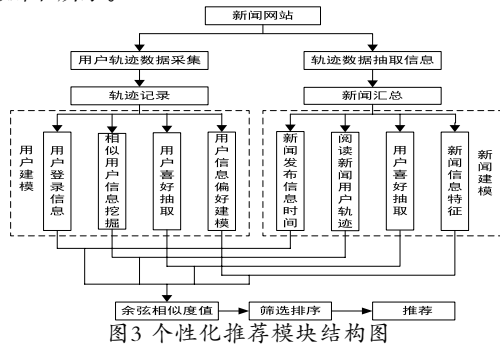


图3 个性化推荐模块结构图

Fig.3 Personalized recommendation module structure

首先,通过用户访问网站的轨迹分析用户的偏好特征,如用户感兴趣的话题、登录时间等信息;然后,使用特征向量描述每一篇新闻,利用余弦相似度度量方法,获取用户与相关新闻的相似性,并对结果进行筛选排序;最后,将相似度值最高的k个信息推荐给用户,k由系统或用户设定,默认值为8。

3.4 信息展示

信息展示部分主要以词云和列表的方式在Web页面中展示所查询到的教育新闻热点信息。系统还支持个性化查询,支持的查询方式包括按照日期(含时间段)、类别、新闻题目、作者、全文、媒体名称等方式。在列表方式展示中,为了突出重点关注新闻的重要性,将按照当天新闻的关注程度智能地进行排序,如多个站点均包含的新闻热点信息将排列在列表的前部。

4 系统实现(System implementation)

本系统采用主题爬虫、个性化推荐、词云等技术,利用Python语言,实现了面向个性化推荐的教育新闻爬取及展示功能。系统界面如图4所示。系统界面左侧为当天新闻热点词云展示页面,展示了当天的热点词语。通过查看此词云,可以获取今日教育新闻的热点信息。系统右侧为当天教育新闻的列表。



图4 PC浏览器操作界面

Fig.4 PC browser interface

热点新闻的具体展示页面如图5所示。



图5 教育新闻展示界面

Fig.5 Education news display interface

上述图4和图5实现了以词云和列表方式展示在Web页面上。本系统具有如下优点:根据用户的需要,实时推送给用户感兴趣的教育新闻;提出一种基于权重模型的爬行策略,提高爬虫的准确性和效率。

5 结论(Conclusion)

本系统以主题爬虫和个性化推荐为核心技术,建立一个面向个性化推荐的教育新闻爬取及展示系统。利用主题爬虫技术,从海量的教育类新闻页面中爬取各类教育新闻信息;利用协同过滤的个性化推荐策略,根据收集的用户爱好信息,实现教育新闻的个性化推荐;通过词云和列表形式,以友好的界面把用户最感兴趣的信息展示给用户。该系统对于公众获取最具时效性的教育新闻信息具有重要的实用价值。

参考文献(References)

- [1] 周德懋,李舟军.高性能网络爬虫:研究综述[J].计算机科学,2009,36(08):26-29;53.
- [2] 于娟,刘强.主题网络爬虫研究综述[J].计算机工程与科学,2015,37(02):231-237.
- [3] 黄仁,王良伟.基于主题相关概念和网页分块的主题爬虫研究[J].计算机应用研究,2013,30(08):2377-2380;2409.
- [4] 吴岳廷,李石君.基于扩展主题特征库的领域主题爬虫[J].计算机工程与设计,2015,36(05):1342-1347.
- [5] 陈一峰,赵恒凯,余小清,等.基于遗传算法的主题爬虫策略改进[J].计算机仿真,2010,27(10):87-90;123.
- [6] 李璐,张国印,李正文.基于SVM的主题爬虫技术研究[J].计算机科学,2015,42(02):118-122.
- [7] 刘建国,周涛,汪秉宏.个性化推荐系统的研究进展[J].自然科学进展,2009,19(01):1-15.
- [8] Balabanovic M,Shoham Y.Fab:content-based,collaborative recommendation[J].Communications of the ACM,1997,40(3):66-72.
- [9] 王晓军.推荐系统中分布式混合协同过滤方法[J].北京邮电大

(下转第34页)