网络首发时间：2018-08-14 16:06:01
网络首发地址：http://kns.cnki.net/kcms/detail/61.1224.TN.20180810.1351.039.html

DOI:10.16652/j.issn.1004-373x.2018.16.039

基于大数据的Web个性化推荐系统设计

张婷婷
（山东管理学院 信息工程学院，山东 济南 250357）

摘 要： 为了解决基于数据挖掘技术的Web个性化推荐系统对Web的推荐结果准确率低，反应时间长的问题，设计基于大数据的Web个性化推荐系统。塑造系统组成框架图，设计系统的总体功能包括源数据采集、数据预处理、用户兴趣分析与实现、个性化推荐以及推荐引擎。源数据采集利用Sqoop工具将数据库中的数据转移到HDFS中以便H-ICRS算法进行数据提取，并获得推荐的历史数据，实现作为系统上层数据支持的功能。针对分析用户长远和当前的Web兴趣度，分别采用语义分析模型和分片聚类的方法，分析用户Web使用兴趣。塑造单个推荐引擎的推荐引擎架构，得到最终的Web个性化推荐列表。实验结果表明，所设计系统的Web个性化推荐结果准确率高，系统的抗压能力强。

关键词： 大数据；Hadoop；Web个性化推荐；系统设计；Sqoop；H-ICRS算法
中图分类号： TN919-34 **文献标识码：** A **文章编号：** 1004-373X(2018)16-0155-04

Design of Web personalized recommendation system based on big data

ZHANG Tingting
(School of Information Engineering, Shandong Management University, Jinan 250357, China)

Abstract: A Web personalized recommendation system based on big data is designed to solve the problems existing in the Web personalized recommendation system based on data mining technology for its low accuracy rate of Web recommendation results and long reaction time. The composition framework of the system is built. The system′s overall functions including source data acquisition, data preprocessing, user interest analysis and implementation, personalized recommendation, and recommendation engine are designed. During source data acquisition, the Sqoop tool is used to transfer data in the database to the HDFS, so as to extract data by using the H-ICRS algorithm, obtain the recommended historical data, and realize the upper layer data support function of the system. By analyzing users′ long-term and current Web interest degree, the semantic analysis model and fragmentation clustering method are adopted respectively to analyze users′ Web interest. The recommendation engine architecture is constructed for a single recommendation engine to obtain the final Web personalized recommendation list. The experimental results show that the designed system has high accuracy rate of Web personalized recommendation results and strong anti-pressure capability.

Keywords: big data; Hadoop; Web personalized recommendation; system design; Sqoop; H-ICRS algorithm

随着经济技术的迅猛发展，产生了大量的数据信息。人们每天都会获取大量的信息，但是信息质量都各有不同。如何确保用户在获得自己感兴趣的Web同时，将外界干扰的Web影响降至最低[1]，是当前推荐系统亟需解决的问题。随着数据的增长，传统基于数据挖掘技术的Web推荐系统向用户推荐的Web准确度较低，已经无法满足用户的个性化需求。针对该问题，本文设计基于大数据的Web个性化推荐系统，提高系统的个性化推荐效果。

1 基于大数据的Web个性化推荐系统

本文基于大数据的Web个性化推荐系统，结合搜索引擎下的推荐系统，Hadoop大数据框架的Web个性化推荐系统。其中Web搜索引擎以系统服务者的身份参与到本文基于Hadoop框架大数据的Web个性化推荐系统中，负责系统进行信息检索和部分数据的供应。

收稿日期：2017-09-22 **修回日期：**2017-11-17
基金项目：国家自然科学基金青年项目(71301086)；山东省社科规划专项基金(17CQXJ11)；山东省高等学校科技计划资助项目(J16LN70)
Project Supported by National Natural Science Foundation of China for Youth (71301086), Special Fund of Social Science Plan of Shandong Province (17CQXJ11); Science and Technology Plan for Colleges and Universities in Shandong Province (J16LN70)

Hadoop大数据框架负责向系统大数据的处理。系统组成框图如图1所示。

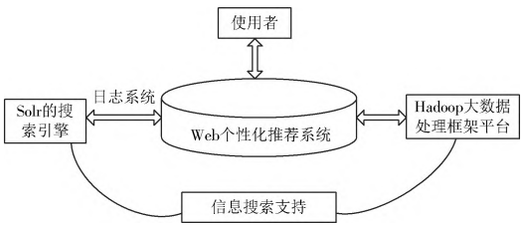

图1 基于大数据的Web个性化推荐系统的组成框图

Fig. 1 Block diagram for composition of Web personalized recommendation system based on big data

基于Hadoop框架的大数据的Web个性化推荐系统的组成框架图可以看出，为保证系统的稳定性和扩展能力，系统应用不同的开源项目[2]，其中包括Solr的搜索引擎和Hadoop大数据处理框架平台。Solr的主要功能是向系统使用者提供信息搜索支持，并将搜索结果经由日志系统传导回本文基于大数据的Web个性化推荐系统。

1.1 系统总体功能设计

基于本文系统的组成框架图，对系统的总体功能进行设计。系统总体功能设计通过分析使用者的Web行为数据，以为使用者推送个性化Web结果为目的进行设计.本文系统的Web个性化推荐分别从源数据采集[3]、数据预处理、用户兴趣分析与实现、个性化推荐以及推荐引擎五个部分进行分析，不同部分实现不同的结构功能，整体协同实现基于大数据的Web个性化推荐。

1.2 源数据采集

源数据采集主要向本文系统进行必要的上层数据支持，系统采集的实时数据经由Web服务器保存在关系型数据库中，H-ICSR算法运行于Hadoop 框架，该算法从HDFS读取数据[4]，需要将数据库中的数据传输到HDFS。同时用户会对系统推荐结果的反馈保存在数据库中。H-ICSR算法利用的源数据涵盖各种Web属性信息、社会关系数据以及过去推荐结果等。通常上述数据被存储于MySQL数据库中，不同的源数据分别与不同的表相对应，上述源数据对应表为tb-userInfo表、tb-action表和tb-follow表。源数据的采集通常采用Sqoop工具将数据库中的数据转移到HDFS中，以便H-ICRS算法进行数据提取，同时获取系统向用户推荐的历史情况。

1.3 数据预处理

本文基于大数据的Web个性化推荐系统的数据预处理是基于Hadoop平台实施的，其中实现源数据与数据预处理的为HDFS，倘若经由HDFS中检索出的源数据未经过格式化[5]，需要先将其进行格式化操作，格式化结果如{String1,String2,…}或{String1String2…}的形式。若获取的数据已经格式化，从中筛选有效信息进行计算或构建模型，并保存在HDFS中。在数据预处理中H-ICSR算法运算Web个性化推荐度与项目聚类[6-7]，由RecommendExtentJob和ClusteringJob两者分别实现。

1.4 用户兴趣分析实现

基于大数据的Web个性化推荐系统针对用户对Web兴趣的分析角度，从两方面考虑：一方面对于用户的一些长远的Web兴趣，本文采用语义分析的方法对用户的Web使用兴趣进行分析；另一方面对用户的当前感兴趣的Web，本文采用分片聚类的方法对该类用户的Web使用兴趣进行分析[8]。对用户兴趣分析的实现代码如下：

```
Open Catalog<Datatypical>LSA (String OwnerData, Route, StringWebDataRoute){
Obtain Owner Number=got Owner Number(OwnerDataRoute);
                                        //得到用户数量
Obtain Web Number=got Web Number (Web Data Route);
                                        //得到Web数量
Obtain Jargon Number=got Jargon Number(Owner Data Route, Web Data Route);
                                        //固定词库/*得到用户文本向量*/
Obtain vector Number=Cut (D); //对待降维的维度进行计算
//采用Jargon进行相似度计算
Owner Similar
Simular=new
Owner Date Esrimate.Owner Jagon Similar (D,vector number)
                                        //利用K-means聚类方法实施聚类
Reentry Consequence
```

对用户的长远兴趣和即时兴趣分析分别采用LSA()和shardCluster()函数。LSA()函数采用分析Web内容与用户之间关系，将两者关系相接近的进行聚类。shardCluster()函数将用户按时间或地点等可以反应用户当前兴趣的Web进行分片[9]，并对相似的用户行为进行分片聚类，以此系统可以针对用户的兴趣内容向用户推荐个性化的Web。

1.5 推荐引擎实现

本文系统的推荐引擎架构主要由三部分组成，分别为推荐引擎的基本组成要素，如下：

1）特征向量，其来源为经数据预处理后得到的数据以及用户的数据特征，或是直接存在的特征向量，特征向量主要是作为向用户进行Web个性化推荐的依据；

2）主要是一些计算出的离线表[10]，依靠特征向量和特征-源数据等得到初始的Web个性化推荐列表；

3）对得到的初始个性化推荐列表进行处理,得到最终的基于大数据的Web个性化推荐列表。

2　实验分析

实验为验证本文系统是否可以高效地向用户进行Web个性化推荐,将本文基于大数据的Web个性化推荐系统进行实际应用,与传统基于数据挖掘技术的Web个性化推荐系统的推荐结果做为对比。为了确保本文系统Web个性化推荐结果具有较高的普遍性,实验分别从某高校财务管理专业、电子商务专业和软件技术专业各随机选取4个学生进行Web的个性化推荐测试,其中各专业男、女学生人数均占1/2。实验分别从系统推荐结果准确度、专业性、页面布局效果以及满意度四个方面进行评判,如表1所示。

表1　传统基于数据挖掘技术的Web个性化推荐系统评判结果

Table 1　Evaluation results of traditional Web personalized recommendation system based on data mining technology

学生专业	学生编号	准确度/%	专业性/分	页面布局效果/分	满意度/%
财务管理	1	62.3	48	68	62.4
	2	56.2	56	67	62.2
	3	65.2	57	75	56.8
	4	57.6	52	74	55.6
电子商务	5	58.5	64	62	57.6
	6	55.2	57	64	64.2
	7	64.5	49	65	65.5
	8	65.5	64	74	64.2
软件技术	9	65.5	64	75	67.5
	10	64.5	54	77	57.6
	11	67.8	57	64	64.5
	12	64.2	48	62	57.6
平均成绩		62.2	56	69	61.2

分析表1数据可知,传统基于数据挖掘技术的Web个性化推荐系统的Web推荐结果准确度不超过70%,说明该系统的个性化推荐效果较差,且推荐Web的专业性能不高,无法向学生用户提供有用的页面效果,评分也较低,整体的用户满意水平较低。综合分析表1和表2中数据可得,采用本文基于大数据的Web个性化推荐系统对不同用户进行实际推荐过程中,在准确度、专业性、页面布局效果和满意度方面均优于传统基于数据挖掘技术的Web个性化推荐系统。在满意度方面,所提方法的满意度为84%,远高于传统方法的61.2%,说明本文方法实际应用性较强,用户满意度高。

实验为分析本文基于大数据的Web个性化推荐系统是否可以快速、稳定地向用户进行个性化的Web推荐。实验以基于数据挖掘技术的Web个性化推荐系统和基于Spark的Web个性化推荐系统为对比,分析三个系统的系统响应时间和最大抗压能力。图2和图3分别为三个系统在系统使用人数不同时的反应时间以及系统最大的抗压结果。

表2　基于大数据的Web个性化推荐系统评判结果

Table 2　Evaluation results of Web personalized recommendation system based on big data

学生专业	学生编号	准确度/%	专业性/分	页面布局效果/分	满意度/%
财务管理	1	84.6	77	75	80.7
	2	79.4	80	78	79.9
	3	86.7	79	82	81.9
	4	76.3	75	87	82.6
电子商务	5	77.9	81	86	85.1
	6	83.2	86	76	80.8
	7	88.4	72	85	87.3
	8	89.6	75	84	88.6
软件技术	9	75.2	76	90	90.4
	10	82.3	73	89	83.9
	11	87.3	80	77	89.8
	12	86.3	70	82	77.6
平均成绩		83.1	77	83	84.0

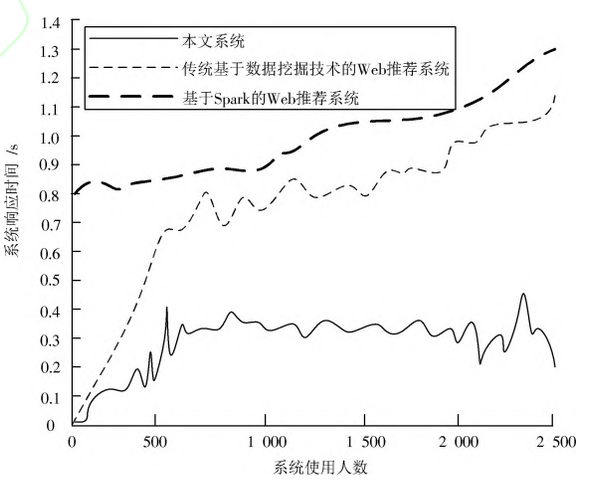

图2　三个系统的系统反应时间测试结果

Fig. 2　Testing results for system response time of three systems

分析图2中数据可知,本文系统在不同的系统使用人数下,时间从0逐渐增加到0.4 s后系统的反应时间趋于稳定,不再变化;基于数据挖掘的Web个性化推荐系统同样从0时刻开始变化,但该系统随着使用人数的增加系统反应耗时一直逐渐增大;分析基于Spark的Web个性化推荐系统从系统运行初始的耗时基数较大且随着系统使用人数越来越多,系统反应时间逐渐加快。综

合三个系统的反应耗时可以得出，本文系统向用户进行Web的个性化推荐时，推荐结果的效率较高。

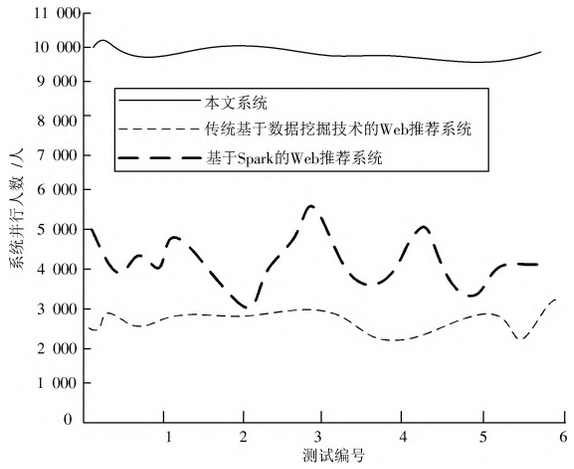

图3 三个系统的系统抗压力测试结果

Fig. 3 Testing results for system anti-pressure capability of three systems

实验分别对三个Web个性化推荐系统的抗压能力进行6次测试。分析图3可得，传统基于数据挖掘技术的Web个性化推荐系统其最大承受人数约为2 500人。说明系统仅能保证2 500正常同时使用，超出该人数系统可能发生崩溃，其抗压能力较弱。基于Spark的Web个性化推荐系统最大承受人数约为2 500且承载人数波动较大，说明该系统稳定性和抗压能力较差；而本文提出的基于大数据的Web个性化推荐系统，最大承受人数约为10 000人，并且承载人数波动幅度较小，说明该系统的稳定性和抗压能力较强。

3 结 论

本文设计的基于大数据的Web个性化推荐系统能提高对Web个性化推荐结果的准确度。系统整体运行效率高，抗压能力强。

参 考 文 献

[1] 尤海浪，钱锋，黄祥为，等.基于大数据挖掘构建游戏平台个性化推荐系统的研究与实践[J].电信科学，2014，30(10)：27-32.
YOU Hailang, QIAN Feng, HUANG Xiangwei, et al. Research and practice of building a personalized recommendation system for mobile game platform based on big data mining [J]. Telecommunications science, 2014, 30(10): 27-32.

[2] 孟祥武，纪威宇，张玉洁.大数据环境下的推荐系统[J].北京邮电大学学报，2015，38(2)：1-15.
MENG Xiangwu, JI Weiyu, ZHANG Yujie. A survey of recommendation systems in big data [J]. Journal of Beijing University of Posts and Telecommunications, 2015, 38(2): 1-15.

[3] 应毅，刘亚军，陈诚.基于云计算技术的个性化推荐系统[J].计算机工程与应用，2015，51(13)：111-117.
YING Yi, LIU Yajun, CHEN Cheng. Personalization recommender system based on cloud-computing technology [J]. Computer engineering and applications, 2015, 51(13): 111-117.

[4] 李文海，许舒人.基于Hadoop的电子商务推荐系统的设计与实现[J].计算机工程与设计，2014，35(1)：130-136.
LI Wenhai, XU Shuren. Design and implementation of recommendation system for E-commerce on Hadoop [J]. Computer engineering and design, 2014, 35(1): 130-136.

[5] 刘其成，冯利光.一种基于MapReduce的微博信息推荐并行算法[J].小型微型计算机系统，2017，38(7)：1518-1522.
LIU Qicheng, FENG Liguang. Parallel microblog information recommendation algorithm based on MapReduce [J]. Journal of Chinese computer systems, 2017, 38(7): 1518-1522.

[6] 陈万志，林澍，王丽，等.基于用户移动轨迹的个性化健康建议推荐方法[J].智能系统学报，2016，11(2)：264-271.
CHEN Wanzhi, LIN Shu, WANG Li, et al. Personalized recommendation algorithm of health advice based on the user′s mobile trajectory [J]. CAAI transactions on intelligent systems, 2016, 11(2): 264-271.

[7] 张时俊，王永恒.基于矩阵分解的个性化推荐系统研究[J].中文信息学报，2017，31(3)：134-139.
ZHANG Shijun, WANG Yongheng. Personalized recommender system based on matrix factorization [J]. Journal of Chinese information processing, 2017, 31(3): 134-139.

[8] 武慧娟，秦雯，孙鸿飞，等.基于标签的个性化信息推荐系统动力学模型与仿真[J].现代情报，2016，36(3)：12-16.
WU Huijuan, QIN Wen, SUN Hongfei, et al. System dynamics model and simulation based on the tag of personalized information recommendation [J]. Modern information, 2016, 36(3): 12-16.

[9] 黄亚坤，王杨，苏洋，等.基于两层社区混合计算的个性化推荐方法[J].计算机科学，2016，43(z1)：440-447.
HUANG Yakun, WANG Yang, SU Yang, et al. Personalized recommendation method based on hybrid computing in two layers of community [J]. Computer science, 2016, 43(S1): 440-447.

[10] 余刚，王知衍，邵璐，等.基于奇异值分解的个性化评论推荐[J].电子科技大学学报，2015，44(4)：605-610.
YU Gang, WANG Zhiyan, SHAO Lu, et al. Singular value decomposition-based personalized review recommendation [J]. Journal of University of Electronic Science and Technology of China, 2015, 44(4): 605-610.

作者简介：张婷婷(1981—)，女，山东泰安人，硕士，讲师。研究方向为网络数据挖掘、推荐系统。