Computer Science and Application 计算机科学与应用, 2013, 3, 257-261
http://dx.doi.org/10.12677/csa.2013.35044 Published Online August 2013 (http://www.hanspub.org/journal/csa.html)

The Research on Book Intelligent Recommendation System Based on Cloud Computing

Chuzhen Li, Xinling Wu

School of Computer Science, Guangdong Polytechnic Normal University, Guangzhou
Email: gdlcz_1006@163.com

Received: May 4th, 2013; revised: May 21st, 2013; accepted: Jun. 2nd, 2013

Abstract: The book intelligent recommendation system, the concept and related technologies of cloud computing are briefly introduced. In view of the fact that the traditional book intelligent recommendation system cannot store massive data and recommend information in time up to now, this paper proposes to construct a book intelligent recommendation system based on cloud computing technology and elaborates its architecture and recommendation process in detail.

Keywords: Book Intelligent Recommendation System; Cloud Computing; Architecture; Association Rules

基于云计算的图书智能推荐系统研究

李楚贞，吴新玲

广东技术师范学院，计算机科学学院，广州
Email: gdlcz_1006@163.com

收稿日期：2013 年 5 月 4 日；修回日期：2013 年 5 月 21 日；录用日期：2013 年 6 月 2 日

摘　要：简要介绍了图书智能推荐系统与云计算相关概念和技术，针对目前传统图书智能推荐系统因海量数据带来的存储及推荐速度问题提出了一个基于云计算环境下的图书智能推荐系统，并对其体系架构和推荐过程进行了详细的阐述。

关键词：图书智能推荐系统；云计算；架构；关联规则

1. 引言

随着 Internet 技术的迅猛发展，人们正在享受着信息共享带来的方便与快捷的同时，面临着“数据丰富，但信息贫乏”的问题，个性化服务逐渐成为一种趋势。无论是在电子商务、保险行业还是在电信业务中，都提倡为用户提供个性化服务，而数字图书馆也不例外。作为学校最重要的信息服务机构，数字图书馆如何在海量的图书信息中发现其背后隐藏的重要信息并快速准确主动地为读者推荐其感兴趣的书籍就显得尤为重要，图书智能推荐系统就是在这样的背景下应运而生。目前，已经有很多国内学者投入到图书智能推荐系统的研究当中。其中，丁雪提出构建基于数据挖掘技术的图书智能推荐系统，黄晓斌提出基于协同过滤的数字图书馆推荐系统，刘飞飞提出基于多目标优化双聚类的数字图书馆协同过滤推荐系统。然而，在这些研究中，图书智能推荐系统都是以互联网为基础提供服务的，难以解决图书馆因海量数据带来的存储及推荐实时性问题。鉴于云计算平台能提供海量的信息资源存储和强大的计算能力，本文提出将

图书智能推荐系统构建在云计算平台上。通过结合目前 IT 领域研究比较热门的云计算和数据挖掘技术来解决数字图书馆的个性化服务效率问题，从而提升用户的满意度。

2. 图书智能推荐系统

个性化服务是数字图书馆提高图书服务质量和信息资源使用效率的一项有利手段，已经成为图书馆重要的服务方式，而图书智能推荐系统则是实现个性化服务的核心系统。图书智能推荐系统是一个高级智能系统，它利用人工智能重要研究领域——数据挖掘技术，对用户的专业、偏好、浏览历史记录及特定的需求等信息进行分析挖掘，从而为用户主动、及时、准确地提供其感兴趣的信息，并根据用户的反馈进一步改进推荐结果。从技术上来看，它实际上是一种对特定类型的数据集进行知识发现和利用的应用系统[1]。图书智能推荐系统大致可以分为三个模块：输入模块、推荐模块和输出模块。用户通过在输入模块输入想要获得的信息，该信息通过 web 服务器传递给推荐模块进行处理，最后再将处理结果传给输出模块。不同的推荐系统其主要区别在于它的核心模块——推荐模块中的推荐算法，它直接影响推荐效率与质量。目前使用的主流推荐算法包括如下几种：1) 基于数据挖掘技术的推荐算法；2) 基于内容的推荐算法；3) 协同过滤推荐算法；4) 混合推荐算法[2]。

3. 云计算概念及相关技术

3.1. 云计算概念

云计算作为一种新型的服务计算模型，是并行计算、分布式计算和网格计算的融合与发展，是 IT 产业继 PC、互联网之后的第三次革新浪潮，是业界、学术界的热点名词与技术之一。本质上，云计算是指用户终端使用简易的设备如 PC、手机、PDA 等通过互联网轻松地获取存储、计算、数据库、服务器等计算资源。其中，这些计算资源是由成千上万服务器组成的“云”端提供的，它们在用户看来是透明的且可以无限扩展。云计算结合了虚拟化、分布存储、海量数据管理等技术，利用互联网将分散的、动态的、异构的信息资源和计算能力有效整合起来，供用户方便地访问与使用，实现按需伸缩、按需使用、按需付费，达到高效率低成本的目的。

3.2. 云计算相关技术

3.2.1. 开源 Hadoop 云平台

目前较为成熟的云计算平台有 Windows Azure、Google Apps Engine、Blue Cloud、Hadoop 等。本文提出的图书智能推荐系统是基于 Apache 开源组织的 Hadoop 框架。它部署在由大规模计算机组成的集群上，具有良好的可扩展性及大容量数据存储能力，是传统的 IDC 服务器所不能比拟的。其核心部分 MapReduce 是对 Google 分布式计算模型的实现，能使计算任务分布到超大集群上的各个计算机并发处理，而 HDFS 是一个可扩展的分布式文件系统，具有较强的容错性和高吞吐率等优点。

3.2.2. MapReduce

MapReduce 是 Hadoop 采用的分布式计算模型，通过将一个大型的任务分成很多更细粒度的子任务，这些子任务能够分布式且并行地在多个节点上进行调度和计算，从而实现处理大规模数据集(大于 1 TB)的能力。MapReduce 的执行由两种不同类型的节点负责：Master 和 Worker[3]。Master 负责分配并监控任务的执行，Worker 负责执行任务。而这一过程可以简单看成是一个分类与汇总的过程，首先 Map 把输入数据分成许多小的模块并分配给大量计算机独立处理，在这过程中会生成大量的 key/value 对，接着 Reduce 任务把分开处理的结果汇总并输出到 master 节点。

3.2.3. HDFS

HDFS 是一个采用主/从结构的分布式文件系统，用于存储海量的非结构化数据。它由一个管理节点(NameNode)和多个数据节点(DataNode)组成[4]，其中管理节点负责管理文件系统所有的元数据，而数据节点负责数据的存储和管理。通常在 HDFS 中，一个文件被分割成一个或多个数据块，存储在一组不同数据节点上[5]。

4. 基于云计算的图书智能推荐系统

基于云计算的图书智能推荐系统在数据的处理方式上与传统的图书智能推荐系统有所不同，其区别主要有：1) 在存储数据时，不再是把所有数据或文件

存储在数据库或数据仓库中，而是把它们分割成大小固定的片段并存储在分布式文件系统中的不同的闲置数据节点上，用户可以通过数据节点直接读取数据，从而解决了传统图书智能推荐系统因图书数据量大而带来的存储问题。当所需的存储容量不够时，只需再添加多个计算机即可。2) 采用流水线复制技术，复制每个数据块的副本并将其保存在不同的存储节点上，避免了因某个节点出错而导致数据丢失的现象，真正确保数字图书馆服务的安全性。3) 在执行某一推荐任务时，由主控节点(master)负责分配子任务给多个计算节点并监控它们的执行，这样可以大大提高图书推荐的速度。

图书智能推荐系统的核心是其推荐算法。尽管基于内容的推荐算法和协同过滤推荐算法有许多的优点，但是它们都需要用户对项目进行评价，而大多数读者在访问数字图书馆系统时很少对图书质量进行评价。文献[3]提出的基于 Apriori 算法的并行关联规则挖掘算法，避免了基于内容的推荐算法和协同过滤推荐算法存在的数据稀疏和冷启动问题，同时能描述某些事件同时出现的规律和模式，因此，更适合于云计算环境下的图书智能推荐系统。基于上述研究，本文在借鉴参考文献[3]的基础上提出了一个基于云计算的图书智能推荐系统的基本架构，如图 1 所示。利用 Hadoop 云计算平台开发实现图书智能推荐系统需要①安装 Sun Java6 JDK ②配置 NameNode 和 DataNode 节点 ③安装并配置 SSH 为公钥认证 ④修改 Hadoop 相关配置文件等等，详细步骤略去。当基于 Hadoop 的云计算平台搭建完后就可以将图书智能推荐系统迁移到该平台上。任何一个授权用户都可以通过客户端访问该系统，享受其强大的计算及存储能力。

基于云计算的图书智能推荐系统主要由两大模块组成，分别是关联规则挖掘模块和推荐模块。下面详细讨论这两个模块的主要功能及处理过程。

4.1. 关联规则挖掘模块

这一模块属于离线操作，它由获取读者信息、数据预处理、知识挖掘三部分组成。由于 MapReduce

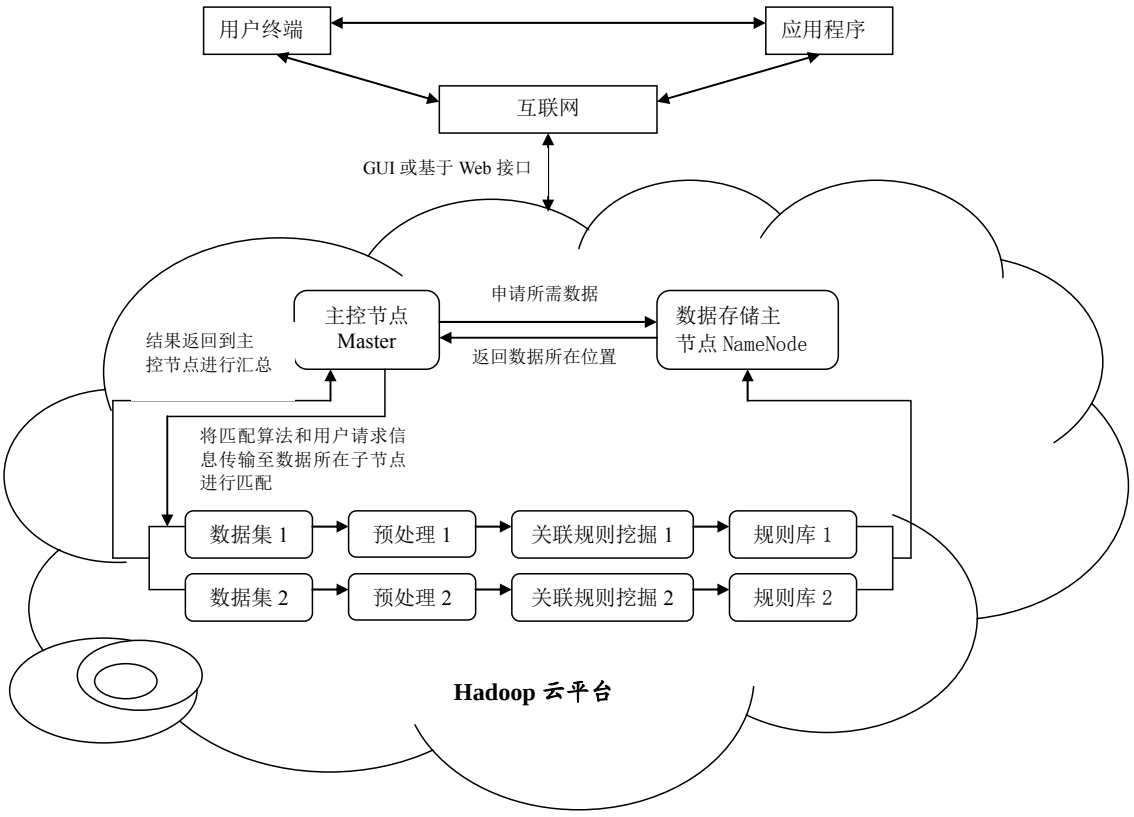

Figure 1. The basic architecture of book intelligent recommendation system based on cloud computing
图 1. 基于云计算的图书智能推荐系统基本架构

框架中的计算节点和HDFS分布式文件系统中的存储节点通常是同一个节点，而且在进行任务分配时采用数据存储在哪台计算机上，就由该计算机上的计算节点进行该部分数据的计算可以有效节约网络带宽资源，因此在此把计算节点和存储节点统称为同一节点即计算存储节点。云计算环境下的关联规则挖掘执行流程如下：每当读者访问图书智能推荐系统时，系统自动记录读者的浏览信息并分布式存储在HDFS不同的节点上；当master接收到挖掘任务后，向NameNode节点获取数据块所在位置信息，接着将相应的预处理方法、知识挖掘算法发送到原始数据所在的节点上并立即启动计算工作。在这过程中，计算节点会每隔一段时间就向master汇报其运行状态。

下面详细介绍关联规则挖掘模块的每一个部分：

1) 获取读者信息

读者信息的获取是图书智能推荐系统进行推荐的基础和前提。读者的信息包括显性信息和隐性信息。其中，显性信息主要是由读者注册提供的，包括用户的学历、学科专业、年龄、性别、兴趣爱好等；而隐性信息主要是读者的浏览行为数据。读者每次访问数字图书馆系统时都会留下一些借阅记录，包括IP地址、读者号、借阅时间、书名等。由于HDFS仅能存储半结构或非结构数据，所以应将Web上获取到的读者信息解析成XML文件或其它半结构、非结构数据。解析后的数据被存储在分布式文件系统中的不同数据节点上，同时为防止某个数据节点出现故障而导致的数据丢失问题，需要对每个数据块都复制几个副本并分别存储在不同的节点上。

2) 数据预处理

图书推荐过程中的一个主要任务是数据预处理，预处理的结果直接影响推荐产生结果的质量，是保证推荐质量的关键。由于收集到的图书馆历史数据和实时数据通常具有不一致性、冗余性及模糊性，所以必须对它们进行预处理，如数据清洗、用户识别、用户会话识别等。

3) 知识挖掘

这一过程将用到文献[3]提出的基于Apriori算法的并行关联规则挖掘算法，它是基于传统Apriori算法改进后的适用于云计算平台的并行算法。经过这一步骤，每个计算节点把挖掘的结果都传输到master上进行汇总排序并存放到HDFS文件系统中的规则库中。

4.2. 推荐模块

这一过程是在前面关联规则挖掘模块的基础上进行的，属于在线即时操作。当读者使用各种设备如PC、手机、PDA访问数字图书馆网站时，该站点的图书智能推荐系统就被触发，系统迅速收集该读者的注册信息和历史访问记录，接着MapReduce API将读者信息和匹配算法复制到存储规则库的每一台计算存储节点上，这时计算存储节点就地执行map程序。当map程序执行完后其中间结果被保存在本地磁盘并把位置信息发送给了Master节点，Master再将位置信息发送给执行Reduce任务的节点，执行Reduce任务的节点把中间结果汇总、排序，最后发送到客户端，这样便产生了图书推荐列表。其中，匹配算法描述如下：

输入：用户已浏览图书D,D={$book^1$, $book^2$,…, $book^n$}；

输出：被推荐图书集合Book_Set。

方法：

1) Book_Set = Φ; //被推荐图书集合初始值为空

2) repeat

3) 查找所有前件为D且后件B没有被用户浏览过的强关联规则Rule_Set；

4) count = |Rule_Set|; //记录强关联规则Rule_Set总数

5) if Rule_Set = null then

6) 按顺序减小用户已浏览图书D；

7) else if 系统设定的最大推荐数量N < count then

8) Book_Set = {B|B是Rule_Set中top N条规则的后件}; //选择Rule_Set中的前N条规则，并规则中的后件B作为被推荐书目保存到Book_Set中 else Book_Set = {B|B是Rule_Set规则中的后件}；//把规则中的后件B作为被推荐书目保存到Book_Set中

9) until Book_Set ≠ Φ or D = Φ

5. 结束语

本文针对传统图书智能推荐系统在海量数据及推荐速度受到很大限制，提出了一种把图书智能推荐系统建构在云计算平台上的设计思想。在介绍了图书智能推荐系统与云计算相关概念和技术的基础上，给

出了基于云计算的图书智能推荐系统架构。该系统在利用互联网的基础上，把云计算作为后台，并采用并行 Apriori 算法作为推荐算法，有效地解决了存储和计算问题[6,7]。

参考文献 (References)

[1] 黄晓斌. 数字图书馆推荐系统研究[J]. 情报资料工作, 2005, (4): 53-56.
[2] 安德智, 刘光明, 章恒. 基于协同过滤的图书推荐模型[J]. 图书情报工作, 2011, 55(1): 35-38.
[3] 程苗. 基于云计算的 Web 数据挖掘[J].计算机科学, 2011, 38(10A): 146-149.
[4] 丁雪. 基于数据挖掘的图书智能推荐系统研究[J]. 情报理论与实践, 2010, 33(5): 107-110.
[5] 林立宇, 陈云海. 基于云计算的电子商务推荐平台的构建分析[J]. 广东通信技术, 2010, 30(1): 7-10.
[6] 季涛. 基于云计算的个性化推荐系统的研究[D]. 对外经济贸易大学, 2011.
[7] 杨引霞, 谢康林, 朱扬勇等. 电子商务网站推荐系统中关联规则推荐模型的实现[J]. 计算机工程, 2004, 30(19): 57-59.

