doi:10.3969/j.issn.1000-436x.2013.z2.026

基于云计算的视频推荐系统的设计

李英壮，高拓，李先毅
（大连理工大学 网络与信息化中心，辽宁 大连 116024）

摘 要：通过对现有视频网站的调查研究，发现大部分都存在信息过载的问题。所以对视频网站来说拥有推荐系统是有必要的。通过对现有视频推荐系统的分析研究，利用开源云计算技术—Hadoop，及其部分相关组件 Hive、Hbase 等，设计了一种基于云计算的个性化视频推荐系统，此系统仅适用于以专业视频为主的网站。
关键词：信息过载；视频；推荐系统；云计算
中图分类号：TP391 **文献标识码**：A **文章编号**：1000-436X(2013)Z2-0138-03

Design of video recommender system based on cloud computing

LI Ying-zhuang, GAO Tuo, LI Xian-yi
(Network and Information Center, Dalian University of Technology, Dalian 116024, China)

Abstract: Through the survey and research of the existing video sites，some video sites face the problem of information overload. So it is necessary for the video sites to have a recommender system. Through the study of the existing video recommender system, using open source cloud computing technologies—Hadoop, and some of its related components such as Hive and Hbase, a personalized video recommendation system based on cloud platform was designed. The system is only applicable to professional video-based website.
Key words: information overload; video; recommender system; cloud computing

1 引言

随着互联网和信息技术的快速发展，渐渐从信息匮乏的时代走向了信息过载的时代。对于每个用户来讲，如何在众多信息中挑选出能让自己感兴趣的信息是非常困难的；对于网站来说，如何使自己产生的信息在大量信息中脱颖而出，从而吸引用户同样也是一件非常困难的事。在这种背景下，就产生了推荐系统。

推荐系统是帮助用户在大量的信息中，筛选出自己喜欢的信息的一种工具。它和搜索引擎不同，不需要用户给出十分明确的需求，而是通过对用户历史行为的分析，根据用户的兴趣进行建模，从而主动地向用户推荐可以满足他们需求和兴趣的信息资源。最早在 20 世纪 90 年代，就出现了几篇关于推荐系统的论文[1~3]，由此推荐系统这个领域就受到了各界的广泛关注，并且发展得很好。

现在的视频网站主要分为 2 种，一种主要是像国外的 YouTube 和国内的土豆和优酷这样主要内容为用户自己上传的网站，即 UGC 网站；另一种则是像国外的 Hulu、Netflix 和国内的搜狐视频、乐视这样主要内容为专业视频的网站。这 2 种视频网站的内容和用户的行为存在一定的差别，因此，对于设计相应的推荐系统也会存在差异。本文主要是针对专业视频内容为主的网站进行设计的。

2 相关技术介绍

2.1 Hadoop

Apache 软件基金会旗下存在很多开源的项目，Hadoop 就是其中之一，它是一个开源分布式计算平台。其核心为 HDFS(hadoop distributed file system)和 MapReduce(Google MapReduce 的开源实现)，它为用户屏蔽了系统底层实现的细节，提供了透明的分布式架构。用户可以通过 Hadoop 利

收稿日期：2013-09-04

用计算机的各种资源，包括集群的计算和存储能力，从而搭建分布式计算平台来完成处理海量数据的目的。现在除了 2 个核心项目外，Hadoop 还拥有很多其他的子项目，包括 Pig、Hive、HBase 等。它们可以为彼此提供服务，或者在核心层之上提供更高级的服务[4]。

2.2　Apache Mahout

Apache Mahout 起源于 2008 年，当时还只是 Apache Lucene 的一个子项目，但是经过两年多的发展，在 2010 年 4 月正式成为了 Apache 的顶级项目。Apache Mahout 的主要目标是针对大规模的数据集实现一些可以伸缩的机器学习算法。Mahout 现在已经包括许多机器学习算法的实现，例如推荐引擎、聚类、分类和频繁子项挖掘等。此外，通过使用 Apache Hadoop 库，可以将其功能有效地扩展到 Hadoop 云平台中，它通过 MapReduce 模式实现。

3　系统分析与设计

本节主要分为两部分，一部分是介绍使用的推荐方法，另一部分是整个系统的结构设计。

3.1　基于 Mahout 的分布式推荐方法

传统的推荐引擎算法多在单机上实现，它们只能处理一定量的数据。如果数据量达到一定的规模，传统的推荐算法就会出现各种问题。所以在本文中采用基于 Mahout 的分布式推荐算法。算法流程如下所述。

1) 通过登录用户行为记录的日志分析发现，原始数据都存在一些相同的特征。因此可以将原始的数据以 UserID、ItemID、Preference 的形式作为一条记录，UserID 用来识别用户，ItemID 用来识别视频，Preference 表示一个用户对看过的某一个视频的评分。

2) 根据用户的原始数据记录，计算视频与视频之间的相似关系，得到视频的相似性矩阵：***ItemAID***、***ItemBID***、***Value***（相似度）。计算相似性时，采用修正的余弦相似性算法，设对视频 i 和视频 j 共同评分的用户集合用 U_{ij} 表示，U_i 和 U_j 分别表示对视频 i 和视频 j 评过分的用户集合。则视频 i 和视频 j 之间的相似性 $sim(i,j)$ 为

$$sim(i,j)=\frac{\sum_{c\in u_{ij}}(R_{c,i}-\bar{R}_c)(R_{c,j}-\bar{R}_c)}{\sqrt{\sum_{c\in u_i}(R_{c,i}-\bar{R}_c)^2}\sqrt{\sum_{c\in u_j}(R_{c,j}-\bar{R}_c)^2}}$$

其中，$R_{c,i}$ 表示用户 c 对视频 i 的评分，$\bar{R}_c$ 表示用户对所有看过的视频的平均评分。

除了相似度矩阵之外，还要计算表示用户喜好的向量，在该向量中，对于用户评过分的视频会对应相应的评分，而没评过分的视频可以选择用 0 代替。例如对某一用户 i 来说，它的喜好向量可以表示为（5.0,0,3.5,4.0,0,…）。通过计算可以得到所有用户的喜爱值，其实这也是一个矩阵，矩阵的行值表示视频编号，列值是用户编号，行列对应的元素值为用户对产品的喜爱值。

3) 计算推荐列表，将 2)中计算得到的相似性矩阵乘以表示用户喜好的列向量，可以得出一个新的列向量，以此为得分，并按照由高到底进行排序；除去用户已经观看过的视频之外，最后返回排名靠前 N 个视频，即为推荐结果。

3.2　结构设计

视频推荐系统从用户的历史观看记录中分析用户的兴趣模型，然后挑选出适合用户并能让用户感兴趣的视频展示给用户。因此本系统包括日志系统、推荐引擎和展示界面三部分[5]。本系统的框架如图 1 所示。

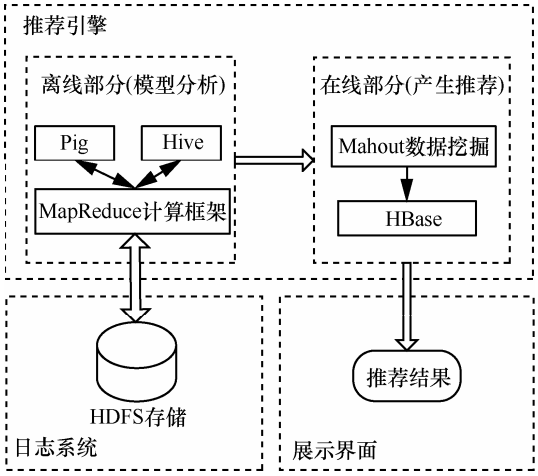


图 1　系统框架

日志系统主要收集用户的历史行为和用户对推荐系统的反馈。其中以用户的观看记录和评论表示用户对视频感兴趣的行为；以感兴趣、收藏推荐的视频表示用户对推荐系统的反馈行为。由于视频网站的流量和用户规模都比较庞大，因此后台的日志系统使用云计算平台进行处理，通过搭建 Hadoop 集群来处理用户的行为日志文件。系统定期地取出一段指定时间间隔内的一

些日志数据，存储到 HDFS 或 HBase 中以供后续使用。

推荐引擎分为离线部分和在线部分。离线部分主要用于模型分析、对用户的兴趣建模，根据在日志系统中收集到的用户行为日志，利用 Hive、Pig、MapReduce 等工具计算出用户和视频之间的评分矩阵和表示用户喜好的向量，并把它们存储在 HBase 中，可以给在线部分提供实时的查询和调用。

在线部分主要是对用户的视频观看请求进行实时响应，把产生的推荐结果展示给用户。这部分主要利用 Mahout 根据离线部分提取和分析到的用户历史行为以及产生分析的结果，对其进行过滤并排序，给出最终的推荐列表，列表保存在 HBase 中供一段时间内使用。

展示界面不在本文的研究范围之内，所以不在这里进行讨论。

4 系统测试

为了测试系统的性能，本文在特定的平台上进行实验，利用测试数据来测试系统的推荐精度。

实验环境：采用 6 台配置相同的机器作为 hadoop 集群：处理器类型：Intel 双核处理器；内存容量：2 GB。

实验数据及内容。在实验中采用 MovieLens 10 MB 的开源数据集模拟真实系统中的数据。实验的内容是根据采用不同推荐结果的数量N来测量系统推荐结果的精度。

评测指标。采用的评价指标是推荐结果的准确率和召回率。准确率描述最终的推荐列表中有多少比例是发生过的用户——物品评分记录，而召回率描述有多少比例的用户——物品评分记录包含在最终的推荐列表中。分别用如下公式计算它们。

$$Precision=\frac{\sum_u |R(u)\cap T(u)|}{\sum_u |R(u)|}$$

$$Precall=\frac{\sum_u |R(u)\cap T(u)|}{\sum_u |R(u)|}$$

其中，$R(u)$表示给用户推荐的视频，$T(u)$表示用户在测试集上感兴趣的视频。

实验结果分析：通过分析实验结果，作者发现推荐结果的精度不与推荐结果的数量成正比或反比，因此选择合适的推荐结果数量 N 是很重要的。结果如图 2 所示。

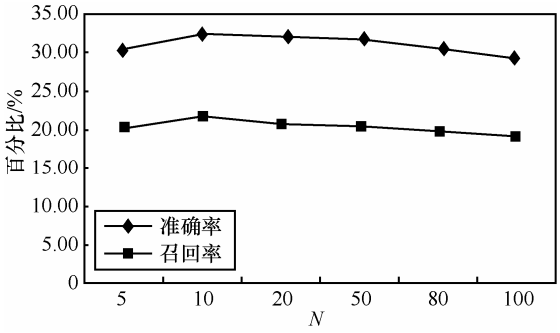


图 2 不同 N 值对应的准确率和召回率的值

5 结束语

视频网站现在的视频量已经超出用户可以观看的限度，并且其中很多都是用户不感兴趣的，同时视频网站希望可以提供给用户他们感兴趣的视频，从而使用户愿意选择它们的网站作为首选，以上因素都决定了视频网站拥有推荐系统是很有必要的。目前国外很多视频网站都已经开发了适合自己的推荐系统，同时国内的一些视频网站也正在加大投入研发成本和人力来部署自己的推荐系统。本文就是在这种大环境下，通过对现有推荐系统的研究，设计了一种基于云计算的视频推荐系统，该系统的性能有待进一步提高。

参考文献：

[1] SHARDANAND U, MAES P. Social information filtering：algorithms for automating “Word of Mouth”[A]. Proc of the Conf on Human Factors in Computing Systems[C]. New York:ACM Press, 1995. 210-217.

[2] HILL W, STEAD L, ROSENSTEIN M, *et al*.Recommending and evaluating choices in a virtual community of use[A]. Proc of the Conf on Human Factors in Computing Systems[C]. New York: ACM Press, 1995. 194-201.

[3] RESNICK P, IAKOVOU N，SUSHAK M, *et al*. GroupLens：an open architecture for collaborative filtering of net news[A]. Proc of the Computer Supported Cooperative Work Conf[C]. New York:ACM Press, 1994.175-186.

[4] 路嘉恒. Hadoop 实战(第 2 版)[M]. 北京：机械工业出版社, 2012.
LU J H. Hadoop in Action(2nd Edition)[M]. Beijing: China Machine Press, 2012.

[5] 郑华. 推荐系统在视频网站中的应用[J]. 程序员, 2011,(5):66-69.
ZHENG H. The application of in recommended system video site[J]. Programmer, 2011,(5):66-69.

（下转第 147 页）

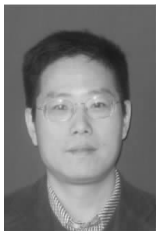
[10] 蒋清锋，陈惠欢，郑建立等. 基于门禁的高校开放式实验教学管理系统[J]. 计算机系统应用，2013, 3: 51-54.
JIANG Q F, CHEN H H, ZHENG J L, *et al*. The open experimental teaching management system based on access control[J]. Application of Computer System, 2013, 3:51-54.

作者简介：

许彩娥（1982-），女，河北邯郸人，硕士，浙江大学信息中心校园卡办公室技术员，主要研究方向为校园卡应用服务等。

徐锋（1976-），男，浙江湖州人，硕士，浙江大学信息中心校园卡办公室主任，主要研究方向为校园信息化应用、校园卡应用服务等。

厉晓华（1975-），男，浙江东阳人，硕士，浙江大学信息中心高级工程师，主要研究方向为数字电路与网络信息安全。



鲁东明（1967-），男，浙江余姚人，博士，浙江大学信息中心教授，主要研究方向为文化遗产数字化保护、虚拟现实与数字博物馆、数字媒体网络系统等。

程艳旗（1967-），男，陕西西安人，浙江大学信息中心副研究员，主要研究方向为校园信息化应用服务。

（上接第 140 页）

作者简介：

李英壮（1955-），男，辽宁大连人，大连理工大学教授，主要研究方向为计算机网络。

高拓（1988-），女，黑龙江齐齐哈尔人，大连理工大学硕士生，主要研究方向为计算机网络。

李先毅（1978-），男，辽宁大连人，大连理工大学博士生，主要研究方向为计算机网络。