基于关联规则的气象服务智能推荐

赵文芳，刘亚楠，余东昌
北京市气象信息中心，北京 100089

摘要
针对气象公众服务存在的专业化程度不高、针对性不强等问题，提出了一种使用关联规则向用户推荐气象服务产品的方法。将用户访问的Web日志数据经过清洗、转换后进行关联规则挖掘，同时设计了一种基于气象产品浏览总次数计算产品相似度的算法，对挖掘出来的关联规则进行筛选。共进行了5组实验，并对用户浏览产品的行为进行了简单预测。实验结果表明，挖掘出的规则在降水天气和雾霾天气中的正确率比较高，结合气象服务产品的相似度能过滤掉10%的无效规则，从而提升规则的正确性。

关键词
智能推荐；关联规则；Web日志；气象公众服务；并行计算；相似度；协同过滤

中图分类号：TP183　　文献标识码：A　　doi: 10.11959/j.issn.2096-0271.2018020

Intelligent recommendation of meteorological service based on association rules

ZHAO Wenfang, LIU Yanan, YU Dongchang
Beijing Meteorological Information Center, Beijing 100089, China

Abstract
To overcome the problem in which public meteorological services are rarely optimized for personalization, a recommendation method incorporating improved association rules and collaborative filtering was presented. First, FP-Growth algorithm was applied on pre-processed Web log data to generate association rules. Meanwhile, a customized collaborative filtering algorithm was used to calculate pairwise similarities between meteorological products based on users' browsing records. A total of five experiments were conducted. The experimental results show that the rules are relatively accurate for precipitation and haze weather, because pairwise similarities between meteorological products which are calculated by the proposed algorithm could reduce the total number of invalid or irrelevant association rules by 10%.

Key words
intelligent recommendation, association rule, Web log, public meteorological service, parallel computing, similarity, collaborative filtering

1 引言

气象与国民经济各行各业关系密切，气象部门在千方百计提高预报预测准确率的同时，一直致力于推动气象与经济社会的融合，更好地发挥气象在经济社会中的作用。为了顺应信息化时代发展趋势，中国气象局提出了“互联网+气象”行动计划，促进气象与经济社会融合发展。“互联网+气象”行动计划对气象服务提出了明确要求，不再局限于单纯地向社会提供气象预报信息，而是为天气高影响行业提供有针对性的专业气象服务，对公众提供精细化、个性化的气象服务。目前，面向公众和行业的气象服务由于缺乏对用户行为数据的分析，难以把握其活动特征，难以主动提供个性化气象服务产品，存在专业化程度不高、针对性不强等问题。因此，如何为用户推荐合适的气象服务产品成为了需解决的重要问题。

推荐系统是帮助人们快速发现有用信息的重要工具，它不仅给用户推荐符合个人兴趣的内容，还为用户节省了在海量数据中查找信息的时间[1]。推荐系统中常用的推荐技术主要有基于内容的推荐、协同过滤推荐及关联规则推荐[2-7]等。其中，基于内容的推荐算法适用于推荐新商品，但不适用于新用户推荐；基于知识的推荐算法适用于用户的行为数据较少，同时又有明确需求的场景；协同过滤算法依据用户评分为用户推荐商品，同时也存在冷启动问题。

目前，现有的气象服务模式主要以手机应用程序、微信公众号、微信小程序和网站为载体，向用户提供气象服务产品的浏览和查询，没有评分功能，无法获取用户评分数据，只能大量记录用户行为的Web日志数据。本文提出了一个根据用户浏览记录向用户推荐气象服务产品的方法，具体贡献如下。

- 提出了一种对气象产品进行挖掘的方法，利用FP-Growth算法进行关联规则挖掘，然后对误导性强关联规则进行排查和过滤，建立暴雨、高温、大风、雾霾等高影响天气下的气象服务推荐模型。
- 提出了一种基于页面浏览总量的气象产品相似度算法，实现用户对每个产品浏览总量的线性变换处理，并映射到区间[1,5]，作为用户对每个产品的评分，最后利用修正后的余弦相似度计算产品之间的相似度。
- 基于Spark框架设计了两种数据预处理并行算法，将原始数据抽象为弹性分布式数据集（resilient distributed datasets，RDD）模型，利用map、groupByKey等函数进行数据的大量迭代计算、归一化处理，快速实现数据清洗、转换。
- 提出了利用Impala实现原始数据冗余的处理方法。
- 通过在Cloudera大数据平台的Spark框架下进行实验，说明本方法的实用性和高效性。

2 相关工作

对于气象服务产品的推荐，目前主要以手机移动软件为载体实现两个方面的研究应用：基于用户地理位置的气象服务；根据用户的订阅计划及时推送相应气象产品。“气象雷达—天气预报、温度和雨量图”手机应用软件提供了便捷的地点搜索，为用户提供最近距离气象站的精准天气数据，并为用户提供3种订阅计划。“中国气象”手机应用软件能为用户提供最近

距离的气象站实况监测信息和站点搜索功能。“智慧气象”手机应用软件为用户提供“一键式专业气象信息订制”功能，让用户一键轻松获取所需信息。“彩云天气”“墨迹天气”“北京气象”等手机应用软件也提供了类似的功能。

针对关联规则个性化推荐的研究主要包括：参考文献[8]对关联规则算法进行改进，并提出了一种基于划分的关联规则并行分层挖掘算法；参考文献[9]提出了Hadoop 平台下的并行Web日志挖掘算法；参考文献[10]首先利用模糊聚类进行数据预处理，在此基础上再进行频繁项集的挖掘；参考文献[11]提出了基于MapReduce的并行关联规则增量更新算法；参考文献[12]提出了一种基于关联规则挖掘的跨网络知识关联及协同应用；参考文献[13]提出了一种结合关联规则的协同过滤推荐算法；参考文献[14]提出了基于业务路径和频度矩阵的关联规则挖掘算法；参考文献[15]采用多重约束进行时序关联规则挖掘，分析得到了与实际情况高度吻合的气候指数与陆地区域异常降水事件间的关系；参考文献[16]采用了改进的FP-Growth算法挖掘出各种气象因子之间可能存在的关联，从而发现气象特点，对近期天气气象做出预报。

以上研究表明，通过挖掘用户行为提取兴趣特征，在兴趣特征基础上再为用户提供个性化推荐服务的方法广泛用于电子商务中。但是，使用包含用户轨迹的Web日志数据分析研究物品之间强关联规则，再进行推荐的方法相对较少。此外，气象服务推荐也局限于基于地理位置的精细化服务，缺少对气象服务产品之间相关性的挖掘和用户对不同产品的偏好研究，无法对用户进行更精准、个性化的推荐。因此，本文提出一种基于气象服务产品关联规则的推荐模型。

3 气象服务产品推荐模型

FP-Growth算法是韩家炜等人在2000年提出的关联分析算法，通过构造一个树结构来压缩数据记录，使挖掘频繁项集只需要扫描两次数据记录，而且该算法不需要生成候选集合，效率比经典Apriori算法高。本文将气象服务网站的Web日志作为基础数据，结合FP-Growth算法和物品相似度算法建立气象服务产品的推荐模型。

本节主要介绍模型概述和相关数据预处理算法。

3.1 模型概述

用户每次访问气象服务网站时，会停留一段时间浏览一些页面的服务产品，点击页面上的菜单按钮，有时候还会下载相关的气象数据和产品，这些被浏览的页面、用户点击页面、下载数据的行为形成用户访问记录。用户浏览轨迹表示用户在这段时间内浏览的气象服务产品序列，该序列可以表示为：$List=\{p_1, p_2, \cdots, p_n\}$。用户访问记录定义为：$User\_Res=\langle U, server\_time, List, time\_spent, actions, actions\_count\rangle$，其中，$U$表示一个用户，$server_time$表示用户访问网站时间，$List$表示浏览过的产品序列，$time_spent$表示用户停留在网站的时间，$actions$表示用户的行为，$actions_count$表示用户行为次数。

气象服务产品作为向用户推荐的对象，其本身具有一些属性。本文划分为静态属性和动态属性，静态属性来自于产品气象本质特征，动态属性来自于对动态数据的统计分类的结果。静态属性包括产品更新频率、产品所属气象数据类别（如观

测、数值模式产品，预报产品）、产品的用途（如实况、预报、统计）、产品来源等，本文以标签的形式描述这些属性特征。气象服务产品的静态属性定义为：$P\_sta=\langle P_{ID}, tag_1, tag_2, \cdots, tag_n\rangle$。$P_{ID}$表示产品的唯一编号，$tag_i$表示产品的第$i$个属性标签，表示某个属性特征。动态属性主要指产品浏览次数和浏览时间，这些数据需要从Web日志数据统计得到。本文根据气象服务产品的静态属性和动态属性分别计算相似度。

FP-Growth 算法中的关键参数min_*support*及min_*confidence*通常都是人为设置的，生成的规则容易受到挖掘者主观因素的影响，有必要对强关联规则做进一步筛选和验证，以提升规则的可靠性和准确性。

基于以上定义，本文提出的基于关联规则的气象产品推荐模型如**图1**所示。模型从网站访问日志中提取出用户浏览记录，筛选出每个用户在有效时间内访问过的气象服务产品序列，形成样本数据集，开始训练模型，使用FP-Growth算法进行关联规则挖掘，得到符合条件的关联规则；计算气象产品的相似度，根据相似度筛选关联规则，构建气象服务产品推荐模型。

3.2 数据预处理方法

气象服务网站一直使用谷歌公司的开源工具PIWIK进行网站流量和用户行为监控，相关Web日志信息存放在MySQL数据库表中。本文使用了两个表的数据：用户访问记录表和页面信息管理表。其中，用户访问记录表用来记录用户每次访问网站的相关页面，包括用户ID、页面ID、用户访问时间、页面停留时间、用户行为（浏览或下载）；页面信息表包括页面ID与页面URL对应关系。具体表结构见**表1**和**表2**，**表1**中的页面ID字段和**表2**中的页面ID字段相关联。

与电子商务网站商品不同的是，气象产品是与时间相关的，当用户请求同一个页面URL并获取不同时间气象产品时，相关的Web日志会赋给这个页面URL一个新页面ID。例如，用户在2017年1月1日20:00和2017年1月1日21:00对命名为aws.index.php的自动站实况页面进行访问，页面URL为aws.index.php?time=2017-01-01 20:00:00和aws.index_

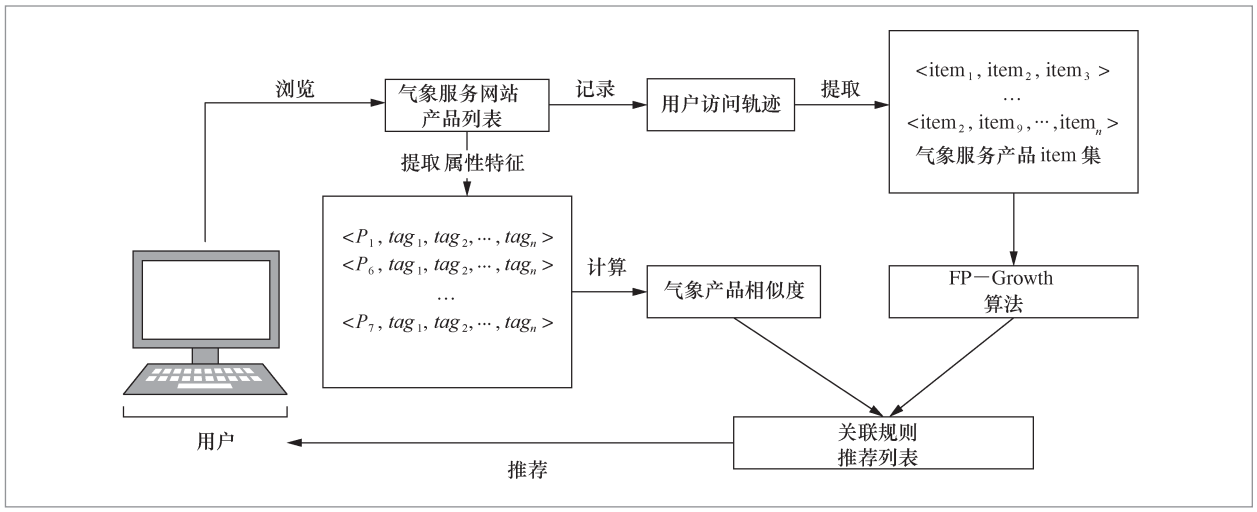


图1 基于关联规则的气象产品推荐模型

php?time=2017-01-01 21:00:00，Web日志会赋予这两个URL不同的页面ID进行记录，造成大量冗余页面，在进行数据挖掘前，必须进行冗余页面处理。

本文对冗余页面的处理利用数据表的关联实现。首先，创建页面元数据信息表，其结构见**表3**。其次，对**表2**的id_re_name字段和**表3**的id_name字段进行关联匹配，得到**表2**的id_url在**表3**中的对应值，同时使用这个对应值对**表1**字段id_url进行更新。

本文中所有的挖掘算法处理都是基于经过冗余处理的数据进行的。

3.3 气象服务产品相似度计算方法

3.3.1 基于静态属性的相似度计算方法

当新的气象服务产品上线时，由于缺少Web访问日志数据，无法开展关联规则推荐。当用户浏览该产品时，可以把和它最相似的产品推荐给用户。

表1 用户行为记录表结构

字段ID	字段含义
idvisitor	用户ID
id_url	页面ID
server_time	用户访问时间
time_spent_ref_action	页面停留时间
action_type	用户行为

表2 页面信息表结构

字段ID	字段含义
id_url	页面ID
id_name	页面URL
id_re_name	去掉时间参数后的URL

表3 页面元数据信息表结构

字段ID	字段含义
id_url	页面ID
id_name	页面URL

两个气象服务产品可能会包含相同标签，例如更新频次、所属气象数据类别，那么这些相同标签就是这两个气象服务产品的相似之处。两个产品（p_i和p_j）中的相同标签占两个产品所包含的所有标签的比值表示这两个产品的相似程度，用StaSim(p_i, p_j)表示，计算方法如式（1）所示：

$$\mathrm{StaSim}(p_i, p_j)=\frac{\mathrm{Tags}(\mathrm{p}_i)\cap\mathrm{Tags}(\mathrm{p}_j)}{\mathrm{Tags}(\mathrm{p}_i)\cup\mathrm{Tags}(\mathrm{p}_j)} \quad (1)$$

其中，Tags(p_i)和Tags(p_j)分别表示p_i和p_j包含的所有标签。对于p_i和p_j，它们包含的标签的交集中元素的个数与并集中元素的个数的比值表示p_i与p_j的相似程度。

3.3.2 基于动态属性的相似度计算方法

基于物品的协同过滤算法适用于物品数量变化小于用户数量变化的场合，本文采用该算法，使用修正后的余弦相似度计算式计算气象服务产品相似度。具体计算方法如式（2）所示：

$$\mathrm{Sim}(i,j)=\frac{\sum_{u\in U}\left(M_{u,i}-\overline{r_i}\right)\left(M_{u,j}-\overline{r_j}\right)}{\sqrt{\sum_{u\in U}\left(M_{u,i}-\overline{r_i}\right)^2}\sqrt{\sum_{u\in U}\left(M_{u,j}-\overline{r_j}\right)^2}} \quad (2)$$

其中，Sim(i, j)表示气象服务产品i和j的相似度。$M_{u,i}$表示用户u对产品i的评分，$M_{u,j}$表示用户u对产品j的评分，$\overline{r_i}$表示用户对产品i的平均评分，$\overline{r_j}$表示用户对产品j的平均评分。

3.3.3 用户访问页面总次数的线性变化方法

对用户访问页面总次数进行线性变化处理，使结果值映射到区间[1,5]，转换函数如式（3）所示：

$$F(x)=\frac{4\times(x-\min x)+(\max x-\min x)}{(\max x-\min x)} \quad (3)$$

其中，$F(x)$表示映射后的值，x表示用户访问某页面的总次数，max x表示某页面访问次数的最大值，min x表示该页面访问次数的最小值。

3.3.4 气象产品相似度的计算方法

算法将用户浏览产品的总次数经过线性变换后将结果映射到区间[1,5]，利用修正后的余弦相似度计算式计算产品相似度。具体的算法描述如下，算法流程如**图2**所示。

算法：气象产品相似度的预处理算法。

输入：去掉冗余后的用户访问记录。

输出：气象服务产品相似度矩阵。

步骤1 从日志信息中获取用户与页面浏览量数据。

步骤2 对每个独立IP用户，迭代计算其浏览过的每个页面总次数。将输入数据map映射为JavaPairRDD<String，Int>的键值对，命名为Res_1。其中，键值对的key表示为“用户ID_页面ID”，value值自动累加，表示用户ID访问页面ID所代表页面的总次数，即<用户ID_页面ID，总次数>。例如<3312_78，56>表示ID为3312的用户，对ID为78的页面一共访问了56次。

步骤3 对用户访问每个页面的总次数进行归一化处理。在数据集Res_1中，找出key中包含相同页面ID的所有键值对，对其中的value进行归一化处理，将页面访问总次数按式(3)映射到区间[1,5]，生成新的数据集Res_2。

步骤4 对Res_2数据集进行map映射计算，从key中提取出用户ID，页面ID和value的组合作为value，生成新的JavaPairRDD<String，String>，得到数据集New_Res。

步骤5 对New_Res数据集进行groupByKey聚合计算，将相同key的多个JavaPairRDD<String，String>映射为JavaPairRDD<String，Iterable<String>>，得到数据集Count_Res。

步骤6 将Iterable<String>从数据集Count_Res提取出来，转换成气象服务产

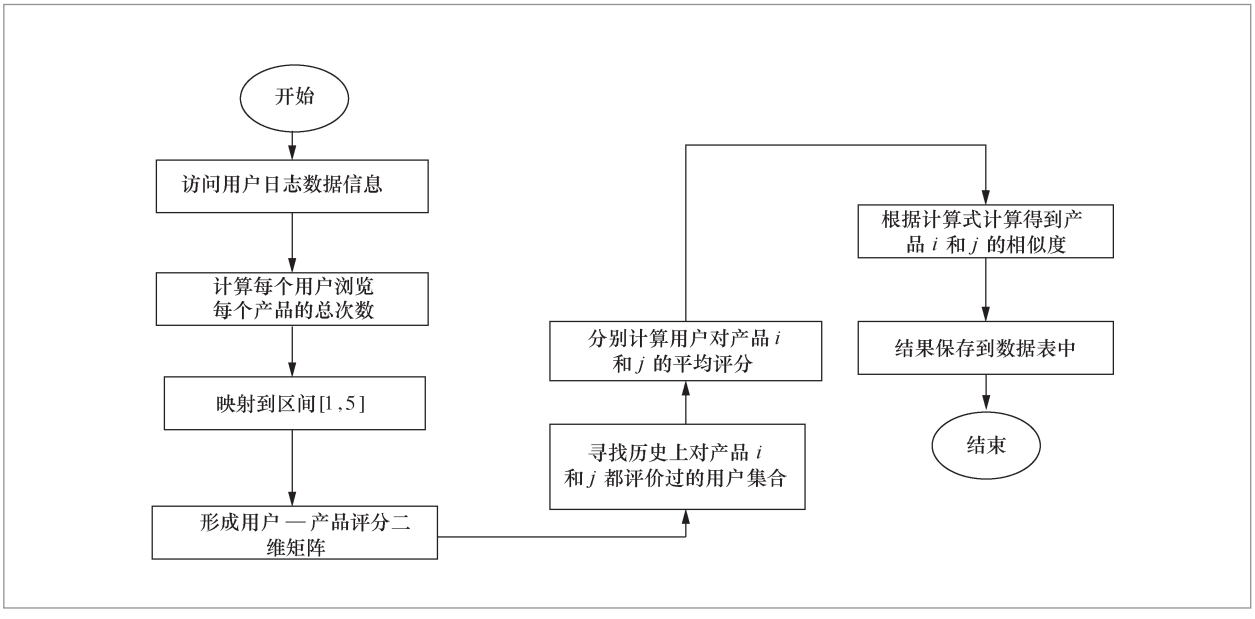


图2 气象产品相似度算法流程

品评分矩阵。

步骤7 对于n个产品，依次计算产品1与其他$n-1$个产品的相似度；再计算产品2与其他$n-2$个产品的相似度。对于其中任何两个产品i和j而言：查找对i和j共同评价过的用户集合U_{ij}；分别计算用户对产品i和产品j的平均评价值$\overline{r_i}$和$\overline{r_j}$；按照式(2)计算产品之间的相似度。

步骤8 将计算得到的相似度结果存在数据库表中，数据库表结构见**表4**。

表 4 气象服务产品相似度

字段ID	字段含义
id_url	页面ID
sim_url	相似度页面ID列表，ID之间用“$”分隔
sim_url_value	相似度列表，值之间用“$”分隔

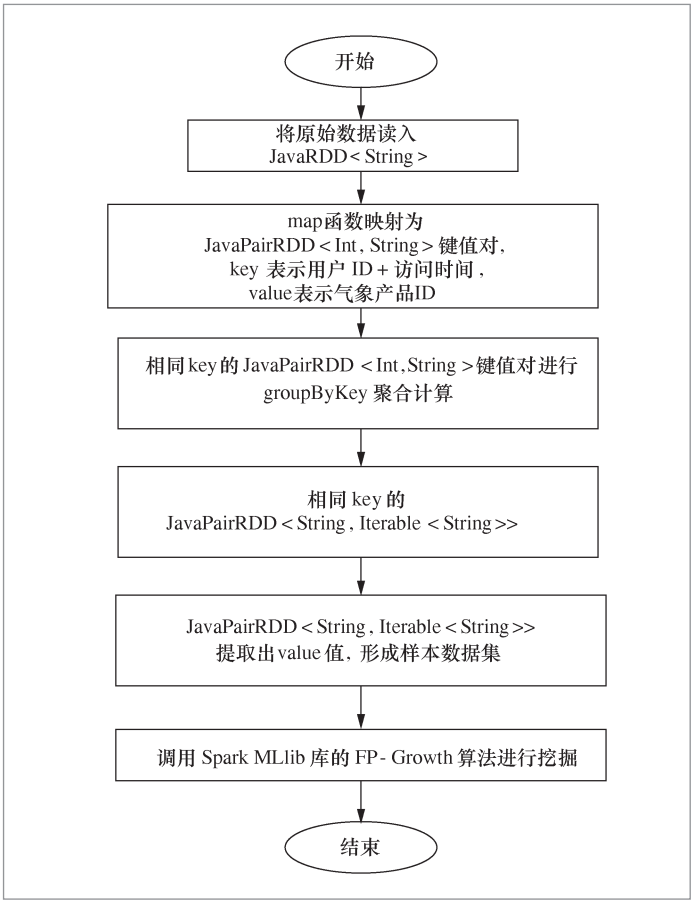


图 3 气象产品 item 提取算法流程

3.4 基于用户访问日志的关联规则算法

该算法用于从用户访问日志生成最终关联规则。首先对用户访问日志进行清洗，转化为用户访问产品的序列数据集，然后调用FP-Growth算法产生关联规则。该算法的关键在于用户访问日志的预处理阶段，需要生成一个数据集，数据集中的每一个数据项包含了用户每次访问网站浏览过的产品ID序列。具体的算法描述如下，算法流程如**图3**所示。

算法：基于用户访问日志的关联规则算法。

输入：去掉冗余后的用户访问记录。

输出：气象服务产品关键规则列表。

步骤1 读入访问记录，通过map映射为JavaPairRDD<Int，String>的键值对Res_1。其中，key表示用户ID+访问时间，访问时间形式为YYYYMMDDHHMM，精确到分；value表示气象产品ID。

步骤2 对Res_1数据集进行groupByKey聚合计算，将相同key的多个JavaPairRDD<Int，String>映射为JavaPairRDD<String，Iterable<String>>，得到数据集Res_2。

步骤3 从数据集Res_2中提取出每一个键值对的格式为Iterable<String>的value，然后逐天保存，作为数据挖掘的数据样本集data_sam。

步骤4 对数据样本集data_sam调用FP-Growth算法，产生关联规则列表。

3.5 强关联规则的筛选方法

关联规则挖掘算法需要预先设定最小支持度和最小可信度参数才能进行计算。而人为主观设置这两个参数容易导致规则过多或者或少，或者把不相关的规则推荐

给用户等问题。为了减少关联规则挖掘的这些不利影响，让规则总数量尽可能在一个适当范围，利用气象服务产品之间相似度来筛选关联规则。本文重点针对2个向量频繁项集和3个向量频繁项集中置信度大于90%的强关联规则进行筛选。具体使用的方法描述如下。

- 对于2个向量频繁项集，查看2个向量表示的气象服务产品的相似度。如果其相似度小于支持度，则剔除，反之则保留。例如有规则“[13] => [4]，置信度为 0.92”，而ID分别为13和4的气象服务产品相似度为0.78，大于最小支持度0.2，标记为有效规则进行保留。
- 对于3个向量的频繁项集，将向量间的相似度加权平均，再与最小支持度进行比较。对于规则“[***A***, ***B***] => [***C***]”，将向量***A***与向量***C***的相似度和向量***B***与向量***C***的相似度加权平均得出的结果与最小支持度相比，如果其相似度小于支持度，则剔除，反之则保留。

4 实验与分析

4.1 实验环境

本文选择Apache Flume 1.5、Spark1.5、Java JDK(Java Development Kit) 1.7、Impala2.0、Cloudera 5.8.3作为系统的运行环境，采用另一种资源协调者(yet another resource negotiator, YARN)模式作为系统的运行方式。其中，Cloudera集群由7台实体服务器构成，每台服务器的硬件配置见**表5**。

在Cloudera环境下，需要将YARN、Impala、Spark服务合理地部署在不同的节点上，以便充分利用集群资源给系统运行提供更好的性能保障。选择运行Cloudera Manager服务的节点作为分布式文件系统的名字节点（namenode），并运行Spark Master、Impala Server和YARN资源管理服务。

4.2 实验数据

本文使用的实验数据来源于气象服务网站2011—2016年的Web日志信息。经过冗余处理后，可以获得39 146 882条用户浏览记录，每条记录包括用户的访问时间、用户ID、IP地址、浏览气象产品ID、浏览时间。其中，独立IP地址1 946条，包含气象产品562个。筛选掉包含错误信息的记录以及伪访问用户的记录（有些用户将客户端浏览器设置为4 s以下频率刷新网站，不能真实反映用户浏览产品情况）后，提取出5 148 691条浏览记录，在此基础上进行本文的所有实验。

4.3 算法实现

4.3.1 冗余页面的处理

本文使用Impala实现对冗余页面的处理。根据文中提到的方法，将**表1**和**表2**数据导入Impala，并在Impala中建**表3**。然后在Impala中进行表的关联操作。

表1记录数为39 146 882条，**表2**记录数为5 548 191条，**表3**记录数为563条。Impala是Cloudera平台下的大规模并行处理（massive parallel processing, MPP）数据库引擎，通过它进行**表2**和**表3**的关联，耗时322 ms，**表1**和**表2**的关联耗时9.13 s。

4.3.2 基于修正的余弦相似度算法实现

基于修正的余弦相似度算法从5 148 691条浏览记中生成1 946个用户对562个产品的

评分，最后得出562个产品相似度矩阵。算法基于Spark用Java语言开发实现，程序选择Spark on YARN运行方式。在运行之前，先调整YARN服务的可用资源。在7个节点的集群中，每个节点有8个核以及32 GB的内存。其中，1个节点运行YARN资源管理服务，6个节点运行YARN节点管理服务；考虑到每个节点上操作系统、Hadoop 的Daemon（守护）进程以及其他组件进程的运行也需要一定资源，在该测试中分配给YARN 75%的资源，即每个节点上YARN服务的可用资源为28 GB内存和5个核。

对该算法运行效率进行测试，考察在数据量、任务提交方式及其他配置不变的情况下，改变num-executors、executor-cores和executor-memory 3个运行参数引起的时效变化。一共进行9组试验，详细的参数配置见**表6**，num-executors、executor-cores和executor-memory在表中分别用参数1、参数2和参数3表示。

表 5 Cloudera 集群中各主机配置

操作系统	Red Hat 7.1
CPU	Intel(R) CPU E5-2680 v3，主频2.60 GHz，8核
内存	32 GB
硬盘	SAS 300 GB×2做的RAIDl，用于操作系统和软件；STAT l TB×3用于Hadoop 数据存储

表 6 Spark Streaming 作业运行参数配置

序号	配置参数				
	参数1	参数2	参数3	spark.default.parallelism	耗时/s
1	2	1	1	未启用	138
2	4	8	2	未启用	126
3	8	8	2	未启用	44
4	12	6	8	未启用	38
5	16	4	8	未启用	33
6	20	4	8	280	27
7	24	3	2	900	19
8	24	2	3	1 000	23
9	24	3	3	1 200	26

在试验中，num-executors从2逐渐增大到24。由**表6**中每组试验的耗时可知，算法耗时随num-executors的增加呈下降趋势。当num-executors取值为2~4、4~8、8~16时，耗时并没有明显的变化。在将num-executors设置为20~24的4组试验中，启用spark.default.parallelism参数，程序运行效率得到提升。当num-executors为24、executor-cores为3、executor-memory为2 GB、spark.default.parallelism为900时，整体性能达到最优。spark.default.parallelism参数用于设置每个阶段的默认任务数量。前5组试验没有启用这个参数，而Spark默认设置任务数为几十个，导致60%~70%的Executor进程没有任务执行，因此尽管Executor参数在增大，但程序的耗时基本没有变化。由此可见，如果这个参数不设置或者设置不当，会直接影响Spark作业性能。另外，Spark作业性能与运行参数、平台分配给YARN的可用资源、spark.default.parallelism等诸多因素有关，在其他参数不变的情况下，性能与num-executors参数也不是简单的线性关系。因此，在实际运行中，需要统筹考虑这些参数的影响，选择一个最优的参数组合。

4.3.3 FP-Growth 算法的关联规则挖掘实现

编写Java程序，从5 148 691条浏览记录中经过预处理生成2 500 000个样本数据，通过调用Spark机器学习库（machine learning library，MLlib）生成FP-Growth 算法，最终生成气象服务产品的关联规则列表。程序选择Spark on YARN运行方式，按照**表6**中耗时最少的参数组合提交作业到集群上。

设定最小支持度为20，取出1 000、10 000、50 000、100 000、500 000、

1 000 000、1 500 000、2 000 000、2 500 000条数据，进行算法效率对比实验，结果如**图4**所示。

由**图4**可知，在给定最小支持度的条件下，FP-Growth算法挖掘所需的时间随记录数增加呈上升趋势，但是并非线性增加。记录数为1 000时，耗时25 s左右，而记录增加到2 500 000，扩大2 500倍时，算法耗时仅增加一倍多。这说明当数据量呈指数级别增加时，算法耗时并没有随之大幅增加，充分体现了分布式并行计算框架的优势。

另外给定2 500 000 条数据，在最小支持度为10、20、30、40、50、60的条件下进行实验，进行规则总条数与最小支持度的对比实验，结果如**图5**所示。

由**图5**可知，随着最小支持度的增加，规则挖掘的总数量在不断减少，当最小支持度达到70时，几乎没有挖掘出符合条件的强关联规则。这种变化说明，最小支持度对挖掘出的关联规则数量有很大影响，在实际应用上要小心设置最小支持度。设置过小，容易造成关联规则过多，给筛选和排查带来困难；设置过大，容易过滤掉很多关联规则，对结果造成影响。

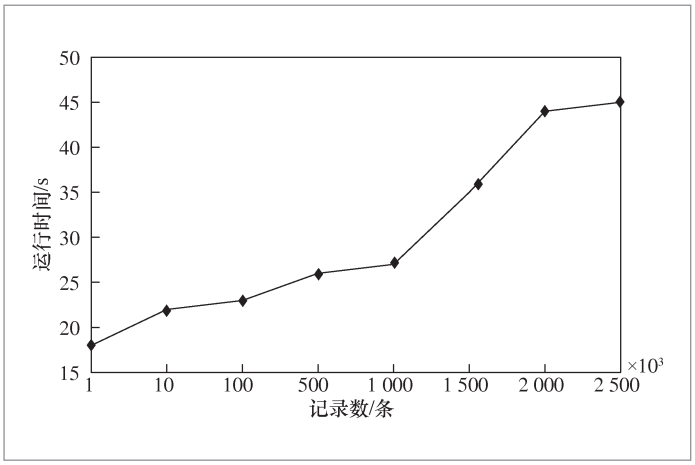


图 4 不同记录数的算法效率对比

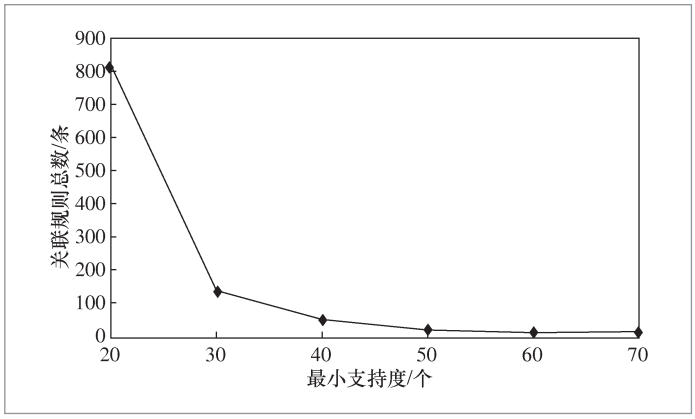


图 5 最小支持度与关联规则总数量对比

4.4 实验结果分析

一共进行5组实验，第一组选择所有样本；第二组选择2012—2016年累计降水超过50 mm的时间来收集样本，作为暴雨天气条件下的样本集；第三组选择2012—2016年超过35℃高温的时间来收集样本，作为高温天气条件下的样本集；第四组选择2012—2016年阵风超过6级的时间来收集样本，作为大风天气条件下的样本集；第五组选择2012—2016年有雾霾的时间来收集样本，作为雾霾天气条件下的样本集；最小支持度为20%，最小置信度为80%。

第一组实验得到819条规则，1个向量频繁项集106个，2个向量频繁项集412个，3个向量频繁项集187个，4个向量频繁项集220个。第二组实验得到237条规则，第三组实验得到137条规则，第四组实验得到50条规则，第五组实验得到96条规则。**表7**给出了每组实验置信度最高的5条规则。

对挖掘结果进行分析可知：第一组大多数规则反映出浏览了自动站实况和预警的用户，会浏览逐小时天气预报；第二组大多数规则反映出浏览了雷达回波、自动站累计降水量的用户，会关注逐10 min累计降水量预报、全球定位系统（global positioning system，GPS）水汽分布、微波辐射计等；第三组大多数规则反映出浏览了温度预报、自动站温度实况的用户

表 7 1~5 组实验最高置信度关联规则

实验组编号	规则	支持度	置信度
1	[10,17]=>[3]	69.78%	98.52%
	[10] =>[3]	62.14%	97.71%
	[11]=>[824]	32.03%	97.34%
	[13]=>[4]	38.65%	96.21%
	[996]=>[824]	28.07%	95.6%
2	[926]=>[3]	65.42%	96.32%
	[16]=>[38]	62.06%	95.41%
	[963, 966]=> 4	38.79%	93.01%
	[966, 17]=> 824	26.41%	88.14%
	[966, 3] => 824	23.01%	83.42%
3	[18, 3] =>[17]	46.13%	98.02%
	[1020] => [3]	37.98%	95.4%
	[10] => [824]	26.3%	92.1%
	[15, 11] => [17]	24.32%	90.12%
	[57, 11] => [17]	21.7%	89.03%
4	[28, 13] =>[17]	36.19%	93.58%
	[1 019] => [3]	33.98%	91.02%
	[1 019, 3] => [824]	27.26%	88.21%
	[1 018, 11] => [17]	24.03%	82.9%
	[672, 11] => [17]	21.07%	79.16%
5	[27, 16] =>[982]	36.19%	93.58%
	[672, 45] => [982]	33.98%	91.02%
	[1 001, 3] => [675]	27.26%	88.21%
	[673, 3] => [671]	24.03%	82.9%
	[672,11] => [17]	21.07%	79.16%

会查看模式预报结果和探空图；第四组大多数规则反映出风和探空图存在较强关系；第五组大多数规则反映出能见度、相对湿度、风、大气成分存在较强关系。

4.5 实验结果预测

4.5.1 命中率预测

强降水天气和雾霾天气容易给公众的出行和工作带来很多影响，本文选择这两种天气下的样本数据，用规则预测用户行为，计算规则的命中率作为客观指标。把每次天气过程当成一个样本集，对规则[A]=>B，计算[A]=>B出现次数与[A]=>[X]出现次数的比值。用pop(A，B)表示，计算方法如式(4)所示：

$$\mathrm{pop}(A,B)=\frac{\mathrm{Sum}(A,B)}{\mathrm{Sum}(A,X)} \qquad (4)$$

其中，Sum(A，B)表示[A]=>B规则出现次数，Sum([A]=>[X])表示所有以[A]为条件的规则总数。

用**表7**中第二组关联规则对北京2017年1—6月的5个强降水过程进行预测。另外用**表7**中第五组关联规则对北京2016年10—12月雾霾天气过程进行预测。结果如**图6**和**图7**所示。

由**图6**可以看出，2个向量频繁项的规则命中率高于3个向量频繁项，而且变化不明显，大致分布在区间[60%，95%]；而3个向量频繁项的命中率随不同降水过程波动相对大一些，规则3、规则4和规则5在降水量过程4中的命中率明显偏低。

由**图7**可以看出，规则的命中率分布在区间[20%，65%]；规则1的命中率趋势平坦，分布在区间[50%，60%]；规则4和规则5的命中率较低，分布在区间[20%，40%]。造成这种现象的原因与雾霾样本数偏少有关，仅有182个样本数据可用；此外，也与雾霾天气特点有关，例如持续时间长、天气系统稳定等导致气象服务产品本身随时间变化较小，用户总体访问量也小，从而能收集的日志数据偏少。

4.5.2 不相关规则的过滤

以第一组实验生成的819条规则为例进行排查。选出置信度超过85%的2个向量频繁项集规则，查看2个向量的相似度，如果相似度小于最小支持度，则判断这条规则有明确不相关性，标记为不可用规则。一共检查了412个规则，结合相似度过滤了76条规则。对置信度超过85%的3个向量频繁项集规则进行处理，一共检查了187个规则，过滤掉49个规则。

5 结束语

本文将关联规则和物品协同过滤方法结合起来运用于气象服务产品的推荐方法中，综合使用大数据平台多种技术实现了算法的并行运行，通过对算法效率的改进和推荐规则的有效排查，使规则挖掘所需的时间以及推荐的准确率都有所改善，初步证明了利用这种方法进行气象服务产品的推荐是可行的。在以后的工作中，可进一步细化用户兴趣的行为因素，建立用户人物画像群，对挖掘得到的推荐结果按强关系和弱关系用户进行筛选，进一步提升推荐的准确性。

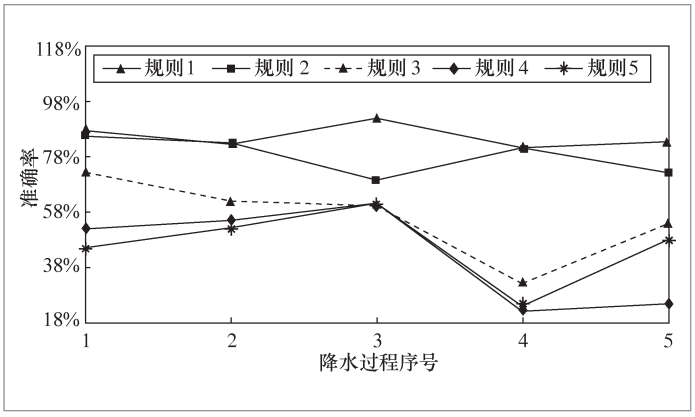


图 6 关联规则在 5 个降水过程中的命中率分布

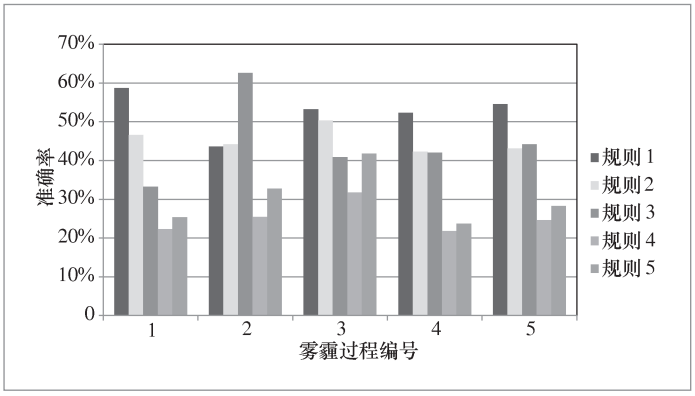


图 7 关联规则在 5 个雾霾过程中的命中率分布

参考文献:

[1] 纪淑娟，王理，梁永全，等. 基于神经网络的用户视频评分自动获取方法[J]. 计算机科学，2015，42(11)：96-100.
JI S J, WANG L, LIANG Y Q, et al. Neural-network-based method for automatic acquisition of user's video rating[J]. Computer Science, 2015, 42(11): 96-100.

[2] 王涛，覃锡忠，贾振红，等. 基于相似度和信任度的关联规则微博好友推荐[J]. 计算机应用，2016，36(8)：2262-2267.
WANG T, QIN X Z, JIA Z H, et al. Association rules recommendation of microblog friend based on similarity and trust[J]. Journal of Computer Applications, 2016, 36(8): 2262-2267.

[3] LIU J G, ZHOU T, WANG B H. Personalized recommender systems: a survey of the state-of-the-art[J]. Progress in Natural Science, 2009, 19(1): 1-15.

[4] LIN W, ALVAREZ S A, RUIZ C. Efficient adaptive-support association rule mining for recommendation systems[J]. Data Mining and Knowledge Discovery, 2002, 6(1): 83-105.

[5] 李杰，徐勇，王云峰，等. 面向个性化推荐的强关联规则挖掘[J]. 系统工程理论与实践，2009，29(8)：144-152.
LI J, XU Y, WANG Y F, et al. Strongest association rules mining for personalized recommendation[J]. System Engineering-Theory &Practice, 2009, 29(8): 144-152.

[6] 易芝，汪琳琳，王练. 基于关联规则相关性分析的Web 个性化推荐研究[J]. 重庆邮电大学学报(自然科学版)，2007，19(2)：234-237.
YI Z, WANG L L, WANG L. Research on Web personalized recommendation based on correlation analysis of association rule[J]. Journal of Chongqing University of Posts and Communications (Natural Science), 2007, 19(2): 234-237.

[7] 鲍玉斌，王大玲，于戈. 关联规则和聚类分析在个性化推荐中的应用[J]. 东北大学学报(自

然科学版), 2003, 24(12): 1149-1188.
BAO Y B, WANG D L, YU G. Application of association rules and clustering analysis to personalized recommendation[J]. Journal of Northeastern University (Natural Science), 2003, 24(12): 1149-1188.

[8] 张忠林, 田苗凤, 刘宗成. 大数据环境下关联规则并行分层挖掘算法研究[J]. 计算机科学, 2016, 43(1): 286-289.
ZHANG Z L, TIAN M F, LIU Z C. Parallel hierarchical association rule mining in big data environment[J]. Computer Science, 2016, 43(1): 286-289.

[9] 周诗慧, 殷建. Hadoop 平台下的并行Web 日志挖掘算法[J]. 计算机工程, 2013, 39(6): 43-46.
ZHOU S H, YIN J. Parallel web log mining algorithm in Hadoop platform[J]. Computer Engineering, 2013, 39(6): 43-46.

[10] 刘帅, 杨英杰, 常德显, 等. 改进的模糊关联规则及其挖掘算法[J]. 计算机工程与设计, 2015, 36(4): 942-947.
LIU S, YANG Y J, CHANG D X, et al. Improved fuzzy association rule and its mining algorithm[J]. Computer Engineer and Design, 2015, 36(4): 942-947.

[11] 程广, 王晓峰. 基于MapReduce的并行关联规则增量更新算法[J]. 计算机工程, 2016, 42(2): 21-26.
CHENG G, WANG X F. Incremental updating algorithm of parallel association rule based on MapReduce[J]. Computer Engineer, 2016, 42(2): 21-26.

[12] 黄晓雯, 严明, 桑基韬, 等. 基于关联规则挖掘的跨网络知识关联及协同应用[J]. 计算机科学, 2016, 43(7): 51-57.
HUANG X W, YAN M, SANG J T, et al. Association rules mining based cross-network knowledge association and collaborative application[J]. Computer Science, 2016, 43(7): 51-57.

[13] 陈平华, 陈传瑜, 洪英. 一种结合关联规则的协同过滤推荐算法[J]. 小型微型计算机系统, 2016, 37(2): 287-293.
CHEN P H, CHEN C Y, HONG Y. Incorporating association rules for collaborative filtering recommendation algorithm[J]. Journal of Chinese Computer Systems, 2016, 37(2): 287-293.

[14] 胡波, 黄宁, 仵伟强. 基于业务路径和频度矩阵的关联规则挖掘算法[J]. 计算机科学, 2016, 43(12): 146-154.
HU B, HUANG N, WU W Q. Algorithm for mining association rules based on application paths and frequency matrix[J]. Computer Science, 2016, 43(12): 146-154.

[15] 石岩, 邓敏, 刘启亮, 等. 海陆气候事件关联规则挖掘方法[J]. 地理信息科学, 2014, 16(2): 182-191.
SHI Y, DENG M, LIU Q L, et al. Discovering sequential association rules between single ocean climate index and land abnormal climate events[J]. Journal of Geo-information Science, 2014, 16(2): 182-191.

[16] 刘娟, 宋安军. 改进FP-growth 算法在气象预报中的应用[J]. 计算机系统应用, 2016, 25(10): 199-205.
LIU J, SONG A J. Application of an improve FP-growth algorithm in meteorological forecast[J]. Application of Computer System, 2016, 25(10): 199-205.

[17] BINZENHÖFER A, TUTSCHKU K, GRABEN B A D, et al. A P2P-based framework for distributed network management[M]. Berlin: Springer Berlin Heidelberg, 2005.

[18] CHEVUL S, BINZENHFER A, SCHMID M, et al. A self-organizing concept for distributed end-to-end quality monitoring[R]. Germany: University of Wurzburg Institute, 2006.

[19] 董长青, 任女尔, 张庆余, 等. 基于HBase+ElasticSearch的海量交通数据实时存取方案设计[J]. 大数据, 2017(1): 80-89.
DONG C Q, REN N E, ZHANG Q Y, et al. Design scheme of massive traffic data real-time access based on HBase and Elastic Search[J]. Big Data Research, 2017(1): 80-89.

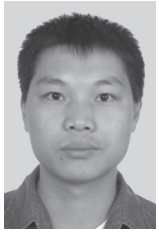
[20] 李大中, 刘剑, 邓景文. 大数据能力开放平台创新和发展[J]. 大数据, 2017(1): 72-80.
LI D Z, LIU J, DENG J W. Innovation and development of big data ability open platform[J].

Big Data Research, 2017(1): 72-80.
[21] 石勇, 孟凡. 信用评分基本理论及其应用[J]. 大数据, 2017(1): 19-27.
SHI Y, MENG F. Credit scoring: basic theory and applications[J]. Big Data Research, 2017(1): 19-27.
[22] 刘岩, 王华, 秦叶阳, 等. 智慧城市多源异构大数据处理框架[J]. 大数据, 2017(1): 51-60.
LIU Y, WANG H, QIN Y Y, et al. Multisource heterogeneous big data processing architecture in smart city[J]. Big Data Research, 2017(1): 51-60.

作者简介

赵文芳(1980-),女,北京市气象信息中心高级工程师、科长,主要研究方向为大数据技术在气象信息领域的应用、机器学习。

刘亚楠(1981-),男,北京市气象信息中心工程师,主要研究方向为计算机网络安全。



余东昌(1978-),男,北京市气象信息中心高级工程师、科长,主要研究方向为软件架构、高性能计算、并行计算。

收稿日期: 2017-09-21
基金项目:中国气象局公益性行业科研专项基金资助项目(No.201206031)
Foundation Item: The Public Welfare Industry Research Funds of China Meteorological Bureau (No.201206031)