Hadoop 平台下海量数据排行榜过滤算法①

黄德才，陈 欢

(浙江工业大学 计算机科学与技术学院，杭州 310023)

摘 要：排行榜作为现代社会很受关注的一项事物深入大家的生活，但对于海量数据的排行，即使在分布式环境下，依然需要耗费大量硬件资源和很长的时间，有时甚至无法产出榜单。首先对贝叶斯方法进行了改进，提出了一种基于 hadoop 分布式环境下的行榜海量数据过滤算法，该方法利用熵值理论对缺损数据进行填补得到完整数据；再利用改进的贝叶斯方法计算某商品当日销量进榜单的概率，并对概率低于概率阈值的商品数据进行过滤使其不参与排行计算，从而在确保排行榜精确度的前提下大大缩短榜单的产出时间。对淘宝网中 400 万条销售记录数据进行实验仿真，结果验证了上述方法的有效性和优越性能。

关键词：排行榜；Hadoop；海量数据；熵；贝叶斯

Rankings Filtering Algorithm of Massive Data Based on Hadoop and its Application

HUANG De-Cai, CHEN Huan

(College of Computer Science & Technology, Zhejiang University of Technology, Hangzhou 310023, China)

Abstract: Rankings as a popular production in modern society has gone deeply into everyone's life. For the rankings on massive data, it costs large consumption of hardware resources and time though running under the distributed environment, even may not be produced sometimes. This paper improves the Bayesian algorithm and proposes a rankings filtering algorithm of massive data based on hadoop. We first fill the missing data by entropy theory for getting the complete data. Then, we compute the probability in the sales volume on the very day by the improved Bayesian algorithm. If the probability is smaller than threshold, the goods would be filtered not to attend the ranking computation. Simulation on four million sales from Taobao shows the effectiveness and excellent property of the proposed algorithm.

Key words: rankings; Hadoop; massive data; entropy; bayes

1 引言

排行榜作为现代社会很受关注的一项事物深入大家的生活，比如学生会关心自己在班级中的成绩排名，工人会关注绩效排名等等，在电子商务领域，该问题尤其重要。卖家需要知道自己店铺内热销的榜单，各类目下的榜单，整个行业中的榜单等等来调整自己的经营策略；买家同样需要知道店铺的信誉排行等等来选择购物。它是决策的重要依据之一，也正因此，排行榜成为了淘宝数据平台最重要的产品之一。

但是随着电子商务平台的迅猛发展，待排序数据的基数呈指数式增加，这对排行榜的顺利产出提出了较大的考验，特别是对一些长周期或细粒度的排行，如一年内的商品日销售排行，由于数据量为 TB 级别，甚至可能达到了 PB 级，因此这几乎是一个不可能完成的任务。原有的排行榜生产对一些不必要的数据进行了排序，大大浪费了系统中有限的硬件资源。这就促使我们对数据进行有效的过滤，在保证准确度的条件下大大缩短榜单产出时间。

现有的排行榜数据过滤方法中，通常采用设定简单的阈值，并将阈值以下的数据进行过滤的方法，例如，对于销量排行榜，将销量为 1 的数据过滤，但事实证明，一些长周期的榜单依然无法产出；另一方面

① 基金项目:浙江省重大科技计划(2009C11024)
收稿时间:2011-07-06;收到修改稿时间:2011-08-24

将销量小于或等于 1 的全部删除，可能出现潜在的误差，因为有些商品本身的总体销量就很低，这就使得即使销量为 1 的商品仍可能进榜单。

文献[1]描述了反垃圾邮件系统中的贝叶斯模型，它通过邮件主题以及正文的关键词建立 TOKEN 串，并统计其频次；每个邮件对应的哈希表，并记 Hashtable_good 为非垃圾邮件集，Hashtable_bad 为垃圾邮件给，表中存储 TOKEN 串到频次的映射关系，从而计算每个哈希表中某一 TOKEN 出现的概率，结合 Hashtable_good 与 Hashtable_bad，计算新邮件为垃圾邮件的概率。但是这种方法依赖于中文分词，而在文本挖掘中分词本身就是一个有较大难度的方法；另外，现有的邮件系统中的过滤方法，并未考虑 TOKEN 串之间的联系，这也会使得增加冗余数据的同时降低邮件过滤的准确性。

阮彤等学者[2]认识到了一般的向量模型、布尔模型、概率模型、最近邻法等均不能描述事件之间的复杂关系，他们企图用贝叶斯网络这种图模型充分考虑事件之间相关性。 它描述能力强，直观、易理解与易交互，并且具有一定的语料无关性与知识可传递性。但是，它与贝叶斯模型一样，在处理文本信息时对文本分词的依赖性太大；同时，在以减小生产时间为目的的排行榜过滤中，显然这种复杂的模型不太适用；另外，已发展的贝叶斯网络学习方法均具有结构不变的假设[3-5]，而在现实中，随着对实际问题认识的深入，贝叶斯网络结构往往需要调整，以使模型具有良好的适应性。

为了解决传统过滤方法由于基于区域划分、采用组播技术而造成的效率低、稳定性差的问题，文献[6]提出了一种适用于大规模分布式虚拟环境的新的数据过滤方法——基于模糊关联空间的数据过滤方法，它将数据过滤问题转化为在模糊关联空间中求取关联实体集的问题。虽然模拟计算和理论分析均说明了该方法弥补了传统的基于路由技术的数据过滤方法的缺陷，但是该方法受初始关联实体的影响较大，它的选取将直接影响最后的过滤结果；另外，只考虑两两实体间的关联性使得数据丢失了部分全局信息。

综上所述，虽然现有的信息过滤模型已经在工程应用中发挥了重要作用，但是那几乎都是基于文本的，而对于排行榜数据过滤的模型研究少之又少。ACM Computing Surveys 的一篇文章[7]这样评价：“贝叶斯方法是最有前景的一种方法，因为其将领域知识引入信息检索领域中”。但是考虑到一般的贝叶斯模型未考虑全局的数据相关性，而贝叶斯网络模型又过于复杂，将严重影响算法的运行时间，本文以淘宝海量数据排行为背景，基于 hadoop 分布式文件系统的云计算平台为依托，在不影响精确性的前提下缩短榜单产出时间为目的，改进了贝叶斯概率模型进行数据过滤。实验结果得到很好的证实了该方法具有很强的可行性。

2 HDFS与Map/Reduce

云计算是分布式计算(Distributed Computing)、并行计算 (Parallel Computing) 、网格计算 (Grid Computing)的发展与延伸，作为当今计算机届最热门的研究方向之一，它不仅引起了学者们的广泛关注，一些大型计算机公司也致力与该研究。其基本原理是利用分布式的并行服务器集群为互联网用户提供各种高性能服务，使得用户可以将资源切换到需要的应用上，从而有效地提高了对软硬件资源的利用效率。

受 Google Lab 开发的 MapReduce[8]和 Google File System[9]的启发，Apache 公司于 2005 年正式引入 Hadoop 项目。作为一个成功的分布式系统基本架构，它能够使用户在不了解分布式底层细节的情况下，开发分布式程序。

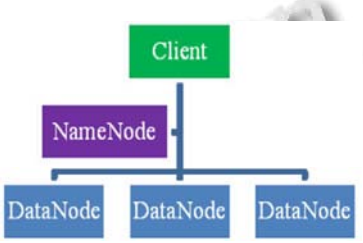

图 1 Map/Reduce 编程框架

Hadoop 最主要的两个部分是 HDFS 和 Map/Reduce 编程框架(图 1 所示)。HDFS 是一个高度容错的分布式文件系统，适合部署在廉价的机器上，并能提供高吞吐量的数据访问。一个 Hadoop 集群由一个 NameNode 的主节点和多个 DataNode 的子节点组成。NameNode 负责管理文件系统名称空间和控制外部客户机的访问，它决定了是否将文件映射到 DataNode 上的复制块上以及存储位置。DataNode 则响应来自 HDFS 客户机的读写请求，它们还响应创建、删除和复制来自 NameNode 的块命令。NameNode 依

赖来自 DataNode 的定期消息反馈， NameNode 可以根据这些消息验证块映射和其他文件系统元数据。如果 DataNode 不能发送消息反馈，那么 NameNode 将采取修复措施，重新复制该节点上错误数据。

图 2 描述了一个基本的 Map/Reduce 过程。

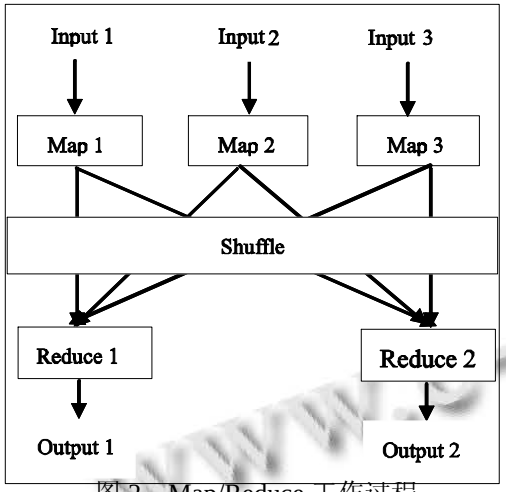

图 2 Map/Reduce 工作过程

(1) Input：根据用户应用程序提供的 Map 和 Reduce 函数，提取一些运行参数（如输入输出路径等）。同时，该阶段还将输入目录下的大数据文件，划分为若干独立的数据块。

(2) Map：Map/Reduce 框架把用户作业的输入看作是一组<key,value>键值对，按照 Map 函数的规则处理每一个<key,value>键值对，并生成一批新的中间<key,value>键值对。

(3) Shume＆Sort：为了保证 Reduce 的输入是有序的，在 Shufle 阶段，框架通过 Http 为每个 Reduce 获得所有 Map 输出中与之相关的<key,value>键值对；而在 Sort 阶段，框架将按照 key 的值对 Reduce 的输入进行分组，通常 Shuffle 和 Sort 两个阶段是同时进行的，Reduce 的输入也是一边被取出，一边被合并的。

(4) Reduce：对每一个唯一 key，执行用户自定义的 Reduce 函数，并按照其中的规则输出新的<key,value>键值对。

(5) Output：将 Reduce 输出的结果写入到输出目录的文件中。

3 基于贝叶斯理论的排行榜数据过滤算法

3.1 基于熵论的缺损数据补值

在电子商务领域中，商品会因为一些原因导致其在销售过程中中断销售行为，比如某商品突然断货了，过了几天进货后商品继续销售，该商品在断货的时间内的销售记录是空的，但是如果我们将空值理解为销售量为 0，那显然是不可取的，因此我们需要通过某种方法计算出如果该商品在断货期内仍有销售，那么销量应该是多少，从而对缺损数据进行有效的补值。

在信息论中，信息熵是一个信源发出消息所含信息量的度量，当某一信源发出的消息越确定时，该信源的信息熵就越小[10]。它是系统无序程度或混乱程度的度量，表示了系统的平均不确定度。而熵值法是一种通过属性数值所提供信息的大小来确定权重系数的一种方法。对于确定的属性 j，各数据第 j 个属性之间的差异越大，则说明该项指标的相对作用就越大，即其信息量就越大，熵值越小。它具有客观性强，评价过程透明性和可再现性好的特点。在以降低时间复杂度为目的的海量数据过滤方法中，基于信息熵的过滤无疑是一种较简介且有效的补值方法。

假设原始交易量数据集

$T=(T_1, T_2, \ldots\ldots, T_m)'$

$$=\begin{bmatrix} t_{11} & t_{12} & \cdots & t_{1j} & \cdots & t_{1n} \\ t_{21} & t_{22} & \cdots & t_{2j} & \cdots & t_{2n} \\ \vdots & \vdots & & \vdots & & \vdots \\ t_{i1} & t_{i2} & \cdots & t_{ij} & \cdots & t_{in} \\ \vdots & \vdots & & \vdots & & \vdots \\ t_{m1} & t_{m2} & \cdots & t_{mj} & \cdots & t_{mn} \end{bmatrix},$$

$Ti=(t_{i1}, t_{i2}, \ldots\ldots, t_{in})'$表示第 i 个商品的历史交易数据，$Pj=(t_{1j}, t_{2j}, \ldots\ldots, t_{nj})$表示第 j 个交易日各商品的交易数据，tij 表示第 i 个商品第 j 个交易日的交易值。

若商品 T_a 的第 b 个交易日的交易值 t_{ab} 缺损，运用熵值理论确定缺损值

$t_{ab}=w_1t_{a1}+w_2t_{a2}+\ldots+w_{(b-1)}ta_{(b-1)}+w_{(b+1)}ta_{(b+1)}+\ldots+w_nt_{an}$

从而达到填补缺损值 tab 的效果。

步骤如下：

1. 计算第 f 个交易日的熵值

$$I_f = -k \cdot p_f \cdot \ln(p_f)$$

$$p_f = \frac{d_f}{\sum_{i=1}^{n} d_i}$$

其中，k=1/ln(n)，

$$d_f = (\sum_{i=1}^{p}(t_{if} - t_{ib})^2)^{\frac{1}{2}}$$

，表示第 f 个交易日数据与所缺损的第 b 个交易日数据的距离，体现了两者的相关性，由

非缺损数据计算得到。

2. 计算第 f 个交易日的差异系数

rf=1-If，　　f=1,2,...,n

差异系数是反应数据作用大小的量，其值越大，数据体的作用越大，反之亦然。

3. 计算第 f 个交易日的权重系数

$$w_f=\frac{r_f}{\sum_{i=1}^{m} r_i}，\ f=1,2,\ldots,m$$

考虑到当某一商品的缺损量较大时，通过较少的已知数据对大量未知数据进行填补，精确性会较低，因此只对缺损值个数小于 n/3 的商品进行缺损值填补。

3.2 改进的 Bayes 模型数据过滤

排行榜数据过滤的目的是将那些不可能进榜的数据过滤，以节省有限的云梯资源从而达到合理化生产的目的，其核心是根据商品历史销量数据判断该商品在当日销量下能否进榜单，若不能则过滤；反之，则不过滤。也就是说，讨论的是该商品当日在该销量下能否进榜单的概率，若概率小于某一给定的值，则过滤；反之，则不过滤。

3.2.1 基于朴素 Bayes 模型的排行榜数据过滤

经过 2.1 节的处理，我们得到了一个完备的数据集 $X=(X_1, X_2, \ldots\ldots, X_m)'=$

$$\begin{bmatrix} t_{11} & t_{12} & \cdots & t_{1j} & \cdots & t_{1n} & x_{11} \\ t_{21} & t_{22} & \cdots & t_{2j} & \cdots & t_{2n} & x_{21} \\ \vdots & \vdots & & \vdots & & \vdots & \vdots \\ t_{i1} & t_{i2} & \cdots & t_{ij} & \cdots & t_{in} & x_{i1} \\ \vdots & \vdots & & \vdots & & \vdots & \vdots \\ t_{m1} & t_{m2} & \cdots & t_{mj} & \cdots & t_{mn} & x_{m1} \end{bmatrix}$$

$x=(x_{11},x_{21},\ldots,x_{m1})'$，为排行榜当天的交易量。

令集合 U 为进榜商品的数据集，V 为未进榜商品的数据集；设事件 A 表示商品进入榜单，事件 B 表示商品的当日销量

根据朴素贝叶斯理论

$$P(A\mid B)=\frac{P(A,B)}{P(B)}=\frac{P(B\mid A)\cdot P(A)}{P(B)}$$

对于任意商品 Xi，当天交易量为 xi1，计算它的进榜概率如下：

1. $P(B\mid A)=\dfrac{\text{在进榜的日期中，日销量小于}x_{i1}\text{的次数}}{\text{该商品总的进榜次数}}=\dfrac{\underset{j=1,2,\ldots,n}{\text{count}}(t_{ij}<x_{i1})}{\underset{j=1,2,\ldots,n}{\text{count}}(t_{ij})}$，$(t_{ij}\in U)$

$\underset{j=1,2,\ldots,n}{\text{count}}(t_{ij}<x_{i1})$ 表示 j 从 1 到 n，tij<xi1 的个数。

2. $P(A)=\dfrac{\text{该商品进榜总次数}}{\text{所选周期长度n}}=\dfrac{\underset{j=1,2,\ldots,n}{\text{count}}(t_{ij})}{n}$，　$(t_{ij}\in U)$

3. $P(B)=\dfrac{\text{日销量小于}x_{i1}\text{的次数}}{\text{所选周期长度n}}=\dfrac{\underset{j=1,2,\ldots,n}{\text{count}}(t_{ij}<x_{i1})}{n}$

因此，$P(A|B=x_{i1})=\dfrac{\dfrac{\underset{j=1,2,\ldots,n}{\text{count}}(t_{ij}<x_{i1})}{\underset{j=1,2,\ldots,n}{\text{count}}(t_{ij})}\cdot\dfrac{\underset{j=1,2,\ldots,n}{\text{count}}(t_{ij})}{n}}{\dfrac{\underset{j=1,2,\ldots,n}{\text{count}}(t_{ij}<x_{i1})}{n}}=\dfrac{\underset{j=1,2,\ldots,n}{\text{count}}(t_{ij}<x_{i1})}{\underset{j=1,2,\ldots,n}{\text{count}}(t_{ij}<x_{i1})}$，

其中，分子的 $t_{ij}\in U$

3.2.2 基于改进的朴素贝叶斯模型排行榜数据过滤

朴素贝叶斯模型充分的考虑了商品不同交易日间的关系，根据单一商品的历史交易数据得出了较为准确的过滤结果。但是，不同商品的交易量之间也存在相互影响，比如某交易日的整体交易量较高使得该商品该日即使高交易量却仍不能将进榜单。朴素贝叶斯模型忽略了这种数据间的纵向相关性信息。

因此，我们作如下改进。

对于任意商品 $X_i=(t_{i1},t_{i2},\ldots,t_{in},x_{i1})$，随机选取 2γ+1 个商品进行考虑。

$C=(X_{i-4}, X_{i-3}, \ldots\ldots, X_{i+4})'$

$$\begin{bmatrix} t_{i-\gamma,1} & t_{i-\gamma,2} & \cdots & t_{i-\gamma,n} & x_{i-\gamma,1} \\ t_{i-\gamma+1,1} & t_{i-\gamma+1,2} & \cdots & t_{i-\gamma+1,n} & x_{i-\gamma+1,1} \\ \vdots & \vdots & & \vdots & \vdots \\ t_{i,1} & t_{i,2} & \cdots & t_{i,n} & x_{i1} \\ \vdots & \vdots & & \vdots & \vdots \\ t_{i+\gamma,1} & t_{i+\gamma,2} & \cdots & t_{i+\gamma,n} & x_{i+\gamma,1} \end{bmatrix}$$

令 r_{ij} 为第 i 个商品在第 j 个交易日的销量指数，它表示了该商品在所选 γ 个商品中的销量地位，$r_{ij}=rank(t_{ij})$，即 tij 在第 j 列的 γ 个销量值中的排名。ri 为商品 i 在所选 γ 个商品中的销量排名。

r_{ij}/r_i 的引进不仅使得贝叶斯模型体现了不同交易日总体销量之间的影响，而且也使得商品销量陡增或骤减问题得到了解决。

$$P(B|A)=\frac{\text{该商品进榜的有效日销量差值}}{\text{该商品进榜的总日销量差值}}$$

$$=\frac{\sum_{j=1,x_{i1}>t_{ij}}^{n}(x_{i1}-t_{ij})\cdot\frac{r_{ij}}{r_i}+\sum_{j=1,x_{i1}<t_{ij}}^{n}\frac{1}{(t_{ij}-x_{i1})}\cdot\frac{r_{ij}}{r_i}}{\sum_{j=1,x_{i1}>t_{ij}}^{n}(x_{i1}-t_{ij})+\sum_{j=1,x_{i1}<t_{ij}}^{n}\frac{1}{(t_{ij}-x_{i1})}},$$

其中，$(t_{ij}\in U)$

$$P(A)=\frac{\text{该商品进榜的总日销量差值}}{\text{该商品总销量差值}}$$

$$=\frac{\sum_{j=1,x_{i1}>t_{ij}}^{n}(x_{i1}-t_{ij})+\sum_{j=1,x_{i1}<t_{ij}}^{n}\frac{1}{(t_{ij}-x_{i1})}}{\sum_{j=1,x_{i1}>t_{ij}}^{n}(x_{i1}-t_{ij})+\sum_{j=1,x_{i1}<t_{ij}}^{n}\frac{1}{(t_{ij}-x_{i1})}+\sum_{j=1,x_{i1}>t_{ij}}^{n}(x_{i1}-t_{ij}')+\sum_{j=1,x_{i1}<t_{ij}}^{n}\frac{1}{(t_{ij}'-x_{i1})}},$$

其中，$(t_{ij}\in U, t_{ij}'\in U\cup V)$

$$P(B)=\frac{\text{该商品总的有效日销量 差值}}{\text{该商品总销量差值}}$$

$$=\frac{\sum_{j=1,x_{i1}>t_{ij}}^{n}(x_{i1}-t_{ij})\cdot\frac{r_{ij}}{r_i}+\sum_{j=1,x_{i1}<t_{ij}}^{n}\frac{1}{(t_{ij}-x_{i1})}\cdot\frac{r_{ij}}{r_i}}{\sum_{j=1,x_{i1}>t_{ij}}^{n}(x_{i1}-t_{ij})+\sum_{j=1,x_{i1}<t_{ij}}^{n}\frac{1}{(t_{ij}-x_{i1})}+\sum_{j=1,x_{i1}>t_{ij}}^{n}(x_{i1}-t_{ij}')+\sum_{j=1,x_{i1}<t_{ij}}^{n}\frac{1}{(t_{ij}'-x_{i1})}},$$

其中，$(t_{ij}\in U, t_{ij}'\in U\cup V)$

所以，当日销量 x_{i1} 条件下的进榜概率

$$P(A|B=x_{i1})$$

$$=\frac{\sum_{j=1,x_{i1}>t_{ij}}^{n}(x_{i1}-t_{ij})\cdot\frac{r_{ij}}{r_i}+\sum_{j=1,x_{i1}<t_{ij}}^{n}\frac{1}{(t_{ij}-x_{i1})}\cdot\frac{r_{ij}}{r_i}}{\sum_{j=1,x_{i1}>t_{ij}}^{n}(x_{i1}-t_{ij})\cdot\frac{r_{ij}}{r_i}+\sum_{j=1,x_{i1}<t_{ij}}^{n}\frac{1}{(t_{ij}-x_{i1})}\cdot\frac{r_{ij}}{r_i}+\sum_{j=1,x_{i1}>t_{ij}}^{n}(x_{i1}-t_{ij}')\cdot\frac{r_{ij}}{r_i}+\sum_{j=1,x_{i1}<t_{ij}}^{n}\frac{1}{(t_{ij}'-x_{i1})}\cdot\frac{r_{ij}}{r_i}}$$

，其中$(t_{ij}\in U, t_{ij}'\in U\cup V)$

3.2.3 改进的朴素贝叶斯过滤算法流程图

步骤 S101，获取固定时间段内的历史数据。

步骤 S102 与 S103，判断原始数据的缺损程度，根据原始数据的缺损程度确定原始数据是否需要进行数据过滤以及是否需要进行缺损数据补值，如果不需要数据过滤，结束；如果需要数据补值，执行步骤 S104，否则，直接将原始数据作为历史数据。

步骤 S104，利用熵值理论根据原始数据模拟缺损数据，将模拟得到缺损值补充进原始数据，得到历史数据。

步骤 S105，根据历史数据和排行日数据，计算商品排行日数据的进榜单概率。

步骤 S106，对进榜单概率和概率阈值进行比较，根据比较结果判断是否过滤商品的排行日数据。具体的，如果进榜单概率大于概率阈值，说明该商品当日可能进榜，不过滤商品的排行日数据，反之，则过滤商品的排行日数据。

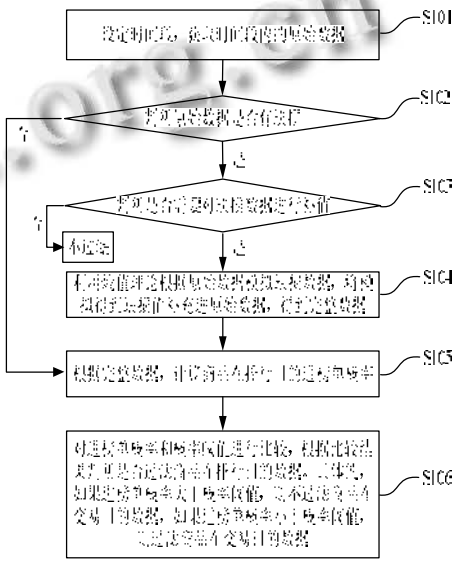

图 3 算法流程图

4 仿真实验

本文从 2011.2.1-2011.3.5 日的销售数据中随机选取部分数据进行实验，将 2.1 日—3.4 日的数据作为训练数据，3.5 日的数据为测试数据，对比过滤后的 3.5 日数据排行与真实的 3.5 日数据排行来验证算法的优劣。其中，每日的销售数据表约有 600 万件商品，44 列，不计重复的约有 400 万件商品，文件总大小约为 0.6TB。

结果如下：

表 1 仿真结果表

	过滤数据比例	排行榜正确率
商品的日排行	10%	93.8%
	20%	85.6%
	30%	75.3%

（下转第124页）

求计划详细清单（其中计划库存量是在不补充库存的情况下的库存量，是一种假设结果）如表 1 所示。

表 1 2011 年 6 月各旬 18#工字钢的物料需求计划详细清单

时间段	6 月上旬	6 月中旬	6 月下旬
总需求量	12.410	15.126	9.586
计划到货量	13.022	0.000	0.000
计划库存量	10.126	-5.000	-14.586
补充后库存量	10.738	10.365	10.365
净需要量	0.000	14.753	9.586
计划接受订货	0.000	14.753	9.586
计划发出订货	14.753	9.586	0.000

5 结语

目前，我国煤业集团通过网络采购来实现生产物料库存控制这一方面的研究很少。因篇幅所限，本文只从理论和算法的角度进行了“基于网络采购的煤业集团库存控制系统核心算法设计”的研究，没有涉及系统的以实现，但本文的研究对煤业集团库存控制具有实际的指导作用，它的应用可以为煤业集团带来很大的经济效益，对促进电子商务在我国的发展也具有积极的社会意义。

参考文献

1 黄梯云.管理信息系统(第 4 版).北京:高等教育出版社,2009.40.

2 郭洪禹,王倩.MRPⅡ系统中 MRP 分析与实践.燕山大学学报,2007,31(4):368–372.

3 唐世文.MRP 材料采购运算失真分析与修正.中国管理信息化,2009,12(1):65–67.

4 陈慧萍,李绍峰,王建东.基于多层 C/S 结构的饲料行业 MRP 系统.计算机工程与设计,2007,28(1):173–175.

5 苗文明,陈泳,陈关龙,等.面向延迟制造的 MRP 动态调整方法研究.自动化学报,2008,34(8):950–955.

6 石瑞红,赵一飞.MRP 库存管理在冷藏箱调运中的应用.物流工程与管理,2011,33(199):47–49,55.

7 付永贵.基于网络招标采购的煤业集团库存控制研究.太原:山西财经大学,2006.1–3,23–24,45–51.

（上接第 115 页）

5 结语

排行榜设有商品粒度下的、性别粒度下的、店铺粒度下的等等，在这么多榜单中共有约 2/3 的数据将出现，因此要将多余的 1/3 无用数据准确的筛选出是非常困难的。另外，海量的数据规律也是比较难发掘的，因为海量的基数使得噪声的基数也被扩大，从而无法发现某些噪声数据。由于上述操作基本可在离线 hadoop 环境下运行，因此大大缩短了实际工程中在线环境中产生榜单的时间，达到了减小在线生产时间的目的，这对于海量数据的排行问题具有一定的贡献。

参考文献

1 戴劲松,白英彩.基于贝叶斯理论的垃圾邮件过滤技术.计算机应用与软件,2006,23(1):108–124.

2 阮彤,冯东雷,李京.基于贝叶斯网络的信息过滤模型研究.计算机研究与发展,2002,39(12):1564–1571.

3 Cooper GF, Herskovits E. A Bayesian method for the induction of probabilistic networks from data. Machine Learning, 1992,9(4):309–347.

4 Lain W, Bacchus F. Learning Bayesian belief networks: an approach based on the MDL principle. Computational Intelligence, 1994,10(4):269–293.

5 王双成,冷翠平.贝叶斯网络适应性学习.小型微型计算机系统,2009,30(4):706–709.

6 钱雪平,赵沁平.基于模糊关联空间的数据过滤方法.计算机学报,2002,25(7):723–729.

7 Crestani F, Lalmas M, et al. Is this document relevant… probably: A survey of probabilistic models in information retrieval. ACM Computing Surveys, 1998,30(4):529–551.

8 Dean J, Ghemawat S. MapReduce: simplified data processing on large clusters. Communications of the ACM, 2008,51(1): 107–113.

9 Ghemawat S, Gobioff H, Leung ST. The Google file system. SOSP’03. 19th ACM Symposium on Operating Systems Principles, 2003,37(5):29–43.

10 王洪春,彭宏.一种基于熵的聚类算法.计算机科学,2007, 34(11):178–200.