基于张量分解的药品个性化推荐

王 龙 王嘉伦 程转丽 李 然 张 引
(华中科技大学计算机学院 武汉 430074)

摘 要 在当前网购越来越流行的趋势下,网上买药也给很多病人带来了极大的便利。但是普通人在网上购买药品时普遍存在盲目购药、无法获得买药指导的问题,针对这一问题,提出首先根据药品的功能描述信息进行聚类,设计了基于用户相似度的协同过滤药品推荐算法;然后针对该算法的冷启动以及数据稀疏性等问题提出了基于张量分解的个性化药品推荐算法来对获取到的药品功能描述信息进行特征分析,构建标签特征向量,利用特征向量与用户对药品的评分值构建三阶张量,再利用张量分解方法对该三阶张量进行分解;最后得到推荐评估值,再利用该推荐评估值进行Top-N药品推荐。通过对真实的药品销售网站数据进行抓取并分析,构建了张量模型,并进行数据建模,与协同过滤的推荐结果相比,其得到了较好的推荐效果。

关键词 药品个性化,协同推荐,K-means聚类,张量分解
中图法分类号 TP391 **文献标识码** A **DOI** 10.11896/j.issn.1002-137X.2015.5.045

Personalized Medicine Recommendation Based on Tensor Decomposition

WANG Long WANG Jia-lun CHENG Zhuan-li LI Ran ZHANG Yin
(School of Science&Technology, Huazhong University of Science and Technology, Wuhan 430074, China)

Abstract As the online shopping is becoming more and more popular, buying medicine online has brought great convenience for many patients. But when ordinary people buy drugs online, they always purchase medicine blindly. There is a big problem that they do not have access to the medicine guidance. In order to solve this problem, firstly, we clustered the drug into several groups according to the functional description information of the drug, and proposed the personalized medicine recommendation based on user collaborative filtering. Then considering the shortcomings of the collaborative filtering algorithm, we used the tensor decomposition methods to model the relationship of the user, symptom and medicine, and recommended the top-N related medicines to the users according to their symptoms. We crawled the real data from the internet and compared the results with collaborative filtering method. The results show good performance.

Keywords Personalized medicine, Collaborative filtering, K-means clustering, Tensor decomposition

随着信息技术的不断发展,越来越多的研究人员开始尝试将最先进的科技运用到医疗健康领域[1,2],医药知识的普及和网上购物的兴起,使现在生病之后自己在网上买药的人也越来越多,国外网上药店[3]及国内金象大药房、九洲网上药店等网上药店受到越来越多的关注。网上买药不受时间、空间、地域的限制,对时间忙或不便于长时间运动的人群来说尤为方便;此外,在网上可以获得大量的药品信息、价格信息以及用户评论等信息,还可以买到当地没有的药品;而且由于网上药店省去租店面、雇员及储存保管等一系列费用,总的来说其价格较一般药店的同类药品更便宜。网上买药在带来极大便利的同时,也存在很多问题,比如在不正规的网站容易买到假药、一些网站通过网上非法发布药品信息、夸大宣传导致患者身体受到伤害、买药出现问题不易解决等,还有最重要的一点是购药的时候无法获得买药指导,在这种情况下,病人在网上买药都具有一定的盲目性,容易出现用错药物、重复用药、忽视药物间相互作用等错误,不能在第一时间买到最适合自己病情的药品。虽然已经有人开始将个性化推荐技术用在健康医疗方面[4],但是目前网上药店个性化推荐的研究相对较少,普遍的做法是按照药品销量进行排序显示,针对这一主要问题,本文试图引入个性化推荐系统的一些算法,分析了利用基于协同过滤算法进行推荐的优劣,然后提出了基于张量分解的个性化药品推荐算法,指导患者在网上找到自己所需的药品。

1 相关研究

随着网络信息的不断增长,信息过载问题也越来越严重,

到稿日期:2014-05-20 返修日期:2014-07-25 本文受基金项目:基于云计算的普适人体传感网关键技术研究(F020809)资助。
王 龙(1989-),男,硕士生,主要研究方向为数据挖掘、推荐系统、健康医疗大数据,E-mail:longwang.epic@gmail.com;王嘉伦(1989-),男,硕士生,主要研究方向为数据挖掘、机器学习、大数据分析、普适计算、移动云计算,E-mail:jialun.cs@gmail.com;程转丽(1990-),女,硕士生,主要研究方向为物联网、云计算以及移动云计算等,E-mail:zhuanlicheng.cs@gmail.com;李 然(1990-),男,硕究生,主要研究方向为大数据、云计算,E-mail:zhuanlicheng.cs@gmail.com;张 引(1986-),男,博士后,讲师,主要研究方向为系统分析与集成、半结构化数据管理、大数据、移动计算、云计算等,E-mail:dr.yinzhang@gmail.com。

在这种背景下，出现了个性化推荐系统[5-7]，以帮助人们最快找到自己感兴趣的内容。人们定义电子商务推荐系统为："利用电子商务网站向客户提供商品信息和建议，帮助用户决定该购买什么商品，模拟销售人员辅助客户购买商品的过程。"而个性化推荐技术则是电子商务推荐系统中的核心技术，决定了推荐系统的效果与性能。

1.1 主要的推荐技术

协同过滤推荐(collaborative filtering recommendation)是目前研究最多的个性化推荐技术，它使用的是基于用户或者项目的相似度来进行推荐，Barragáns-Martínez A B等人[8]利用协同过滤推荐与奇异值分解混合算法对电视节目进行推荐，达到了前所未有的效果；Cai X[9]等人利用协同过滤算法对社交网络中的用户进行个性化推荐，对某一人的交友倾向进行了准确的预测。

基于内容的推荐(content based recommendation)基于用户评价对象的特征学习用户的兴趣，根据用户的资料与待预测项目的匹配程度进行推荐，比如新闻推荐系统。Pasquale Lops等人[10]对基于内容的推荐系统进行了概览介绍。

基于知识的推荐(knowledge based recommendation)通常需要对某一特定领域的知识和领域内对象相互关系信息支持，Carrer-Neto W等人[11]就利用了在电影方面的领域专家知识对用户可能感兴趣的电影节目进行个性化推荐，达到了很好的效果。

基于关联规则的推荐(association rule based recommendation)通常出现在事务数据分析中，通过把推荐对象作为规则体，发掘推荐对象中的内在联系，进行个性化推荐。关联规则的发现是一件非常耗时的事，但是可以通过离线数据分析来解决。Duan L等人[4]通过利用关联数据挖掘算法对患者的信息进行分析，并对患者进行最佳治疗方案的推荐，这是推荐算法在医疗领域的新尝试，具有很大的潜力。

2 技术简介

2.1 K-means聚类算法

聚类是数据挖掘[12]中常用的数据分析方法，K-means[13]是一种常用的聚类方法，它的目标是将数据点划分为 k 个cluster，找到这每个 cluster 的中心，并且最小化函数

$$\arg\min_{s}\sum_{i=1}^{k}\sum_{x_j\in s_i}\|x_j-u_i\|^2 \tag{1}$$

其中，u_i 是第 i 个 cluster 的中心。上式要求每个数据点要与它们所属 cluster 的中心尽量接近。

为了得到每个 cluster 的中心，K-means 迭代地进行两步操作。首先随机地给出 k 个中心的位置，然后把每个数据点归类到离它最近的中心，这样就构造了 k 个 cluster。但是，这 k 个中心的位置显然是不正确的，所以要把中心转移到得到的 cluster 内部的数据点的平均位置。实际上也就是在每个数据点的归类确定的情况下，计算上面函数取极值的位置，然后再次构造新的 k 个 cluster。这个过程中，中心点的位置不断地在改变，构造出来的 cluster 也在变化。通过多次的迭代，这 k 个中心最终会收敛并不再移动。

2.2 张量及张量分解

张量(Tensor)是多维数组的空间表示，是 N 维向量空间，一个三维张量由 3 个方向的坐标系构成，其中每一个坐标系代表一维数据的表示[14]。矩阵由两维数据表示，也可以理解为二维张量。三维以上的数据张量可以称为高阶张量。

利用三阶张量中的值，我们可以针对不同的应用进行张量分解，在预测和个性化推荐领域进行应用。

张量定义：设 $V_1,\cdots,V_N$ 为维数分别为 $I_1,\cdots,I_N$ 的 N 个有限维欧几里得空间，对于 N 个向量 $u_1\in V_1,\cdots,u_n\in V_N$，定义 $V_1\times\cdots\times V_N$ 上的线性映射 $(u_1\circ\cdots\circ u_N)$ 为

$$(u_1\circ\cdots\circ u_N)(x_1\circ\cdots\circ x_N)=\langle u_1,x_1\rangle_{V_1}\cdots\langle u_N,x_N\rangle_{V_N} \tag{2}$$

其中，$\langle u_i,x_i\rangle_{V_i}$ 为 V_i 上的标量积，x_i 为 V_i 上的任意向量($i=1,\cdots,N$)，则全体 $(u_1\circ\cdots\circ u_N)$ 构成的空间称为 $V_1,\cdots,V_N$ 的张量空间，此张量空间上的元素称为 $V_1\times\cdots\times V_N$ 上的 N 阶张量。特别地，如果 $V_n=R^{I_n}$，$n=1,\cdots,N$，则此张量空间称为 N 阶 $I_1,\cdots,I_N$ 维实张量空间，记为 $R^{I_1\times\cdots\times I_N}$；类似地，定义 N 阶 $I_1,\cdots,I_N$ 维复张量空间，记为 $C^{I_1\times\cdots\times I_N}$。对于 N 阶张量 $A\in R^{I_1\times\cdots\times I_N}$，其矩阵展开形式 $A_{(n)}\in R^{I_n\times\prod I_k}$，张量的元素 $(i_1,i_2,\cdots,i_N)$ 映射到矩阵元素 (i_n,j)，映射关系为 $j=1+\sum_{k=1}^{N}(i_k-1)J_k$。

张量 tucker 分解：张量 tucker 分解是一种高阶的主成分分析方法，它将原始张量分解为核心张量和一系列矩乘积的形式，张量分解在求解核心张量的过程中可以进行降维处理。

tucker 分解定理：一个张量 $A\in R^{I_1\times\cdots\times I_N}$ 可以表达为

$$A=B\times U^{(1)}\times U^{(2)}\times\cdots\times U^{(N)} \tag{3}$$

其中，B 为核心张量；$U^{(1)},U^{(2)},\cdots,U^{(N)}$ 为一系列正交矩阵。由于投影矩阵 U 的正交性，可以通过式(3)求得核心张量 B 为

$$B=A\times U^{(1)T}\times U^{(2)T}\times\cdots\times U^{(N)T} \tag{4}$$

张量模型一个成功的应用领域是个性化标签推荐系统，在这些系统中，原先的协同过滤方法在面对三维及以上数据时，缺乏扩展性，不能有效地反映多模态数据之间的复杂关系，而采用张量模型则能够很好地解决这一问题。

3 药品个性化推荐模型

药品的个性化推荐过程中，首先根据药品的描述信息采用向量空间模型(VSM)[15]将药品特征格式化，利用 K-means 算法对药品进行聚类，然后利用用户评价进行协同过滤推荐，之后根据协同推荐的不足，提出利用张量分解进行建模，得到个性化药品推荐结果。

3.1 药品聚类

我们首先根据药品治疗功效对其进行聚类，聚类的目的是为了试图在用户输入病症的时候提供一些符合病情的基本药品列表，便于下一步的个性化推荐，同时也可以对药品进行分类。本文采用信息检索中常用的向量空间模型来表示每种药品，它的主要思想是：将每一种药品都映射为由一组规范化正交词条矢量张成的向量空间中的一个点。对于所有的药品，都可以用此空间中的词条向量 $(T_1,W_1,T_2,W_2,\cdots,T_n,W_n)$ 来表示，其中，T_i 为特征向量词条，W_i 为 T_i 的权重。

首先统计出所有药品描述中的词汇，去掉停用词后将所有词汇作为词典，然后利用 tf-idf[16]方法计算出每种药品在词典中出现的词汇的权重，得到每种药品的向量特征表示。第 j 个药品的描述中与词典里第 k 个词对应的 tf-idf 为：

$$TF\text{-}IDF(t_k,d_j)=TF(t_k,d_j)\cdot\log\frac{N}{n_k} \tag{5}$$

其中，$TF(t_k,d_j)$是第 k 个词在药品 j 的描述中出现的次数，而 n_k 是所有药品的描述中包括第 k 个词的药品数量。

得到每种药品的特征向量后，利用 k-means 算法对药品进行聚类，设定将样本聚类成 k 个簇，具体算法如下：

1. 随机选取 k 个聚类质心点为 $u_1,u_2,\cdots,u_k\in R^n$。

2. 重复下面过程直到收敛 {

对于每一个样例 i，计算其应该属于的类

$$c^{(i)}:=\arg\min_j\|x^{(i)}-u_j\|^2 \tag{6}$$

对于每一个类 j，重新计算该类的质心

$$u_j:=\frac{\sum_{i=1}^{m}1\{c^{(i)}=j\}x^{(i)}}{\sum_{i=1}^{m}1\{c^{(i)}=j\}} \tag{7}$$

}

3.2 基于用户评分的药品协同推荐

上面的聚类实际上是为下面进行的协同推荐做数据准备，每次用户针对自己的病情输入病情描述时，就可以根据上一步的聚类匹配出最接近的一些类别，得到一个基本的适合用户需求的药品列表，在这个药品列表的基础上，我们提取出所有对这些药品进行评价过的用户评分，构建用户-药品评分矩阵，利用这个评分矩阵进行协同推荐。这里采用基于用户相似度的协同推荐，具体算法如下。

3.2.1 用户相似度计算

计算用户之间的相似度采用 Pearson 距离，公式如下：

$$sim(a,b)=\frac{\sum_{i\in I}(R_{a,i}-\overline{R_a})(R_{b,i}-\overline{R_b})}{\sqrt{\sum_{i\in I}(R_{a,i}-\overline{R_a})^2}\sqrt{\sum_{i\in I}(R_{a,i}-\overline{R_b})^2}} \tag{8}$$

其中，I 为所有物品的集合，$R_{a,i}$和 $R_{b,i}$分别表示用户 a 和 b 对物品 i 的评分，$\overline{R_a}$和$\overline{R_b}$分别表示用户 a 和 b 的平均评分。

3.2.2 预测值计算

根据之前算好的用户之间的相似度，接下来对用户未打分的物品进行预测，这里采用加权求和的方式。对给定用户 u 的相似用户给物品 i 已打的分数进行加权求和，权值为用户 u 与各个用户的相似度，计算得到给定用户 u 对物品 i 的打分，公式如下：

$$P_{u,i}=\frac{\sum_{j}^{N}(s_{u,j}*R_{j,i})}{\sum_{j}^{N}(|s_{u,j}|)} \tag{9}$$

其中，N 为相似用户数目，$s_{u,j}$为用户 u 与用户 j 的相似度，$R_{j,i}$为用户 j 对物品 i 的打分。计算出当前用户没有评分过的所有药品，推荐出预测评分数最高的前 N 个药品。

但是协同推荐算法无法解决新用户的推荐问题，即冷启动问题，并且随着互联网中用户数量的急剧增长，推荐系统的输入数据集规模也显著增长，海量数据规模也给推荐算法带来极大挑战。

3.3 基于张量分解的药品推荐

为了解决上述提到的问题，本文提出基于张量分解的药品推荐算法，该算法通过对“药品-用户-标签”三元关系进行三阶张量建模与分解，抽取核心张量并根据药品预测评分来对用户进行个性化药品推荐。主要的算法流程如图 1 所示，其中数据预处理过程在前面的阶段完成。

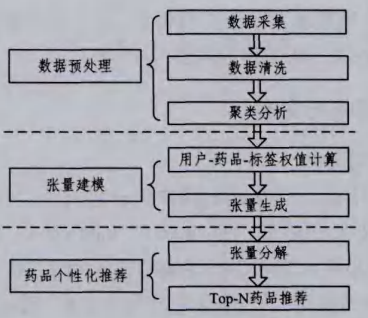

图 1 模型流程图

3.3.1 张量建模

本文提出基于用户评分的权值张量表示形式，综合考虑用户、药品、药品标签的权值与用户评分，得到一个“用户-药品-标签”三元元组的权值 $w_{u,i,t}$，然后对三元元组建立张量模型时，将权值 $w_{u,i,t}$作为张量中的元素进行建模。这里的权值 $w_{u,i,t}$通过如下方法进行计算：首先利用上节提到的 K-means 算法对药品进行聚类，然后将每一种药品都映射为由一组规范化正交词条矢量张成的向量空间中的一个点，对于给定的用户、药品、药品标签，使用标签的权值乘以用户对与该药品的评价分数(0－5)，得到的即为 $w_{u,i,t}$。定义 U 表示用户集合，I 表示药品集合，T 表示某一药品的标签集合，$R_{i,j}$表示用户 i 对药品 j 的评分，$T_{i,j,k}$表示对于给定的用户 i、药品 j、标签 k 在标签向量中的权重值，且$\sum_k T_{i,j,k}=1$，$w_{i,j,k}$表示三阶张量中对于给定用户、药品以及药品的某一个标签的权值，$w_{i,j,k}=R_{i,j}\cdot T_{i,j,k}$，$i\in U,j\in I,k\in T$。这样我们得到的张量模型如图 2 所示。

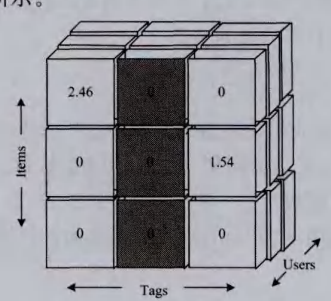

图 2 根据“用户-药品-标签”构建的三阶张量

3.3.2 基于张量分解的推荐

按照张量分解法进行张量分解后得到一些新的元素，其权重表示用户根据特定的标签对药品给出的评分估值，这样我们利用三阶张量中的用户-标签这两个维度，就可以推荐出评分最高的前 K 个药品。

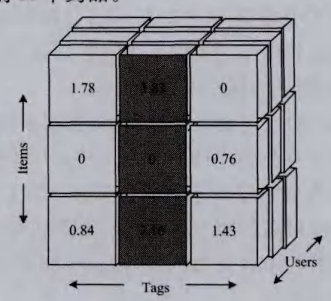

图 3 分解后的张量

如图 3 所示，经过张量分解后一些新的元素由 0 变为非

0，根据图中所得到新元素的权值大小，我们可以针对某一用户产生 Top-N 的药品推荐列表。

4 实验结果与分析

4.1 数据来源

本文中的数据需要用到用户评分进行个性化推荐，因此采用爬虫技术在国外的网上药店 walgreens[1] 上面爬取了 21559 条药品评价信息，将其存储在 Mysql 数据库中。

其中每条数据记录的属性为药品名称、药品描述、用户名、性别、年龄、用户评分。数据首先进行预处理，将信息不全的数据去掉后统计如表 1 所列。

表 1 实验数据集

数据集	药品评价数	用户数	药品数
原始数据	21559	6071	8163
预处理后	20397	5827	7901

4.2 实验评价标准

本文利用推荐系统中最常用的准确率和召回率做为评估指标。首先根据药品描述信息得到每个药品的特征向量后可以提取出药品所对应的标签，这里我们认为是该药品可以治疗的疾病症状标签。实验过程中我们将数据集中部分用户对应的“用户-药品-标签”三元组作为测试集，利用张量分解得到 Top-N 的药品推荐列表后，看该用户的所有标签对应的药品是否出现在推荐列表中，准确率的计算公式如下：

$$Precision=\underset{i=1}{\overset{M}{\text{avg}}}\frac{hitmedicine_i}{N} \tag{10}$$

其中，M 为预测的用户数目，$hitmedicine_i$ 为每个用户的三元组（即相应标签对应的药品）在推荐列表的数目。

召回率的计算公式如下：

$$Recall=\underset{i=1}{\overset{M}{\text{avg}}}\frac{hitmedicine_i}{H_i} \tag{11}$$

其中，M 为预测的用户数目，$hitmedicine_i$ 为每个用户的三元组（即相应标签对应的药品）在推荐列表的数目，H_i 为每个用户的实际三元组数目。

最后我们用 $F1$ 评价方法来综合权衡评价推荐结果的效果。

$$F1=\frac{2*Precision*Recall}{Precision+Recall} \tag{12}$$

4.3 实验结果分析

本文主要利用 Python 语言和 Matlab 平台实现协同过滤和张量分解算法。

首先根据药品描述得到药品特征向量后进行聚类，取聚类数为 10，选取每个簇的中心向量中按权重从大到小前 5 个特征作为该簇的特征值，这些特征值后面都表示为每个用户和药品对应的症状标签。我们事先随机挑选 2000 条评价数据作为测试集，剩下的数据作为训练集。

在进行用户协同过滤推荐过程中，根据测试集用户的症状标签选出其所在聚类中的所有用户药品评价数据，构建所有评价过这些药品的用户-药品评分矩阵。我们先计算出与当前用户相似度最高的前 20 个用户，然后利用这 20 个用户计算出当前用户没有评分过的所有药品，推荐出预测评分数最高的前 N 个药品。

在张量分解中选取 core=(6 * 6 * 6)的核心张量进行分解，在得到的近似张量中保留单元格中大于 0.1 的单元格。

根据不同的推荐列表长度 1、3、5、10、15、20、25、30，得到协同过滤(CF)和张量分解(Tensor)的实验结果对比如下：

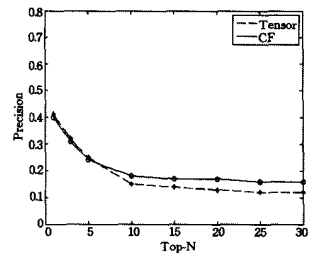

图 4 Tensor 和 CF 的准确率对比

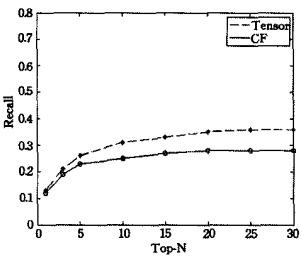

图 5 Tensor 和 CF 的召回率对比

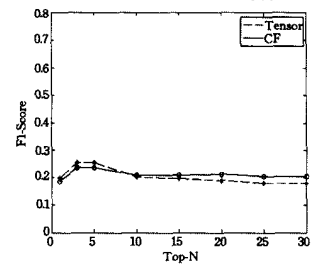

图 6 Tensor 和 CF 的 $F1$ 指标对比

图 4－图 6 的结果表明，在推荐长度小于 10 时张量分解的推荐准确度比协同过滤高。在推荐长度大于 10 之后，计算准确度时由于分母增加明显，导致准确率下降。

从实验结果可以看出，基于张量分解的推荐算法在处理稀疏性大量数据的时候，得到的推荐列表长度比协同过滤要好，在推荐列表大于 10 的时候准确率下降较快，这和张量分解后对于新元素都比较接近，区分相关度不明显。

结束语 本文针对现实生活中用户网上买药时缺乏指导这一问题，提出利用聚类和基于用户协同推荐的算法对用户进行个性化药品推荐；之后考虑协同过滤在处理海量数据和数据稀疏性问题时的不足，提出利用张量分解进行个性化推荐。实验结果显示：这一个性化药品推荐过程模拟了病人买药的过程，先选出一部分符合病情的药品列表，然后根据其他人的评分数据找到当前用户没有使用的药品，具有良好的效果。本文提出的模型具有很重要的现实意义，如何更好地结合用户自身的特征比如年龄、地域等因素，进一步提高个性化药品推荐的准确度是我们下一步研究的重点工作。

参 考 文 献

[1] Hripcsak G, Albers D J. Next-generation phenotyping of electronic health records[J]. J AmMed InformAssoc, 2013, 20(1): 117-121

[2] Deshpande R, Thuptimdang W, DeMarco J, et al. A collaborative framework for contributing DICOM RT PHI (Protected Health Information) to augment data mining in clinical decision support [C] // SPIE Medical Imaging. International Society for Optics and Photonics, 2014

[3] http://www.drugstore.com

[4] Duan L, Street W, Xu E. Healthcare information systems: data mining methods in the creation of a clinical recommender system

1) http://www.walgreens.com/

[J]. Enterp Inf Syst,2011,5(2):169-181
[5] Jøsang A,Guo G,Pini M S,et al. Combining Recommender and Reputation Systems to Produce Better Online Advice[M]. Modeling Decisions for Artificial Intelligence. Springer Berlin Heidelberg,2013:126-138
[6] Cho Young-Sung, et al. Clustering Method using Item Preference based on RFM for Recommendation System in u-Commerce [M]. Ubiquitous Information Technologies and Applications. Springer Netherlands,2013:353-362
[7] Yuan X,Lee J H,Kim S J,et al. Toward a user-oriented recommendation system for real estate websites[J]. Information Systems,2013,38(2):231-243
[8] Barragáns-Martínez A B,Costa-Montenegro E,Burguillo J C,et al. A hybrid content-based and item-based collaborative filtering approach to recommend TV programs enhanced with singular value decomposition[J]. Information Sciences,2010,180(22):4290-4311
[9] Cai X, Bain M, Krzywicki A, et al. Collaborative filtering for people to people recommendation in social networks[M]. AI 2010: Advances in Artificial Intelligence. Springer Berlin Heidelberg,2011:476-485
[10] Lops P,de Gemmis M,Semeraro G. Content-based recommender systems:State of the art and trends[M]. Recommender Systems Handbook. Springer US,2011:73-105
[11] Carrer-Neto W,Hernández-Alcaraz M L,Valencia-García R,et al. Social knowledge-based recommender system. Application to the movies domain[J]. Expert Systems with Applications,2012,39(12):10990-11000
[12] Dunham M H. Data Mining:Introductory and Advanced Topics [M]. Pearson Education,2007
[13] Zhang Y J, Cheng E. An optimized method for selection of the initial centers of k-means clustering[M]// Integrated Uncertainty in Knowledge Modelling and Decision Making. Springer Berlin Heidelberg,2013:149-156
[14] Liu J,Musialski P,Wonka P,et al. Tensor completion for estimating missing values in visual data[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence,2013,35(1):208-220
[15] Werner D,Cruz C. A Method to Manage the Precision Difference between Items and Profiles[C]// 2013 International Conference on a Context of Content-Based Recommender System and Vector Space Model. Signal-Image Technology & Internet-Based Systems (SITIS). IEEE,2013:337-344
[16] Huang X,Wu Q. Micro-blog commercial word extraction based on improved TF-IDF algorithm[C]// 2013 IEEE Region 10 Conference on TENCON. IEEE,2013:1-5

(上接第 224 页)

文评估了集群中数据分布的均匀性:写入读取测试完成后,分别计算了两种策略下集群事件密集度的变差系数。变差系数为所有数据节点的事件密集度标注差与平均数的比值,用以衡量各数据节点存储文件量的离散程度。得到的默认策略下系统的变差系数为 0.0418,而本文策略下的仅为 0.0038,说明文件分布较前者更加均匀。以上实验表明本文提出的基于事件密集度的数据放置策略相比于 HDFS 默认的机架感知策略减少了块的写入时间和读取时间,有效地提高了系统的吞吐量,并且能够将文件更均匀地分布在各个节点上,从而提供了更好的负载均衡。读块时间的减少也有利于系统更及时地响应用户的访问请求。

结束语 随着智能交通系统的普及,Hadoop 等分布式架构将越来越广泛地应用于交通监控系统以应对海量交通监控视频的高效存储以及开放交通监控查询平台等应用对于系统性能的需求。在将 HDFS 用于交通监控视频存储的背景下,本文分析了 HDFS 中默认的机架感知策略的不足,根据交通监控视频的事件分类特征,提出了一种 HDFS 中基于事件密集度的交通监控视频数据放置策略。实验表明本文策略相对于默认策略能够更好地平衡节点负载,提高系统的吞吐量。

参 考 文 献

[1] 王国锋,宋鹏飞,张蕴灵. 智能交通系统发展与展望[J]. 公路,2012(5):217-222
[2] 张庆华. 云存储技术在视频监控中的发展与应用[J]. 中国安防,2013(8):53-58
[3] Borthakur D. The hadoop distributed file system: Architecture and design [J]. Hadoop Project Website,2007,11:21
[4] Shvachko K,Kuang H,Radía S,et al. The hadoop distributed file system [C]// 2010 IEEE 26th Symposium on Mass Storage Systems and Technologies (MSST). IEEE,2010:1-10
[5] Ghemawat S,Gobioff H,Leung S T. The Google file system [J]. ACM SIGOPS Operating Systems Review. ACM,2003,37(5):29-43
[6] 蔡斌,陈湘萍. Hadoop 技术内幕[M]. 北京:机械工业出版,2013:216-217
[7] 武文斌. 视频监控云存储模型设计[J]. 山西科技,2012(3):35-37
[8] Xie J,Yin S,Ruan X,et al. Improving mapreduce performance through data placement in heterogeneous hadoop clusters [C]// 2010 IEEE International Symposium on Parallel & Distributed Processing, Workshops and Phd Forum (IPDPSW). IEEE,2010:1-9
[9] 林伟伟. 一种改进的 Hadoop 数据放置策略[J]. 华南理工大学学报:自然科学版,2012,36(1):152-158
[10] 刘琨,钮文良. 一种改进的 Hadoop 数据负载均衡算法[J]. 河南理工大学学报:自然科学版,2013,32(3):332-336
[11] 徐骁勇,潘郁,丁燕艳. 基于灰色马尔可夫链预测模型的 HDFS 云存储副本选择策略[J]. 计算机应用,2012,31(A02):39-42
[12] Ananthanarayanan G, Agarwal S, Kandula S, et al. Scarlett: coping with skewed content popularity in mapreduce clusters [C]// Proceedings of the sixth conference on Computer systems. ACM,2011:287-300
[13] Abad C L,Lu Y,Campbell R H. DARE: Adaptive data replication for efficient cluster scheduling [C]// 2011 IEEE International Conference on Cluster Computing (CLUSTER). IEEE,2011:159-168
[14] 吴萌. 交通监控视频中的异常事件检测[D]. 北京:北京邮电大学,2010
[15] Massie M L,Chun B N,Culler D E. The ganglia distributed monitoring system: design, implementation, and experience[J]. Parallel Computing,2004,30(7):817-840