基于大数据的用户学习偏好建模及应用

单留举 王晓东 马英运
(河南师范大学计算机与信息工程学院 河南 新乡 453007)

摘 要 针对自主研发的“睿训课堂”移动学习教学软件中长期积累海量数据分析的需要,将学生学习兴趣分为短期学习兴趣和长期学习兴趣,提出一种基于“大数据”的学生学习偏好模型。并根据学生浏览行为和日志记录数据,挖掘学生短期学习兴趣。利用后台服务器数据库中的数据,对初始学生用户注册信息进行挖掘,提取出学生长期学习兴趣。实验结果表明,基于“大数据”的学生用户学习偏好模型方法是合理和有效的。

关键词 移动学习 大数据 短期学习兴趣 长期学习兴趣 学生学习偏好 日志挖掘

中图分类号 TP311.131 **文献标识码** A DOI:10.3969/j.issn.1000-386x.2016.01.020

USER LEARNING PREFERENCE MODELLING AND APPLICATION BASED ON BIG DATA

Shan Liuju Wang Xiaodong Ma Yingyun
(*College of Computer and Information Engineering*, *Henan Normal University*, *Xinxiang 453007*, *Henan*, *China*)

Abstract For the need of analysing the long-term accumulated massive data in self-developed "core training class" mobile learning educational software, we divided the students' interest in learning into short-term interest and long-term interest, and proposed a "big data"-based student learning preference model. Furthermore, according to students' browsing behaviours and the log record data, we mined students' short-term interest in learning. By the use of the data in background server database, we mined the registration information of initial student users, and extracted students' long-term interest in learning. Experimental result proved that the "big data"-based learning preferences model of student users is reasonable and effective.

Keywords Mobile learning Big data Short-term interest in learning Long-term interest in learning Students learning preferences Log mining

0 引 言

近些年,随着云计算、物联网、三网融合、互联网等通信和IT技术的迅速发展,数据的爆炸性增长成为了很多业界共同关注的热点,因此,信息技术社会已经进入了“大数据”[1-3]时代。大数据的出现不光改变了企业的运作模式以及人们的工作和生活方式,甚至还改变了科学研究模式。大量资源信息不但充裕了人们的生活,也带来信息过载的问题。在这种情况下,信息过滤技术已经普遍运用到目前个性化推荐范畴[4]。而个性化推荐的关键就是构建用户偏好模型,其目的就是帮助用户从大量的信息中挑选出感兴趣的信息。因此,在大数据时代下构建一个基于“大数据”的学生用户学习偏好模型对于辅助老师教学具有非常重要的现实意义。

1 相关研究综述

关于本文用户学习偏好建模的研究,主要是从用户的学习兴趣的提取与学习偏好建模两个方面进行展开。

1.1 用户学习兴趣的提取

在大数据移动学习的信息检索与过滤中,建立用户学习偏好文档使跟踪用户学习的行为和兴趣成为可能,有助于为学生学习提供个性化信息及学习资源服务[5]。对于心理学来讲,兴趣是人们倾向于熟悉、钻研得到某种知识的情绪特性,是能够促进我们求知的一种动力。如果一个学生对某课程感兴趣的话,他就会连续地全心全意的研究它,那么他就达到了提高学习效率的目的。

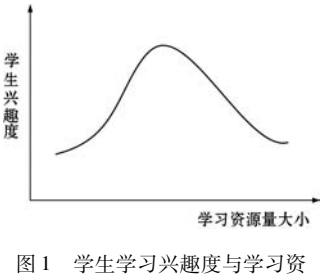

图1 学生学习兴趣度与学习资源量关系示意图

可以认为,用户学习兴趣的大小和用户对这个兴趣相关的学习信息资源需求量是相关的。尤其是在大数据移动学习平台中,学生学习兴趣的大小对相关学习资源的接受程度是相关的,它们之间的关系可如图1[6]所示。所以,识别并提取用户学习兴趣成为了亟待解决的问题。

学习兴趣大体上可以分为长期学习兴趣与短期学习兴趣两

收稿日期:2014-07-02。河南省科技攻关计划项目(102300410198);河南师范大学与郑州第二中学合作项目“网络创新教育”项目助资。**单留举**,硕士生,主研领域:数据挖掘,语义网络和本体。**王晓东**,教授。**马英运**,硕士生。

种。长期学习兴趣是因所学资源或者学习过程自身直接产生的,而短期学习兴趣是因学习活动的结果产生的。短期学习兴趣具有明显的自觉性。当学生意识到学习的重要性或者关切到自己的利益时,他们对学习的兴趣自然而然就产生了。譬如,为了获取教师、父母的赞赏,朋友、同学的尊重,在学校举行考试中获得更好的成绩等等,也可以激发他们对学习的兴趣。短期学习兴趣相对来说容易消失、不稳定,然而对学生的当前学习偏好起着非常重要的作用,成为了教师最为关注的部分。

1.2 用户学习偏好的建模

原型法、交互法、多模型方法构建了三种常用的用户偏好模型方法。其中原型法是三种构建方法中最常用的方法。在过滤的消息范畴中,原型法不但能够用于建构对用户开始感兴趣方面的文档,而且还能够用于增添额外的知识以增加初始兴趣的文档[7]。用户承继了其所属的单个或多个原型的偏好文档[8]。张卫丰等人认为的交互方法[9]是按照特征信息提取来进行用户和用户交互。对于推荐系统而言,信息过滤系统中常用的一种方法就是交互法。其中提交的关键字方法是相互作用的、最简单的方法,即以达到用户的兴趣倾向为目的对用户进行填写表单。对于用户偏好模型的构建而言,文献[10]基于 Cetintemel 等提出的一种多模型 MM(Multi—Model)的方法。用户偏好文档使用多模型表示方法,该文档不能被表示为一个单一的偏好向量,并表示作为一个集合的用户兴趣相关的类。多模型法是在目前的用户偏好建模中最为常用的方法,该方法不但能通过多个向量的集合来反映用户的偏好,而且还能够有效地处理用户的行为倾向和用户的反馈信息,然后把这些信息反馈给偏好模型,对偏好模型进行调整。

综合上面的分析,用户偏好建模的理论已有相关研究,并获得了一定的应用。但是基于"大数据"的用户学习偏好建模相关研究很少,特别是从心理学角度看,用户学习偏好受两方面影响——短期学习兴趣和长期学习兴趣,并且它们有一定的相互关联。而目前的用户学习偏好建模没有考虑这一点,从而导致对用户学习兴趣的转换和提取很难来描述。文章将介绍使用用户学习兴趣来进行用户偏好挖掘提取过程。

2 用户学习偏好挖据的理论模型

用户学习偏好受到的两方面影响为短期学习兴趣和长期学习兴趣,因此用户学习的兴趣偏好文档可简单的表示为:

$$D = \{M, N\} \tag{1}$$

式中:M 表示短期学习兴趣,N 表示长期学习兴趣。由于学习兴趣的繁多性,所以 M 和 N 可以分别表示为:

$$M = \{S_1, S_2, \cdots, S_n\}, N = \{L_l, L_2, \cdots, L_n\}$$

且 $U = \{S_1, S_2, \cdots, S_n, L_l, L_2, \cdots, L_n\}$ 表示为用户学习兴趣偏好。

针对用户各种短长期学习兴趣而言,为了更详细地区分用户感兴趣的程度与它们的类别,学习兴趣向量应该蕴涵大量的资源信息。于是,针对每一个 S_i, L_j($i = 1,2,\cdots,m; j = 1,2,\cdots,n$)来说,引进类别属性变量 E_i, E_j 与权重属性变量 F_i, F_j,因此 S_i, L_j 能够表示成

$$\begin{aligned} S_i &= <S_i, F_i, E_i> \quad i = 1,2,\cdots,m \\ L_j &= <L_j, F_j, E_j> \quad j = 1,2,\cdots,n \end{aligned} \tag{2}$$

结合式(1)与式(2)能够得出用户学习兴趣偏好文档可以通过一个二维表的形式来表示。如下:

$$D = \begin{Bmatrix} S_1 & S_2 & \cdots & S_m & L_1 & L_2 & \cdots & L_n \\ F_1 & F_2 & \cdots & F_m & F_{m+1} & F_{m+2} & \cdots & F_{m+n} \\ E_1 & E_2 & \cdots & E_m & E_{m+1} & E_{m+2} & \cdots & E_{m+n} \end{Bmatrix} \tag{3}$$

为了表达方便,简写为

$$D = \{ <S_1, F_1, E_1>, <S_2, F_2, E_2>, \cdots, <S_m, F_m, E_m>, <L_1, F_{m+1}, E_{m+1}>, <L_2, F_{m+2}, E_{m+2}>, \cdots, <L_n, F_{m+n}, E_{m+n}> \}$$

式中:S_m, L_n 分别为短期学习兴趣与长期学习兴趣的某个属性值;E_{m+n} 代表用户学习兴趣所对应的学习资源所属资源类别;F_{m+n} 是代表属性值词汇的学习兴趣权重,表示学生对某个类别学习资源的感兴趣程度,它的特点为随着学生反馈信息值不断的变化,且有 $F_1 + F_2 + \cdots + F_{m+n} = 1$。为了能够达到较好的理解采用短期或长期学习兴趣表示偏好文档的目的,下面例举了郑州教育局合作研发的教学软件项目——睿训课堂中近三年积累的教学资源的资源分类标准,学生用户在睿训课堂上搜索学习资源的时候也是采用该标准来选择学习资源的。

例如,我们开发的教学软件——睿训课堂使用对象可分为学生、管理员、教师;对象属性包括性别、姓名、学号、教工号、年龄、偏好等;对象的具体属性值词汇包括男、1208180631、30 ~ 40 Years、数学等。所以,如图 2 所示,可以构建出一个"对象-属性-属性值"的树形结构图。

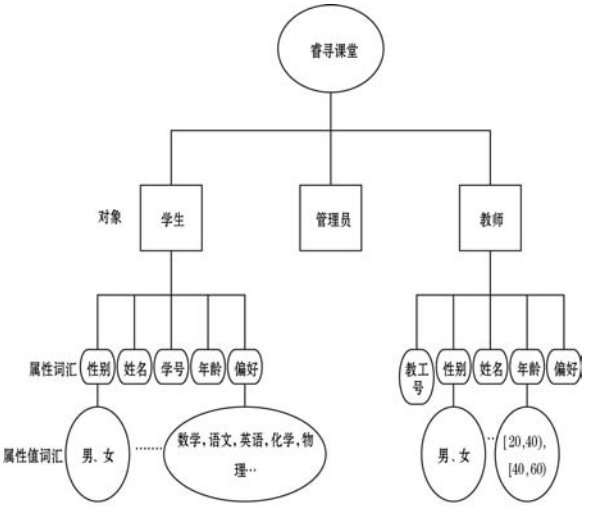

图 2 对象-属性-属性值树形结构图

3 用户学习偏好获取

对于我们研发的教学软件"睿训课堂"来讲,从近三年积累的大数据中获取用户学习偏好信息采用方式为:显式获取与隐式挖掘。显式获取,即用户学习偏好信息直接获取。这种方式需要用户直接参与,并且它是通过学习者注册信息或者填写偏好信息表单完成的。隐式挖掘,即通过挖掘学生用户浏览 Web 历史行为来获取学生学习偏好信息。在睿训课堂大数据环境下,为了准确地获取用户学习偏好信息,采用隐式挖掘方式更为合理。

根据学生的浏览行为与浏览内容来构建学生的学习偏好模型。首先,通过学生注册数据和填写偏好信息表单数据来判断学生类别,采用显示获取方式来获得学生长期学习兴趣向量。其次,通过跟踪和观察学生的浏览行为(包括 cookies 记录、Web

服务器日志以及收藏夹)隐式挖掘出学生短期学习偏好向量。最后,通过学生对推荐学习资源信息和学习资源信息的反馈来修正学生学习偏好。建立基于大数据的学生用户学习偏好模型框架,如图 3 所示。

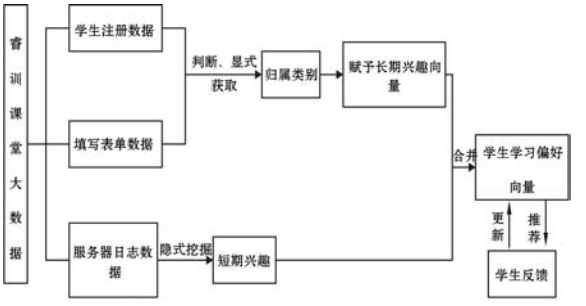

图 3　基于大数据的学生用户学习偏好模型框架

3.1　短期学习兴趣偏好获取

由于随着学生的兴趣衰减,睿训课堂产生的学生短期学习兴趣偏好向量逐渐趋于 0,所以,用户偏好矩阵就变得非常稀疏。为了能解决这样的问题,采取的措施为:通过长期学习兴趣关联推荐为短期学习兴趣提供新的学习资源。

获取学生之间长期学习兴趣的关联性来发现学生的短期学习兴趣即为长期学习兴趣关联推荐。例如,学生 C 和 D 都喜欢英语,学生 C 喜欢浏览听力方面的题,学生 D 喜欢浏览阅读方面的题。但是在某个时间里,学生 C 也会去浏览阅读方面的题,这样阅读方面的题对 C 来说可以看作 C 的短期学习兴趣。因此可以作为学生 C 的短期学习兴趣补充。

根据学习兴趣类别的相似度我们可以判断学生长期学习兴趣是否关联,即可以设定两个阈值 A 与 B,如果 $0 < A < B < 1$,那么这两个学习兴趣具有相互关联性。为了学生能够得到更多的短期学习兴趣,我们借助于长期学习兴趣的关联推荐,这样就解决了短期学习兴趣矩阵稀疏的问题。

3.2　基于聚类分析的学生用户长期学习偏好获取

对于我们开发的教学软件“睿训课堂”来说,无论学生要浏览学习资源信息还是要做各个学科方面的题,都需要学生首先注册账号,目的是为了存储学生基本信息情况,进一步了解学生学习需求,从而更利于提供更好的个性化服务。一般学生注册提供的信息有:姓名、年龄、性别、学号、班级、学科偏好、入学成绩、学生自我评价等。其中有几个基础变量对学生的长期学习偏好起到决定性的作用。下面详细研究基于聚类分析获取学生长期学习偏好。

3.2.1　建立学生用户注册信息向量

“睿训课堂”注册学生信息包括:姓名、年龄、性别、学号、班级、学科偏好、入学成绩、学生自我评价。此时,可用一个向量 Y =(姓名、年龄、性别、学号、班级、科目的偏好、入学考试成绩、学生的自我评价)表示学生基本信息,进而转换为向量分量数值的形式即 $Y=(y_1,y_2,\cdots,y_8)$。例如,对于性别来说,1 为女,0 为男。

3.2.2　基于 K-MEANS 算法的学生用户聚类

针对学生基本信息向量,K-MEANS 聚类算法[11-13]是目前最为有效的用户聚类算法,因为该算法比较简洁和速度比较快。运用该算法能把学生聚类为 K 类稳定用户集合。设数据点的集合 $P=(Y_1,Y_2,\cdots,Y_m)$,其中 $Y_i=(y_{i1},y_{i2},\cdots,y_{i8})$,$i=1,2,\cdots,m$。将其分为 K 个组 $D_1,D_2,\cdots,D_K$,具备的性质有:①$D_i\neq\phi$,$i=1,2,\cdots,K$;②$D_i\cap D_j=\phi$ 且 $\bigcup_{i=1}^{k} D_i = P, i,j=1,2,\cdots,K$,且 $i\neq j$。具体算法步骤如下:

(1) 令 $M=1$,从 m 个点集合$(Y_1,Y_2,\cdots,Y_m)$随机选取 k 个点 $(Q_{1(M)},Q_{2(M)},\cdots,Q_{k(M)})$作为 k 个簇的中心。

(2) 当且仅当满足 $\|Y_i-Q_j\|<\|Y_i-Q_t\|$ $(t=1,2,\cdots,k$,且 $j\neq t)$,则将 $Y_i(i=1,2,\cdots,m)$归入簇 $D_j(j=1,2,\cdots,k)$。

(3) 计算簇的新中心点 $Q_{1(M+1)},Q_{2(M+1)},\cdots,Q_{k(M+1)}$,计算公式为:

$$Q_{i(M+1)}=\frac{1}{m_i}\sum_{Y_j^{(i)}\in D_i} Y_j^{(i)} \quad i=1,2,\cdots,k$$

式中:m_i 是处于簇 D_i 的点的数量,且令平均误差准则函数:

$$F(M+1)=\sum |Y_j^{(i)}-Q_{i(M)}|^2/\frac{1}{m_i}$$

(4) 给定的算法精度 ∂,如果 $|Z(M+1)-Z(M)|<\partial$ 则算法结束,否则 $M=M+1$,返回步骤(2)继续。

3.2.3　聚类获取学生的长期学习兴趣偏好

采纳 3.2.2 节 K-MEANS 算法,将学生用户样本集聚为 K 类。每类兴趣采用向量 <类别、关键字、权值> 的形式来表示,看着每一类学生的总体特征。根据 K-MEANS 算法具体步骤,最后得出 K 类学生共同偏好,即:

$$N_L=\{<c_1,f_1,w_{11}>,<c_1,f_2,w_{12}>,\cdots,<c_1,f_h,w_{1h}>,<c_2,f_1,w_{21}>,<c_2,f_2,w_{22}>,\cdots,<c_2,f_g,w_{2g}>,\cdots,<c_i,f_j,w_{ij}>\} \quad (4)$$

式中:h、g、i 为聚类获取的每一类学生偏好的关键字个数,$j=1,2,\cdots,k$ 为偏好类别,w_{ij}表示权值,且 $\sum W_{ij}=1$。

4　应用与分析

4.1　模型应用

将基于大数据的用户学习偏好模型应用到教学中,数据样本采用郑州二中最近 3 年来积累的大量数据。硬件 MAC OS 平台:英特尔酷睿 i5,4 GB 内存,1 TB 硬盘。后台选用数据库 sql server 2005,开发工具 JAVAEE。

4.2　实验数据与标准

实验采用 SQL Server 2005 后台数据库中 200 个学生样本的数据,即由睿训课堂系统中 200 位学生所提供的学生注册个人信息,这些信息包括:姓名、年龄、性别、学号、班级、科目的偏好、入学考试成绩、学生的自我评价等。兴趣组是:数学,汉语,物理,英语,化学,生物,美术,历史,地理,政治,音乐。同时选取每个学生最近 10 天 Web 服务器客户端日志数据,生成学生学习偏好之后,为学生推荐学习资源,收集学生的学习资源评分表。

通过查阅大量的文献可知获取学生学习偏好信息的评估标准是查准率(p)。查准率是检测学习资源之中真正符合学生兴趣的学习资源所占的比率,定义如下:

$$p=\frac{m}{n} \quad (5)$$

式中:m 为符合学生兴趣的、准确的学习资源个数,n 为教学软

件睿训课堂系统实际推荐出的学习资源个数。

4.3 实验结果分析

为了分析我们所构建的基于“大数据”的学生用户长短期学习偏好综合模型的效果,把该模型与仅基于 Web 日志数据的短期学习偏好旧模型进行比较分析,分析它们关于“查准率”与“决策时间”这两个方面的优劣。

我们随机抽取 50 名学生进行实验分析后,各个对应的查准率结果如表 1 与图 4 所示。

表 1 查准率对比

学生	p	学生	p	学生	p
01001	0.9	01018	0.6	01035	0.7
01002	0.7	01019	0.8	01036	0.5
01003	0.7	01020	0.9	01037	0.8
01004	0.6	01021	0.7	01038	0.6
01005	0.8	01022	0.8	01039	0.7
01006	0.9	01023	0.8	01040	0.9
01007	0.8	01024	0.8	01041	0.8
01008	0.7	01025	0.7	01042	0.6
01009	0.6	01026	0.6	01043	0.6
01010	0.7	01027	1.0	01044	0.8
01011	0.9	01028	0.5	01045	0.7
01012	0.8	01029	0.9	01046	0.6
01013	0.8	01030	0.8	01047	0.9
01014	0.7	01031	0.8	01048	0.9
01015	0.9	01032	0.9	01049	0.7
01016	1.0	01033	0.9	01050	0.8
01017	0.5	01034	0.7		

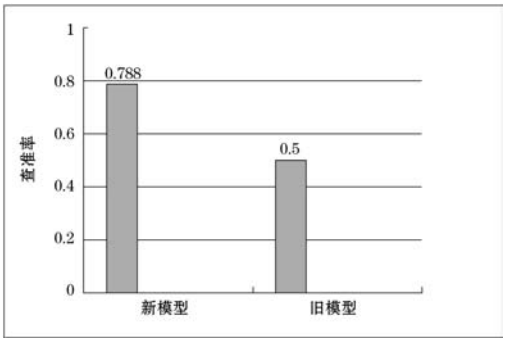

图 4 新旧模型查准率对比

由上面实验结果可以得出,采用基于“大数据”的学生用户长短期学习偏好综合模型能获得比较理想的查准率。比只有短期学习偏好模型的查准率高出 28.8%,提高了了解学生学习偏好的准确率。

下面为两种模型的学生决策时间的对比图,如图 5 所示。从图中我们可以明显地看出在应用了新模型之后的学生决策时间有了很大缩短。在我们开发的教学软件睿训课堂系统背景下,推荐学习资源后学生的决策时间大部分在 3 ~ 9 分钟,应用新的模型后,学生的决策时间大多数在 1 ~ 7 分钟。因此,可以证明基于“大数据”的学生用户长短期学习偏好综合模型是合理的和有效的,该模型明显地提高了学生的感受度。

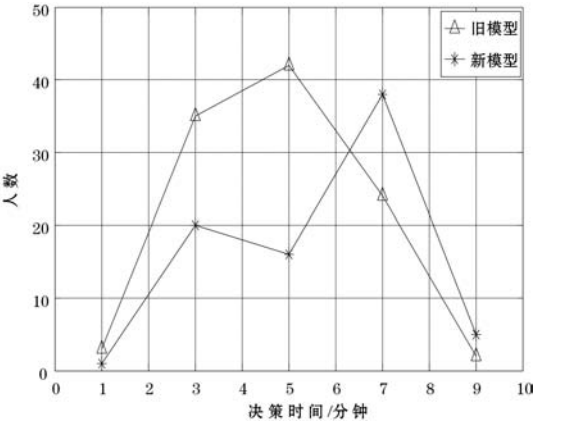

图 5 新旧模型决策时间对比图

5 结 语

通过对学生使用教学软件“睿训课堂”中积累的大量数据进行数据挖掘,在大数据的基础上,建立了一种基于“大数据”的学生长短期学习偏好模型。该偏好模型不但可以反映学生学习资源信息需求的变化,而且还能表现出学生对不同学科偏好的重视程度,最大深度地表现出学生的学习需求和偏好,并且能很好地应用于我们开发的睿训课堂系统推荐中。本研究具有良好的教育和教学的理论和实践价值。

参 考 文 献

[1] Chen M, Mao S W, Liu Y H, et al. Big Data: A Survey[J]. Mobile Networks and Application, 2014, 22(19): 171 - 209.

[2] 施聪莺,徐朝军,杨晓江,等. 电子书包中基于大数据的学生个性化分析模型构建与实现路径[J]. 中国电化教育,2014,326(3):63 - 69.

[3] 王元卓,靳小龙,程学旗,等. 网络大数据:现状与展望[J]. 计算机学报,2013,36(6):1125 - 1138.

[4] 李肇明. 基于个人兴趣的用户偏好建模[D]云南:云南大学,2013:953 - 960.

[5] Zhou X L. User prefence modeling and designing based on web data mining[J]. Journal of Xiangtan Normal College, 2009, 42(6): 55 - 59.

[6] 王洪伟,邹莉. 考虑长期与短期兴趣因素的用户偏好建模[J]. 同济大学学报:自然科学版,2013(6):953 - 960.

[7] 李贵林,杨禹琪,高星,等. 企业搜索引擎个性化表示与结果排序算法研究[J]. 计算机研究与发展,2014,51(1):206 - 214.

[8] Tsvi K, Bracha S, Peretz S, et al. Stereotype-Based versus Personal-Based Filtering Rules in Information Filtering System[J]. Journal of The American Society of Information Science and Technology, 2003, 54(3): 243 - 250.

[9] 张卫丰,徐宝文,徐蕾,等. 利用 Agent 个性化搜索结果[J]. 小型微型计算机系统,2001,22(6):724 - 727.

[10] Cetintemel U, Franklin M J, Giles C L, et al. Self-adaptive user profiles for large-scale data delivery[C]//IEEE, 2000: 622 - 633.

[11] Wang X, Cao J, Liu Y, et al. Text clustering based on the improved TFIDF by the iterative algorithm[C]//IEEE, 2012: 140 - 143.

[12] 施聪莺,徐朝军,杨晓江,等. TFIDF 算法研究综述[J]. 计算机应用,2009,29(6):167 - 170.

[13] 贾瑞玉,管玉勇,李亚龙,等. 基于 MapReduce 模型的并行遗传 k-means 聚类算法[J]. 计算机工程与设计,2014,35(2):657 - 660.