【信息科学与工程】

DOI:10.14168/j.issn.1673－4939.2021.02.10

基于改进 Apriori 算法的毕业生就业智能推荐系统

刘丽娜①

（广州工商学院 计算机科学与工程系，广东 广州　510850）

摘　要： 针对高校毕业生就业盲目性、签约速度回笼时间长的问题，研究毕业生行为与就业岗位需求的关系。以广东省若干高校的历届毕业生为研究对象，依托 Spark 计算框架提出对原 Apriori 算法的改进算法 Apriori_S，提高算法的执行效率和分析复杂数据集的适用性，并应用于挖掘历届毕业生数据和就业信息之间的关联规则。实验分析表明，改进算法的执行效率远高于原算法，且适用于分析多维度的复杂数据集。最后利用挖掘结果构建就业岗位“黏性”模型，构建应届毕业生就业智能推荐管理系统，实现就业岗位智能推荐。

关键词： Apriori 算法；复杂数据集；算法优化　‘黏性”模型；智能推荐

中图分类号： TP393　　**文献标志码：** A　　**文章编号：** 1673－4939（2021）02－0130－06

改革开放以来，中国经济不断向好。1999 年我国高校扩招[1]，一定程度上解决了经济过速发展带来高素质人才供不应求的难题，但高校毕业生却很难找到适合自己的工作。《2019 年中国大学生就业报告》指出 2018 届高校毕业生就业率比 2017 届降低 1.1 个百分点，初次就业率连续 4 年呈下降趋势[2]，2018 年末全国毕业生未就业人数约 70 万人，而 2019 年末全国高校毕业生未就业人数比 2018 年多约 17 万人[3]。广东省高校毕业生数量一直呈现稳步上升趋势，2016 至 2018 年初次就业率依次为 95.11%、95.10%、94.18%，整体呈下降趋势。2018 年广东省高校毕业生 57.14 万人，未就业率为 5.82%，约 3 万人[4]。

每年毕业季各种大型现场招聘会琳琅满目，五花八门的招聘网站让人眼花缭乱。招聘信息的杂乱，使得应聘人员基本是“海投式”求职，即通过“广撒网”的方式来增加就业机会，很难准确、高效地找到匹配的工作。由于信息的不对称，毕业生对用人单位的真实可靠性无法准确辨别，出现上当受骗的情形时有发生；再加上近两年的疫情影响，使得就业形势更加严峻，毕业生如何找到适合自己的工作，用人单位如何找到自己需要的人才，变得更加困难。

为此，笔者考虑以网络计算机为媒介，根据待业毕业生和用人单位的需求，实现就业岗位个性化智能推荐，使得待业毕业生和用人单位足不出户，即可初步确认合适的工作岗位和员工。待业毕业生可通过网络计算机智能推荐，线上选择适合自己的就业岗位，而用人单位也可同时获取符合要求的应聘人员信息。

综上所述，笔者采用数据挖掘技术寻找高校毕业生的综合信息与就业单位的岗位需求之间的内在联系，实现高校毕业生及用人单位双方需求的精准引导。笔者依托 Spark 计算框架提出一种可分析多维复杂数据集的关联规则改进算法——Apriori_S（Apriori Algorithm Based on Spark）挖掘广东省若干高校历届毕业生的数据，分析挖掘结果，探索学生行为与就业岗位需求之间互相影响的联系，进而实现就业智能推荐，为高校毕业生就业及教育管理工作提供有意义的参考。

1　Apriori 算法相关研究

数据挖掘算法之关联规则 Apriori 算法，因其简单易用且能挖掘数据的有趣规律使其应用非常广

① 收稿日期：2020－02－07
基金项目：广东高校优秀青年创新人才培养计划资助项目（2018KQNCX309）；教育部 2020 年第一批产学合作协同育人项目（202002191035）
作者简介：刘丽娜（1987—），女，广东汕尾人，硕士，网络工程师，讲师，研究方向：数据挖掘技术与应用、网络技术。

泛，但由于其数据处理效率低下造成研究瓶颈，近几年对 Apriori 算法的改进屡见不鲜。张维国以转换数据集矩阵达到压缩数据量的目的从而提高算法执行效率，并应用于学生数据挖掘从而为学生选课提供个性推荐，但改进后算法对大数据量的挖掘存在一定约束，且无处理复杂数据集的分析说明[5]。郭鹏等结合改进的聚类算法和改进的 Apriori 对学生成绩进行挖掘，通过调整聚类中心使聚类更加稳定，并将得到的聚类结果作为增加兴趣度指标 Apriori 算法的输入，从而得到更准确的挖掘结果，但其大量的前期工作和增加兴趣度指标的改进对 Apriori 算法效率的提高有限[6]。

随着新技术的引入，并行计算框架已是大数据处理的大势所趋。Hadoop 计算框架就是其中之一，通过 Map 和 Reduce 两个算子简单易用的分布执行方式和节点数与运行效率成正比的特性，使其成为近几年应用研究的热点。刘木林等提出基于 Hadoop 平台的 Apriori 算法并行化改进，明显提高了原 Apriori 算法的执行效率，但未明确指出对多维复杂数据集的分析设计[7]。数据量的不断增加且数据处理的复杂性加大，基于内存的 Spark 并行计算框架应运而生，高速计算且支持查询交互的处理特性使其在大数据处理中更占优势。许德心等依托 Spark 计算框架改进 Apriori 算法，其执行效率是基于 Hadoop 平台的十倍以上，但该改进算法同样未指出如何分析复杂数据集[8]。

Apriori 算法执行效率低且一般分析的数据多为类似于“是”与“非”的二维数据类型，针对以上问题笔者提出一种基于 Spark 计算框架、适用于分析多维复杂数据集的改进算法 Apriori_S。然后利用 Apriori_S 算法挖掘往届毕业生信息分析毕业生行为与就业岗位需求的关系，最后构建就业岗位智能推荐系统，实现高校毕业生就业岗位需求智能推荐。

2　基于 Spark 的改进算法设计

笔者依托 Spark 平台对关联规则挖掘算法 Apriori 进行并行设计，同时对 Apriori 算法处理多维复杂数据集进行优化。

设数据集 $D=\{P_1, P_2, \cdots, P_n \mid , n>0$ 且 $n\in N\}$，其中 n 为事务总数，事务项 $P=\{v_1, v_2, \cdots, v_m \mid m>0$ 且 $m\in N\}$，$I\subset D$，m 为数据集的总维度数，v 为事务项中的元素，多维复杂数据集的元素 $v=\{a_1, a_2, \cdots, a_i, b_1, b_2, \cdots, b_j, \cdots, t_1, t_2, \cdots, t_s \mid i, j, s\in N\}$。其中 i、j 和 s 为自然数，可相等也可不等。

2.1　Apriori 算法介绍

Apriori 算法是一种寻找频繁项集的基本算法，通过设置最小支持度 $S_{\min}$，剪枝数据集中支持度小于 $S_{\min}$的元素生成频繁 1 项集 L_1，然后通过自连枝 $L_1\bowtie L_1$ 生成候选频繁 2 项集 C_2，然后重复支持度的对比剪枝步骤，直至候选频繁项集为空。在此基础上，计算由频繁项集产生的非空子集之间的置信度，将其与最小置信度 $C_{\min}$对比，进而产生强关联规则。

支持度 S 和置信度 C 是 Apriori 算法中的两个重要指标:

$$S=\frac{\sum_{x=1}^{n} AR(v_x\rightarrow v_k)}{n}\times 100\%, \tag{1}$$

$$C=\frac{\sum_{x=1}^{n} AR(v_x\rightarrow v_k)}{\sum_{x=1}^{n}(v_x)}\times 100\%。 \tag{2}$$

式（1）和式（2）中，n 为数据集 D 中事务项总数，$\sum_{x=1,k=1}^{m} AR(v_x\rightarrow v_k)$ 表示事务项中的元素 v_x 与 v_k 的同时出现的次数，其中 x，$k=1$，2，…，m，$x\neq k$，m 为数据集的总维度数。

1.2　Apriori_S 算法设计及实现

Spark 是专门分析大数据设计的基于内存计算的并行计算框架，在该计算框架中，master 节点程序启动很多 workers 节点，然后 workers 节点读取数据后转化为 RDD，最后在内存中对 RDD 进行缓存和计算，如图 1 所示。

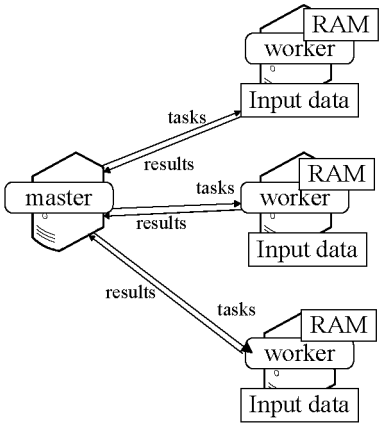

图1 Spark并行计算框架

Apriori_S 算法依托 Spark 环境，数据存储依赖于 HBase，多维复杂数据存储的一般模式如表1所示。

表 1 多维复杂数据一般存储模式

事务项	元素			
	v_1	v_2	…	v_m
P_1	a_1	b_1	…	t_1
P_2	a_2	b_2	…	t_2
…	…	…	…	…
P_n	a_i	b_j	…	t_s

在表 1 中，P_n 为事务项，v_m 为事务项中的维度元素，而 a_i，b_j，…，t_s 为各个维度的不同属性值。

Apriori_ S 的支持度和置信度公式设计见式（3）和式（4）。

$$S_s = \frac{\sum_{x=1}^{n} AR((a_x, b_x, \cdots, t_x) \rightarrow (a_k, b_k, \cdots, t_k))}{n} \times 100\%, \tag{3}$$

$$C_s = \frac{\sum_{x=1}^{n} ((a_x, b_x, \cdots, t_x) \rightarrow (a_k, b_k, \cdots, t_k))}{\sum_{x=1}^{n} (a_x, b_x, \cdots, t_k)} \times 100\%。 \tag{4}$$

式（3）和式（4）中，$AR((a_x, b_x, \cdots, t_x) \rightarrow (a_k, b_k, \cdots, t_k))$ 为多维复杂数据的事务项 p_x 与 p_k 的同时出现，n 为数据库中事务总数，a、b、t 代表各个维度中的不同属性值，$(a_x, b_x, \cdots, t_x)$ 为多维度的事务 p_x，$(a_k, b_k, \cdots, t_k)$ 为多维度的事务 p_k，$x \in N$ 且 $x \leqslant m$，$k \in N$ 且 $x \leqslant m$，m 为数据集的总维度数。

定义 推荐度是支持度与置信度的平均值。

本实验中的推荐度 R 计算公式见式（5）：

$$R = \frac{((n + \sum_{x=1}^{n} (a_x, b_x, \cdots, t_x)) \sum_{x=1}^{n} ((a_x, b_x, \cdots, t_x) \rightarrow (a_k, b_k, \cdots, t_k))) \times 100\%}{2n \sum_{x=1}^{n} (a_x, b_x, \cdots, t_x)} \tag{5}$$

式（5）中的符号定义与式（3）和式（4）相同。

Apriori_S 算法主要设计如下：

输入 多维复杂数据集 D，最小支持度 $S_{\min}$，最小置信度 $C_{\min}$

输出 频繁项集、支持度和置信度

步骤 1. 遍历所有事务项，计算每个维度值的统计数，对于统计数小于最小支持度的属性值进行第一轮剪枝。

步骤 2. 构建 Self - linked（）函数对事务项元素进行自连枝，产生候选频繁项集 C_φ。

步骤 3. 构建 Pruning（）函数对 Self - linked（）函数产生的 C_φ，根据最小支持度 $S_{\min}$（或最小置信度 $S_{\min}$）进行剪枝。

步骤 4. 检查符合最小支持度和置信度并进行支持度与置信度计算，分别用一维数组存储频繁项集的支持度与置信度。

步骤 5. 最后输出 L_φ 频繁项集、支持度、置信度及推荐度。

3 Apriori_S 算法性能评估

笔者结合工作实际，以广东若干院校 2014 届至 2018 届共 5 届的毕业生为研究对象，收集教务管理系统、学生就业信息、校外考证数据、学生选课系统、图书管理数据、一卡通管理系统、人事管理 OA 系统以及评教数据中的相关数据构建数据集。经过清洗数据得到约 15 万个往届毕业生的数据记录，涉及 37 个维度，通过数据概化[9]后得到 148 个属性，总大小约 16 Mb，实验数据扩充至 200 万条、213 Mb。

实验以校企合作企业青软创新科技集团股份有限公司的 QST 青软实训平台提供的虚拟机构建 Spark 集群。集群节点配置均采用 64 位的 CentOS7 操作系统，软件环境变量配置为 Scala2. 11 + Spark2. 2. 1 + Hadoop2. 6. 1 + JDK1. 8. 0_ 151 + Hbase1. 4. 9 以及相关插件。

在最小支持度 $S_{\min}$ = 0. 05 时，对优化前的 Apriori 算法与优化后的 Apriori_S 算法进行弹性测试以评估其性能。通过相同节点数据量逐步递增的弹性对比运行时间如图 2 所示。

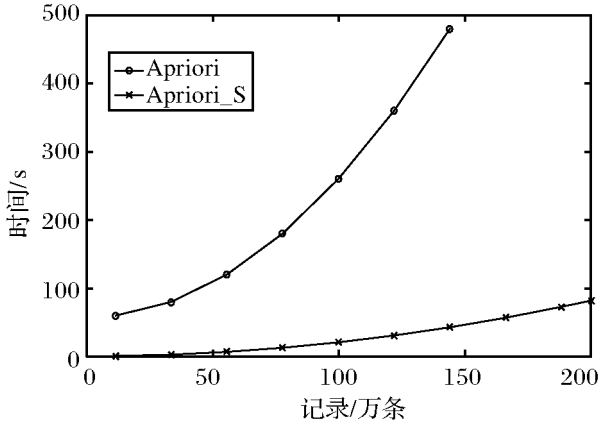

图2 不同算法运行时间对比

从图 2 中可以看出在 Spark 计算框架下的 Apriori_S 算法随着数据量递增其运行效率远比 Apriori 算法高，执行时间几乎是 Apriori 算法的 9%。且随

着数据量的不断增大，原 Apriori 算法的运行时间将不可估量甚至宕机，而 Apriori_S 算法的执行优势却更加明显。

4　Apriori_S 算法的系统应用及验证

教育数据的指数增长以及数据挖掘技术的应用为我们提供一个利用历史数据推动未来发展的契机。往届毕业生的成绩数据、学籍数据、一卡通消费数据、图书借阅数据、选课数据、教师评教数据、学生校外考证数据以及学生的就业数据等等都是来自不同数据源的历史数据。如深入分析这些数据，挖掘出影响学生就业的因素，将会为学校的就业管理决策带来积极影响。

将改进后的 Apriori_S 算法所挖掘出来的结果（抽取部分推荐度大于等于 0.15 的挖掘结果如表 2 所示）应用于高校应届毕业生的就业推荐中，并构建应届毕业生就业岗位“黏性”模型，如图 3 所示。该模型中 H 表示就业单位，S 表示应届毕业生，H_0，H_1，…，H_n 表示不同的就业单位，而 S_0，S_1，…，S_m 表示不同的应届毕业生，在就业单位与学生的“黏性”模型中存在四种情况：学生选择该单位，则由学生向就业单位发出一条虚线；就业单位选择了学生，则由就业单位向学生发出一条点线；学生与就业单位互相选择，则黑色虚线与点线叠加变为线，说明相互之间“黏性”强；而学生与就业单位互无意向则无“黏性”，学生与就业单位无需连线。

表 2　挖掘结果

规则	推荐度/%
专业（建筑工程）→推荐岗位：建筑工程技术人员	83.55
专业（建筑工程），籍贯（广东东莞）→东莞市××建筑工程公司	56.02
专业（建筑工程）→广东××建筑工程有限公司	45.34
获奖情况（有），在校职务（有）→推销、展销人员	35.21
获奖情况（有），在校职务（有）→深圳市××时贸易有限公司	34.33
平均成绩（优），图书浏览记录（频繁）→就业签约时间（快）	75.88

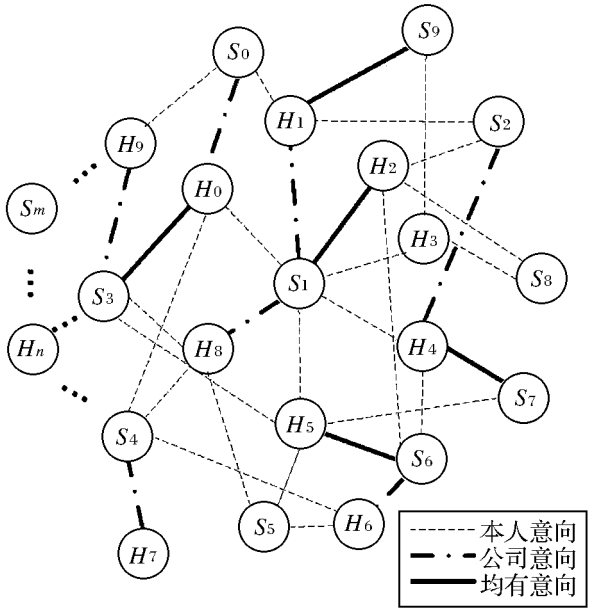

图3 就业岗位“黏性”模型

根据就业岗位“黏性”模型，构建应届毕业生就业智能推荐管理系统。在该系统中导入应届毕业生的各项数据和往届毕业生的就业单位数据，同时，就业单位也可入驻该系统。学生在系统中搜索就业推荐功能，系统即可根据就业岗位“黏性”模型，即挖掘出来的规律和推荐度推荐相应的就业单位及职业类型岗位。学生可以查看就业单位的详细信息并投递简历，此则说明该学生对投递简历的公司有就业意向，在“双方意向”中设置为“就业意向”。就业单位则根据推荐度选择查看学生的信息，并对符合要求的学生传递公司意向，在“双方意向”中设置为“单位意向”。如学生也同时对该公司有意向则可确定双方均有意向，在“双方意向”中设置为“双向选择”，而双方均无意向则无需设置，系统设计流程图如图 4 所示。

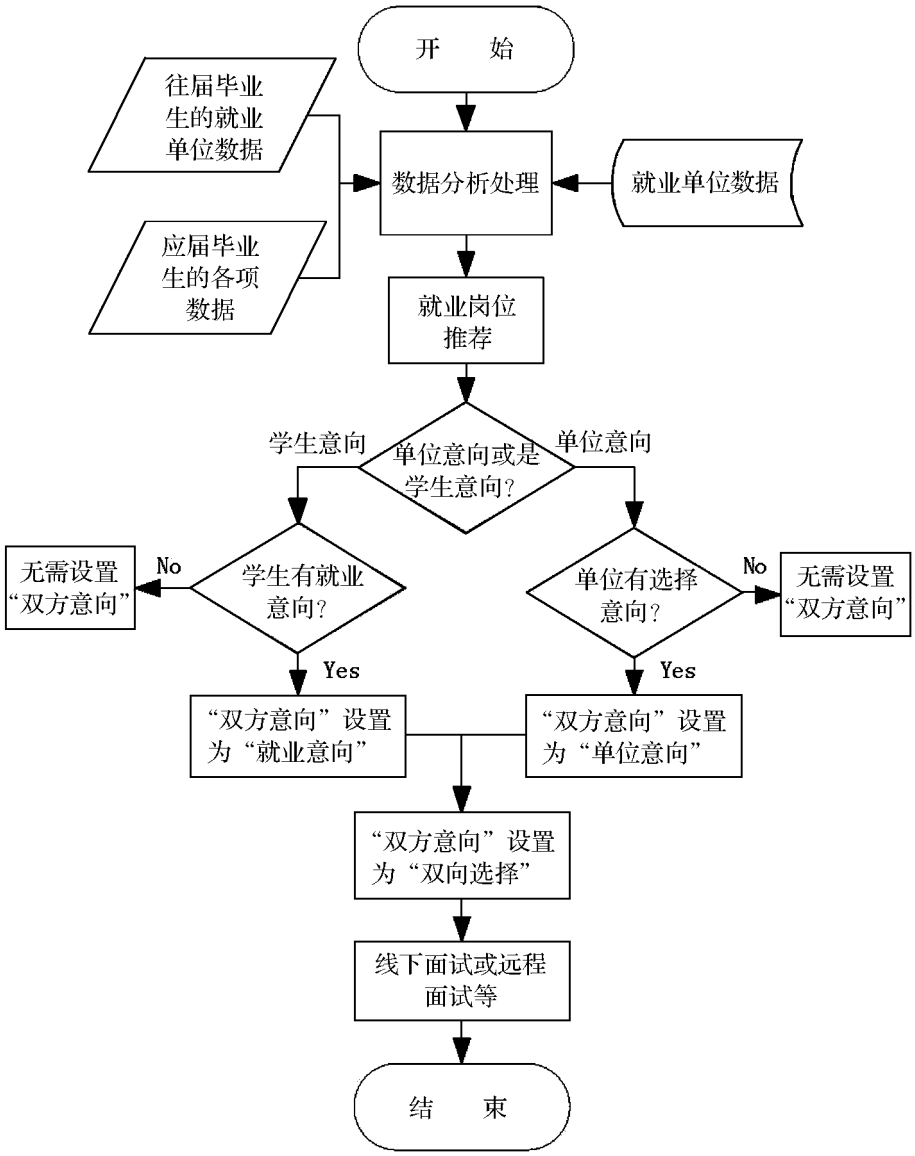

图4 系统设计流程图

该系统在一定程度上能让学生有针对性地快速就业，解决应届毕业生的盲目就业问题，缩短就业签约时间：同时用人单位也可在系统推荐的应届毕业生中查询应聘学生的信息，实现精准定员，缩短人才招揽时间。

结 语

针对高校应届毕业生就业难问题，提出利用数据挖掘技术分析海量往届毕业生数据以挖掘就业规律，从而实现就业智能推荐。本文在基于内存计算的 Spark 并行计算框架下，研究了 Apriori 算法在 Spark 计算框架下处理多维复杂数据的并行优化，通过实验验证了 Apriori_S 算法在大数据分析中存在明显的优势。然后，利用 Apriori_S 算法对广东若干院校 2014 届至 2018 届共 5 届的往届毕业生数据进行挖掘。最后，根据挖掘结果构建就业岗位“黏性”模型，创建应届毕业生就业智能推荐管理系统。通过该系统指导应届毕业生快速就业，缩短就业签约时间，节约用人单位的招聘成本，并为高校就业及教学管理工作提供有益的参考。

参考文献：

[1] 周茂，李雨浓，姚星，等. 人力资本扩张与中国城市制造业出口升级：来自高校扩招的证据 [J]. 管理世界，2019（5）：64－77，198－199.

[2] 麦可思研究院. 数说 [J]. 中国就业，2019（9）：16.

[3] 欧媚. 2020 年高校毕业生总体就业率超 90% [N]. 中国教育报，2021－02－27（01）.

[4] 广东省教育厅就业中心. 2018 年广东省高校毕业生就业质量年度报告 [R]. 广州：广东省教育厅，2018.

[5] 张维国. 面向知识推荐服务的选课决策 [J]. 计算机科学，2019，46（6A）：507－510.

[6] 郭鹏，蔡骋. 基于聚类和关联算法的学生成绩挖掘与分析 [J]. 计算机工程与应用，2019，55（17）：169－179.

[7] 刘木林，朱庆华. 基于 Hadoop 的关联规则挖掘算法研究——以 Apriori 算法为例 [J] 计算机技术与发展，2016，26（7）：1－5.

[8] 许德心，李玲娟. 基于 Spark 的关联规则挖掘算法并行化研究 [J]. 计算机技术与发展，2019，29（3）：30－34.

[9] 刘艳. 基于协同过滤算法实现高校个性化就业推荐系统研究 [J] 现代信息科技，2019，3（15）：10－11，14.

[10] 李昊. 基于深度神经网络的大学生就业情况分析与推荐 [J] 信息通信. 2020（8）：196－199.

An Intelligent Recommendation System for Graduate Employment Based on Improved Apriori Algorithm

LIU Li－na

(*Department of Computer Science and Engineering*, *Guangzhou College of Technology and Business*, *Guangzhou* 510850, *China*)

Abstract: To solve the blindness of college graduates in employment, the relationship between graduate behavior and position demand was studied. Taking the past graduates of several universities in Guangdong Province as the object and with the Spark computing framework as the basis, the authors proposed an improved Apriori_S algorithm to improve the execution efficiency and the applicability of complex data set analysis of the algorithm. The association rules between the graduate data and the employment information were analyzed. The experimental elasticity analysis shows that the execution efficiency of the improved algorithm is much higher than that of the original algorithm, and it is suitable for analyzing multi－dimensional complex data sets. Finally, the results of the mining are used to construct the "stickiness" model of employment positions, and the intelligent recommendation management system for fresh graduates' employment is constructed to realize the intelligent recommendation of jobs.

Key Words: Apriori algorithm; complex data set; algorithm optimization; "stickiness" model; intelligent recommendation

（责任编辑：龙海波）