基于 Apriori 算法的 Web 数据挖掘技术研究

孙德刚 [1]
SUN Degang

摘 要 首先从开发优势与应用原理两个角度阐述数据挖掘技术，进一步分析如今可用的数据挖掘算法；同时，提出一种优化改进的 Apriori 算法。该算法在完成常规数据源选择、数据采集与存储、数据预处理等操作后、构建了一种近邻和决策树两种数据挖掘分类模型，进而在发现频繁项集过程中，将数据建模和模型评估，并对其每个子元素进行计数操作，对子元素的计算数值小于当前的项集阶数时的则进行特征描述标记；最后，对两种模型的混淆矩阵数据进行计算，比较模型的预测准确率，最终得出准确率较高的数据模型，从而提高了算法的整体效率。

关键词 Apriori 算法；Web 数据挖掘；关联规则分析；协同过滤

doi：10.3969/j.issn.1672-9528.2022.07.021

0 引言

Python 属于高级编程语言，其语法结构并不复杂，并自带数据库和 API。Python 各类工具库对数据挖掘有明显的推动作用，依托于 Python，使用经过优化后的 Apriori 算法来提取及挖掘 Web 数据，可运用到多个领域中，促使信息分析及搜集速度得以优化，有效增强数据挖掘技术的效率和精确度。

1 算法优势

Apriori 关联规则挖掘算法目前是数据挖掘领域的一个有效手段，其主要原理是通过特定的关联规则，首先找出所有支持度大于等于最小支持度间值的频繁项集，然后由频繁模式生成满足可信度阀值的关联规则 [1]。

Web 的数据挖掘是从在网页的海量数据内容中，提取出有价值的信息，而后利用挖掘处理，提炼出更深层次的内容，基于对已有信息的分析，完成预测判断。在网络信息科技的持续更新中，数据挖掘使用前景逐渐扩大。虚拟网络中的数据信息，无统一的结构形式，而通过大范围浏览网页的方式，筛选出有用资料，会耗用较长的时间 [2]。在互联网技术逐渐普及中，通过利用数据挖掘，把分散的数据集中起来，按照用户所需功能选择进行信息筛选挖掘。此项技术的数据整合效率极高，并且适应性优秀，借此可反映出各挖掘内容之间的统筹效果，从不同视角发现问题。基于 Python，使用经过优化后的 Apriori 算法，设置网络系统，可控制传送信息期间

1. 山东华宇工学院信息工程学院 山东德州 253000
[基金项目] 山东华宇工学院模式识别应用工程技术研发中心研究基金（No. 201905）

[7] 林子雨 , 郑海山 , 赖永炫 . Spark 编程基础（Python 版)[M]. 北京 : 人民邮电出版社 , 2020.
[8] 郭景瞻 . 图解 Spark: 核心技术与案例实战 [M]. 北京：电子工业出版社 , 2017.
[9] 孙莎莎 . 基于 J2EE 技术的医院互联网管理系统的设计与实现 [J]. 计算机测量与控制 , 2020, 28(8): 177-181.
[10] 孙逢春 . 新能源汽车车联网大数据平台关键技术及应用 [J]. 智能网联汽车 , 2019(1):73-74.
[11] 李怡霖 . 新能源汽车充电管控平台数据挖掘研究 [D]. 大连 : 大连理工大学 , 2019.
[12] 傅筱 , 韩俊毅 , 曹阔 . 大数据驱动的钢铁工业智能故障诊断技术综述 [J]. 计算机测量与控制 , 2020, 28(11):1-5.
[13] 郑振 . 新能源汽车设计中的车联网技术运用之研究 [J]. 时代汽车 , 2018(4):62-63.
[14] 钟佳伶 , 黎茂锋 , 黄俊，等 . 气动数据可视化系统的设计与实现 [J]. 计算机测量与控制 , 2021, 29(2):155-160.
[15] 何平 . 国内新能源汽车监控平台概况 [J]. 河南科技 , 2017(12): 80-82.
[16] 郭先超，林宗缪，姚文勇 . 互联网 + 质量检测平台设计 [J]. 计算机技术与发展 , 2016,26(5):120-124.

【作者简介】
魏晓艳（1978—），女，副教授，陕西蒲城人，研究方向：计算机软件技术、大数据技术方面的教育教学研究。

（收稿日期：2022-05-04 修回日期：2022-06-05）

的错误率，支持筛查遗漏，自动完成补充[3]。特别在面对大量的待处理任务中，可迅速实现归类对接，按照用户动作，改变运行程序。数据挖掘期间，可选出最优的信息传送路径，大幅度缩短传送时间。所以说，数据挖掘技术拥有显著的开发优势，通过今后的技术研究，会达到更加理想的高度，灵活运用各类汇编语言，提高应用效果[4]。

2 应用原理

数据挖掘技术是在实际使用中实现，网络条件下对用户浏览内容的脚本锁定，将各类可搜索到的有用资料进行排序，基于用户的功能申请，组成有用资料的集合，同时，访问系统中的相关信息[5]。在查看 Web 上的有关资料后，会通过信息验证，进一步筛选出有用资料的范围，把页面中的相关信息归集起来，完成对有效信息的归类及整合。其中的信息归类及整合属于执行模块化管理的前提，同时，还是数据挖掘技术的运行原理。在挖掘期间，牵涉到多种爬虫算法，并且最终的算法决策会影响挖掘动作、筛选效率和信息集合的可靠性。对于挖掘功能的实现，在用户进入 Web 页面时的起始页到最后，不间断地进行信息提取，并逐步延伸到外层，以此形成多层次的数据获取链接，完成对数据的捕捉[6]。其中数据挖掘和提取应当属于相对应的动作，在挖掘完成的同时，明确信息来源，基于此，才能开展后续的功能建设。而在提取数据后，把信息传送至既定功能层。页面操作开展中，筛选出有用资料，同时，完成结构性的整合，利用搜索和分析，锁定最后的挖掘数据。

发现频繁项集、挖掘关联规则是经过优化后的 Apriori 算法两大核心步骤。其核心思想是既要找到原始数据中所有符合最低支持度的频繁项集，同时，由于符合标准的频繁集可能为一个，也可能为多个，所以还要尽最大可能找到待挖掘数据最大项数的频繁集[7]。其次，对待挖掘数据进行预处理工作，一是要优化对挖掘最终结果影响不大的字段属性；二是要对字符串值过长的字段进行简化或者使用编码替代；三是要将使用百分制表示的数值型字段转化为分区间的等级制字段[8]。以上三个预处理操作步骤不但可以有效降低数据挖掘过程中内存资源的使用消耗，而且还能有效提高计算比对效率；再次，依据经过优化后的 Apriori 算法的基本挖掘步骤，依此将经筛选后的参数传递方案导入之前已经设计好的算法函数，并需要进行多次优化输入参数值，指导获取最优的挖掘效果[9]；最后，参考此 Apriori 算法关联规则所得到的挖掘结果进行深入细致的梳理与检验，以期达到将最终得到的结论应用于更多的现实数据挖掘的判断与过程决策中去，切实提高数据挖掘的效率和可靠性。

3 数据挖掘技术的爬虫算法

3.1 优先算法

优先算法具体包含广度与深度两项。其中，广度优先算法，会由起始页开始到最后页面，开展从内向外的运算过程。同时汇总多链接数据，数据挖掘期间可直接到达下层，对各项目录实施深度统计分析，保障挖掘对象实现目录的全覆盖。此种爬虫算法的优势是运算精准性较高，但同时由于需要分析大量目录，使得运算时间延长。广度优先算法可精确筛选目录，完成对链接的挖掘，同时，借助并行处理，挖掘 Web 数据，实际提取效果会得到提高。但倘若挖掘对象牵扯到深层目录，会对结果有影响。在深度优先算法中，根据排序从浅至深依次访问，而后开展在其他分支上采取同样的动作，在遍历所有页面链接后，才能判定爬虫工作结束。利用该算法实施深层的数据挖掘，势必会耗用大量的系统资源[10]。

第一步：产生频繁项集，发现满足最小支持度阀值的所有项集，其原理是：如果一个项集是频繁的，则它的所有子集一定也是频繁的，反之，如果一个项集是非频繁项集，那么它的所有超集也是非频繁的。

第二步：产生规则，从上一步发现的频察项集中提取所有高置信度的规则，这些规则称作强规则，通常，须禁项集产生所需的计算资源远大于产生规则所需的计算资源。其具体原理为：由频繁项集生成 k+1 候选项集，即如果一个项集是频繁的，则它的所有子集一定也是频繁的。反之，如果一个项集是非频繁项集，那么它的所有超集也是非频繁的。特别重要，体现算法原理，节约计算资源，有 3 个关键点。

第三步：从 k 候选项集生成 k+1 候选项集。目标是频繁项集，此时，为节约计算资源，非频繁项集能不参与计算，而是让频繁 k 项集两两组合，保留 k+1 候选项集，逐步计算。

3.2 实验准备

爬虫从网络首页开始，计算 Web 页面上的值，评估对应页面的可用性。基于此，先进入更大的页面上，使得爬虫效率得以优化，并能保障遍历成效。但容易出现实际和预计遍历情况有明显偏差的情况，导致挖掘信息结果的准确性下降。该种运算方法可以看成的改良版。在计算开始之前，所有 Web 页面都拥有同等值。在下载结束后，处理页面中，相应的最大值会被均匀分配到各个链接页面，此时爬虫会根据值大小，锁定优先级，先下载值更大的页面。在该运算方法中，不涉及迭代计算的问题，可用在具体运算中。

在数据挖掘中，结构化储存属于比较常用的技术方法，其能分类整合原本毫无章法可言的信息，以呈现明显的结构化特征。从无结构数据中提取信息，处理成一种规则化的链接形式，保存在系统本地。借此可规范数据的存储格式，实际执行期间能借助人工整合处理，形成良好的场景模式。储存信息动作中，实行结构化处理，应保证速率和精准性，不仅需适应多链接信息的挖掘要求，还应按照存储结构进行调节，确保所有链接能迅速整合。通过结构化存储，保障信息综合处理的速度，在 Web 空间中，可自动改变数据结构，利用结构完成有效转换，控制人工操作引起的失误问题。在网

络环境中，结构只是保存信息的形式，在规范结构中，还应注重数据本身的类型。在系统运行时，选择最为迅速的存储形式，确保信息在执行该程序中的完整安全、应用效率。若想实现结构化存储，需确保数据挖掘具备极高的准确性，不仅稳定性及速率需满足应用要求，还应支持自动分类，继而在归类期间，提高综合控制水平，自动化完成信息的结构化存储，形成二维的统计表格，实现有效的功能整合[11]。

3.3 正则表达

在 Web 中，一般是“html”的形式，对应页面包含不同的语义对象，各自带有标识。基于对 html 页面的分析，搭配适宜正则表达，可完成对相应字符串信息的查找与提取。例如，爬取的 Web 页面上涉及“is”的数据源代码，能运用，完成数据提取，基于此自动匹配输出和“is”有关的资料。其中的便属于正则表达式匹配，是一种提取信息供应的可用方法。同时，为满足网站页面不断升级的现象，确保匹配过程的可靠性，用户能直接利用 Python 自带的数据库和模型，完成内容分析及提取。探索用机器学习算法对评论标注 type 的可能性；依据情感词库匹配情感词，计算每条评论的情感值，进而机器标注每条评论的正负类型 type, 用词云图直观呈现正负评论的关键词，初步获得用户的反馈意见。最后利用 gensim 库构建主题挖掘模型，深入了解用户的意见、购买原因、产品的优缺点等。

关联分析（association analysis）是一种在大规模数据集中寻找有趣关系的非监督学习算法，是利用一些有趣性的量度来识别数据库中发现的强规则[12]。这种关系可以有两种形式：频繁项集、关联规则。频繁项集（frequent item sets）是经常出现在一块的物品的集合，关联规则（association rules）暗示两种物品之间可能存在很强的关系[12]。

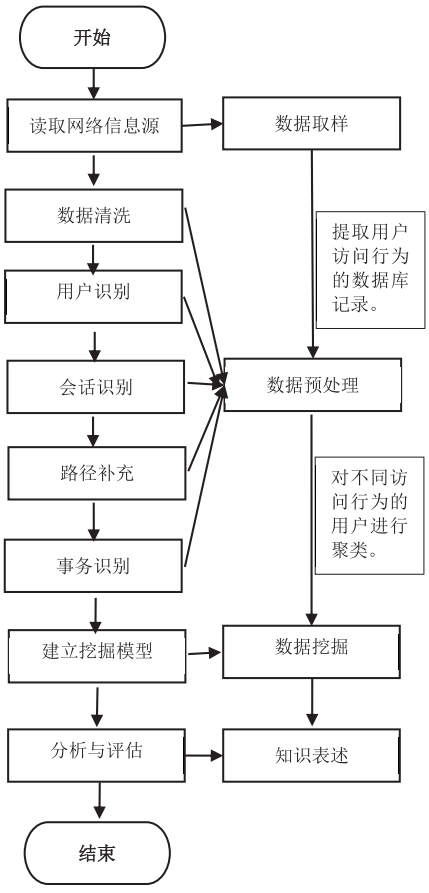

图 1 Apriori 算法测试执行驱动流程

Apriori 算法测试执行驱动流程如图 1 所示。参考知网发布的情感分析用词语集，统计评论数据的正负情感指数，然后进行情感分析，通过词云图直观查看正负评论的关键词，比较“机器挖掘的正负情感”与“人工打标签的正负情感”，采用 LDA 主题模型提取评论关键信息，以了解用户的需求、意见、购买原因、产品的优缺点等。

4 Python 语言下 Web 数据挖掘的规划设计

4.1 爬虫功能

在 Python 的基础上，使用经过优化后的 Apriori 算法，对 Web 页面进行数据挖掘，先要确定具体运行的爬虫功能，按照用户操作使用习惯，结合具体的系统功能诉求以及现有可用爬虫算法的特点。利用 Python 的基础分析，计算期间合理扩大信息的涵盖广度，同时按照各类信息和关键词，在相应浏览访问操作中出现的频率，自动完成信息定位，以此选择后续的语言扩增方法[13]。对于信息结构的规划设计，既应经过稳定性评估比较，又需按照数据抓取期间的链接统计分析情况，完成链条匹配环节。而系统中的爬虫功能，需按照用户浏览访问页面的现实状况，重构页面脚本信息，利用应答服务机制和重构期间取得的超链接，对海量信息资料实行迅速筛查。并且在最后提取与整合数据内容中，要按照分析信息，完成整合操作。数据挖掘时，应当对牵涉到的所有功能实施合理化调整，使数据挖掘能被不断增强，具备多样性化的信息整合水平。在进行信息分析中，需要提取出页面对应的源代码，基于对此的分析，进一步提高信息整合的稳定性。按照不同场合下得到的结果，利用多样性的调整方式，增强数据间的配合效力。

4.2 数据表达

在确定数据表达方式时，应当强调两个方面的问题，即数据挖掘和最后的实际应用要保持稳定；结合数据实际的表达水平。在整体规划中，筛选出最优的表达方式，有效改进信息构建形式。另外，对于数据表达，应当关注调整各页面浏览过程的问题，可基于元数据体系确定，需要从表达方式上，反映出各操控指令间的对接效果。同时，在筛选与构建数据表达期间的方法理念，还应从不同的融合视角出发，彰显其中的综合控制水平，特别在数据表达的规划设计环节中，不同功能的有效融合，发挥出元数据在多样性控制方面的作用，同时能体现出信息挖掘的状态。对于各功能页面，可利用不同结构形式完成数据表达，基于此，捕捉及保存有用的内容。数据表达的规划设计环节，也应利用使用经过优化后的 Apriori 算法实施整体模拟，以落实爬虫功能和提取信息挖掘提取中各功能间的有效控制。通过多样性的整合形式，优化信息的分层结构，完成结构化存储，实现信息挖掘、提取与形式构建的充分整合。在数据表达的规划设计中，增

强信息综合控制的能力，并有效调整系统运转的状态，与此同时在管理环节中，保障信息的综合表达效果，进一步优化表达步骤与形式，最终构成集数据表达和提取为一身的运行模式。

4.3 具体实现

使用经过优化后的 Apriori 算法来实现网站用户评论数据的爬虫编写、后续模型的构建运行，以及数据可视化等操作获取指定音乐下用户的评论数，并运用结巴分词模型进行初步处理 通过 snowNLP 库进行情感分析，对用户的评论进行统计以及积极消极分析，来证实用户评论对音乐情感类别区分的可靠性。其核心实现代码如下所示。

```
C1=[ ]
for transaction in dataSet;
 for item in transaction;
  if not [item] in C1;
 # 遍历所有的元素，如果不在 C1 出现过，那么就 append
    C1.append([item])
 # 对数组进行从小到大的排序
Print 'sort 前 =', C1
C1.sort()
```

frozenset 表示冻结的 set 集合，元素无改变；可以把它当字典的 key 来使用。

```
Print 'sort 后 =', C1
Print 'frozenset=',  map(frozenset,C1)
return map(frozenset,C1)
```

替换语料集 , 通过替换先前版本的数据集，可以提高对商品评论情绪检测的精确性训练数据集对新数据集进行训练，而 snowNLP 是一个处理中文的类库，有中文分词、词性标注、情感分析、文本分类、提取关键词等功能，具体混淆矩阵模型实验对照结果如表 1 所示。

表 1 混淆矩阵模型实验对照表

用户	Url-a1	Url-a2	Url-b1	Url-b2	Url-c1	Url-c2
A、F	5.9	4.0	4.0	40.1	16.0	6.9
B、D、E、G	4.9	5.9	9.8	9.8	6.3	16.0
C	11.0	5.0	0	0	13.0	4.3

爬取评论数据 , 在指定页面后，通过 URL 中的网站代码，对页面的评论数据进行爬取并分类，经初步清洗后 , 对爬取下来的数据进行初步清洗分词，去除无意义的数据，并简化数据，便于接下来的操作，绘制词云图片后，利用清洗后的数据绘制词云图进行情感分析 , 运用自然语言处理模型对数据进行处理进行情感分析。

5 结论

结束语：现如今的网络环境下，对 Web 的数据挖掘，逐渐显现出其不容忽视的工具作用，其提供的各类数据库与计算，使数据挖掘的推进效率不断提高，并保障结果的准确性。本案例首先利用优化改进后的 Apriori 算法应用于文本挖掘，采用决策树算法构建情感分类模型，对于非结构化、碎片化网站数据进行清洗，经优化处理，进而转化为结构化数据，最后对文本数据进一步挖掘与分析，通过词云图直观查看正负评论的关键词，精度达到 89%。

参考文献：

[1] 杨迎 . 基于 Python 语言的 Web 数据挖掘与分析研究 [J]. 现代信息科技 , 2019(23):63-65.

[2] 韦建国 , 王建勇 . 基于 Python 的 Web 数据挖掘应用 [J]. 浙江水利水电学院学报 , 2019(4):79-82.

[3] 何远宏 . 基于 Python 语言的 Web 数据挖掘研究 [J]. 计算机产品与流通 , 2019(1):112.

[4] 曾展挺 . 面向云计算环境下 Web 数据挖掘技术 [J]. 智能计算机与应用 , 2021(1):167-169.

[5] 胡涛 . 基于关联规则的数据挖掘算法 [J]. 电子技术与软件工程 , 2018(2):186.

[6] 牛磊 . 基于数据挖掘的 Web 负载测试用户模型研究 [D]. 哈尔滨：哈尔滨工程大学 , 2019.

[7] 罗芳 , 徐阳 , 蒲秋梅 , 等 . 基于 PageRank 的多维度微博用户影响力度量 [J]. 计算机应用研究 , 2020(5):1354-1358.

[8] 齐慧 . 基于 python 的 WEB 数据挖掘技术实现与研究 [J]. 软件工程 , 2019(8):21-23.

[9] 黄雪华 . 基于 Python 的决策树算法在学生招生录取数据中的应用研究 [J]. 电脑知识与技术 , 2018(29):16-17.

[10] 马振宇 , 张威 , 吴纬 , 等 . 基于 Web 数据挖掘的个性化网络教学平台的研究 [J]. 计算机时代 , 2020(1):84-86+90.

[11] 尚京威 , 陈平 , 韩邢健 . 融合 LDA 的卷积神经网络主题爬虫研究 [J]. 计算机工程与应用 .2019(11):123-128+178.

[12] 景冰 . 基于主题网络爬虫思想的 Web 数据挖掘算法探讨 [J]. 景德镇学院学报 , 2020(3):66-68.

[13] 郁益斌 . 基于训练集聚类的 KNN 算法及其应用研究 [D]. 青岛：山东科技大学 ,2017.

【作者简介】

孙德刚 (1978—)，男，山东聊城人，副教授，高级工程师，研究方向：视频、图像信息处理、人工智能。

（收稿日期：2022-03-12 修回日期：2022-04-09）