# 2002年~2003年关联规则研究情况调研

成员：黄艺杰（组长，25%），金玉洲（25%），左士海（25%），胡叶民（25%）

## 完成调研任务的主要步骤及其截图

（1）观察dblp网页，可以在每个论文的view图标处找到被引信息，如图 1 被引信息入口

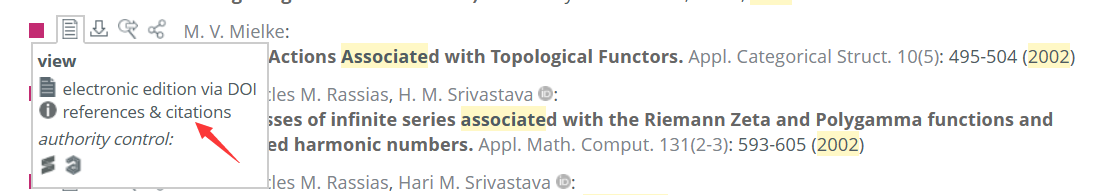


图 1 被引信息入口

（2）点进view图标下的references & citations，找到cited by，可以看到共有三种数据来源可以找到被引信息，如图 2 被引信息来源

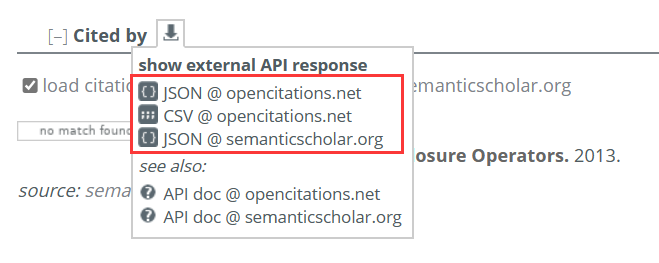


图 2 被引信息来源

（3）通过开发者模式工具找到cited by的下载按钮，观察链接，多找几个就可以发现规律<https://opencitations.net/index/api/v1/citations/10.1023/A:1020534620672?format=json>，只需要<https://opencitations.net/index/api/v1/citations/>+论文编号+?format=json就可以通过爬虫下载到每一篇论文的被引信息，如图 3 链接格式

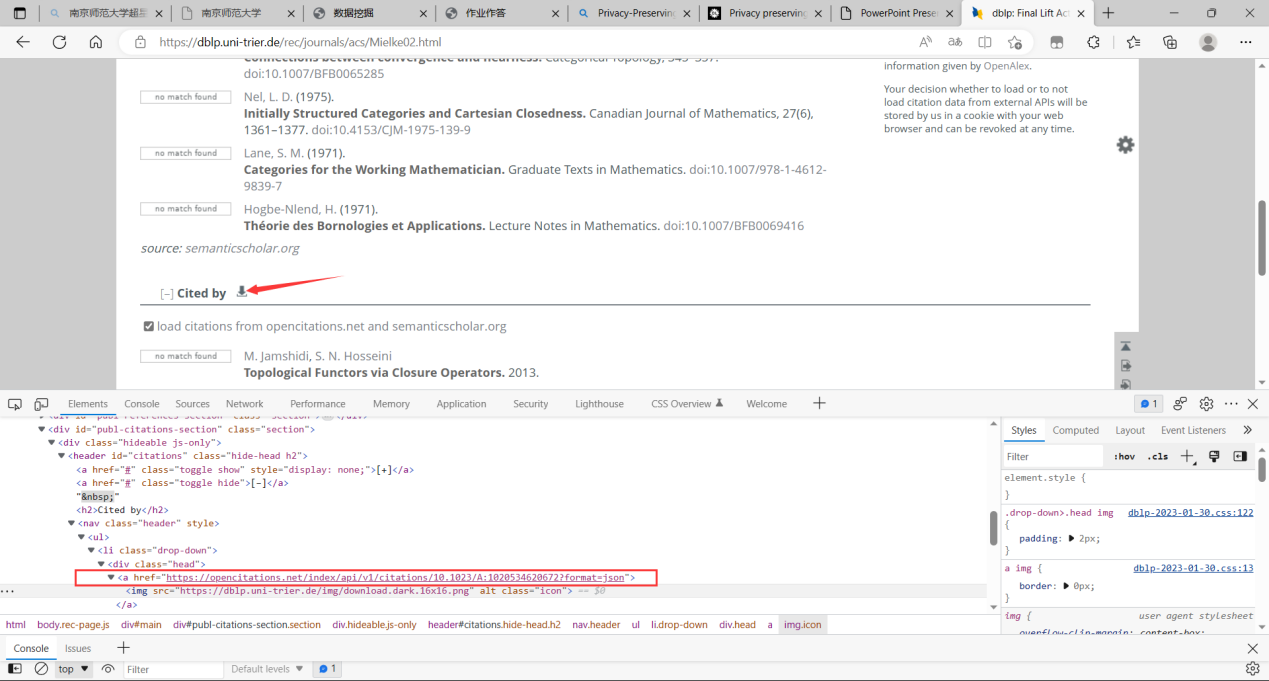


图 3 链接格式

（4）如何找到论文的编号？在主页上用开发者工具找到view图标的element，可以从中获取到论文的编号，如图中所示view链接为<https://doi.org/10.1023/A:1020534620672>，论文编号为10.1023/A:1020534620672，如图 4 获取论文编号

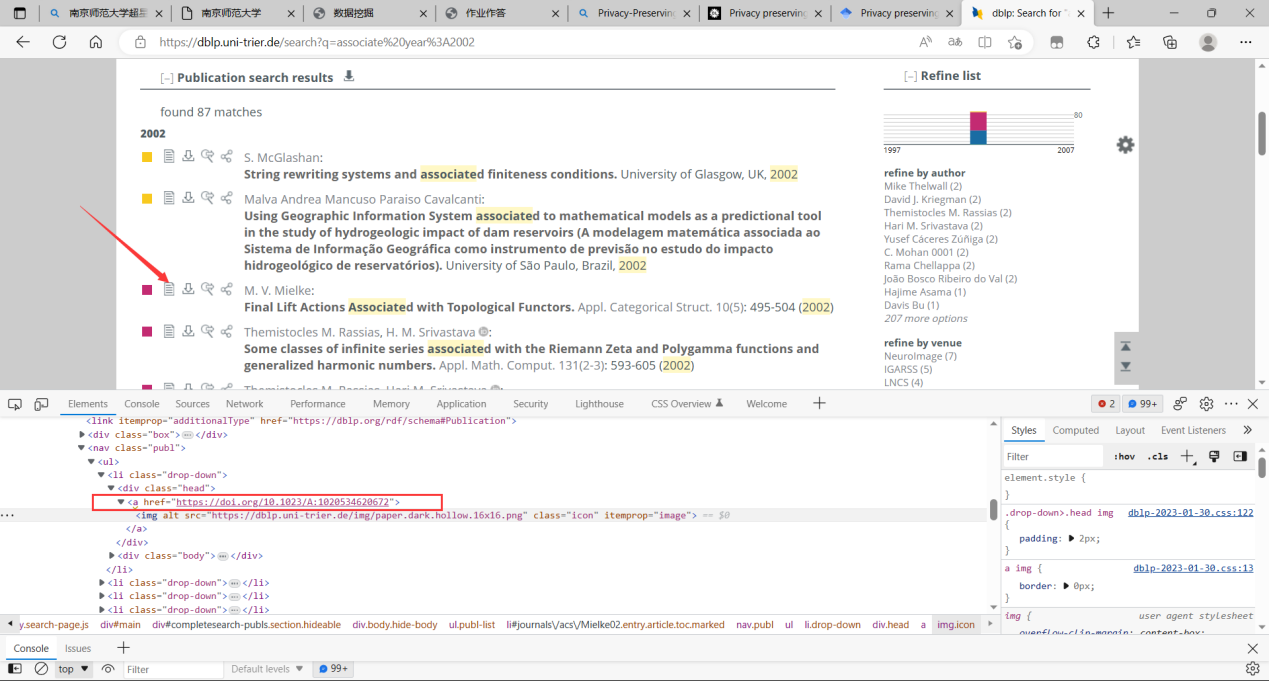


图 4 获取论文编号

遇到的问题：dblp网页是懒加载，需要往下滚动才会翻页

解决方法：打开开发者工具，发现在往下翻页时会触发一个新的get请求，通过比对，请求的链接只需要在最后添加一些参数，其中有一个就是页数，页数为当前论文条数/30，据此可以通过更改页面号来解决懒加载，如图 5 懒加载。

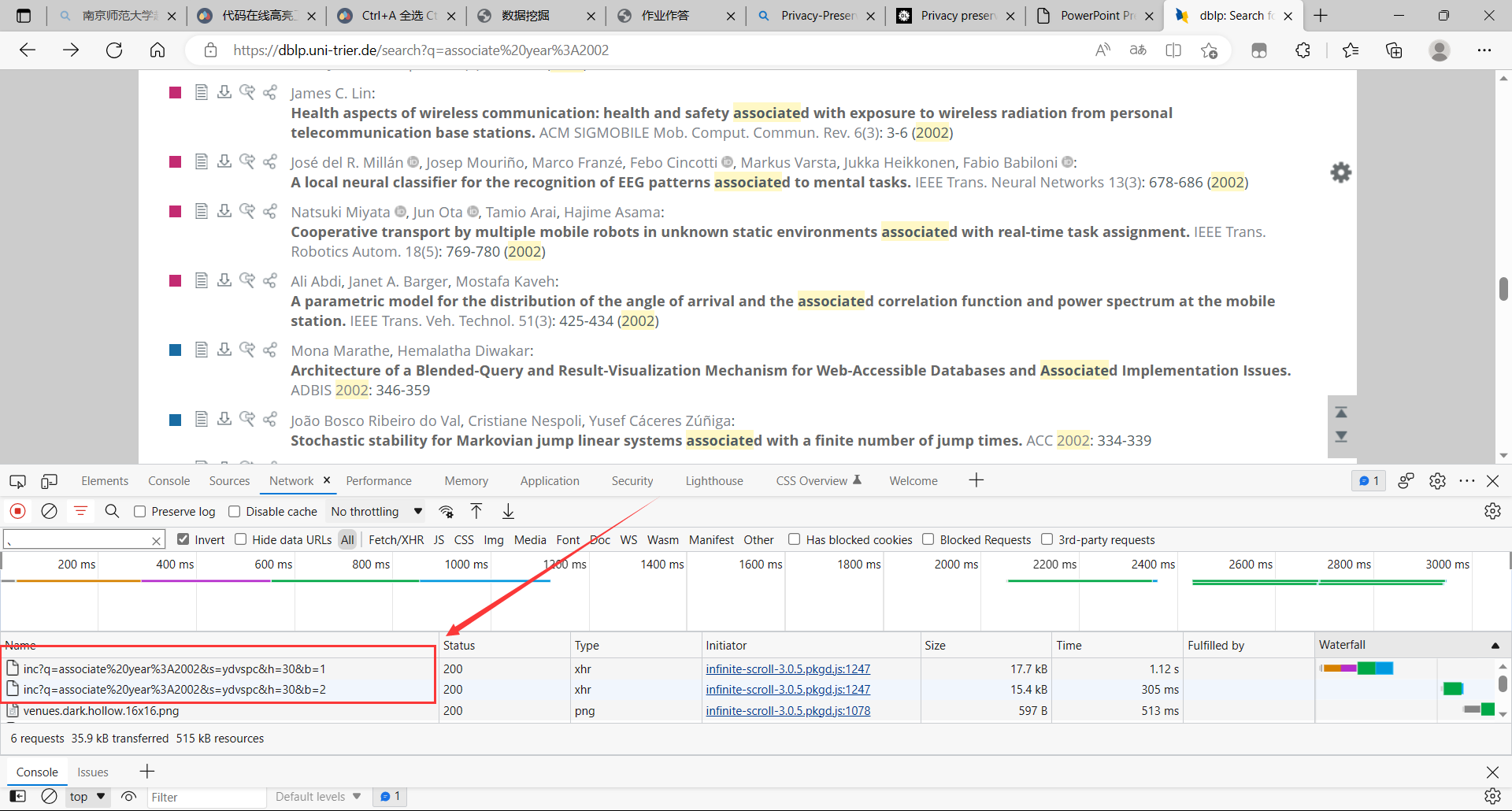


图 5 懒加载

爬虫的流程为：对网页发请求——找到每一个论文编号——按照一定的格式装载后下载被引信息的json文件——统计json文件内信息条数得到被引次数

爬虫代码（爬取2003年association rule相关论文）：

1. import requests
2. import json
3. import csv,unicodecsv
4. from bs4 import BeautifulSoup
5. *# 1. 创建文件对象*
6. with open('2003\_res.csv', 'w',encoding='utf-8-sig', newline='') as f:
7. *# f = open('res.csv','w',encoding='utf-8-sig',newline='')*
8. *# 2. 基于文件对象构建 csv写入对象*
9. *# csv\_writer = csv.writer(f)*
10. csv\_writer=csv.writer(f)
11. *# 3. 构建列表头*
12. csv\_writer.writerow(["序号","论文标题","相关链接","论文编号","被引次数"])
13. lazy\_load=["https://dblp.org/search?q=association%20rule%20year%3A2003",
14. "https://dblp.org/search/publ/inc?q=association%20rule%20year%3A2003&s=ydvspc&h=30&b=1",
15. "https://dblp.org/search/publ/inc?q=association%20rule%20year%3A2003&s=ydvspc&h=30&b=2",
16. "https://dblp.org/search/publ/inc?q=association%20rule%20year%3A2003&s=ydvspc&h=30&b=3",
17. "https://dblp.org/search/publ/inc?q=association%20rule%20year%3A2003&s=ydvspc&h=30&b=4",
18. "https://dblp.org/search/publ/inc?q=association%20rule%20year%3A2003&s=ydvspc&h=30&b=5"]
19. *# lazy\_load=["https://dblp.uni-trier.de/search?q=associate%20year%3A2002"]*
20. num=0
21. title\_list = []
22. rows=[]
23. for lazy\_url in lazy\_load:
24. response=requests.get(lazy\_url)
25. response=response.text
26. *# print(response)*
27. soup = BeautifulSoup(response, 'html.parser')
28. links = soup.findAll("nav", {"class": "publ"})
29. name = soup.findAll("span", {"class": "title", "itemprop": "name"})
30. for each in name:
31. title=each.get\_text()
32. print(title)
33. title\_list.append(title)
34. for link in links:
35. each\_row = []
36. num=num+1
37. print("第"+str(num)+"个论文")
38. each\_row.append(num)
39. print("论文标题:")
40. print(title\_list[num-1])
41. each\_row.append(title\_list[num-1])
42. li=link.findAll("li", {"class", "drop-down"})[0]
43. *# print(li)*
44. a=li.findAll("a")[0]
45. a=str(a)
46. print("相关链接："+a)
47. each\_row.append(a)
48. left=a.find("doi.org/")+8
49. a=a[left:a.find(">")-1]
50. *# print(response.text)*
51. print("论文编号："+a)
52. each\_row.append(a)
53. if '9' >= a[0] >= '0':
54. citedByURL="https://opencitations.net/index/api/v1/citations/"+a+"?format=json"
55. res\_json = json.loads(requests.get(citedByURL).text)
56. print("被引信息：",end="")
57. print(res\_json)
58. *# each\_row.append(res\_json)*
59. print("被引次数"+str(len(res\_json)))
60. each\_row.append(str(len(res\_json)))
61. else:
62. print("被引信息暂无")
63. *# each\_row.append("被引信息暂无")*
64. print("被引次数暂无")
65. each\_row.append("被引次数暂无")
66. csv\_writer.writerow(each\_row)
67. *# rows.append(each\_row)*
68. *# csv\_writer.writerows(rows)*
69. f.close()

最终爬虫处理结果整合进csv文件，如图 6 爬虫结果

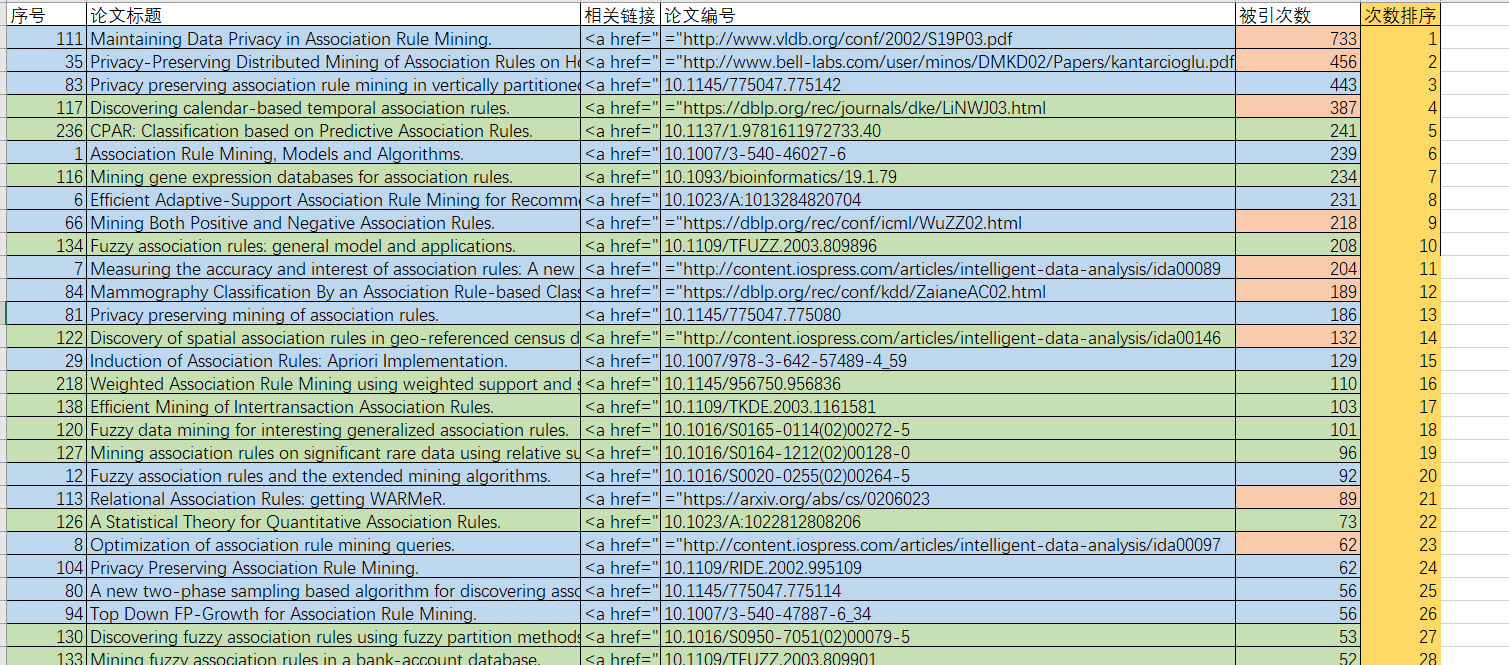


图 6 爬虫结果

## 调研结果

### 引用最多的10篇论文列表（包含作者、题目、期刊或者会议名称、被引用次数）

#### 1.Maintaining Data Privacy in Association Rule Mining

链接：https://www.vldb.org/conf/2002/S19P03.pdf

作者：Shariq J. Rizvi, Jayant R. Haritsa

期刊或会议名称：VLDB '02: Proceedings of the 28th international conference on Very Large Data Bases

引用次数：733

#### 2.Privacy-Preserving Distributed Mining of Association Rules on Horizontally Partitioned Data.

链接：http://www.bell-labs.com/user/minos/DMKD02/Papers/kantarcioglu.pdf

作者：Murat Kantarcioglu, Chris Clifton

期刊或会议名称：2002 {ACM} {SIGMOD} Workshop on Research Issues in Data Mining and Knowledge Discovery, Madison, Wisconsin, USA, June 2, 2002

引用次数：456

#### 3.Privacy preserving association rule mining in vertically partitioned data.

链接：https://doi.org/10.1145/775047.775142

作者：Jaideep Vaidya, Chris Clifton

期刊或会议名称：Proceedings of the Eighth {ACM} {SIGKDD} International Conference on Knowledge Discovery and Data Mining, July 23-26, 2002, Edmonton, Alberta, Canada

引用次数：443

#### 4.Discovering calendar-based temporal association rules.

链接：https://doi.org/10.1109/TIME.2001.930706

作者：Yingjiu Li, Peng Ning, Xiaoyang Sean Wang, Sushil Jajodia

期刊或会议名称：Eigth International Symposium on Temporal Representation and Reasoning, TIME-01, Civdale del Friuli, Italy, June 14-16, 2001

引用次数：387

#### 5.CPAR: Classification based on Predictive Association Rules

链接：https://doi.org/10.1137/1.9781611972733.40

作者：Xiaoxin Yin, Jiawei Han

期刊或会议名称：Proceedings of the 2003 SIAM International Conference on Data Mining

引用次数：241

#### 6.Mining gene expression databases for association rules

链接：https://doi.org/10.1093/bioinformatics/19.1.79

作者：Chad Creighton, Samir Hanash

期刊或会议名称：期刊 CELL: Bioinformatics, Volume 19, Issue 1, January 2003, Pages 79–86

引用次数：234

#### 7.Efficient Adaptive-Support Association Rule Mining for Recommender Systems

链接：https://doi.org/10.1023/A:1013284820704

作者:W Lin, SA Alvarez, C Ruiz

期刊或会议名称：:Data mining and knowledge discovery, 2002 - Springer

引用次数:231次

#### 8.Mining both Positive and Negative Association Rules

链接：https://dblp.org/rec/conf/icml/WuZZ02.html

作者:Xindong Wu, Chengqi Zhang , Shichao Zhang

期刊或会议名称：:ICML-2002

引用次数:218次

#### 9.Fuzzy Association Rules: General Model and Applications

链接：https://doi.org/10.1109/TFUZZ.2003.809896

作者:Miguel Delgado, Nicolás Marín, Daniel Sánchez, María-Amparo Vila

期刊或会议名称：:IEEE TRANSACTIONS ON FUZZY SYSTEMS, VOL. 11, NO. 2, APRIL 2003

引用次数:208次

#### 10.Measuring the accuracy and interest of association rules: A new framework

链接：http://content.iospress.com/articles/intelligent-data-analysis/ida00089

作者：Berzal, Fernando Blanco, Ignacio, Sánchez, Daniel

出处：期刊 Intelligent Data Analysis, vol. 6, no. 3, pp. 221-235, 2002

引用次数：204

### 主要研究内容及其分类

#### 主要研究内容

##### 1.Maintaining Data Privacy in Association Rule Mining

链接：https://www.vldb.org/conf/2002/S19P03.pdf

作者：Shariq J. Rizvi, Jayant R. Haritsa

期刊或会议名称：VLDB '02: Proceedings of the 28th international conference on Very Large Data Bases

引用次数：733

主要内容：在关联规则挖掘中，保护用户隐私是一个重要的问题。本文提出了一种新的保护用户隐私的关联规则挖掘算法，称为MDL(Maintaining Data Privacy)算法。MDL算法基于数据扰动技术，通过给原始数据添加噪声的方式来保护敏感数据。具体地，MDL算法中，在保证数据准确性的前提下，通过改变数据分布来提高数据隐私性。本文还在公开数据集上进行了实验，实验结果表明，MDL算法能够有效地提高数据隐私性，同时保持较高的准确性。因此，MDL算法具有在关联规则挖掘领域应用的潜力。

关键词：关联规则挖掘、数据隐私、差分隐私、隐私保护、数据保护

##### 2.Privacy-Preserving Distributed Mining of Association Rules on Horizontally Partitioned Data.

链接：http://www.bell-labs.com/user/minos/DMKD02/Papers/kantarcioglu.pdf

作者：Murat Kantarcioglu, Chris Clifton

期刊或会议名称：2002 {ACM} {SIGMOD} Workshop on Research Issues in Data Mining and Knowledge Discovery, Madison, Wisconsin, USA, June 2, 2002

引用次数：456

主要内容：数据挖掘可以从大型数据集中提取重要知识，但有时这些数据集被分散在不同的方。出于隐私考虑，各方可能无法直接共享数据和某些类型的关于数据的信息。我们解决了横向划分数据的安全关联规则挖掘问题。该方法采用加密技术最小化共享的信息，同时为挖掘任务增加很少的开销。

关键词：隐私保护 、分布式计算、关联规则挖掘、水平分区数据、加密技术、安全多方计算

##### 3.Privacy preserving association rule mining in vertically partitioned data.

链接：https://doi.org/10.1145/775047.775142

作者：Jaideep Vaidya, Chris Clifton

期刊或会议名称：Proceedings of the Eighth {ACM} {SIGKDD} International Conference on Knowledge Discovery and Data Mining, July 23-26, 2002, Edmonton, Alberta, Canada

引用次数：443

主要内容：隐私方面的考虑通常会限制数据挖掘项目。本文解决了事务跨源分布的关联规则挖掘问题。每个站点都持有每笔交易的一些属性，并且这些站点希望协作来识别全局有效的关联规则。但是，网站不得透露个人交易数据。我们提出了一种两方算法，用于有效地发现具有最低支持水平的频繁项集，而没有任何一个站点透露单独的交易价值。

关键词：隐私保护、关联规则挖掘、垂直分区数据、加密技术、Apriori算法、数据拆分

##### 4.Discovering calendar-based temporal association rules.

链接：https://doi.org/10.1109/TIME.2001.930706

作者：Yingjiu Li, Peng Ning, Xiaoyang Sean Wang, Sushil Jajodia

期刊或会议名称：Eigth International Symposium on Temporal Representation and Reasoning, TIME-01, Civdale del Friuli, Italy, June 14-16, 2001

引用次数：387

主要内容：一个时间关联规则是在特定时间间隔内成立的关联规则。例如，早上鸡蛋和咖啡的销售经常同时发生。本论文研究了用户给定的日历模式下的时间关联规则。通常使用日历模式可以使发现的时间关联规则更易理解。一个日历模式的例子是 (年，月，日)，它可以产生形如 (d/sub 3/, d/sub 2/, d/sub 1/) 的一组基于日历的模式，其中每个d/sub i/ 可以是一个整数或符号 \*。例如，(2000, \*, 16) 就是这样一个模式，对应于时间间隔，每个间隔都由 2000 年某个月的第16天组成。本文定义了两种类型的时间关联规则：精确匹配关联规则要求关联规则在每个时间段内都成立，模糊匹配要求关联规则在大多数时间段内成立。本文扩展了著名的Apriori算法，并开发了两种优化技术来利用基于日历的模式的特殊属性。实验表明，算法和优化技术是有效的。

关键词：关联规则、日历、交易数据库、信息系统、数据挖掘、数据分析、合同

##### 5.CPAR: Classification based on Predictive Association Rules

链接：https://doi.org/10.1137/1.9781611972733.40

作者：Xiaoxin Yin, Jiawei Han

期刊或会议名称： Proceedings of the 2003 SIAM International Conference on Data Mining

引用次数：241

主要内容：提出了一种基于预测关联规则的分类算法，即CPAR（Classification based on Predictive Association Rules）。CPAR算法将关联规则挖掘技术与分类任务相结合，旨在提高分类准确性和关联规则的可解释性。它首先从训练数据集中提取出与目标变量（类别）相关的关联规则，然后使用这些规则来进行分类预测。CPAR算法还具备对抗噪声和处理高维数据的能力，适用于各种分类问题。

关键词：数据挖掘、分类算法

##### 6.Mining gene expression databases for association rules

链接：https://doi.org/10.1093/bioinformatics/19.1.79

作者：Chad Creighton, Samir Hanash

出处：期刊 CELL: Bioinformatics, Volume 19, Issue 1, January 2003, Pages 79–86

引用次数：234

主要内容：使用 Hughes 等人的数据集展示了一种从基因表达数据中有效挖掘关联规则的算法。使用该算法，在数据中发现了许多规则。对其中一些规则的粗略分析揭示了某些基因之间的许多关联，其中许多在生物学上是有意义的，其他基因则提出了可能值得进一步研究的新假设。在源自酵母数据集的数据集中，但每个转录本的表达值相对于实验随机移动，没有发现任何规则，表明从实际数据集中挖掘的大多数规则不太可能具有偶然发生的。

关键词：数据挖掘、计算机生物学、算法

##### 7.Efficient Adaptive-Support Association Rule Mining for Recommender Systems

链接：https://doi.org/10.1023/A:1013284820704

作者:W Lin, SA Alvarez, C Ruiz

期刊或会议名称：:Data mining and knowledge discovery, 2002 - Springer

引用次数:231次

主要内容:文章实际上介绍了一种不同于常见的推荐系统中应用关联规则挖掘的常用方式类似于之前实验所设计的超市购物篮的例子，不存在前者需要提前指定最小支持度以及效率低下等问题，作者所叙述的算法不设定最小支持度而是给出规则数量的目标范围，同时算法可以调整不同用户的最小支持度以获得大小在期望范围的规则集。即算法所挖掘的关联规则不再具有普遍性，而是针对特定用户，即推荐系统的使用者或特定项目。这不仅保证了为高质量推荐提供足够的规则，同时也避免了传统关联规则挖掘预先指定最小支持度所带来的巨大的时间代价。并在最后对该算法性能进行评估以展示其确实由于传统关联规则挖掘算法的事实。

关键词:关联规则挖掘在推荐系统中的应用

##### 8.Mining both Positive and Negative Association Rules

链接：https://dblp.org/rec/conf/icml/WuZZ02.html

作者:Xindong Wu, Chengqi Zhang , Shichao Zhang

期刊或会议名称：:ICML-2002

引用次数:218次

主要内容:文章重点围绕负关联规则介绍与定义，不频繁项集，提取正负面规则以及相关的工作和实验结果进行叙述。通过举出相关例子展开叙述。例如对于两个频繁集t和c，t∪c不是频繁集但t∪~c的支持度可能会很高。这就给关联规则挖掘提供了新的思路。在上面的例子中t=>~c就是一个有效的规则，称为具有潜在意义而t∪c则称为潜在相关项集。找到不频繁项集才能挖掘出负关联规则。同样参照关联规则中的置信度，定义Interest（X,Y）=|supp（X∪Y）-supp（X）\*supp（Y）|以及最小interest mi。那么这样就可以参照关联规则定义出负关联规则A=>~B和不频繁项集.同时文章也给出了负关联规则的挖掘方法，并通过程序和算法来演示提取负关联规则的过程及方法。

关键词:负关联规则挖掘

##### 9.Fuzzy Association Rules: General Model and Applications

链接：https://doi.org/10.1109/TFUZZ.2003.809896

作者:Miguel Delgado, Nicolás Marín, Daniel Sánchez, María-Amparo Vila

期刊或会议名称：:IEEE TRANSACTIONS ON FUZZY SYSTEMS, VOL. 11, NO. 2, APRIL 2003

引用次数:208次

主要内容:文章重点关注模糊关联规则的通用模型和实际应用。文章主体分为两大部分，除去开头对模糊集，模糊关联规则，模糊事务等的概念的介绍，分别依照模糊集的通用模型以及具体化这个通用模型后的一些应用示例展开。通用模型的开发以发现模糊事务中的模糊关联规则的挖掘。通用模型中具体展示了相当于关联规则中的支持度，置信度，准确性和重要性等概念在模糊关联规则挖掘中的具体含义和实现方法，以及如何对关联规则中常用算法如Apriori，FPGrowth等的修改以期其可以同样运用在挖掘模糊关联规则的事务当中。如在支持度和置信度一节中将原来的框架进行推广：采用基于量化句子评估语义的方法。其中量化句子是“Q of F are G”形式的表达式。F,G是有限集X的两个模糊子集，而Q是相对模糊量词。而应用部分主要介绍了关系数据库中的模糊关联规则和模糊和近似函数依赖两个部分，前者由关系数据库中关联规则挖掘的定量关联规则的局限和弊端引出模糊关联规则这一解决办法，通过在定量属性中由模糊集表达有意义的语义标签，并使其成为新的域，解决传统解决方法—聚类的值含义不清晰的弊端以及规则对边界微小变化敏感的问题。

关键词:模糊关联规则

##### 10.Measuring the accuracy and interest of association rules: A new framework

链接：http://content.iospress.com/articles/intelligent-data-analysis/ida00089

作者：Berzal, Fernando Blanco, Ignacio, Sánchez, Daniel

出处：期刊 Intelligent Data Analysis, vol. 6, no. 3, pp. 221-235, 2002

引用次数：204

主要内容：支持度和置信度作为重要性和准确性的衡量标准来评估关联规则的常用框架有几个缺点，该论文提出了一个新的的框架，该框架基于 Shortliffe 和 Buchanan 的确定性因素以及非常强规则的新概念，讨论了新框架的一些直观属性。进行的理论特性和实验都表明，可以避免发现误导性规则，提高结果的可管理性和质量。

关键词：数据挖掘、机器学习

#### 论文分类

将这10篇论文按照研究主题进行分类，可以分为以下5类：

##### 1.数据隐私保护和关联规则挖掘

Maintaining Data Privacy in Association Rule Mining

Privacy-Preserving Distributed Mining of Association Rules on Horizontally Partitioned Data

Privacy preserving association rule mining in vertically partitioned data

这三篇论文的研究都围绕着如何在保护数据隐私的前提下进行关联规则挖掘。它们介绍了不同的方法来实现数据安全挖掘，包括对数据进行加密、分区和转换等。这些方法能够保护敏感信息，并增强数据安全性。

##### 2.时间关联规则挖掘

Discovering calendar-based temporal association rules

该论文介绍了一种基于日历的时间关联规则挖掘方法，它使用日历模式和基于日历的模式，对给定数据中存在的时间间隔进行过滤和筛选，以提高挖掘效率并减少结果中的噪声。该方法可以应用于推荐系统或其他相关领域中时间序列分析和事件流分析的场景。

##### 3.基于关联规则的分类

CPAR: Classification based on Predictive Association Rules

该论文介绍了一种基于预测性关联规则的分类方法，即使用概率映射函数来将交易映射到预测的类别。该方法可以应用于许多具有高维离散属性的数据集，并且可以大大减少分类错误率。

##### 4.关联规则挖掘和算法

Association Rule Mining, Models and Algorithms

Mining gene expression databases for association rules

Efficient Adaptive-Support Association Rule Mining for Recommender Systems

Mining Both Positive and Negative Association Rules

这四篇论文介绍了关联规则挖掘和算法及其在不同领域的应用，包括基因表达数据挖掘、推荐系统和消费市场等。它们分别讨论了如何选择合适的数据集、特征提取和建模等方面的问题，并提出了相应的解决方案。其中的算法和模型也能够帮助研究者更好地理解并应用关联规则挖掘。

##### 5.模糊关联规则挖掘

Fuzzy association rules: general model and applications

该论文讨论了关联规则挖掘中的模糊性质，并提出了一种通用模糊关联规则模型。该模型可以应用于各种数据集和领域，并且可以找到不同的方式来定义支持和置信度等关联规则参数。