计算机系统应用 ISSN 1003-3254, CODEN CSAOBN
Computer Systems & Applications,2017,26(7):50−55 [doi: 10.15888/j.cnki.csa.005877]

E-mail: csa@iscas.ac.cn
http://www.c-s-a.org.cn
Tel: +86-10-62661041

基于关联规则的自然灾害预测系统①

李汉巨, 梁浩波

(广东电网有限责任公司东莞供电局, 东莞 523008)

摘　要: 建立自然灾害预测模型, 对自然灾害进行预测和分析, 有利于提升防灾减灾的技术水平. 基于关联规则和Web文本挖掘技术提出自然灾害预测系统的设计方案及实现方法. 该系统利用成熟开源的爬虫框架从权威的灾害信息发布平台中定向抓取非结构化的自然灾害信息, 通过中文分词技术进行数据清理将其整理成结构化的自然灾害数据库, 并利用改进的关联规则算法从中挖掘出自然灾害事件的关联规则, 进而可通过实时监控关联规则的前端信息, 实现对自然灾害事件的预测. 试运行结果表明该系统能有效挖掘出自然灾害信息的关联规则, 并具有较高置信度.

关键词: Web文本; 自然灾害; 灾害预测; 关联规则; 文本挖掘

引用格式: 李汉巨,梁浩波.基于关联规则的自然灾害预测系统.计算机系统应用,2017,26(7):50–55. http://www.c-s-a.org.cn/1003-3254/5877.html

Natural Disaster Forecasting System Based Association Rules

LI Han-Ju, LIANG Hao-Bo

(Guangdong Power Grid Co. Ltd., Dongguan Power Supply Bureau, Dongguan 523008, China)

Abstract: The establishment of natural disaster prediction model to predict and analyze the occurrence of natural disasters is conducive to enhance the technical level of disaster prevention and mitigation. We present a design and implementation of natural disaster forecasting system based on association rules and the Web text mining technology. The system uses a mature open source crawler framework to capture the unstructured natural disaster information from the authoritative disaster information release platforms. By using the Chinese word segmentation technique the data are cleaned up and organized into a structured natural disaster database, and mining association rules of natural disaster events are worked out from improved association rule algorithm. And then by monitoring the front-end information of association rules in real time, the prediction of natural disasters can be achieved. Experimental results show that the system can effectively mine the association rules of natural disaster information with high confidence.

Key words: Web text; natural disaster; disaster prediction; association rules; text mining

1　引言

Web文本中蕴含丰富的以自然语言描述的非结构化自然灾害信息[1-4]. 通过Web文本挖掘技术为灾害数据源, 自动抽取并整理包含灾害事件类型、时间、空间位置以及影响范围等结构化的自然灾害信息, 是对传统结构化灾害数据库的重要补充[5,6]. 而基于Web文本自动提取和构造结构化、综合性灾害信息, 是灾害信息领域研究的前沿问题[7], 目前国内外利用Web文本挖掘技术在灾时与灾后的应急响应与救援, 灾害的早期预警和风险分析等方面开展应用研究.

从自然灾害发生机理研究发现, 同一地区不同类型自然灾害的发生、不同地区同一类型自然灾害的发生以及不同地区不同类型自然灾害的发生之间存在着联系[8]. 一方面, 通过对区域范围内大量积累的、文本

① 基金项目: 广东电网有限责任公司职工创新项目(GDZC-031920160256)
收稿时间: 2016-11-18; 收到修改稿时间: 2017-01-04

语言记录的历史灾害信息进行分析和挖掘, 有利于发现灾害事件存在的联系, 进而对灾害发生的类型、时空分布特征进行分析, 为不同地域空间的自然灾害事件发生的关联性提供决策支持. 另一方面, 在突发灾害事件下, 需要针对事件可能发生的前兆和演化过程的数据进行快速收集获取、整理, 以实现对自然灾害的预警、预测, 提升自然灾害的应急处理能力.

基于上述背景, 本文设计并研发了自然灾害预测系统, 该系统利用成熟开源爬虫框架(WebMagic)从权威的灾害信息发布平台中定向抓取非结构化的自然灾害信息, 并利用中文分词技术进行数据清理将其转换成结构化的自然灾害数据库. 接着通过关联分析算法从中挖掘出不同区域灾害事件发生的关联规则, 最后通过实时监控某灾害事件发生的前兆, 结合关联规则, 进而实现自然灾害事件的预测.

2 自然灾害预测系统设计

2.1 系统功能及设计

自然灾害预测系统实现如下功能:

(1) 定向抓取自然灾害Web文本信息, 通过数据清理, 形成结构化数据, 并存储在MySQL数据库.

(2) 利用改进关联规则算法对MySQL数据库的数据进行挖掘, 产生关联规则库.

(3) 利用关联规则库和实时抓取的自然灾害Web文本信息监测关联规则的前端信息, 实现对关联规则的后端信息的预测.

系统设计如图1所示.

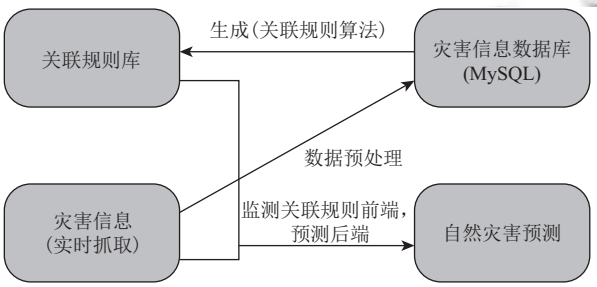

图1 系统设计

用户界面设计如图2所示.

2.2 Web技术架构

采取如图3所示的Web技术架构.

2.3 自然灾害预测技术原理

关联规则采取形如X→Y(时间间隔)的蕴涵式, X和Y分别称为关联规则的前端和后端, 该关联规则表示X发生, 经过时间间隔后, Y将发生. 比如关联规则:

桂林_暴雨→东莞_台风(15天)

表示桂林发生暴雨15天后东莞将发生台风. 因此假设上述关联规则成立, 那么只要监测到桂林发生暴雨, 那么就可以预测15天后东莞发生台风.

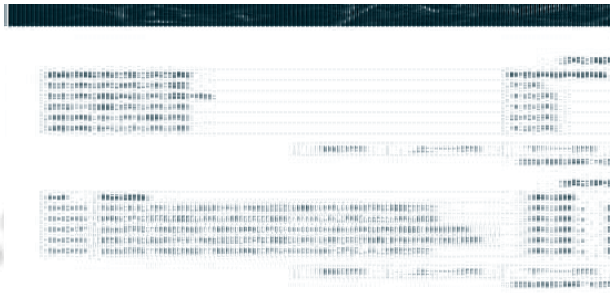

图2 用户界面设计

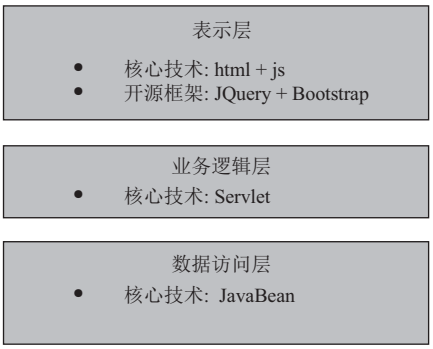

图3 Web技术架构

因此自然灾害预测系统实现的关键点是:

(1) 自然灾害Web文本信息抓取, 并通过数据清洗, 形成结构化的数据;

(2) 如何从结构化数据中挖掘关联规则;

(3) 实时抓取关联规则的前端信息.

3 自然灾害预测系统实现

3.1 Web文本数据收集

3.1.1 数据收集的难点

目前国内还没有统一的结构化自然灾害数据信息库, 因此存在灾害数据标准不同、数据来源的可靠性与广泛性难以界定、缺乏统一的收录数据标准界定和数据管理范式(包括灾害特征类、字段名称、对应数据类型等规范的确定)等种种数据质量问题, 很难实现灾害信息的应用层面共享. 此外, 由于灾害信息发布来

源在区域尺度、时间尺度、信息的精度、信息的时效性、信息条目的全面性等方面往往存在较大的偏差, 因此如何得到统一标准的、规范的、可统计分析的结构化数据成为本系统实现的技术难点.

3.1.2 数据源选取

为确保灾害数据来源的可靠性以及能够覆盖地震、洪涝等十余种主要灾害类型, 经过调查分析研究, 最终选择下面国家权威机构的灾害信息数据发布平台作为本系统的灾害历史数据来源. 具体网站信息如下:

(1) 中国农业部种植业管理司历史灾害查询网站(www.zzys.moa.gov.cn);

(2) 中国森林防火网(www.slfh.gov.cn);

(3) 国家减灾网(www.jianzai.gov.cn);

(4) 中国地震信息网(www.csi.ac.cn).

3.1.3 数据收集方式

由于系统的数据源来自不同资料平台, Web文本数据形式以及内容均不一样, 因此系统使用第三方爬虫软件进行历史数据收集, 根据不同的数据源定制化采集网页中指定的文本信息. 最终完成原始数据的采集.

3.2 数据清洗

3.2.1 数据清洗目标

灾害信息的原始数据均是Web文本中非结构化的自然语言, 如何从大段的Web文本中提取结构化的满足需求的有用灾害信息成为本系统实现的难点. 对文本灾害信息的语义理解和抽取, 重点是解决文本语言信息的形式化问题, 建立模糊的、定性的语言与定量化的计算机模型之间的联系, 实现从大段叙述性的Web文本中整理抽取成形如“时间+地点+灾害类型”的结构化数据. 因此需要按照一定的规则从文本中抽取匹配有关灾害事件的命名实体, 如灾害类型、时间、地点, 确定实体之间的关系, 进而实现非结构化灾害信息向结构化灾害信息的转换.

3.2.2 基于中文分词的数据清洗方法

根据数据来源复杂、数据内容杂乱无序等特点, 采用了机械匹配法(又称为字符串匹配法)的自然语言分词方法, 实现从文本数据中提取有效灾害关键信息(灾害类型、时间、地点).

机械分词方法又叫基于字符串匹配的分词方法, 它是按照一定的策略将待分析的字符串与一个“充分大的”机器词典中的词条进行匹配, 若在词典中找到某个字符串, 则匹配成功(识别出一个词). 该方法是目前Web文本数据提取的主流实现方法之一, 具有易实现、可维护、可扩展等优点. 但该方法也存在难以处理未登录词, 无法有效克服歧义切分的缺点. 由于本系统中Web文本不涉及语义分析, 同时“时间”、“地点”、“灾害类型”的匹配词库量较小, 因此综合考虑最终使用机械匹配法来实现数据提取清洗.

3.2.3 数据清洗规则

目前收集的Web文本数据存在如下数据质量问题:

(1) 原始数据针对时间要素相关的描述存在多种格式, 没有统一的规范格式. 以1990年1月1日为例, Web文本数据中存在“1990年1月1日”、“1990-01-01”、“19900101”等多种形式.

(2) 原始数据针对地点要素相关的描述存在描述地域粒度不同, 缺乏统一的唯一标示. 在Web文本中存在类似“广东省”、“东莞市”、“珠江三角洲流域”、“华南地区”等不同级别不同粒度的地域描述说明, 无法形成统一的结构化要素.

(3) 原始数据针对灾害类型要素相关的描述存在缩略语或者同义词, 缺乏统一的定义. 例如Web文本中“雪灾”灾害类型可能存在被描述成“暴雪”、“大雪”等同义词.

(4) 原始数据针对时间要素相关的描述存在模糊缺省的情况, 例如“1990年1月, ...”.

(5) 原始数据来源复杂, 数据排列无序.

(6) 数据中存在由于录入错误等行为导致的违背常识错误无效数据, 例如(1月56日)等.

根据对以上原始数据质量问题进行归纳整理, 得出如下数据清洗规则:

(1) 将时间要素提取成格式统一的结构化要素, 为方便后续关联分析算法使用, 时间要素格式定位为4位数字表示年份、2位数字表示月份、2位数字表示日期的格式, 即“19900101”的形式.

(2) 将地点要素提取成格式统一的结构化要素, 将地区性的地点描述、省级的地点描述转换成相应城市的地点描述, 将地点要素统一成以行政市为单位的数据.

(3) 将灾害类型要素统一定义, 最终形成地震、洪灾、干旱等十类灾害.

(4) 针对时间要素缺省日期的情况, 在当月时间内实现随机日期补全, 针对时间要素缺省月份的情况则

视该条数据为无效数据.

(5) 将各条原始数据提取信息后按照时间顺序进行排序, 形成结构化、有序的数据集.

(6) 针对清洗后的结构化数据进行常识性容错检查, 发现错误后将该条无效数据剔除.

按照以上规则进行数据清洗, 最终获取33717条结构化数据(表1所示), 并存入MySQL数据库中.

表1 MySQL数据库数据集示例

时间	地区	灾害
20160722	铜仁地区	雷电
20160722	铜仁地区	暴雨
20160722	铜仁地区	洪涝
20160722	毕节地区	雷电
20160722	毕节地区	暴雨
20160722	毕节地区	洪涝
20160722	遵义	雷电
20160722	遵义	暴雨
20160722	遵义	洪涝
20160722	黔东南苗族侗族自治州	雷电
20160722	黔东南苗族侗族自治州	暴雨

3.3 自然灾害Web文本信息实时抓取

3.3.1 实现思路

针对成熟开源爬虫框架(WebMagic)进行二次开发, 定制化实现“标题+发布时间+灾害类型”的原灾害Web文本信息采集. 对采集到的非结构化文本信息数据按照上节方法进行数据清洗, 检索其文本内容, 解析生成“时间+地点+灾害类型”的结构化前端信息.

3.3.2 WebMagic框架介绍

WebMagic框架介绍内容来源于网络参考文献《WebMagic爬虫框架学习》, 本文摘取和修改其中重要内容方便读者阅读. WebMagic的框架由四大组件Downloader、PageProcessor、Scheduler和Pipeline组成, 而Spider负责将它们组织起来. 这四大组件对应爬虫生命周期中的下载、处理、管理和持久化等功能. Spider是WebMagic内部流程的核心. 四大组件都是Spider的一个属性, 可以自由设置这些属性, 从而实现不同的功能. Spider也是WebMagic操作的入口, 它封装了爬虫的创建、启动、停止、多线程等功能. WebMagic总体架构图如下:

(1) Downloader负责下载页面, 供后续处理. WebMagic默认以Apache HttpClient作为下载工具.

(2) PageProcessor负责解析页面, 抽取有用信息, 以及发现新的链接. WebMagic使用Jsoup作为HTML解析工具, 并基于其开发了解析XPath的工具Xsoup. PageProcessor对于每个站点每个页面都不一样, 是需要使用者定制的部分.

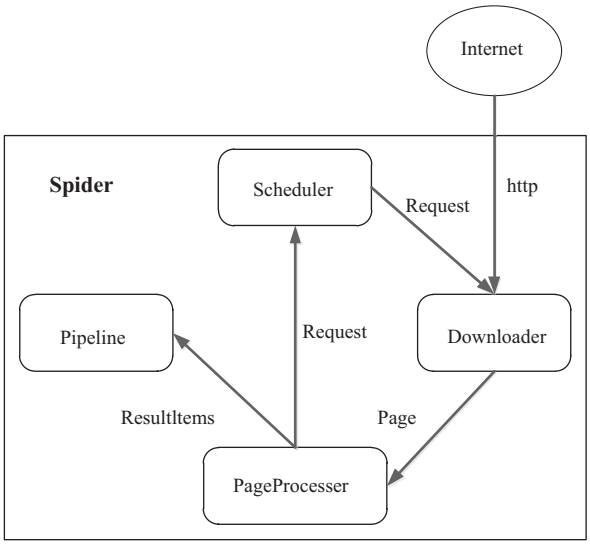

图4 WebMagic总体架构图

(3) Scheduler负责管理待抓取的URL和去重工作. WebMagic默认提供了JDK的内存队列来管理URL, 并用集合来进行去重. 也支持使用Redis进行分布式管理. 除非有一些特殊的分布式需求, 否则无需自己定制Scheduler.

(4) Pipeline负责抽取结果, 包括计算、持久化到文件、数据库等. WebMagic默认提供了“输出到控制台”和“保存到文件”两种结果处理方案. Pipeline定义了结果保存的方式, 如果你要保存到指定数据库, 则需要编写对应的Pipeline. 对于一类需求一般只需编写一个Pipeline.

3.4 关联规则算法介绍及改进

3.4.1 算法介绍及选择

R.Agrawal等[9]于1993年提出了关联规则的概念, 用于挖掘顾客交易数据的频繁模式. 关联规则挖掘算法最常用的就是Apriori和FP-Growth算法. 严格地说Apriori和FP-Growth都是寻找频繁项集的算法. 其中最经典的算法是Apriori[10], 但是其致命的缺点是需要多次扫描事务数据库. FP-Growth算法是韩家炜等人在2000年提出的关联分析算法[11,12], 它采取如下分治策略: 将提供频繁项集的数据库压缩到一棵频繁模式树(FP-tree), 但仍保留项集关联信息. 该算法和Apriori算

法最主要不同点有: 第一, 不产生候选集; 第二, 只需要两次遍历数据库, 大大提高了效率. 因此我们选择FP-Growth算法挖掘关联规则.

FP的全称是Frequent Pattern, 在算法中使用了一种称为频繁模式树(Frequent Pattern Tree)的数据结构. FP-tree是一种特殊的前缀树, 由频繁项头表和项前缀树构成. 所谓前缀树, 是一种存储候选项集的数据结构, 树的分支用项名标识, 树的节点存储后缀项, 路径表示项集.

FP_growth算法描述如下[13](伪代码):

输入: 事务数据库D; 最小支持度阈值min_sup;

输出: 频繁模式的完全集;

方法:

(1) 按以下步骤构造FP树:

(a) 扫描事务数据库D一次. 收集频繁项的集合F和它们的支持度计数. 对F按支持度计数降序排序, 结果为频繁项列表L;

(b) 创建FP树的根结点, 以“null”标记它. 对于D中每个事务Trans, 执行: 选择Trans中的频繁项, 并按L中的次序排序. 设Trans排序后的频繁项列表[p|P], 其中p是第一个元素, 而P是剩余元素的列表. 调用insert_tree([p|P], T). 该过程执行情况如下. 如果T有子女N使得N.item-name=p.item-name, 则N的计数增加1; 否则, 创建一个新结点N, 将其计数设置为1, 链接到它的父结点T, 并且通过结点链结构将其链接到具有相同item-name的结点. 如果P非空, 则递归地调用insert_tree(P, N).

(2) FP树的挖掘通过调用FP_growth(FP_tree, null)实现. 该过程procedure FP_growth(Tree, α)实现如下:

(a) if Tree包含单个路径P then;

(b) for路径P中结点的每个组合(记作β);

(c) 产生模式$\beta \cup \alpha$, 其支持度计数support_count等于β中结点的最小支持度计数;

(d) else for Tree的头表中每个a_i;

(e) 产生一个模式$\beta = a_i \cup \alpha$, 其支持度计数support_count=a_i. Support_count;

(f) 构造β的条件模式基, 然后构造β的条件FP树$Tree_\beta$;

(g) if $Tree_\beta \neq \emptyset$ then;

(h) 调用FP_growth($Tree_\beta$, β).

3.4.2 算法辅助改进

系统使用Hadoop平台mahout库中自带的FP-Growth算法进行频繁模式的挖掘. 由于算法的特性, 根据本系统涉及问题进行算法辅助改进.

问题1. 关联分析的输入数据类型为标称型数据, 而从数据采集结果得到的是具有时间、地点、灾害类型三个属性的数据.

辅助改进: 编写shell脚本将数据采集结果整理成算法需要的标称型数据, 首先把地点与灾害类型合并, 再将同一时间发生的灾害作为一个事务进行处理.

问题2. FP-Growth算法的结果得到的是关于灾害预测的频繁项集, 而本系统需要的是带有特定结构的关联规则.

辅助改进: 编写python脚本, 将频繁项集整理输出为本系统需要的关联规则.

问题3. 本系统需要分析出具有时间间隔的关联规则.

辅助改进: 首先编写python脚本将数据采集结果整理为具有时间间隔的标称型数据, 再利用大平台环境进行后续的频繁模式挖掘, 最后利用对解决问题2而编写的python脚本整理输出为具有时间间隔的关联规则.

3.4.3 关联规则库设计及示例

关联规则作为重要的发现知识, 被单独存在MySQL数据库, 与图1中的灾害信息数据库是分开的. 关联规则库(或关联规则数据库)的数据集结构设计为“前端+时间间隔(天)+后端+置信度”, 其中字段“前端”和“后端”由地点和灾害类型合并而成, 比如前端“贵港_洪涝”是指关联规则的前端信息, 表示贵港发生洪涝灾害. 置信度是指前端发生的条件下, 后端发生的概率.

对表1所示的数据集用改进后的FP-Growth算法进行挖掘, 并按照阈值(置信度为60%)对关联规则进行筛选, 得到137620条有效关联规则(见表2).

3.5 自然灾害预测实现

预测任务由系统的预测程序负责, 操作接口如图5所示.

表2 关联规则库数据集示例

前端	时间间隔(天)	后端	置信度(%)
贵港_洪涝	15	安顺_洪涝	60.81
绵阳_洪涝	15	迪庆藏族自治州_暴雨	60.81
百色_洪涝	15	毕节地区_洪涝	60.81
玉林_洪涝	15	遵义_洪涝	60.81
钦州_洪涝	15	贵阳_洪涝	60.81
贵港_洪涝	15	铜仁地区_洪涝	60.81
防城港_洪涝	15	毕节地区_洪涝	60.81

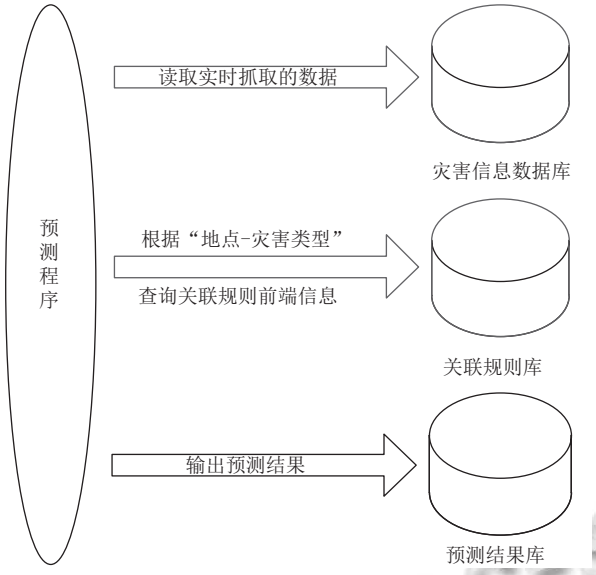

图5 自然灾害预测

对预测效果的评估存在一些困难, 因为预测准确性受到很多因素影响, 一是灾害信息报道不及时, 导致抓取实时灾害信息延期, 影响最后预测结果; 二是预测结果的验证存在困难, 如预测结果确实发生了, 但在指定的网络上不存在相关的报道(或许在别的网站上有相关报道), 导致无法抓取到真实信息. 目前在系统上设计自动验证程序存在技术难点, 因此预测效果的评估靠人工核实.

4 结论

自然灾害严重威胁着人民生命和国家财产的安全, 随着国家经济发展和人口增长, 自然灾害所造成的巨大损失正在日益加重. 本文基于关联规则和Web文本挖掘技术提出了一种自然灾害预测系统设计方案及设计方法, 该系统可定向抓取自然灾害的Web文本信息, 通过中文分词技术将非结构化的Web文本信息转化为结构化数据, 并利用改进的关联规则算法从结构化数据中挖掘出自然灾害关联规则库, 最后通过实时抓取自然灾害Web文本信息监测关联规则的前端信息, 即某特定自然灾害事件发生的前兆, 结合关联规则库以实现该自然灾害事件的预测. 试运行结果表明该系统能挖掘出有效的关联规则, 有效提升自然灾害的防灾减灾能力. 该系统还存在一些不足, 比如缺少有效的预测效果评估程序, 这也是本研究后期努力改进的方向.

参考文献

1 Brunt J. Using the world wide web to advance data management in LTER. LTER Network News, 1998, 11(1): 18–19.
2 周宁. 信息资源数据库. 2版. 武汉: 武汉大学出版社, 2006. 233–235.
3 韦方强, 崔鹏, 胡凯衡, 等. 泥石流灾害信息共享的方法与实现. 灾害学, 2002, 17(3): 60–64.
4 林孝松, 赵纯勇. GIS在重庆市地质灾害信息管理系统中的应用. 灾害学, 2003, 18(1): 71–76.
5 Dunbar PK. Increasing public awareness of natural hazards via the Internet. Natural Hazards, 2007, 42(3): 529–536. [doi: 10.1007/s11069-006-9072-3]
6 Peduzzi P, Dao H, Herold C. Mapping disastrous natural hazards using global datasets. Natural Hazards, 2005, 35(2): 265–289. [doi: 10.1007/s11069-004-5703-8]
7 李卫江, 温家洪. 基于Web文本的灾害信息挖掘研究进展. 灾害学, 2010, 25(2): 119–123, 128.
8 任振球. 关于加强特大自然灾害预测新途径新方法研究的讨论. 地球信息科学, 2000, 2(2): 76–77.
9 Agrawal R, Imieliński T, Swami A. Mining association rules between sets of items in large databases. Proc. 1993 ACM SIGMOD International Conference on Management of Data. Washington DC, USA. 1993. 207–216.
10 Agrawal R, Srikant R. Fast algorithms for mining association rules. Proc. 20th International Conference on Very Large Data Bases. Santiago, Chile. 1994. 487–499.
11 Han JW, Pei J, Yin YW. Mining frequent patterns without candidate generation. Proc. 2000 ACM SIGMOD International Conference on Management of Data. Dallas, Texas, USA. 2000. 1–12.
12 杨勇, 王伟. 一种基于MapReduce的并行FP-growth算法. 重庆邮电大学学报(自然科学版), 2013, 25(5): 651–657, 670. [doi: 10.3979/j.issn.1673-825X.2013.05.016]
13 Han JW, Kamber M, Pei J, 等. 数据挖掘: 概念与技术.范明, 孟小峰, 译. 北京: 机械工业出版社, 2012.