数据融合技术在环境监测领域的应用①

刘卫萍[1,2], 王 宁[2], 周晓磊[2], 张 镝[2]

[1](中国科学院大学, 北京 100049)
[2](中国科学院 沈阳计算技术研究所, 沈阳 110168)

摘 要: 在省级环境监测系统中随着时间的累积有大量数据, 并且监测数据种类繁多, 不同环境指标的描述信息也有很大差别, 同时监测数据又是每分每秒不断增加的, 这样就增加了数据的复杂性. 而数据融合技术则是大数据技术中处理数据的一种方式, 可以将来自多传感器的数据通过数据转换、数据相关和融合计算过程,对数据处理、分析并进行态势分析. 同时利用到了大数据 ETL 技术、MapReduce 处理, 使用 D-S 证据推理算法进行融合推理, 这样就可以增加数据的相关性, 降低数据的规模.

关键词: 数据融合; 数据转换; 数据相关; 融合计算; MapReduce

Application of Data Fusion to Environment Monitoring

LIU Wei-Ping[1,2],WANG Ning[2], ZHOU Xiao-Lei[2], ZHANG Di[2]

[1](University of Chinese Academy of Sciences, Beijing 100049, China)
[2](Shenyang Institute of Computing Technology, Chinese Academy of Sciences, Shenyang 110168, China)

Abstract: With the accumulation of time, the data is sharply increased in provincial environment system. And there are many different kinds of monitoring data and the description of the different environmental indicators varies. As the monitoring data is growing every minute, the complexity of the data is increasing. The data fusion technology is a method of processing data in big data technologies, which can process data, analysis data and make trend analysis through transfer, correlation and fusion compute data from multiple sensors. The using of ETL, MapReduce and D-S evidence reasoning algorithm can increase the correlation data, so reduce the size of the data.

Key words: data fusion; transfer; correlation; fusion compute; MapReduce

近年来, 人类不断地加强对环境问题的关注, 环境质量已经成为生活质量的重要判断标准, 也成为社会文明程度的重要体现, 而环境监测是人类了解环境质量的重要的途径. 随着科技的不断发展, 卫星、遥感、GIS、射频技术和传感网等现代信息技术在环境监测领域的广泛应用, 使数据爆炸式增长, 并逐渐呈现出多源、多维、大量和多态的特性[1], 2013 年全省空气质量监测数据就达到了 1T, 预估每年以数百 G 速度增长, 并且增长速度不断加快, 水环境监测更是如此. 而对于大气、水文等监测数据又来自不同数据源, 传统的控制系统中通常会使用大量传感器对影响环境的每个参数进行检测、分别处理, 各个影响参数之间都是独立的, 这样使得检测数据的处理器工作量很大, 忽略数据之间的相关性, 也就会丢失数据之间组合起来所表现的特征, 因而数据融合技术的出现可以通过选用适当的方法对多传感器的数据进行融合, 降低数据的复杂性, 增加数据描述的准确性, 最终得到更充分的信息.

1 数据融合技术的研究

1.1 数据融合的定义

数据融合是一种数据处理技术, 最早出现在 20 世纪 70 年代初期, 当时主要应用在军事领域. 数据融合的基本原理源于人脑处理信息的方式, 通过中枢神经

① 基金项目:国家水体污染控制与治理科技重大专项课题(2012ZX07505003)
收稿时间:2015-10-29;收到修改稿时间:2015-11-25 [doi:10.15888/j.cnki.csa.005202]

将感知信息传送到大脑进行综合处理, 然后对外部环境进行判断和控制. 数据融合把多传感器在空间或时间上可溶或互补的数据根据特定的算法进行组合, 以获得对被测对象的一致性解释或描述, 从而得出更为准确、可信的结论[2].

1.2 数据融合的关键技术

整个数据融合的过程如图 1 所示[3], 由于来自多数据源的数据格式、描述信息也都不相同, 首先会采用合适的工具将数据的格式进行统一即数据转换, 数据转换之后的数据便可以通过 MapReduce 找出来自不同数据源信息之间的相关性, 基于相关性分析的结果便可以作为融合计算的输入, 由 DS 证据论证方法对来自不同数据源因素进行信息、关系的融合, 最终把融合的结果输出并存储到数据库中.

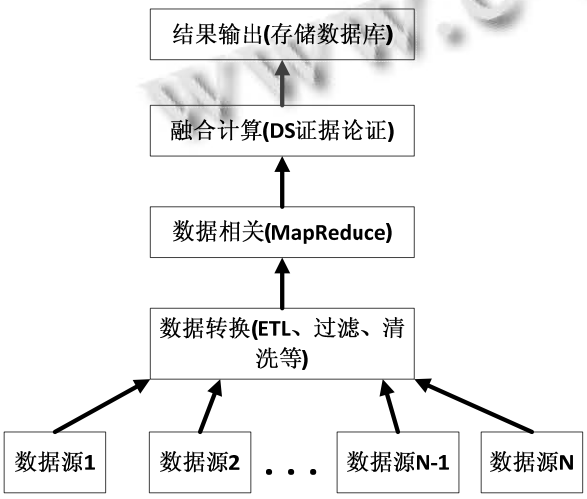

图 1 数据融合过程

1.2.1 数据转换

在多传感器数据融合的系统里, 各个传感器得到的数据极有可能是来自不同维度的数据, 如有些来自 TXT 格式的文件, 有些则是 XML 格式的文件, 并且有关系型数据库和 NoSQL 数据库, 它们对于数据的描述方式是不同的, 即使同样是关系型数据库, MySQL 和 Oracle 也会因数据库规范不同而存在差异, 所以在做数据相关之前要进行数据转换操作, 以达到是数据具有相同形式的目的. 通过大数据技术 ETL 进行数据抽取(Extract)、转换(Transform)和加载(Load)到数据仓库, 过程中涉及到不同数据标准和不同数据格式之间的转换, 需要注意的是数据融合存在时间性和空间性, 所以需要按照特定的规则进行坐标准换[4].

1.2.2 数据相关

数据相关是特征提取的关键技术, 主要是对采集的数据进行相关性分析, 使基础数据建立必要联系, 形成关联数据库, 从而能消除数据的二义性. 一般会使用面向分析型应用的基于 Key Value 存储的 noSQL 技术, 而 MapReduce 是 noSQL 常用的一种能够解决大规模非结构化数据快速批量处理的并行技术框架,能够将任务分成多个更细的子任务, 并且利用这些子任务进行进度调用来实现, 最后按照特定的规则, 合并成最终结果.

1.2.3 融合计算

融合计算是把经过预处理的结果进行验证、综合、补充、推理, 对于一些不相关的因素进行分析和综合, 并最终做出态势估计和最终决策的过程. 根据对数据的处理程度可以分为三个级别: 数据级融合、特征级融合和决策级融合. 数据级是对采集的数据没有进行预处理, 是最低层次的融合; 特征级融合是指对原始数据进行特征提取, 然后会对提取出来的特征信息进行分析和处理; 决策级融合属于高层次的融合, 对数据进行预处理、特征提取, 并进行初步的决策判断, 主要是面向应用的.

2 环境监测系统设计

2.1 省级环境监测系统

结合省级环境监测站点描述数据融合处理过程如图 2 所示. 在环境监测系统中水环境监测站点分为自动监测站点和手工监测站点, 其中自动监测站点有 24 个, 手工监测需要对 297 个断面进行监测, 主要负责采取水环境中的数据, 如 PH 值、氨氮排放量、石油类排放量等. 有 24 个大气监测站点, 对 CO2、NOx、颗粒物等进行监测, 除此之外还有污染源信息, 以及对污染进行应急处理的交通信息, 如应急车辆、应急物资状况, 这些都是非常庞大的数据. 对于这些采集到的信息首先需要进行本地存储,防止出现网络中断, 造成数据丢失的情况. 如果出现网络中断, 则下次网络恢复时, 由数据补送机制将期间未发送给数据中心的数据从本地取出发送给数据中心[5]. 在本地存储之后可以通过网络传输至远程数据中心.

对数据预处理是首先要检验采集数据是否合理, 是否在合理的范围之内, 比如 PH 一般在 0~14 之间, 对于合理的数据要存在数据库中, 而对于不合理的数

据要丢弃, 以防在处理过程中由于数据不合理而导致系统出现异常, 影响安全性. 对于这些数据存储在客户端的数据库中, 然而对于这些数据库以及数据库中的数据可以通过数据库服务器来进行管理, 数据库服务器具有查询、更新、高速缓存等功能, 而且可以在多用户存取时保证数据安全.

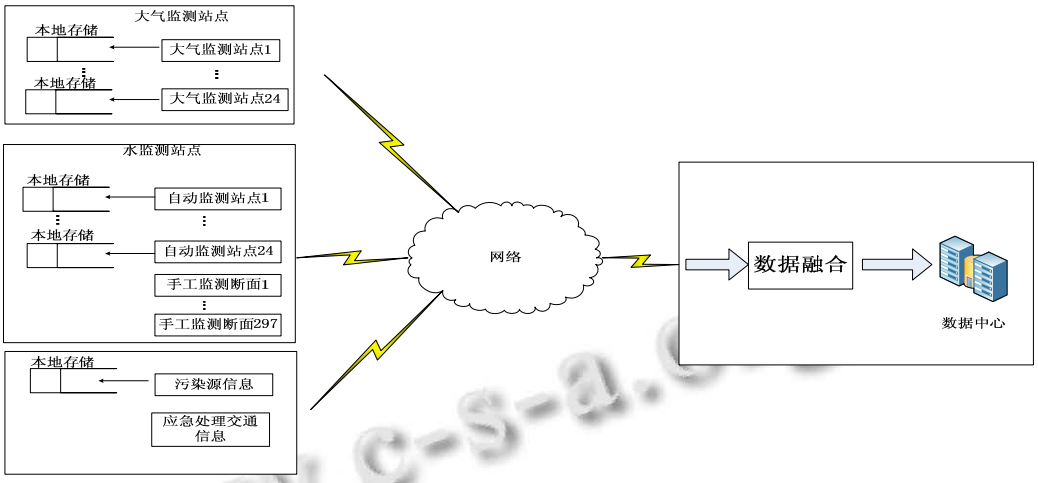

图 2　环境监测系统

2.2 结合数据融合的环境监测系统设计

整体架构图如图 3 所示, 从上到下分为应用层、核心功能层和平台层三层, 应用层根据数据来源分为水环境和大气环境数据融合两部分. 平台层中主要依赖了大数据的技术 MapReduce 框架, 使用 Hive 和 Sqoop 实现从关系型数据库到 NoSQL 的转换和传输, 使用 HDFS 文件系统存储, 所有的机器分布在集群中并且通过 Zookeeper 来对集群进行管理.

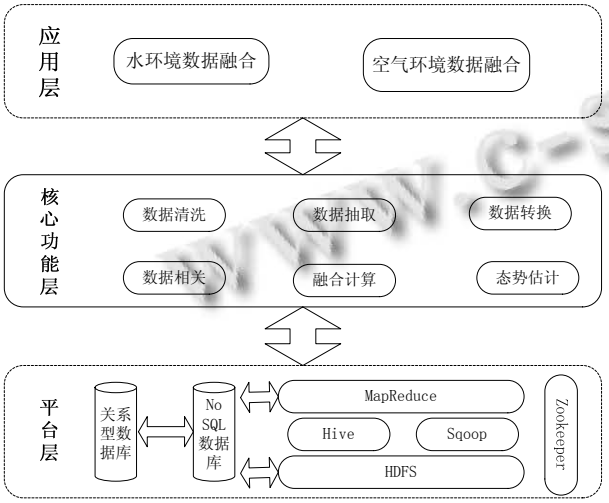

图 3　整体架构图

核心功能层作为数据融合的关键部分, 包含了数据融合所有的关键技术, 既依赖于底层的平台层的技术, 又能够实现应用层模块, 这些技术同时也是与系统的模块是相对应的.

在系统的数据采集模块中对于采集的数据进行合理性验证, 即在采集数据时对于超过规定数据范围的非法数据进行过滤, 保证数据的合法性. 数据预处理模块主要由数据清洗、数据抽取和数据转换技术完成, 通过 Hive 完成 ETL, 在关系型数据库中可以由 SQL 完成数据的抽取, 但是在 NoSQL 中还要通过 Hive 的 HQL 完成. 还要通过 Sqoop 实现在关系数据库和 Hadoop 集群之间的数据传输, 完成从二维表到 HDFS 之间的转换. 特征提取模块主要由数据相关实现, 按照约定好的特征提取规则对预处理的结果进行综合分析, 主要通过 MapReduce 模型实现, 在 Map 中根据监测数据字段进行分片, 在 Reduce 中对所有分片进行一次化简, 这些工作可以由多个进程组成的 Master 和 Worker 执行. 融合计算模块则是由融合计算中按照 DS 证据理论算法进行推理验证, 提高环境监测的准确度. 对于数据融合的结果可以直观的表现出当地的环境状况, 便于进行态势估计和决策支持.

3　模块设计

3.1 数据采集模块

环境监测系统的数据采集可以分为两个部分, 一部分是在监测站点通过数据采集仪来记录某一时间点

水环境中 PH 值、盐度和 COD 等水文数据或空气监测中关于温度、湿度、CO2 含量等气象监测值, 并且通过程序存储在本地数据库中(采用的是 Oracle 关系型数据库). 此外, 各站点还要通过网络编程将这些数据传送至远程数据中心, 远程数据中心就是一个 NoSQL 的数据存储方式.

从网络中传输来的数据在远程数据中心是 TXT 或 XML 文件格式, 所以采用的是 HDFS, 分布在一个由多台机器组成的集群上的, 并且所有的机器采用轮询算法保证一个负载均衡的环状, 对于通过网络传输过来的监测数据会以监测站编号作为 key 值, 通过 key 值的 hash 计算来确定保存在具体哪个机器上, 这样就能基本保证来自同一监测站的数据能够保存在同一台服务器上.

3.2 规则构建模块

规则构建模块的结果是数据预处理、相关以及融合过程中都会涉及的, 是之后几个模块进行工作的一个准则.

数据预处理阶段的数据清洗、转换和抽取都是依赖准则进行的, 而准则则是根据需求制定的. 清洗准则是根据国家相关准则或业务准则制定的, 比如关于时间的格式满足 yyyy-mm-dd hh:mm:ss, PH 值的范围、COD 含量等, 避免影响数据处理过程的安全性, 对于不满足规则的数据进行过滤. 转换规则主要用来进行数据统一, 如会对 XML 或 TXT 文件进行统一使用 TXT 文件, 对于命名有差异的污染物的处理, 或者统一污染物含量的单位等, 会按照相关标准进行统一. 抽取规则则是可以按照监测站点编号、监测时间段维度、污染物类型等处理.

特征提取规则主要使用作用在 MapReduce 的 Map 函数上, 按照不同粒度划分, 如监测数据类型、监测站编号(MONITORINGID)、监测时间(MONITORINGTIME)、监测物(POLUTANTID)划分, 此处的监测数据类型是指水监测数据或空气监测数据.

3.3 数据预处理模块

数据预处理模块是对于从关系型数据库中提取出的数据在进行相关性分析之前的操作, 既可以保证进入数据相关模块的数据的有效性和合理性, 并且可以提高整个过程的安全性.

如图 4 所示, 从关系型数据库中可以通过 SQL 提取数据到预处理模块时要经过数据清理、转换和抽取过程来完成数据的统一和标准化处理, 这三个操作都是要基于已制定好的相应的准则执行, 清洗时对于时间表示为 2014-14-14 12:30:30 的数据项就会因为不满足时间格式中对于月份的要求而被过滤. 水环境监测数据包括自动监测数据和手动断面监测数据, 对于相同字段在两种监测中可能会存在不同的命名, 如监测站点命名可能为 MONITORING 和 MONITORINGID, 氨氮和 NH3-N 之类的命名其实是相同含义, 就可以按照约定进行统一消除歧义. 抽取时可以使用 SQL 根据 MONITORINGID, MONITORINGTIME 或污染物含量(如 COD)进行.

过程中要使用 MapReduce 框架, 在 MapReduce 里是 JobTracker 和 TaskTracker 组成的主从结构, JobTracker 负责启动任务和分配任务给 TaskTracker, 而 TaskTracker 是用来执行任务的.

为防止数据丢失, 对于预处理的结果要进行存储, 因为 Hive 提供了强大的 HQL, 所以可以完成到 HDFS 的转换, 在 HDFS 中有 NameNode 和 DataNode 之分, NameNode 所在的机器就是文件系统的管理者, 主要负责协调对用户文件的访问, 一般会有两个 NameNode, 另一个为备份, 在此没有画出. 而 DataNode 就是实际数据的管理者, 在不同的 DataNode 上会存在相同的信息, 使用冗余机制来保证系统的容错性[6].

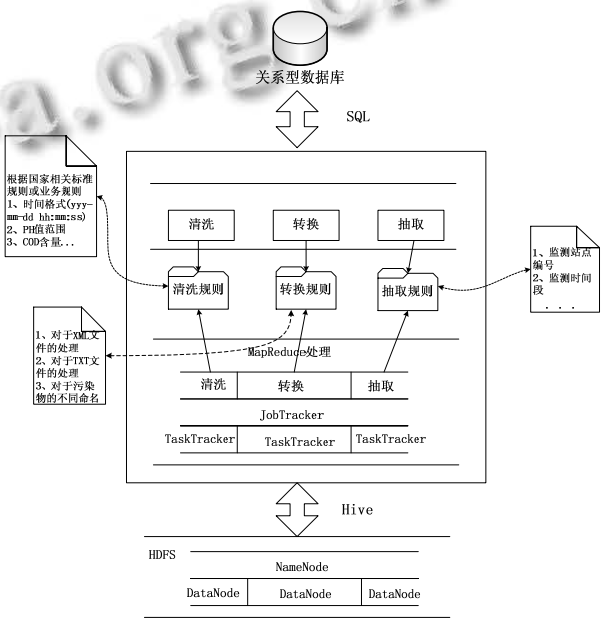

图 4 数据预处理模块流程分析

3.4 特征提取模块

在数据相关模块中会使用 MapReduce, 将大作业根据用户需求拆分为多个小作业, 并且拥有很强的并行处理能力, 对于MapReduce的工作原理如图5所示[7].

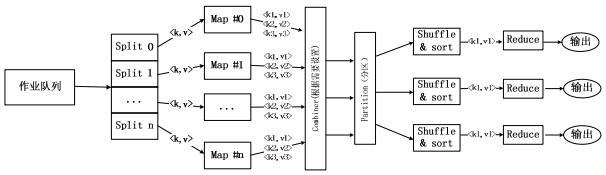

图 5　MapReduce 工作原理

MapReduce 把要处理的任务分成 Map 任务和 Reduce 任务, 主要采用分而治之的思想, 所有的作业形成作业队列后, 对于每个作业多都会进行分片, 原始数据用键值对<k,v>表示, 经过 Map 阶段已经写好的 map 函数之后就会生成 list<k,v>的中间键值对, 此处的 map 函数也是基于规则实现的, 对于多规则会以规则为维度进行 Map 分片. 经过 Map 处理的中间键值对可以先用 Hadoop 自带的合并器 Combiner 做初步合并, 这样可以减少中间数据的传输而带来的网络流量. 此时的数据会被写在缓存中, 对于缓存的数据会被定期刷在磁盘上, 在写在磁盘上之后会把键值对应的位置通知给 Master, Master 就会把相应的信息传给负责 Reduce 的 Worker, 为了确保中间键值对中 key 相同的会被分在同一个 Reduce 中, 需要使用 Partition 类将中间结果进行分区, 这里每个分区就是一个 Reduce 作业.

在每个Reduce任务中又包括shuffle(组合)、sort(排序)和 reduce(聚集)三个阶段, 进入 reduce 后要先对相同 key 值的进行组合, 然后进行排序, Reduce worker 程序遍历排序后的中间数据, 对于每一个唯一的中间 key 值, Reduce worker 程序将这个 key 值和它相关的中间 value 值的集合传递给用户自定义的 Reduce 函数. Reduce 函数的输出被追加到所属分区的输出文件.

当所有的 Map 和 Reduce 任务都完成之后, master 唤醒用户程序. 在这个时候, 在用户程序里的对 MapReduce 调用才返回.

数据特征提取是使用 MapReduce 的原理完成的, 以监测站点 MONITORINGID 作为 key 进行分片, 同进行时间维度的分片, 即根据输入时间范围<T1,T2>, 就会在 MONITORINGTIME>=T1&& MONITORINGTIME<T2 中分别过滤, 进行污染物维度的分片就是以 POLUTANTID 为 key 分片. 这就是对应的 map 的过程.

但是此时存在一个问题, 由于 Map 是基于规则进行的, 当存在的规则很多时, 就会引起很多的 Map 碎片, 在进行 Reduce 时就会大大降低效率. 所以在此做一个优化, 就是对于提出的规则进行归并, 同时作用于 Map, 并进行 Reduce, 规则类 1 结束后进行规则类 2, 如图 6 所示, 构成一个串行的结构.

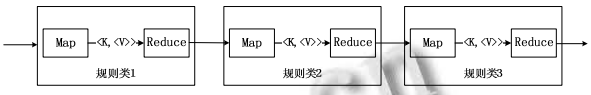

图 6　MapReuce 结构优化

3.5 数据融合计算模块

数据融合计算模块主要对来自多传感器的相关后的结果进行验证、分析、取舍和修改, 对新发现的不相关观测结果进行分析和综合, 对综合态势进行修改. 过程中主要使用D-S证据论证方法, 如图7所示, 对于温度、COD、PH 值等传感器传来的值首先要做预处理, 然后对各个证据的基本概率分配函数、可信度、似然度进行计算, 再根据合成规则把所有证据总和作用下的概率分配函数、可信度、似然度进行计算, 最后根据一定的判断规则, 把拥有最大可信度和似然度的结果作为最终融合结果提供系统使用[8].

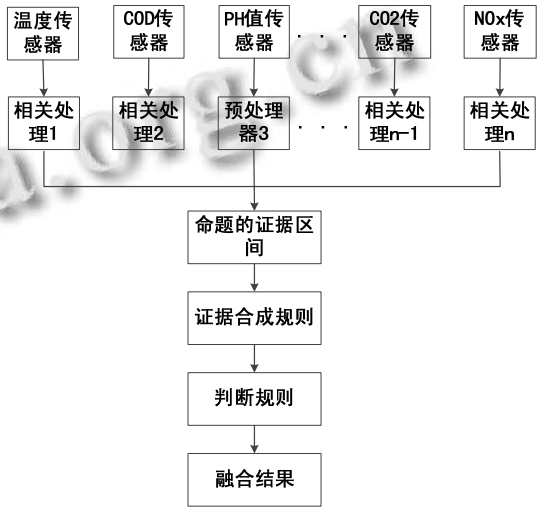

图 7　基于 D-S 证据论证的数据融合

4　展示结果

图 8 所示, 在环境监测系统中选中温度、湿度和风速三个参数对大气监测值进行融合的结果, 对于大气监测值为每 5 秒进行更新, 在此计算的是每两小时

得平均概率分配值, 根据温度、湿度、风速监测值, 经过数据预处理、相关性分析过程得出基本概率分配值, 最终进行 DS 验证计算. 对于适中的环境条件中温度范围是 20-24℃, 湿度适中范围 60%-65%, 风速适中范围是 0.5-0.8m/s, 由于温度、湿度和风速三个参数共同决定环境, 所以融合结果的概率分配值越大则表示当时环境状态越接近适中状态, 这样的结果更能够更加准确的展示环境状态.

监测站点	监测时间	温度（℃）	温度基本概率分配（%）	湿度（%）	湿度基本概率分配（%）	风速(m/s)	风速基本概率分配（%）	融合结果
北陵	201508110600	19	0.392	63	0.288	0.3	0.050	0.4320
北陵	201508110800	20	0.643	63	0.288	0.5	0.063	0.6139
北陵	201508111000	23	0.532	55	0.253	0.7	0.076	0. 3541
北陵	201508111200	28	0.126	40	0.133	0.8	0.160	0.1013
北陵	201508111400	32	0.064	30	0.401	0.9	0.051	0.0190
北陵	201508111600	30	0.059	42	0.150	0.85	0.083	0.1539
北陵	201508111800	24	0.542	60	0.361	0.6	0.103	0.5825

图 8 环境监测系统中数据融合结果图

5 结语

本文主要结合数据融合技术针对环境监测领域中的海量数据处理过程中遇到的问题进行叙述, 对于来自于多传感器的监测值需要通过数据融合技术处理, 能够更加准确的判断监测指标之间存在的关系, 通过各个传感器之间的协调和性能互补来提高整个系统的性能和准确性, 以便做出更准确的决策.

参考文献

1 Wong YC, Sundareshan MK. Data fusion and tracking of complex target maneuvers with a simplex-trained neural network-based architecture. Proc. 1998 IEEE International Joint Conference on Neural Networks, 1998. IEEE World Congress on Computational Intelligence. IEEE. 1998, 2. 1024−1029.

2 莫荣强,艾萍,吴礼福,岳兆新,冯鹏.一种支持大数据的水利数据中心基础框架.水利信息化,2013,(3):16−20.

3 李恒灿,李权才.数据融合技术在环境监测中的应用.中国农机化学报,2011,(4):110−113.

4 童国强,陈前.基于数据融合技术的多模型状态监测与故障预报.南京航空航天大学,2005.

5 李安增,王宁,王常权,邱燕.大数据技术在环境信息中的应用.计算机系统应用,2015,24(1):60−64.

6 尹立松.基于 MapReduce 和编程方式的 ETL 框架研究与应用[学位论文].上海:东华大学,2013.

7 王秀磊,刘鹏.大数据关键技术.中兴通讯技术,2013,19(4):17−21.

8 何彤宇.大数据时代网络学习环境的数据融合.现代教育技术,2013,23(12).