(19)中华人民共和国国家知识产权局

(12)发明专利

(10)授权公告号 CN 109885782 B
(45)授权公告日 2020.05.15

(21)申请号 201910164115.7

(22)申请日 2019.03.05

(65)同一申请的已公布的文献号
申请公布号 CN 109885782 A

(43)申请公布日 2019.06.14

(73)专利权人 重庆工商大学融智学院
地址 401320 重庆市巴南区龙洲湾尚文大道906号

(72)发明人 陈国彬

(74)专利代理机构 成都顶峰专利事务所(普通合伙) 51224
代理人 李崧岩

(51)Int.Cl.
G06F 16/955(2019.01)
G06F 16/29(2019.01)
G06F 16/957(2019.01)
G06F 21/62(2013.01)

(56)对比文件
US 2017/0293635 A1,2017.10.12,全文.
CN 103390039 A,2013.11.13,说明书第[0002]-[0104]段.
陈崇成 等.生态环境空间数据的多尺度集成方法.《环境科学研究》.2000,第13卷(第4期),第34-38页.
林海.基于GIS的测绘地理数据脱密方法及应用.《中国水运》.2014,第14卷(第7期),第336-337、340页.
韦强申.领域关键词抽取:结合LDA与Word2Vec.《中国优秀硕士学位论文全文数据库信息科技辑》.2016,(第12期),第1-23页.
谢克武.大数据环境下基于python的网络爬虫技术.《电子制作》.2017,第44-45页.

审查员 李欢

权利要求书3页 说明书7页

(54)发明名称

一种生态环境空间大数据集成方法

(57)摘要

本发明公开一种生态环境空间大数据集成方法,生态环境数据获取步骤如下:S1、利用网络爬虫获取的网页中与生态环境主题相关的数据;S2、将数据进行清洗和整理,并建立索引存入数据库中;空间地理数据获取步骤如下:L1、直接从地理信息服务器中获取空间地理数据;L2、通过GIS脱密处理方法对L1中获取的空间地理数据进行脱密处理;生态环境空间大数据集成步骤如下:从存入数据库中提取目标数据,并加载到脱密处理后的空间地理数据中,得到集成的生态环境空间大数据。本发明通过脱密处理的空间地理数据既可以满足信息化的实用价值又可以保证涉及国家安全的数据保密,数据集成实现网络化环境信息系统间不同尺度数据交换、传输和共享、数据互操作。

CN 109885782 B

1.一种生态环境空间大数据集成方法,其特征在于:包括生态环境数据获取步骤、空间地理数据获取步骤,和生态环境空间大数据集成步骤:

生态环境数据获取步骤如下:

S1、利用网络爬虫从多个初始URL开始下载网页内容,通过搜索策略获取网页中与生态环境主题相关的数据,同时不断从当前页面提取新的URL,根据网页抓取策略放入待抓取URL队列中,循环执行,直至停止,循环结束,其中,生态环境主题相关的数据包括水质监测数据、大气环境监测数据、土壤环境检测数据或/和生态环境污染源信息;

S2、将S1中获取的网页中与生态环境主题相关的数据进行清洗和整理,并建立索引存入数据库中;

空间地理数据获取步骤如下:

L1、直接从地理信息服务器中获取空间地理数据;

L2、通过GIS脱密处理方法对L1中获取的空间地理数据进行脱密处理;

生态环境空间大数据集成步骤如下:

从步骤S2中的数据库中提取目标数据,并加载到脱密处理后的空间地理数据中,得到集成的生态环境空间大数据;

所述生态环境主题为生态环境关键词集,生态环境关键词集的获取过程为:

S001、定期地从学术论文网站上获取论文,并将新获取的论文保存在语料库中;

S002、对语料库中各篇论文的词语集进行数值化处理;

S003、将数值化处理结果作为训练样本导入Word2Vec模型进行训练,得到各个词语的词向量;

S004、针对每个词语,分别根据词向量计算该词语与其它各个词语的欧式距离或向量夹角,然后根据计算结果选取欧式距离最近或向量夹角最小的前N个其它词语作为相关词语,得到相关词语集,其中,N为介于20～100之间的自然数;

S005、将词语为“生态环境”的相关词语集作为生态环境关键词集;

在步骤S005后,对生态环境关键词集进行补充,包括以下步骤:

S006、将步骤S002的数值化处理结果作为训练样本导入LDA主题模型进行训练,得到主题-词语矩阵和论文-主题矩阵,其中,所述主题-词语矩阵表示每个主题中出现每个词语的概率,所述论文-主题矩阵表示每篇论文中出现每个主题的概率;

S007、针对所述主题-词语矩阵中的各个主题,选取在对应主题中出现概率最大的前M个词语作为特征词语,得到特征词语集合,同时根据所述论文-主题矩阵获取各个主题的关联主题,其中,M为介于20～100之间的自然数,所述关联主题是指与某个主题共同出现在同一篇论文中的另一共现主题;

S008、针对现有生态环境关键词集合中的各个词语,将其分别作为目标词语,查找到在特征词语集合中包含该目标词语的所有目标主题,以及查找到与各个目标主题对应的目标关联主题;

S009、将所有目标主题和所有目标关联主题中的且不在现有生态环境关键词集合中的词语,添加到现有生态环境关键词集合中;

其中,获取各个主题的关联主题 过程为:

S0071、针对所述论文-主题矩阵中的各个主题,分别计算其与另一主题共同出现在各

篇论文中的共现概率之和；

S0072、针对所述论文-主题矩阵中的各个主题，选取对应于其共现概率之和最大的前K个另一主题作为关联主题，其中，K为介于3～10之间的自然数；

其中，步骤S002中对各篇论文的词语集进行数值化处理包括以下步骤：

S0021、对语料库中的所有论文进行切词处理，获取不同的词语，得到所述语料库的总词语集和各篇论文的词语集；

S0022、对所述总词语集中的所有词语进行唯一编码，得到包含有所有词语和与各个词语对应的数字唯一标识的词典；

S0023、根据所述词典将论文词语集中的词语转换为对应的数字唯一标识。

2.根据权利要求1所述的一种生态环境空间大数据集成方法，其特征在于：所述S1的实现过程为：

S101、启动网络爬虫程序；

S102、选取多个初始URL，并将其放入待抓取URL队列中；

S103、从待抓取URL队列中取出某个URL，并下载该URL对应的网页内容，然后将该URL放入到已抓取URL队列；

S104、通过搜索获取网页内容中与生态环境主题相关的数据，判断获取的数据是否已被历史获取，若是则丢弃该数据和网页内容，若否则缓存该数据，丢弃网页内容，同时解析该网页中所包含的URL，判断该URL是否是已抓取URL队列中所包含的URL，若是，则丢弃，若否，则将其放入待抓取URL队列中；

S105、循环执行S103-S104，直至待抓取URL队列中的所有URL被完全抓取，或系统命令停止抓取，循环结束。

3.根据权利要求2所述的一种生态环境空间大数据集成方法，其特征在于：所述S2的实现过程为：将S104缓存的数据进行清洗和整理，并建立索引存入数据库中。

4.根据权利要求1所述的一种生态环境空间大数据集成方法，其特征在于：所述GIS脱密处理方法包括地理要素数据及其属性脱密步骤，和空间位置精度脱密步骤，地理要素数据及其属性脱密步骤包括：使用GIS软件提供的编辑工具，删除涉密的空间地理数据及其属性数据；

空间位置精度脱密步骤包括：利用GIS软件提供的投影转换和空间校正工具对源空间地理数据的空间位置进行精度干扰。

5.根据权利要求4所述的一种生态环境空间大数据集成方法，其特征在于：所述空间位置精度脱密的具体实现过程为：

L201、投影前新建一个投影坐标系用于输出图幅，设置投影坐标系名称及参数，完成投影坐标系定义；

L202、在GIS工具中选择“投影”工具，输入参数，选择新建立的投影坐标系作为原始图幅的输出坐标系；

L203、重复L202，选择需要的投影坐标系为最终坐标系；

L204、加载待变换的原始数据，打开“空间校正”工具；

L205、选择需校正的数据，选择射影校正方法，使用工具对图幅创建移位连接；

L206、进行校正、保存，完成数据变换。

6.根据权利要求1所述的一种生态环境空间大数据集成方法，其特征在于：所述生态环境空间大数据的集成包括同要素空间数据集成和不同要素的空间数据集成；

同要素空间数据集成包括：

a、对各精度较高的小区域中的数据进行综合，提取其主要特征并归并，形成精度较低但空间范围较大的大区域的数据；

b、确定精度较低的大区域的数据精度，将大区域覆盖各精度较高小区域的部分提取出来，对提取出来的区域进行数据综合使其精度满足要求，最后进行接边处理形成完整的区域特定尺度数据提取；

c、由某区域同尺度不同时间的多个数据集推断出在单个数据集中分辨不出的生态环境或利用数据表达地学过程或特征的相关性来提高原来数据的精度；

d、使用多分辨率模型来表达空间实体，实现数据更新；

不同要素的空间数据集成包括：

h、不同要素空间尺度可比时，利用环境要素的相关性由某一或某些数据对某特定要素进行质量检测、数据综合、数据细化、辅助推导进行空间相关分析；利用不同要素之间的相关性生成新数据；

i、不同要素空间尺度不可比时，利用不同数据集进行背景参照分析和要素加权分析。

一种生态环境空间大数据集成方法

技术领域

[0001] 本发明属于空间数据集成领域，具体涉及一种生态环境空间大数据集成方法。

背景技术

[0002] 大数据是以容量大、类型多、存取速度快、应用价值高为主要特征的数据集合，正快速发展为对数量巨大、来源分散、格式多样的数据进行采集、存储和关联分析，从中发现新知识、创造新价值、提升新能力的新一代信息技术和服务业态，全面推进大数据发展和应用，加快建设数据强国，已经成为我国的的国家战略。国务院《促进大数据发展行动纲要》等文件要求推动政府信息系统和公共数据互联共享，促进大数据在各行业创新应用；运用现代信息技术加强政府公共服务和市场监管，推动简政放权和政府职能转变，构建“互联网+”绿色生态，实现生态环境数据互联互通和开放共享。

[0003] 合理的资源利用与良好的区域生态环境是实现社会经济可持续发展的重要基础条件，我国许多地区生态环境脆弱，近年来生态环境恶化的趋势尚未得到有效遏制，以灾害为主要表现形式的的各种生态环境问题对于我国国民经济的长期、稳定发展的制约作用愈加明显，资源环境利用效益降低、区域环境受到不利影响、各种灾害频繁发生等问题受到普遍重视。随着我国人口的持续增长和社会经济的快速发展，对于资源的利用强度不断增大，水土流失、荒漠化土地扩大、天然植被和生物多样性遭到破坏、灾害加剧等一系列资源环境问题仍然存在，在部分地区甚至有加剧趋势。为了逐步实现可持续发展战略，急需对于全国生态环境背景开展多方位数据支持下的空间特征综合研究。

发明内容

[0004] 为了解决现有技术存在的上述问题，本发明目的在于提供一种生态环境空间大数据集成方法。

[0005] 本发明所采用的技术方案为：

[0006] 一种生态环境空间大数据集成方法，包括如下生态环境数据获取步骤、空间地理数据获取步骤，和生态环境空间大数据集成步骤：

[0007] 生态环境数据获取步骤如下：

[0008] S1、利用网络爬虫从多个初始URL开始下载网页内容，通过搜索策略获取网页中与生态环境主题相关的数据，同时不断从当前页面提取新的URL，根据网页抓取策略放入待抓取URL队列中，循环执行，直至停止，循环结束，其中，生态环境主题相关的数据包括水质监测数据、大气环境监测数据、土壤环境检测数据或/和生态环境污染源信息；

[0009] S2、将S1中获取的网页中与生态环境主题相关的数据进行清洗和整理，并建立索引存入数据库中；

[0010] 空间地理数据获取步骤如下：

[0011] L1、直接从地理信息服务器中获取空间地理数据；

[0012] L2、通过GIS脱密处理方法对L1中获取的空间地理数据进行脱密处理；

[0013] 生态环境空间大数据集成步骤如下：

[0014] 从步骤S2中的数据库中提取目标数据，并加载到脱密处理后的空间地理数据中，得到集成的生态环境空间大数据。

[0015] 所述生态环境主题为生态环境关键词集，生态环境关键词集的获取过程为：

[0016] S001、定期地从学术论文网站上获取论文，并将新获取的论文保存在语料库中；

[0017] S002、对语料库中各篇论文的词语集进行数值化处理；

[0018] S003、将数值化处理结果作为训练样本导入Word2Vec模型进行训练，得到各个词语的词向量；

[0019] S004、针对每个词语，分别根据词向量计算该词语与其它各个词语的欧式距离或向量夹角，然后根据计算结果选取欧式距离最近或向量夹角最小的前N个其它词语作为相关词语，得到相关词语集，其中，N为介于20～100之间的自然数；

[0020] S005、将词语为“生态环境”的相关词语集作为生态环境关键词集；

[0021] 在步骤S005后，对生态环境关键词集进行补充，包括以下步骤：

[0022] S006、将步骤S002的数值化处理结果作为训练样本导入LDA主题模型进行训练，得到主题-词语矩阵和论文-主题矩阵，其中，所述主题-词语矩阵表示每个主题中出现每个词语的概率，所述论文-主题矩阵表示每篇论文中出现每个主题的概率；

[0023] S007、针对所述主题-词语矩阵中的各个主题，选取在对应主题中出现概率最大的前M个词语作为特征词语，得到特征词语集合，同时根据所述论文-主题矩阵获取各个主题的关联主题，其中，M为介于20～100之间的自然数，所述关联主题是指与某个主题共同出现在同一篇论文中的另一共现主题；

[0024] S008、针对现有生态环境关键词集合中的各个词语，将其分别作为目标词语，查找到在特征词语集合中包含该目标词语的所有目标主题，以及查找到与各个目标主题对应的目标关联主题；

[0025] S009、将所有目标主题和所有目标关联主题中的且不在现有生态环境关键词集合中的词语，添加到现有生态环境关键词集合中；

[0026] 其中，获取各个主题的关联主体过程为：

[0027] S0071、针对所述论文-主题矩阵中的各个主题，分别计算其与另一主题共同出现在各篇论文中的共现概率之和；

[0028] S0072、针对所述论文-主题矩阵中的各个主题，选取对应于其共现概率之和最大的前K个另一主题作为关联主题，其中，K为介于3～10之间的自然数；

[0029] 其中，步骤S002中对各篇论文的词语集进行数值化处理包括以下步骤：

[0030] S0021、对语料库中的所有论文进行切词处理，获取不同的词语，得到所述语料库的总词语集和各篇论文的词语集；

[0031] S0022、对所述总词语集中的所有词语进行唯一编码，得到包含有所有词语和与各个词语对应的数字唯一标识的词典；

[0032] S0023、根据所述词典将论文词语集中的词语转换为对应的数字唯一标识。

[0033] 在上述技术方案的基础上，所述S1的实现过程为：

[0034] S101、启动网络爬虫程序；

[0035] S102、选取多个初始URL，并将其放入待抓取URL队列中；

[0036] S103、从待抓取URL队列中取出某个URL，并下载该URL对应的网页内容，然后将该URL放入到已抓取URL队列；

[0037] S104、通过搜索获取网页内容中与生态环境主题相关的数据，判断获取的数据是否已被历史获取，若是则丢弃该数据和网页内容，若否则缓存该数据，丢弃网页内容，同时解析该网页中所包含的URL，判断该URL是否是已抓取URL队列中所包含的URL，若是，则丢弃，若否，则将其放入待抓取URL队列中；

[0038] S105、循环执行S103-S104，直至待抓取URL队列中的所有URL被完全抓取，或系统命令停止抓取，循环结束。

[0039] 在上述技术方案的基础上：所述S2的实现过程为：将S104缓存的数据进行清洗和整理，并建立索引存入数据库中。

[0040] 在上述技术方案的基础上，所述GIS脱密处理方法包括地理要素数据及其属性脱密步骤，和空间位置精度脱密步骤，地理要素数据及其属性脱密步骤包括：使用GIS软件提供的编辑工具，删除涉密的空间地理数据及其属性数据；

[0041] 空间位置精度脱密步骤包括：利用GIS软件提供的投影转换和空间校正工具对源空间地理数据的空间位置进行精度干扰。

[0042] 在上述技术方案的基础上，所述空间位置精度脱密的具体实现过程为：

[0043] L201、投影前新建一个投影坐标系用于输出图幅，设置投影坐标系名称及参数，完成投影坐标系定义；

[0044] L202、在GIS工具中选择“投影”工具，输入参数，选择新建立的投影坐标系作为原始图幅的输出坐标系；

[0045] L203、重复L202，选择需要的投影坐标系为最终坐标系；

[0046] L204、加载待变换的原始数据，打开“空间校正”工具；

[0047] L205、选择需校正的数据，选择射影校正方法，使用工具对图幅创建移位连接；

[0048] L206、进行校正、保存，完成数据变换。

[0049] 在上述技术方案的基础上，所述生态环境空间大数据的集成包括同要素空间数据集成和不同要素的空间数据集成；

[0050] 同要素空间数据集成包括：

[0051] a、对各精度较高的小区域中的数据进行综合，提取其主要特征并归并，形成精度较低但空间范围较大的大区域的数据；

[0052] b、确定精度较低的大区域的数据精度，将大区域覆盖各精度较高小区域的部分提取出来，对提取出来的区域进行数据综合使其精度满足要求，最后进行接边处理形成完整的区域特定尺度数据提取；

[0053] c、由某区域同尺度不同时间的多个数据集推断出在单个数据集中分辨不出的生态环境或利用数据表达地学过程或特征的相关性来提高原来数据的精度；

[0054] d、使用多分辨率模型来表达空间实体，实现数据更新；

[0055] 不同要素的空间数据集成包括：

[0056] h、不同要素空间尺度可比时，利用环境要素的相关性由某一或某些数据对某特定要素进行质量检测、数据综合、数据细化、辅助推导进行空间相关分析；利用不同要素之间的相关性生成新数据；

[0057] i、不同要素空间尺度不可比时，利用不同数据集进行背景参照分析和要素加权分析。

[0058] 本发明的有益效果为：

[0059] 本发明基于GIS软件提供的数据和图像处理工具，从地理数据内容和空间地理数据空间精度两个方面，进行了空间地理数据的脱密处理，通过脱密处理的空间地理数据既可以满足信息化的实用价值又可以保证涉及国家安全的数据保密。

[0060] 本发明通过数据集成，数据集成实现网络化环境信息系统间不同尺度数据交换、传输和共享、数据互操作，是利用空间数据进行环境空间分析和决策的重要环节。

具体实施方式

[0061] 下面结合具体实施例对本发明作进一步阐述。

[0062] 实施例：

[0063] 本实施例的一种生态环境空间大数据集成方法，包括生态环境数据获取步骤、空间地理数据获取步骤，和生态环境空间大数据集成步骤：

[0064] 生态环境数据获取步骤如下：

[0065] 第一步、利用网络爬虫从多个初始URL开始下载网页内容，通过搜索策略获取网页中与生态环境主题相关的数据，同时不断从当前页面提取新的URL，根据网页抓取策略放入待抓取URL队列中，循环执行，直至停止，循环结束，其中，生态环境主题相关的数据包括水质监测数据、大气环境监测数据、土壤环境检测数据或/和生态环境污染源信息。具体的实现过程如下：

[0066] S101、启动网络爬虫程序；

[0067] S102、选取多个初始URL，并将其放入待抓取URL队列中；

[0068] S103、从待抓取URL队列中取出某个URL，并下载该URL对应的网页内容，然后将该URL放入到已抓取URL队列；

[0069] S104、通过搜索获取网页内容中与生态环境主题相关的数据，判断获取的数据是否已被历史获取，若是则丢弃该数据和网页内容，若否则缓存该数据，丢弃网页内容，同时解析该网页中所包含的URL，判断该URL是否是已抓取URL队列中所包含的URL，若是，则丢弃，若否，则将其放入待抓取URL队列中；

[0070] S105、循环执行S103-S104，直至待抓取URL队列中的所有URL被完全抓取，或系统命令停止抓取，循环结束。

[0071] 抓取策略包括广度优先搜索策略、深度优先搜索策略和最佳优先搜索策略。

[0072] 广度优先搜索策略的主要思想是：由根节点开始，首先遍历当前层次的搜索，然后才进行下一层的搜索，依次类推逐层的搜索。

[0073] 深度优先搜索策略的主要思想是：从根节点出发找出叶子节点，在一个网页中，选择一个超链接，被链接的网页将执行深度优先搜索，形成单独的一条搜索链，当没有其他超链接时，搜索结束。

[0074] 最佳优先搜索策略，通过计算URL描述文本与目标网页的相似度，或者与主题的相关性，根据所设定的阈值选出有效URL进行抓取。

[0075] 系统包括爬虫主控模块、网页下载模块、网页解析模块、URL调度模块、数据清洗模

块和数据显示模块。

[0076] 爬虫主控模块，生成初始URL，并将这些初始URL放入待抓取URL队列，启动网页下载器下载网页内容，然后解析网页内容，提取需要的数据和URL地址，进入工作循环，控制各个模块工作流程，协调各个模块之间的工作。

[0077] 网页下载模块，对于可匿名访问的网页，可以直接下载，对于需要身份验证的，需要模拟用户登录后再下载，对于需要数字签名或数字证书才能访问的网站，需要获取相应证书，加载到程序中，通过验证之后才能下载网页，数据下载完成后，将下载的网页数据传递给网页解析模块，将URL地址放入已抓取URL队列。

[0078] 网页解析模块，从网页中提取满足要求的信息传递给清洗模块，提取URL地址传递给URL调度模块，另外还通过正则表达式匹配的方式或直接搜索的方式来提取满足特定要求的数据，将这些数据传递给数据清洗模块。

[0079] URL调度模块，接收网页解析模块传递来的URL地址，然后将这些URL地址和已抓取URL队列中的URL地址比较，如果URL存在于已抓取URL队列中，就丢弃这些URL地址，如果不存在于已抓取URL队列中，就按系统采集的网页抓取策略，将URL放入待抓取URL地址相应的位置。

[0080] 数据清洗模块，接收网页解析模块传来的数据，网页解析模块提取的数据，然后对这些数据进行清洗整理，整理为满足一定格式的数据，然后存入数据库中。

[0081] 生态环境主题为生态环境关键词集，生态环境关键词集的获取过程为：

[0082] S001、定期地从学术论文网站上获取论文，并将新获取的论文保存在语料库中；

[0083] S002、对语料库中各篇论文的词语集进行数值化处理；

[0084] S003、将数值化处理结果作为训练样本导入Word2Vec模型进行训练，得到各个词语的词向量；

[0085] S004、针对每个词语，分别根据词向量计算该词语与其它各个词语的欧式距离或向量夹角，然后根据计算结果选取欧式距离最近或向量夹角最小的前N个其它词语作为相关词语，得到相关词语集，其中，N为介于20～100之间的自然数；

[0086] S005、将词语为“生态环境”的相关词语集作为生态环境关键词集，然后对生态环境关键词集进行补充；

[0087] S006、将步骤S002的数值化处理结果作为训练样本导入LDA主题模型进行训练，得到主题-词语矩阵和论文-主题矩阵，其中，所述主题-词语矩阵表示每个主题中出现每个词语的概率，所述论文-主题矩阵表示每篇论文中出现每个主题的概率；

[0088] S007、针对所述主题-词语矩阵中的各个主题，选取在对应主题中出现概率最大的前M个词语作为特征词语，得到特征词语集合，同时根据所述论文-主题矩阵获取各个主题的关联主题，其中，M为介于20～100之间的自然数，所述关联主题是指与某个主题共同出现在同一篇论文中的另一共现主题；

[0089] S008、针对现有生态环境关键词集合中的各个词语，将其分别作为目标词语，查找到在特征词语集合中包含该目标词语的所有目标主题，以及查找到与各个目标主题对应的目标关联主题；

[0090] S009、将所有目标主题和所有目标关联主题中的且不在现有生态环境关键词集合中的词语，添加到现有生态环境关键词集合中。

[0091] 获取各个主题的关联主题的过程为：

[0092] S0071、针对所述论文-主题矩阵中的各个主题，分别计算其与另一主题共同出现在各篇论文中的共现概率之和；

[0093] S0072、针对所述论文-主题矩阵中的各个主题，选取对应于其共现概率之和最大的前K个另一主题作为关联主题，其中，K为介于3～10之间的自然数。

[0094] 对各篇论文的词语集进行数值化处理：

[0095] S0021、对语料库中的所有论文进行切词处理，获取不同的词语，得到所述语料库的总词语集和各篇论文的词语集；

[0096] S0022、对所述总词语集中的所有词语进行唯一编码，得到包含有所有词语和与各个词语对应的数字唯一标识的词典；

[0097] S0023、根据所述词典将论文词语集中的词语转换为对应的数字唯一标识。

[0098] 第二步、将第一步中获取的网页中与生态环境主题相关的数据进行清洗和整理，并建立索引存入数据库中。

[0099] 具体是：将S104缓存的数据进行清洗和整理，并建立索引存入数据库中。

[0100] 空间地理数据获取步骤如下：

[0101] 第一步、直接从地理信息服务器中获取空间地理数据。

[0102] 第二步、通过GIS脱密处理方法对L1中获取的空间地理数据进行脱密处理。

[0103] GIS脱密处理方法包括地理要素数据及其属性脱密步骤，和空间位置精度脱密步骤，地理要素数据及其属性脱密步骤包括：使用GIS软件提供的编辑工具，删除涉密的空间地理数据及其属性数据；

[0104] 空间位置精度脱密步骤包括：利用GIS软件提供的投影转换和空间校正工具对源空间地理数据的空间位置进行精度干扰。

[0105] 具体的实现过程为：

[0106] L201、投影前新建一个投影坐标系用于输出图幅，设置投影坐标系名称及参数，完成投影坐标系定义；

[0107] L202、在GIS工具中选择“投影”工具，输入参数，选择新建立的投影坐标系作为原始图幅的输出坐标系；

[0108] L203、重复L202，选择需要的投影坐标系为最终坐标系；

[0109] L204、加载待变换的原始数据，打开“空间校正”工具；

[0110] L205、选择需校正的数据，选择射影校正方法，使用工具对图幅创建移位连接；

[0111] L206、进行校正、保存，完成数据变换。

[0112] 基于GIS软件提供的数据和图像处理工具，从地理数据内容和空间地理数据空间精度两个方面，进行了空间地理数据的脱密处理，通过脱密处理的空间地理数据既可以满足信息化的实用价值又可以保证涉及国家安全的数据保密。

[0113] 生态环境空间大数据集成步骤如下：

[0114] 从步骤S2中的数据库中提取目标数据，并加载到脱密处理后的空间地理数据中，得到集成的生态环境空间大数据。

[0115] 数据集成是对数据空间、时间和属性的统一处理，但由于计算机数据表达的离散化、人们处理事务的思维方式和已有地球空间数据的静态特征，导致在数据集成中常把时

间作为一个常量或参数对待，结果使不同空间尺度数据集成为数据集成中最经常的形式，由于的同种地学环境现象或过程在不同空间尺度上表现出不尽相同的性质，需要用到各尺度的数据才能完全反映一种物理过程，在多要素分析中，在某一个尺度上的数据要用到另一尺度上的其他要素数据时也涉及到多尺度数据的集成。

[0116] 生态环境空间大数据的集成包括同要素空间数据集成和不同要素的空间数据集成。

[0117] 同要素空间数据集成，空间实体和地学过程在时间上有一定的稳定性，因而在地学分析中常把时间处理为常数，同种要素多尺度数据集成在使用中可以表现为不同的形式，同要素空间数据集成包括：

[0118] a、对各精度较高的小区域中的数据进行综合，提取其主要特征并归并，形成精度较低但空间范围较大的大区域的数据；

[0119] b、确定精度较低的大区域的数据精度，将大区域覆盖各精度较高小区域的部分提取出来，对提取出来的区域进行数据综合使其精度满足要求，最后进行接边处理形成完整的区域特定尺度数据提取；

[0120] c、由某区域同尺度不同时间的多个数据集推断出在单个数据集中分辨不出的生态环境或利用数据表达地学过程或特征的相关性来提高原来数据的精度；

[0121] d、使用多分辨率模型来表达空间实体，实现数据更新；

[0122] 不同要素的空间数据集成包括：

[0123] h、不同要素空间尺度可比时，利用环境要素的相关性由某一或某些数据对某特定要素进行质量检测、数据综合、数据细化、辅助推导进行空间相关分析；利用不同要素之间的相关性生成新数据；

[0124] i、不同要素空间尺度不可比时，利用不同数据集进行背景参照分析和要素加权分析。

[0125] 本发明通过数据集成，数据集成实现网络化环境信息系统间不同尺度数据交换、传输和共享、数据互操作，是利用空间数据进行环境空间分析和决策的重要环节。

[0126] 本发明不局限于上述可选实施方式，任何人在本发明的启示下都可得出其他各种形式的产品，但不论在其形状或结构上作任何变化，凡是落入本发明权利要求界定范围内的技术方案，均落在本发明的保护范围之内。