计算机系统应用 ISSN 1003-3254, CODEN CSAOBN
Computer Systems & Applications,2018,27(6):209−213 [doi: 10.15888/j.cnki.csa.006399]

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

地质灾害信息存储技术及检索方法①

姚梦辉, 刘军旗, 封瑞雪, 陈根深, 赵剑雄

(中国地质大学 (武汉) 教育部长江三峡库区地质灾害研究中心, 武汉 430074)
通讯作者: 刘军旗, E-mail: liujqg@126.com

摘　要: 地质灾害调查、勘查及防治等工作过程中, 获取了大量的多源异构数据, 其中的文本数据多以文件名检索或大字段形式整体存储, 这种传统的存储方式不能满足文本信息中有用信息的快速检索与提取, 是当前地质灾害数据存储和检索所面临的一个重要问题. 本文基于非结构化数据库技术、中文分词技术、关键词提取技术, 实现了地质灾害文本数据中任意有用信息的快速检索及与统计, 可以为灾害数据的深层挖掘与融合提供有力支持.

关键词: 地质灾害; 非结构化数据库; 中文分词; 段落切分; 信息检索

引用格式: 姚梦辉,刘军旗,封瑞雪,陈根深,赵剑雄.地质灾害信息存储技术及检索方法.计算机系统应用,2018,27(6):209–213. http://www.c-s-a.org.cn/1003-3254/6399.html

Geological Hazard Information Storage Technology and Tetrieval Method

YAO Meng-Hui, LIU Jun-Qi, FENG Rui-Xue, CHEN Gen-Shen, ZHAO Jian-Xiong

(Three Gorges Research Center for Geo-hazards (Ministry of Education), China University of Geosciences (Wuhan), Wuhan 430074, China)

Abstract: In the process of investigation, exploration, and prevention about geologic hazard, a large number of heterogeneous data including text data is obtained. The method to storage the text data in file name search or large field is traditional, cannot meet rapidly retrieve and extract the useful information in the text data. It is an important problem faced by the geological hazard data storing and retrieving. In this study, based on the NoSQL, Chinese word segmentation, and Chinese keyword extraction technology, fast retrieval and statistics of any useful information are realized in geological hazard text data. It can provide strong support for the deep mining and fusion of hazard data.

Key words: geologic hazard; NoSQL; Chinese words segmentation; paragraphs segmentation; information retrieve

在长期的地质灾害调查、勘查及防治等工作中, 积累了大量的数据, 既有结构化的, 也有非结构化的, 通常呈碎片化状态以文本、图形和图像方式堆积着[1], 其中包含了大量有效信息, 同时也掺杂了很多无效信息, 如何从这些数据中剔除无效信息, 提取有用信息, 为分析、决策提供支持是当前地质灾害研究工作中遇到的一个难点, 同时也是当前研究的热点. 严光生等[2]对地质调查大数据研究中的主要问题进行了分析; 李超岭等[3]对地质调查主流程信息化的方向和目标进行了讨论; 李靖等[4]讨论了大数据环境下适合多源异构地质数据的存储技术; 李超岭等[5]利用 Hadoop HDFS 与 HBASE 等工具针对非结构化数据的存储、阅读和搜索进行了研究; 吕鹏飞等[6]提出了基于文献的知识发现应用于成矿预测领域的研究思路, 并进行了实验分析; 张戈一等[7]针对地质图书馆书籍多, 数据资料庞大, 数据资料增长快而无法发现兴趣点的问题, 提出了基于大数据分析挖掘的地质文献推荐方法.

在地质灾害非结构化数据中含有大量的具有专题性质和总结性质的报告文档以及地质文献类文本数据. 作为一种常见的非结构化数据, 这些文本数据大多散

① 基金项目: 国家自然科学基金 (41572336)
收稿时间: 2017-10-08; 修改时间: 2017-11-01; 采用时间: 2017-11-27; csa 在线出版时间: 2018-05-28

落在本地硬盘或者以大字段的形式存储在结构化数据库中[8], 导致了针对这些非结构化数据的存储、发现和挖掘效率十分低下. 针对这一问题, 本文基于 MongoDB, 采用分类、段落切分、中文分词、关键字提取等文本预处理技术和关键字检索技术, 实现了地质灾害文本数据中任意有用信息的快速检索, 以及含有相关检索信息的文件名及文件中相关段落的智能提取, 为地质灾害不良信息的数据挖掘提供技术支持.

1 基于自然语言处理的文本预处理

文本预处理包括中文分词、词性标注、去停留词、关键词提取等操作, 是中文自然语言处理的基础.

1.1 中文分词

本文面向的地质灾害文本主要是中文文本. 中文是以字为基本单位, 字与字组合成词, 表达具体的含义, 中文的词与词之间是没有空格, 在对中文文本进行处理时, 计算机无法直接识别中文词语, 需要进行中文分词操作.

本文使用“结巴”分词对文本进行中文分词操作[9]. “结巴”分词是基于词典的中分分词工具, 其基本原理是: (1) 基于字典构建 Trie 树, 即字典树, 以词组集<地质灾害, 地质数据, 地学信息>为例, 构成 Trie 的一棵子树, 如图 1 所示, 从该子树的根节点到任一节点的路径上所有节点按顺序排列的字符串都是该节点对应的词组; (2) 基于最大概率法分词, 通过搜索字典树, 列出所有字典中出现的词组构成有向无环图, 转换成单源最短路径问题, 其中最短的路径就是分词结果. 以“地学信息”为例, 可以分成<地, 地学, 学, 学信, 信, 信息, 息>, 其中每一个词都是一个节点, 如图 2 所示, 分词结果为“地学, 信息”, 其中数字表示两个节点之间的权重. 设 w 表示某词的权重, n 表示总词频, n_i 表示该词的词频, 计算公式:

$$w = log\frac{n}{n_i} \tag{1}$$

1.2 词性标注、去停留词

词的词性包括名词、动词、形容词等各种词性, 真正能够表达具体含义的一般只有名词或者动词. 因此, 需要将没有具体含义的词过滤掉, 即去停留词, 保留能够表达具体含义的词. 以“地质灾害数据中包含了大量的文本信息, 其中蕴含了丰富的信息”, 分词后词性标注的结果为: “地质灾害/n, 数据/n, 中/f, 包含/v, 了/ul, 大量/n, 的/uj, 文本/n, 数据/n,, /x, 其中/r, 蕴含/v, 了/ul, 丰富/a, 的/uj, 信息/n”; 去停留词之后的结果为: “地质灾害/n, 数据/n, 大量/n, 文本/n, 数据/n, 蕴含/v, 信息/n”.

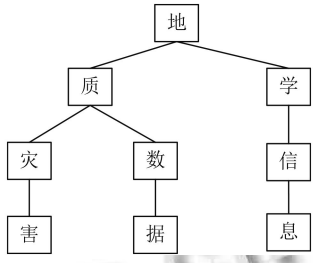

图 1 Trie 树示例

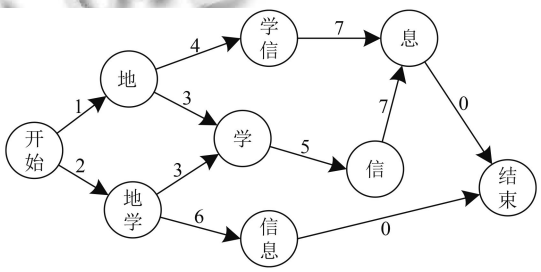

图 2 “地学信息”构成的有向无环图

1.3 关键词提取

关键词是指能够表示一篇文档的基本的特征的关键词语, 关键词提取就是从文档中抽取一些词语作为这篇文档的关键词. 本文采用 TF-IDF 算法提取关键词. TF-IDF 是一种用于信息检索与数据挖掘的常用加权技术, 通过统计的方法来评估一个词对一篇文档的重要程度, 具体来说就是如果一个词在本篇文档中出现的次数多, 而在其他文档中出现的次数小, 则该词更能代表该文档的主题[10]. TF-IDF 计算公式如下:

$$TF-IDF = TF * IDF = \frac{d_x}{d}\log\frac{N}{N(x)} \tag{2}$$

其中, TF 表示某个词在文档出现的频率, IDF (Inverse Document Frequency) 表示语料库中包含该词的文档的数目的倒数. d_x 表示在文档 d 中词 x 出现的次数, d 表示文档 d 中词的总数, N 表示语料库中文档的总数, $N(x)$ 表示语料库中包含词 x 的文档个数.

使用 TF-IDF 算法提取关键词, 在中文分词、词性标注和去停留词的基础上, 计算每个词的 TF-IDF 值, 然后按照每个词的 TF-IDF 值降序排列, 输出指定个数的词作为关键词.

2 基于非关系型数据库的文件存储及有用信息检索

2.1 基于非关系型数据库的文本存储

对于非结构化数据和半结构化数据的传统的存储方式存在很多的弊端, 而非关系型数据库对于非结构化数据存储, 则具有关系型数据库所不能比拟的优势. 当前许多学者对于矢量数据[11,12]、时空数据[13–15]的非关系数据库存储方面的研究为我们对大量文本数据的存储提供了借鉴. 文本数据作为一种常见的非结构化数据, 适合使用非关系型数据库进行存储.

本选取 MongoDB 作为地质灾害文本的存储工具. MongoDB[16]是一种面向文档的非关系型数据库, 以文档 (Document) 作为最基本的存储单位, 文档中以松散的“key/value”对形式表达, 数据结构灵活, 并内置文件存储规范 GridFS.

GridFS 的存储机制是将文件分割成多个小的 chunk(文件片段), 片段大小一般为 256 k, 每个 chunk 是一个文档, 这些 chunk 存储在 chunks 中构成一个集合. 因此 GridFS 在本质上仍然是使用集合对文件进行存储. GridFS 使用两个集合存储一个文件, 即 fs.files 和 fs.chunks. 这样每个子区域就包含了两个集合, 文件内容存储在 chunks 中, 与文件相关的信息如文件名、文件类型、入库时间等存储在 files 中.

2.2 信息检索方法

信息检索是指用户从大量包含各种信息的文档集合中查找所需要的信息或知识的过程. 常用的信息检索模型有 4 种: 布尔模型[17]、空间向量模型[18]、概率模型[19]以及语言模型[20]. 布尔模型即基于给定查询词的检索, 通过“and”、“or”、“not”3 种运算符来连接, 构成一个布尔表达式. 空间向量模型是将查询和文档作为两个空间向量, 查询和文档的相关度由向量的余弦值决定. 概率模型则是根据文档和给定查询相似的概率大小判断文档的相关性. 语言模型将给定的查询看作是由文档生成的, 查询是转化为计算文档生成查询的概率. 通过对比以上 4 种模型, 布尔模型是最简单、最直接的一种查询方法.

3 验证分析

如图 3 所示, 本文针对地质灾害文本数据的挖掘包括两个模块: 第一个模块是数据上传, 包括数据的搜集整理、文件名解析, 文件内容读取, 中文分词, 词性标注、去停留词、提取关键词等操作; 第二个模块是信息的提取, 包括文件段落切分, 中文分词, 关键词查询等. 通过以上两个模块实现了地质灾害文本数据的分类存储及有用信息的检索与统计.

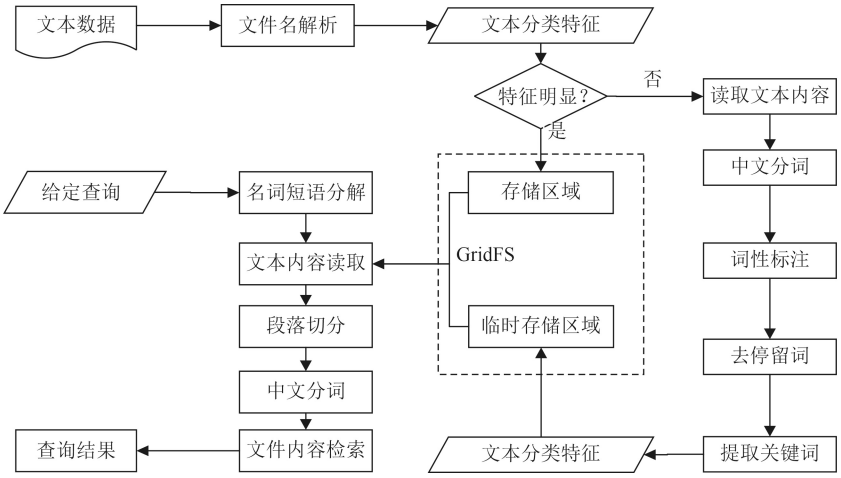

图 3 地质灾害文本存储、查询实现流程

3.1 文本分类存储

本文通过两种方式判定文本的类型: 文件名、文件内容. (1) 通常情况下, 在地质灾害类的文本名称中都包括明显的关键字, 通过关键字, 就可以确定文本的类型. (2) 对于文本名称不清晰或者含有多个关键字而无法区分的情况下, 则不能通过文本名称进行判定. 因此, 本文在数据库中划分两个文本的存储区域, 存储区域和临时存储区域. 每个区域又按照常见的地质灾害类型划分为崩塌、滑坡、泥石流、地裂缝、地面沉降、地面塌陷六个子区域, 如表 1 所示. 存储区域存储文件

名称特征明显的文本文件, 临时存储区域存储文件名特征不明显的文本.

表 1 地质灾害数据存储区域划分

灾害类型	存储区域	临时存储区域
崩塌	collapse	collapse_temp
滑坡	landslide	landslide_temp
泥石流	mud_rock_flow	mud_rock_flow_temp
地裂缝	ground_fissure	ground_fissure_temp
地面沉降	land_subsidence	land_subsidence_temp
地面塌陷	surface_collapse	surface_collapse_temp

如图 3 所示, 文本存储部分说明了文件存储的过程, 对于文本文件名特征明显的文本, 根据其特征与分类列表对比, 找到文本所属分类, 存入数据库. 对于文本文件名特征不明显的文本, 需要对文本内容进行解析. 其大致过程为: 1) 读取文件内容; 2) 中文分词、词性标注、去停留词; 3) 提取关键词. 在提取关键词之后, 对文件类型进行匹配. 例如, 关键词中同时含有滑坡和泥石流两个词, 那么, 就将 TF-IDF 较大的那个词作为文件的类型存入临时区域中对应的子区域内.

本文通过中国知网获取地质灾害类文献 169 篇作为测试数据, 整理成 word 文档后, 按照文本的存储规则存入数据库, 依据文件名中的关键词和文件内容关键词进行分类, 169 篇文献的分类结果如图 4 所示. 其中滑坡类 34 篇, 泥石流类 30 篇, 崩塌类 24 篇, 地裂缝 13 篇, 地面沉降 20 篇, 地面塌陷 15 篇.

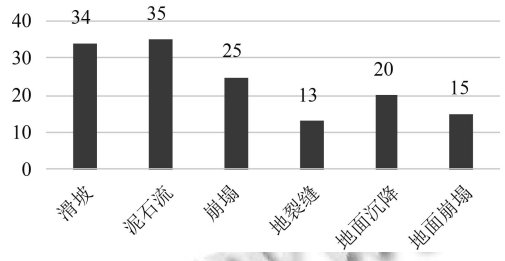

图 4 分类后文本分布

3.2 有用信息的检索

本文使用布尔模型对文本进行检索, 给定的查询词可以是单个词也可以是多个词, 并通过“and”来连接布尔表达式, 即检索目标中同时含有所有给定查询词时才会被记录到查询结果中.

通过布尔模型对文本内容对文本内容进行检索的大致过程是: 文件段落划分、中文分词、查询词检索. 最后, 将符合布尔表达式的文本名及段落写入查询结果, 并统计文件和段落个数.

1) 文件段落切分. 在文本检索的过程中, 我们关心的不仅仅是整个文件, 而是深入到了文件内部的具体段落, 目标是从大量的文本数据中快速的提取含有有用信息的段落. 因此, 在读取文件内容之后, 对文件内容按照段落进行切分, 以段落为基本单位查询检索. 在 word 文档中, 以换行符为段落标志, 一个换行符代表一个段落. 但是也可能出现的情况是空行, 只有换行符, 没有内容, 对于这样的段落将被排除在外.

2) 中文分词. 中文分词是对中文文本进行检索的基础. 此处分词, 按照“结巴”分词的搜索引擎模式进行, 这种模式将对文本进行所有可能的切分, 以避免歧义造成的无法检索到结果的问题. 以“地质灾害数据中包含了大量的文本数据, 其中蕴含了丰富的信息”为例, 按照搜索引擎模式切分, 切分结果为“地质/灾害/地质灾害/数据/中/包含/了/大量/的/文本/数据/, /其中/蕴含/了/丰富/的/信息”, 其中, “地质灾害”被切分成了“地质”、“灾害”、“地质灾害”三个词, 包含了能够切分的所有情况.

4) 给定查询检索. 在完成文件段落切分、中文分词之后, 通过“and”连接关键词, 构建布尔表达式. 将关键词依次与分词列表进行对比, 当查询目标中含有所有的关键词时, 将段落记录到查询结果中, 并将文件信息加入统计结果.

本文通过“汶川地震”、“降雨诱发”和“滑坡”三个关键词对测试数据进行检索, 检索结果如图 5、图 6 所示. 图 5 是检索的统计结果, 在测试数据中, 符合查询条件的有 3 个文件, 26 个段落. 图 6 是符合查询条件的段落信息, 可以看到段落的具体内容, 以及段落所在的文件.

通过上述处理, 实现了地质灾害文本数据中任意有用信息的快速检索及统计, 可以从一个含有大量文档的混杂数据集中, 快速得到含有任意检索信息的文件名、该文件存储位置、含有检索信息的文件中的相关段落等有效信息.

4 结论与展望

本文针对当前传统关系型数据库无法对大量文本数据进行有效存储的问题以及地质文本数据中有用信息提取所面临的困难, 基于 MongoDB, 使用段落切分、中文分词、关键字提取、信息检索等技术实现了地质灾害文本信息的快速检索与统计.

(1) 采用对文件名进行关键字提取、对比识别等方法对混杂数据集中的文件进行合理分类和提取;

关键词：汶川地震、降雨诱发、滑坡

检索结果统计：3个文件，26个段落，结果如下：

序号	文件名	存储位置	时间
1	汶川地震强震区地震诱发滑坡与后期降雨诱发滑坡控制因子耦合分析	landslide	2017-07-1401:19:58
2	汶川地震滑坡灾害研究综述	landslide	2017-07-14 01:19:59
3	汶川地震区暴雨滑坡泥石流活动趋势预测	landslide_temp	2017-07-14 01:19:58

图 5 基于关键词检索结果统计

1、汶川地震强震区地震诱发滑坡与后期降雨诱发滑坡控制因子耦合分析

段落1：摘要：本文以汶川地震强震区北川县典型研究区为例，利用高分辨率航片、SPOT5 卫星图像对北川县典型研究区进行了“5.12”地震之后和“9.24”降雨之后诱发的滑坡解译，解译结果显示：“5.12”地震诱发滑坡 1999 个“9.24”强降雨诱发滑坡828个“9.24”强降雨导致原有地震滑坡面积扩大的滑坡 150 个。研究表明：地震和强降雨都是诱发滑坡的动力成因“9.24”强降雨诱发的滑坡面积是“5.12”地震诱发滑坡面积的 25%，强降雨诱发滑坡的数量增加了41.4%；强降雨不仅诱发新的滑坡，而且促使原来地震滑坡复活，并扩大其面积，强降雨导致地震诱发的滑坡面积扩大了原面积的 68.7%。同时，在遥感解译数据基础之上，开展地震诱发滑坡与降雨诱发滑坡规模对比和控制因子耦合分析及地震与降雨耦合灾害链模式研究，为进一步分析研究地震灾区滑坡的产生、发展趋势、危险性和风险评价等预测预报提供科学依据，也为汶川震区恢复重建中的减灾防灾提供决策参考。

……

图 6 基于关键词检索的部分检索结果

(2) 对文件名特征不明显的文本, 采用中文分词、TF-IDF 关键词提取等方法对文件内容进行处理、分类, 也可以通过这个方法对 (1) 中分类的合理性进行验证;

(3) 通过任意检索关键词对混杂数据集进行有用信息的快速检索和统计, 获得含有检索信息的文件及文件中具体的段落.

参考文献

1 吴冲龙, 刘刚, 张夏林, 等. 地质科学大数据及其利用的若干问题探讨. 科学通报, 2016, 61(16): 1797–1807.

2 严光生, 薛群威, 肖克炎, 等. 地质调查大数据研究的主要问题分析. 地质通报, 2015, 34(7): 1273–1279.

3 李超岭, 李丰丹, 李健强, 等. 智能地质调查体系与架构. 中国地质, 2015, 42(4): 828–838.

4 李婧, 陈建平, 王翔. 地质大数据存储技术. 地质通报, 2015, 34(8): 1589–1594.

5 李超岭, 李健强, 张宏春, 等. 智能地质调查大数据应用体系架构与关键技术. 地质通报, 2015, 34(7): 1288–1299.

6 吕鹏飞, 王春宁, 周峰, 等. 基于文献的知识发现在成矿预测领域的应用研究. 中国矿业, 2017, 26(9): 85–91.

7 张戈一, 胡博然, 常力恒, 等. 基于大数据分析挖掘的地质文献推荐方法研究. 中国矿业, 2017, 26(9): 92–97.

8 王存宇, 李珂, 许锦才, 等. 面向云存储的非结构化数据存储研究. 计算机时代, 2015, (5): 13–15, 18.

9 FXSJY: Python 中文分词组件 jieba. http://www.oschina.net/p/jieba/, 2017-07-10.

10 牛萍, 黄德根. TF-IDF 与规则相结合的中文关键词自动抽取研究. 小型微型计算机系统, 2016, 37(4): 711–715.

11 雷德龙, 郭殿升, 陈崇成, 等. 基于 MongoDB 的矢量空间数据云存储与处理系统. 地球信息科学学报, 2014, 16(4): 507–516.

12 陈崇成, 林剑峰, 吴小竹, 等. 基于 NoSQL 的海量空间数据云存储与服务方法. 地球信息科学学报, 2013, 15(2): 166–174.

13 闫玮. 基于 MongoDB 与 Hadoop 的地学遥感大数据管理系统的设计[硕士学位论文]. 兰州: 兰州大学, 2016.

14 廖理. 基于 NoSQL 的时空数据模型构建和存储方案研究[硕士学位论文]. 重庆: 重庆大学, 2015.

15 阙翔. 面向动态过程模拟和实时表达的地质时空数据模型研究[博士学位论文]. 武汉: 中国地质大学, 2015.

16 MongoDB: What is MongoDB? https://www.mongodb.com/what-is-mongodb. [2017-07-09].

17 Dvorský J, Pokorný J, Snášel V. Word-based compression methods and indexing for text retrieval systems. Proceedings of the 3rd East European Conf. on Advances in Databases and Information Systems. Maribor, Slovenia. 2009. 76–84.

18 Salton G, Wong A, Yang CS. A vector space model for automatic indexing. Communications of the ACM, 1975, 18(11): 613–620. [doi: 10.1145/361219.361220]

19 Gudivada VN, Raghavan VV, Grosky WI, *et al*. Information retrieval on the world wide web. IEEE Internet Computing, 1997, 1(5): 58–68. [doi: 10.1109/4236.623969]

20 Ponte JM, Croft WB. A language modeling approach to information retrieval. Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. Melbourne, Australia. 1998. 275–281.