基于互联网分布式爬取构建自然灾害知识库专家系统

一. 成果完成人及所在单位： 陈秋瑾，电气工程与自动化学院　　指导教师：何原荣

二. 成果简介：

本系统以互联网分布式爬取网络数据为基础技术支撑，结合知识管理理念构建知识树体系，以自然灾情环境为研究背景，收集来自境内外关于灾情环境的相关知识文献、专业化的灾情数据，以及整合实时监测台风路径图便于系统监控预测台风；将要爬取的自然灾害对象相关联的关键字分配到知识树各个子节点下，从而系统服务器将根据设置的关键字爬取网络相关数据信息，而使用者也可根据输入已有关键字检索到所需新闻信息或学术文献；系统将爬取下来的数据信息保存到本地数据库中，数据库中根据知识树结构构建相关表，包括总表（根节点），各个子表（子节点）。本系统的最终成果分为两个界面展示，分别为系统前端与系统后台，前端即客户端，面向用户为各个高校老师、学生等使用者，向用户提供了简洁大方的检索界面，用户可在此网站中检索到想要搜寻的相关自然灾害的新闻讯息及学术文献，网站将根据知识树或数据来源不同（知网、中国科技论文在线、微博、微信、搜狐新闻、百度新闻、知乎等网站）将爬取的数据进行分类展示。除此之外，前端还具有可以声音形式播放文献摘要、显示被引用次数、作者详细信息等功能，这极大地方便了用户的查询时间，也提高了学者、研究员、科研人员的学习及工作效率。系统后台即管理员端，管理员拥有设置知识树节点、增删关键字、设置强制爬取天数、隐藏爬取下的各条数据信息等权限。本系统定位于我校计算机信息工程学院的数字福建自然灾害监测大数据研究所，提供科研协作服务系统和文件数据存储共享平台，融合遥感、无人机、三维点云、GNSS、社交网络等多源数据，帮助研究所更好的集中存储和全面管理包括个人、研究组、内部流转以及对外分享的各类非结构化数据。本项研究在理论研究及实际应用中均有学术价值。

三. 主要研究内容及创新点：

自然灾害知识库专家系统的建设，旨在将互联网上关于灾情信息、遥感等技术信息用网络爬虫技术爬取下来存储在数据库中，通过独立的网站平台将爬取到的数据分类展示以及管理爬取到的数据，并且使不同用户（普通用户和管理员）查看权限不一样，从而设计搭建了两种平台模式分别针对两类权限不同的用户；系统方便用户快速一览整体，旨在为数字福建自然灾害监测研究所及相关实验室的研究人员提供准确、时效、有条理、有依据的信息资讯、学术文献进行浏览查阅及下载保存，并且进行管理；该系统的建设目的主要是为了帮助科研人员了解近期灾情动态、分析自然灾害数据，从而能够做出相关防范计划及研究。

本系统主要针对三个模块进行研究，包括知识管理模块、外部知识获取模块、知识智能模块，其中知识管理模块包含知识库、搜索引擎、知识激励、知识社区、专家系统、个人知识管理等系统管理；外部知识获取模块采用分布式网络爬虫采集技术获取知识，自动去重等清洗技术，自动化归类实现知识获取的功能；知识智能模块通过知识采集、知识树知识图谱模型和提取技术实现知识的挖掘从而达到知识智能利用功能。本系统在实现基础功能之上还已将台风实施路径监测平台加入为子功能栏，该栏目平台详细展示了历年各个台风的风力大小、经过路径等信息。本系统的研究创新点在于以下几个方面：(1)使用互联网爬虫方式获取灾情专家知识，形成灾情数据的积累；(2)使用分布式的架构方式，实现国内外的数据的快速爬取和处理；(3)使用结构化语义分析处理机制，快速进行数据的预处理，提取有价值的信息；(4)使用成熟的多来源的自动化翻译接口，自动翻译整理来自不同国家的灾情相关的研究资料，提高灾情环境专家知识库的情报能力水平；(5)使用协同工作技术，通过知识管理系统，为实验室创造可以分享知识，随时随地查阅知识的系统，方便大家的信息共享和交流，建立实验室在灾情环境专业领域上的专家知识的积累。后期系统还将实现自动灾情数据分析的功能，同时将会继续研究实现自动通过公众号、App 等方式让使用者快速方便的移动互联，从而