课题名称	京津冀安全大数据获取与分析平台		
学院名称	软件学院	专业名称	软件工程
学生姓名	李天鹏	指导教师	王文俊

一、 原始依据

1.1 课题的来源及意义

随着科学、技术和工程的迅猛发展，近 20 年来，许多领域(如光学观测、光学监控、健康医护、传感器、用户数据、互联网和金融公司以及供应链系统)都产生了海量的数据。大数据的数据集大小以难以想象的速度增长，给数据处理带来了极大的挑战[1-3]。

而随着互联网的普及，大众舆论主要聚集在尤其以新浪微博为代表的社交网络上。这使得新浪微博等新媒体在舆论导向上的作用至关重要。2015 年 8 月 12 日凌晨，天津滨海新区塘沽开发区的瑞海国际物流有限公司所属危险品仓库发生爆炸。该重大事故发生后，应急处理工作稳步推进，然而该事件却在互联网上持续发酵，谣言信息，虚假内容漫天飞，负面网文火上浇油引起大众的恐慌。

重大事件发生后，总会受到各种新闻媒体的广泛关注，而大众也会在阅读新闻后发表评论或进行转载，进而加速其传播。新闻的性质以及其影响力也在其扩散的过程中发生变化，其中变化的性质，扩散的速度以及对大众的影响总会呈现一定的规律。

1.2 本课题的研究目的

研究大数据获取及分析技术，爬取京津冀安全相关的开源数据，包括京津冀所有路网、地块、单位的数据，以及京津冀所有突发事件新闻及其舆情数据（数据格式见 2.2 具体要求），爬取的数据存在无用数据以及被污染的数据，所以需要进行清洗。清洗后对数据进行聚类，并在 Mediawiki 和 MVC 上实现这些事件的语义浏览与统计分析，形成地域-事件-人-行为的多层异质网络，并进行多层异质复杂网络分析，挖掘安全事件背后的驱动机制，为京津冀的安全提供策略保障。

1.3 本课题的工作基础以及国内外研究现状

1.3.1 本课题的工作基础

随着大数据的发展，大数据分析平台构架也是层出不穷，这使得信息分析方法的探索不断有新的突破，[4]通过基于工作流的数据挖掘框架和云计算相关的工具（ Apache Hadoop、SciDB）相结合进行大数据的信息提取，可以使得我们对于网络舆情分析成为可能，同时，许多学者们对大数据时代的情报分析技术与方法进行了研究，包括研究大数据预测建模[5]，大数据处理和分析的终极目标是借助对数据的理解辅助人们在各类应用中作出合理的决策.在此过程中,深度学习、知识计算、社会计算和可视化起到了相辅相成的作用.

(1) 深度学习提高精度:如前所述,要挖掘大数据的大价值必然要对大数据进行内容上的分析与计算,而传统的数据表达模型和方法通常是简单的浅层模型学习,效果不尽人意.深度学习可以对人类难以理解的底层数据特征进行层层抽象,凝练具有物理意义的特征,从而提高数据学习的精度.因此,深度学习是大数据分析的核心技术;

(2) 知识计算挖掘深度:每一种数据来源都有一定的局限性和片面性,只有对各种来源的原始数据进行融合才能反映事物的全貌,事物的本质和规律往往隐藏在各种原始数据的相互关联之中.而借助知识计算可以将碎片化的多源数据整合成反映事物全貌的完整数据,从而增加数据挖掘的深度.因此,基于大数据的知识计算是大数据分析的基础.如何基于大数据实现新知识的感知,知识的增量式演化和自适应学习是其中的重大挑战;

(3) 社会计算促进认知:IT 技术的发展使得社会媒体成了一类重要的信息载体,承载着对事物的客观或主观描述信息.因此,通过基于社会媒体数据的社会计算可以促进人们对事物的认知.但是,社会媒体大数据往往蕴含着一个体量庞大、关系异质、结构多尺度和动态演化的网络,对它的分析既要有效地计算方法,更需要支持大规模网络结构的图数据存储和管理结构,以及高性能的图计算系统结构和算法;

(4) 强可视化辅助决策:对大数据查询和分析的实用性和实效性对于人们能否及时获得决策信息非常重要.而强大的可视化技术,不仅可以对数据分析结果进行更有效的展示,而且可以在大数据分析过程中发挥重要作用。

构建基于大数据的数据可视化方法，开发处理大数据的高效和安全的云存储系统[6]-------SAMOA 基于云的在线挖掘平台。对大数据的研究告诉我们数据量大不是困难所在，信息分析的关键而是在于对海量、复杂、非结构化数据的分析，不借助于专业的分析工具很难在规定的时间内完成分析任务，或者很难在较短的时间范围内更多地发现大数据里潜藏着的情报价值[7,8]。运用深度学习、知识计算、社会计算和可视化的辅助能够让我们对大数据计算与分析更加到位。

1.3.2 国内外研究现状

国外对于网络舆情的研究起步相比国内要早，发展也较为成熟。关于舆情的软件主要有：由 Dave 等人研发情感分析工具 Review Seer，它是世界上第一个针对既定产品判别褒贬的；Liu 等人研发了一个名为 Opinion Observer 的系统，它可以处理在线用户对产品的评价信息，通过统计显示出特定产品的用户对其评价的优缺点，而且还可以将多种产品的用户评价放到一起进行对比；由 Gamon 等人研发的 Pulse 系统，该系统主要是用来自动挖掘网络舆情中关于汽车的评价信息；Niblack,Yi 等学者则开发出一个开放领域意见挖掘、多类型数据挖掘的意见挖掘器 Wilson 等人研发的一个可以自动识别主观性语句及语句中与主观性成分相关的系统 Opinion Finder；英国科波拉软件公司研发一款名为“感情色彩”的软件，能够对所有报纸文章对某个政党政策持否定抑或肯定态度的判断，也可实现网上评论文章对某种产品的褒贬，而且该软件的运行速度较快，普通人需要花费 1 小时浏览的文章，该软件 1 秒钟就可实现。同时国外基于网络的舆情分析也比国内开始早，相关研究多：日本海啸与地震对于 twitter 用户话题影响的研究[9]表明日本的地震及海啸报道对日语系用户话题影响巨大;利用有监督的学习构建 twitter 的好友推介算法[10];通过聚类方法进行基于行为的 twitter 社区重叠度调查[11];对 twitter 的标签网络进行的多峰事件侦测[12]但是对于中文的语义化和舆情分析方面可参考的资料较少。

当前我国舆情监测机构的数量在逐年递增，发展也日益壮大。1999 年 10 月，天津市社会科学院舆情研究所在原天津市社会科学院舆情调查研究中心的基础上成立,是国内成立较早并长时间作为国内惟一一家以“舆情研究”为名称的研究机构。2005 年该所承担了国家社科基金项目“建立社会舆情汇集和分析机制研究”。2005 年 10 月,陕西省社会舆情研究中心在西北大学挂牌成立，挂靠该校应用社会科学系。2007 年 7 月,辽宁石油化工大学舆情信息研究基地成立，挂靠该校文学院。舆情分析方面的研究也已经做了许多研究，包括基于微博的舆情分析[13-14],但实验所用数据集较小。随着社会对重大事故关注度的显著提高，不少国内的学者也研究了重大突发事故的舆情演变规律[15]，同样大多都使用了较少的数据进行了统计分析或者仅仅进行了理论分析[16]，因此结论并不能令人完全信服。真正能够在重大突发事件大数据的基础上，利用现有的数据处理与语义化、复杂网络分析的算法进行统计分析并展示到在线平台上的并不多。

2006 年至 2012 年，中国社会科学院以及人民网连续六年发布《社会蓝皮书》和《年度互联网舆情分析报告》，在 2012 年的分析报告中回顾了网络民意诉求、社会热点和网络舆论生态的演变；网民进一步的年轻化，从“80 后”向“90 后”转变；随着互联网的高度透明，政府公信力面临“塔西佗陷阱”的挑战，亟待通过社会化媒体建立公众对政府的信任，提升政府的公信力。同时对于媒体人微博言论新闻化带来道德传播等新问题，也提出了相应的意见。

在政府舆情应对方面，出现了很多以研究网络舆情，提供专业的舆论分析报告为主要服务的机构，对政府舆情的监测与应对均起到了一定的积极作用。人民网舆情检测室是国内最早开始提供这类服务的机构，从微博、论坛、门户网站等

个网络媒介上梳理网络热点，按照“政府响应”“信息透明”“政府公信力”“动态反应”“官员问责”“网络技巧”等具体的指标进行统计分析，对当地政府的应对能力做出评价分析，发布《地方网络舆情能力排行榜》，对地方政府的舆情热点把握，舆情应对，公信力的提高起到了积极的作用。中国传媒大学与中国人民大学相继成立了网络舆情研究所，定期发布网络舆情周报与月报。

此外，还有中国传媒大学公关舆情研究所、华中科技大学的舆情信息研究中心、复旦大学的传媒与舆情研究中心。发展较为成熟的有 1、天津社科院。天津社科院拥有大量舆情监测方面的专业人才，全国四分之一以上的舆情监测专业论文来自这个机构，该机构的研究人员较为成熟稳定，出版了我国第一部网络舆情研究的专著《网络舆情研究概论》2、中国人民大学舆情研究所。中国人民大学舆情研究所、人民网舆情监测室（人民日报社网络中心舆情监测室）是国内最早从事互联网舆情监测、研究的专业机构之一，在舆情监测和分析研究领域处于国内领先地位。

二、设计（研究）内容和要求

2.1 本课题的研究内容及目标

研究大数据爬取技术和分析技术，获取京津冀安全相关的数据，对这些数据进行信息抽取并分析，具体按照以下几步实施：

（1） 从网上爬取京津冀所有路网、地块、单位的数据；

（2） 从网上爬取京津冀所有突发事件新闻及其舆情数据；

（3） 对这些数据进行信息抽取，形成地域-事件-人-行为多层异质网络，并进行多层异质复杂网络分析，挖掘安全事件背后的驱动机制；

（4） 在 MVC 与 Mediawiki 上实现这些事件的语义浏览与统计分析。

2.2 具体要求

（1） 熟悉大数据获取及分析技术及其过程（基于 python 的爬虫技术，如 html 解析器 BeautifulSoup、python 网络爬虫框架 grab、scrapy 等），并能够从理论上理解消化；

（2） 对京津冀安全大数据进行信息爬取中（在新浪、搜狐等网站），爬取的数据格式为 json，具体格式如下：

{‘id(发布消息者的 id)’:,’ userhref(用户主页)’:,’ text(发布消息的内容)’:,’ feedtime(发布时间)’:,’geodata(地理信息)’:,’comment(评论)’:[‘commentCount(评论数)’:,’ comments(评论)’:[{ ‘comment_id(评论者 id)’:,’ comment_href(评论

者主页)’:,’comment_teme(评论时间)’:,’comment_geodata(评论者地理信息)’:,’comment_text(评论内容)’:}]}。

同时需要爬取京津冀所有路网、地块、单位的数据，主要形式为地理位置信息+单位名称。爬取同时要去除爬取过程中的无效数据，例如信息缺失数据，网页无关数据等。

（3） 将京津冀的舆论数据与路网、地块、单位数据进行关联，然后进行语义分析，提取用户发布的信息以及评论信息的主题，主要用到自然语言处理的相关技术，进而形成地域-事件-人-行为的复杂网络，通过聚类分类进行梳理得出同一类中的事件主旨并试着解释所体现的现象，最好能够形成高水平刊物期刊论文；

（4） 在 semantic mediawiki 上实现京津冀安全相关数据爬取、数据处理、语义浏览、统计分析、可视化等功能，故需要对 semantic mediawiki 进行系统学习掌握。

2.3 技术路线和研究手段

(1) 可行性分析：本课题的数据来源主要是微博和新闻（新浪、搜狐），现有的软硬件基础（基于 python 的爬虫技术）可以解决数据爬取相关的问题。其次，根据对国内外研究现状的调查，有许多前沿的技术可以用在重大突发事件的数据构成的网络上，可以解决复杂网络的舆情分析相关的问题。同时正在学习语义化和复杂网络分析相关的论文，与实验室相关老师学生进行交流初步构建了课题的整体框架，因此课题的可行性可以保证。

(2) 已具备的实验条件：实验室已具备本课题相关的一部分数据和爬取代码实现，本人也具备一定的数据爬取、数据处理相关的技术和经验。现已具备了实验的硬件和技术等多方面的实验条件，并对每一步做了详细的可行性分析。

(3) 技术路线：爬取微博、新闻数据->整理数据->数据清洗->特征选择->数据表示->使用 NetworkX 构建复杂网络->分类聚类->统计分析->得出结论->二次开发 Mediawiki,设计展示页面->成果展示

(4) 技术手段：爬取数据使用 python 即可，爬取数据的数据格式为 json。具体如下：

{‘id(发布消息者的 id)’:,’userhref(用户主页)’:,’text(发布消息的内容)’:,’feedtime(发布时间)’:,’geodata(地理信息)’:,’comment(评论)’:[‘commentCount(评论数)’:,’comments(评论)’:[{‘comment_id(评论者 id)’:,’comment_href(评论者主页)’:,’comment_teme(评论时间)’:,’comment_geodata(评论者地理信息)’:,’comment_text(评论内容)’:}]}

由于爬取的数据中包含大量的无效数据，因此要用 python 进行数据清理，去除显而易见的无效数据、停用词等实验无关的内容。由于本课题的数据来源是

中文，故需要对中文分词，使用现有开源的中文分词库即可较好的分词。然后使用合适的方法进行特征提取、特征表示。

对数据的语义化方面，目前还没有想到更好的策略，初步确定使用开源的情感词典，以词频加权统计的方法进行语义化。

复杂网络方面：研读论文学习使用合适的方式构建网络，使用 python 的 Networkx 在计算机中构建基于大数据的复杂网络，并编写算法进行统计、分析得出结论。结果集成到 Mediawiki：Mediawiki 支持嵌入 html 页面，所以可以很方便的进行二次开发，当然有部分功能的实现需要修改源代码实现。

三、进度安排

2016.12 月中旬-2017.01 月 阅读相关资料文献，完成开题报告

2017.01 月-2017.02 月 学习数据爬取相关技术，配置与课题相关的软硬件，并进行数据爬取。

2017.02 月-2017.02 月中旬 进行数据预处理，学习统计分析算法、舆情分析相关内容。

2017.02 月中旬-2017.03 月根据数据建立合适的多层异质网络，计算网络属性。

2017.03 月-2017.04 月算法研究、统计分析、得出结论

2017.04 月-2017.04 月中旬 对 mediawiki 进行二次开发，将实验成果集成到 mediawiki 上。并开始编写毕业设计论文。

2017.04 月-2017.05 月中旬 与指导教师交流并修改完善毕业设计论文

2017.05 月中旬 - 2017.05 月下旬 答辩准备

2017.05 下旬- 答辩，提交毕业论文及相关材料，按规定打印装订

四、参考文献

[1] 程学旗，靳小龙，王元卓等。大数据系统和分析技术综述[J]. 软件学报, 2014(9):1889-1908.

[2] 李学龙, 龚海刚. 大数据系统综述[J]. 中国科学:信息科学, 2015, 45(1):1-44.

[3] 肖源, 郝杰, 刘莹,等. 信息分析视角下的大数据分析平台构架研究[J]. 情报科学, 2016, V34(9):83-89.

[4] Talia D. Clouds for scalable big data analytics[J]. Computer, 2013,46(5):98-101 .

[5] 程学旗，靳小龙，王元卓．大数据系统和分析技术综述[J]．软件学报，2014, (9):1889-1908.

[6] Morales GDF. SAMOA: A platform for mining big data streams. In: Proc. of the 22th Int' l World Wide Web Conf. (WWW 2013). Rio de Janeiro: ACM Press.[EB/OL]. http://www.engineeringvillage.com/search/doc/detailed.url?SEARCHID=M3862b207144

[7] 顾君忠．大数据与大数据分析[J]．软件产业与工程, 2013, (4):17-21.

[8] 李广建，化柏林．大数据分析与情报分析关系辨析[J]．中国图书馆学报，2014,(5):14-22.

[9] Lu X, Brelsford C. Network structure and community evolution on twitter: human behavior change in response to the 2011 Japanese earthquake and tsunami[J]. Scientific reports, 2014, 4: 6773.

[10] C. Ahmed, A. ElKorany, R. Bahgat, A supervised learning approach to link prediction in Twitter, Social Network Analysis and Mining, 6 (2016) 1-11.

[11] L. Guo, Z. Ding, H. Wang, Behavior-Based Twitter Overlapping Community Detection, in: Database Systems for Advanced Applications, Springer, 2016, pp. 371-376.

[12] Y. Yilmaz, A. Hero, Multimodal Event Detection in Twitter Hashtag Networks, arXiv preprint arXiv:1601.00306, (2016).

[13] 唐晓波，宋承伟．基于复杂网络的微博舆情分析[J]．《情报学报》，2012(11): 1153-1162.

[14] 王伟，许鑫．基于聚类的网络舆情热点发现及分析[J]．现代图书情报技术, 2009(3):74 – 79.

[15] 刘怡君，陈思佳，黄远，马宁，王光辉，牛文元．重大生产安全事故的网络舆情传播分析及其政策建议——以“8•12 天津港爆炸事故"为例[J]．管理评论, 2016, 28(3).

[16] 夏火松，甄化春．大数据环境下舆情分析与决策支持研究文献综述[J]．管理评论, 2016, 28(3).

选题是否合适： 是□ 否□

课题能否实现： 能□ 不能□

指导教师（签字）

年 月 日

选题是否合适： 是□ 否□

课题能否实现： 能□ 不能□

审题小组组长（签字）

年 月 日