“十二五”国家重点图书出版规划项目
“十一五”国家科技支撑计划重点项目

综合风险防范关键技术研究与示范丛书

综合风险防范

搜索、模拟与制图

王静爱 武建军 王 平
周明全 王 瑛 刘连友 等 著

科学出版社

综合风险防范关键技术研究与示范丛书

综合风险防范

搜索、模拟与制图

王静爱 武建军 王 平
周明全 王 瑛 刘连友 等 著

科学出版社
北京

总　　序

综合风险防范（integrated risk governance）的研究源于21世纪初。2003年国际风险管理理事会（International Risk Governance Council，IRGC）在瑞士日内瓦成立。我作为这一国际组织的理事，代表中国政府参加了该组织成立以来的一些重要活动，从中了解了这一领域最为突出的特色：一是强调从风险管理（risk management）转移到风险防范（risk governance）；二是强调“综合”分析和对策的制定，从而实现对可能出现的全球风险提出防范措施，为决策者特别是政府的决策者提供防范新风险的对策。中国的综合风险防范研究起步于2005年，这一年国际全球环境变化人文因素计划中国国家委员会（Chinese National Committee for the International Human Dimensions Programme on Global Environmental Change，CNC-IHDP）成立，在这一委员会中，我们设立了一个综合风险工作组（Integrated Risk Working Group，CNC-IHDP-IR）。自此，中国综合风险防范科技工作逐渐开展起来。

CNC-IHDP-IR成立以来，积极组织国内相关领域的专家，充分论证并提出了开展综合风险防范科技项目的建议书。2006年下半年，科学技术部经过组织专家广泛论证，在农村科技领域，设置了“十一五”国家科技支撑计划重点项目“综合风险防范关键技术研究与示范”（2006～2010年）（2006BAD20B00）。该项目由教育部科学技术司牵头组织执行，北京师范大学、中国科学院地理科学与资源研究所、民政部国家减灾中心、中国保险行业协会、北京大学、中国农业大学、武汉大学等单位通过负责7个课题，承担了中国第一个综合风险防范领域的重要科技支撑计划项目。北京师范大学地表过程与资源生态国家重点实验室主任史培军教授被教育部科学技术司聘为这一项目专家组的组长，承担了组织和协调这一项目实施的工作。与此同时，CNC-IHDP-IR借2006年在中国召开国际全球环境变化人文因素计划（IHDP）北京区域会议和地球系统科学联盟（Earth System Science Partnership，ESSP）北京会议之际，通过CNC-IHDP向IHDP科学委员会主席Oran Young教授提出，在IHDP设立的核心科学计划中，设置全球环境变化下的“综合风险防范”研究领域。经过近4年的艰苦努力，关于这一科学计划的建议于2007年被纳入IHDP新10年（2005～2015年）战略框架内容；于2008年被设为IHDP新10年战略行动计划的一个研究主题；于2009年被设为IHDP新10年核心科学计划之开拓者计划开始执行；于2010年9月被正式设为IHDP新10年核心科学计划，其核心科学计划报

告——《综合风险防范报告》（*Integrated Risk Governance Project*）在 IHDP 总部德国波恩正式公开出版。它是中国科学家参加全球变化研究 20 多年来，首次在全球变化四大科学计划［国际地圈生物圈计划（International Geosphere-Biosphere Program，IGBP）、世界气候研究计划（World Climate Research Programme，WCRP）、国际全球环境变化人文因素计划（IHDP）、生物多样性计划（Biological Diversity Plan，DIVERSITAS）］中起主导作用的科学计划，亦是全球第一个综合风险防范的科学计划。它与 2010 年启动的由国际科学理事会、国际社会科学理事会和联合国国际减灾战略秘书处联合主导的“综合灾害风险研究”（Integrated Research on Disaster Risk，IRDR）计划共同构成了当今世界开展综合风险防范研究的两大国际化平台。

《综合风险防范关键技术研究与示范丛书》是前述相关单位承担“十一五”国家科技支撑计划重点项目——“综合风险防范关键技术研究与示范”所取得的部分成果。丛书包括《综合风险防范——科学、技术与示范》、《综合风险防范——标准、模型与应用》、《综合风险防范——搜索、模拟与制图》、《综合风险防范——数据库、风险地图与网络平台》、《综合风险防范——中国综合自然灾害救助保障体系》、《综合风险防范——中国综合自然灾害风险转移体系》、《综合风险防范——中国综合气候变化风险》、《综合风险防范——中国综合能源与水资源保障风险》、《综合风险防范——中国综合生态与食物安全风险》与《中国自然灾害风险地图集》10 个分册，较为全面地展示了中国综合风险防范研究领域所取得的最新成果（特别指出，本研究内容及数据的提取只涉及中国内地 31 个省、自治区、直辖市，暂未包括香港、澳门和台湾地区）。丛书的内容主要包括综合风险分析与评价模型体系、信息搜索与网络信息管理技术、模拟与仿真技术、自动制图技术、信息集成技术、综合能源与水资源保障风险防范、综合食物与生态安全风险防范、综合全球贸易与全球环境变化风险防范、综合自然灾害风险救助与保险体系和中国综合风险防范模式。这些研究成果初步奠定了中国综合风险防范研究的基础，为进一步开展该领域的研究提供了较为丰富的信息、理论和技术。然而，正是由于这一领域的研究才刚刚起步，这套丛书中阐述的理论、方法和开发的技术，还有许多不完善之处，诚请广大同行和读者给予批评指正。在此，对参与这项研究并取得丰硕成果的广大科技工作者表示热烈的祝贺，并期盼中国综合风险防范研究能取得更多的创新成就，为提高中国及全世界的综合风险防范水平和能力作出更大的贡献！

国务院参事、科技部原副部长

刘燕华

2011 年 2 月

目　录

总序
第 1 章　综合风险网络信息搜索技术 …… 1
1.1　综合风险网络信息概述 …… 1
1.2　综合风险网络信息分布模型 …… 4
1.3　综合风险网络信息搜索与服务 …… 7
1.4　领域搜索引擎研究进展与现状 …… 10
1.5　面向领域的语义检索研究 …… 17
第 2 章　中文综合风险网络信息搜索系统（勾勾风险——Gogo Risk） …… 28
2.1　综合风险网络信息搜索系统架构 …… 28
2.2　网络风险信息智能采集子系统 …… 36
2.3　综合风险信息智能处理子系统 …… 48
2.4　风险信息索引及检索服务子系统 …… 58
2.5　综合风险智能搜索引擎结果展示与功能 …… 63
第 3 章　综合风险网络信息服务技术 …… 74
3.1　网络地理信息服务技术国内外研究进展情况 …… 74
3.2　网络地理信息服务的异步交互技术 …… 79
3.3　基于异步交互的 WebGIS 体系结构 …… 84
3.4　综合风险信息发布系统详细设计 …… 87
3.5　结论和展望 …… 94
第 4 章　旱灾风险的模拟与仿真 …… 96
4.1　计算机模拟与仿真技术 …… 96
4.2　旱灾风险模拟与仿真研究进展 …… 98
4.3　旱灾风险模拟理论研究 …… 108
4.4　旱灾风险模拟系统技术开发 …… 129
4.5　研究展望与建议 …… 143
第 5 章　风沙灾害风险模拟系统 …… 144
5.1　风沙灾害风险模拟技术国内外进展 …… 144
5.2　下垫面与风沙灾害天气 …… 146
5.3　沙尘暴与城市空气质量、大气降尘 …… 156
5.4　风沙灾害模拟系统的开发 …… 168
5.5　风沙灾害模拟与仿真系统的应用 …… 181

第 6 章 中国自然灾害风险地图符号设计 …… 209
6.1 灾害地图符号的国内外研究进展 …… 209
6.2 自然灾害地图符号设计的理论与方法 …… 214
6.3 致灾因子图形符号系统构建 …… 223
6.4 致灾因子/孕灾环境色彩符号系统构建 …… 235
6.5 综合风险地图符号库 …… 245
6.6 结论、讨论与展望 …… 255
第 7 章 中国省区面积模式的分区分级符号自动实现 …… 258
7.1 专题地图符号自动实现的研究进展 …… 258
7.2 基于省区面积模式的比值分级法自动实现 …… 263
7.3 基于中国省区面积模式的分区统计图法自动实现 …… 279
7.4 应用案例 …… 291
第 8 章 中国自然灾害风险地图制图综合理论与方法 …… 301
8.1 地图制图综合的国内外研究进展 …… 301
8.2 自然灾害风险制图的自动综合理论构建 …… 312
8.3 自然灾害风险制图的自动综合方法研究 …… 319
第 9 章 中国自然灾害风险地图自动综合实现 …… 344
9.1 点状自然灾害风险地图的自动综合应用研究——以滑坡致灾因子为例 …… 344
9.2 线状自然灾害风险地图的自动综合应用研究——以铁路承灾体为例 …… 362
9.3 面状自然灾害风险地图的自动综合应用研究——以地貌孕灾环境为例 …… 377
第 10 章 中国自然灾害自动制图软件研究 …… 389
10.1 自然灾害地图制图软件的发展 …… 389
10.2 综合风险自动制图软件的设计 …… 408
10.3 综合风险自动制图软件的开发 …… 419
参考文献 …… 423
后记 …… 438

第 1 章　综合风险网络信息搜索技术*

综合风险网络信息涵盖了综合风险互联网信息、地理信息，信息具有明显的地理和行业特征。针对综合风险互联网信息，通用搜索在一段时期内主导了人们对搜索的看法，随着互联网应用更趋向于面向领域细分，领域搜索应运而生。由于通用搜索不可能在每个行业都做到深入、全面，因而搜索技术只有与产业结合才能发挥更大的作用。全面了解领域搜索引擎的历史、发展及研究现状，充分调研搜索技术与灾害风险防范技术中的难点，将搜索技术与传统灾害风险防范产业的认知、资源进行融合，使得面向综合灾害风险防范的领域搜索更有“智慧”，进而为面向灾害防范领域的专业用户提供更精确、更全面的服务。

1.1　综合风险网络信息概述

1.1.1　研究背景

综合风险的传统获取渠道主要是专业的领域数据库和政府报告，这些信息渠道多年来一直是综合风险防范领域、防灾救灾领域的可靠信息依据，为开展相关工作提供了有力的数据保障。这类信息的可靠性非常高，覆盖了各方面的专业资料：一部分是专业数据库的内容，这部分内容涵盖了历史上主要相关信息资料库，存放了一些非常专业、全面的信息，如历史降水量、历史气温等；另一部分就是从全国灾害监测系统、气象部门等相关部门得来的灾害实时监控信息，信息的及时性和可靠性都比较高，如实时卫星云图、风力等数据。

随着信息技术的发展，现如今，互联网已成为电视、广播、报纸之外的第四大媒体，同时也被称为“信息的海洋”。互联网上出现了大量的数字化的灾害信息、预防信息，这部分信息与专业的信息来源相比，缺乏专业性和权威性，但是在综合风险防范及研判中也非常重要。

互联网上的信息体现了当灾害发生时，社会中的群体和个人的反应，有很强的社会性和实时性。减灾救灾的对象除了经济损失外，社会这种影响，包括对普通民众、灾区，以及跟灾区有关的这些民众的影响，也很重要。互联网上的信息对减灾救灾的一些措施，包括媒体部门、政府部门一些对策和决策造成影响。

互联网上的数字化灾害相关信息究竟有多大影响呢？我们可以看一组数字。据统计，中国电视观众将近 10 亿，几乎覆盖整个城乡，而现在上网的网民有 1 亿左右，从数字上

---

* 本章执笔人：北京师范大学的周明全、付艳、王学松。

看覆盖率低于电视观众。但是网民这个群体比较特殊，因为电视观众大部分在晚上或者特定时间段看新闻，而中国网民上网，有50%以上的时间会看新闻，用搜索引擎看一些即时的信息，网民每天有大量时间在网上，他随时可以通过网络获取实时信息。那么网上的信息内容大概有多大呢？2009年，据估计，我们中国网页有45亿～100亿，网站100万～200万个，每天都会有成千上万的信息出现。

1.1.2 互联网上的综合风险信息特点

互联网上的综合风险信息与通常媒体相比，具有很多独特的特点。

（1）信息发布速度非常快，此外，信息膨胀的基数大，在灾害发生后的很短时间之内，它就有相关的信息出现了。当然这不一定是政府的权威媒体，而是网民自主发布的信息，曾经发生过在一天内出现几万条地震信息的情况。

（2）信息质量良莠不齐，既有正面的，也有负面的。正面的信息有政府官方的一些气象预警等信息，如地震部门发布的信息；还有一些负面的，如网民的悲观情绪、不可信的谣言等，比如地震发生以后，多个地区曾经一度流传当地晚上11点要有地震的传言。这些不同渠道发布的信息在互联网上被不同程度地放大，并对整个社会造成了非常巨大的负面影响。

（3）网上信息还有一个比较典型的特点——交互性，任何人都可以发布信息。这就造成对于灾害信息，除了减灾部门的正规监测数据以外，还有大量非正式信息通过一些别的渠道，把现场人的感受或者跟其相关的消息，或放大或非正规泛滥性地流传到网上来。

（4）网络信息具有明显的地理和行业特征，无论是灾害信息还是风险预防，防灾减灾信息都具有明显的地理位置或灾害领域特征，信息更为专业。

互联网上综合风险信息的特点还包括信息发布及时、信息来源多样化、信息结构类型复杂、信息行业性特点突出、时间分布突发性强、时效性强、网站分布范围相对狭窄等。

1.1.3 综合风险网络信息研究的重要性

综合风险网络信息对政府的官方信息、新闻媒体的信息起到很大的补充作用，对于综合风险防范和救灾有很大的辅助性作用；但非官方信息中对灾害的负面描述，同时也会造成一些负作用。对于救灾的专家来说，综合官方信息和网络信息两方面内容进行分析会获得一些更加有效的资料，这些信息正是地震等灾害现场的一些直观信息。

例如，雪灾及地震灾害发生后，相关专家需要分析当时的伤亡和损失状况。数据分析一方面基于历史的资源数据库，另一方面是基于当时监测的实际信息，同时还将监测到的网络相关信息作为重要的辅助数据。

综合风险网络信息的重要性有几个典型的例子可以证明。一个是灾害发生的时候，如汶川“5·12”地震发生后，互联网上的一些报道，包括阿坝州政府的所有通信中断以后，网站为大家不断提供现场的受灾情况；还有网民提供直升机空降地点等。从2008年5月12日地震灾害发生到19日，也就是在一周时间内，据有机构统计，互联网上大概发布了有十几万条新闻，相关的浏览数100多亿次，其中，网民自发的交互和评论、发表自己看

法的信息达1000多万条。伴随而生的也包括到处流传的谣言，包括北京、黑龙江、山西等地都出现了跟地震、灾害有关的谣言信息。

在国际上，非官方互联网信息的应用也有一些成功的案例，曾经有一部分人员即是通过使用网络的双向性作预警或者及时获取信息的。在综合风险和自然灾害的应用上，美国宇航局（NASA）等网站即提供了灾害信息的检索系统。更为成功的案例是影响巨大的全球公共卫生情报网（GPHIN），该网站提供关于公共卫生安全的大量信息，虽然该网站内容跟本课题研究的自然灾害有一定差别，但其使用非官方互联网信息监控社会灾害信息状况的模式非常有效，值得学习。例如，SASI病毒爆发的时候，GPHIN在通过正式渠道获取到信息之前，在内部就已监测到该事件，并启动了内部预警机制监控整个事件。

总体而言，互联网上有丰富的信息，同时这些信息具有很大的价值，对这些信息的监测和整合，应该作为整个综合风险防范和社会应急响应系统基础环境的一部分，不容忽视；另外，也急需将互联网信息与面向不同领域的专业信息等进行整合利用，这样才会使数字化的信息有最大的应用价值。

1.1.4　构建综合风险网络搜索引擎的必要性

面对海量的互联网信息，我们需要有一种对互联网监控和信息分析的有效办法，以期实现专业用户或政府部门等对信息的跟踪和了解。一般通用搜索引擎虽提供了大量的信息检索服务，对现场信息的报道实现了汇集及一些结构的聚合，但这些信息无法直接应用于决策层面，也无法实现对有效和无效信息、正面及负面信息的辨别。

面对快速增长的海量综合风险网络信息，构建综合风险防范领域的领域搜索引擎，能够在很大程度上解决人们从互联网上查找风险信息的困难。面向综合风险防范建立相关网站，也可实现信息的深入分析整理与加工应用。

大家可能要说，已经有百度、谷歌这种数据量很大的通用搜索引擎了，我们自己构建的搜索引擎有什么必要，有什么优势，好在哪儿呢？不可否认的是，现阶段的通用搜索引擎已经整合了互联网上众多的网页资源，也提供了较全面的信息导航和信息查询服务。但是面对综合风险防范领域的应用，通用搜索引擎也存在一系列问题，此时，领域搜索引擎就有了存在的价值。具体对比如下所述。

第一，综合风险防范领域的领域搜索引擎比通用搜索引擎更专业化，因为系统专注于2000～3000个专业网站数据以及几万个相关网站的信息，因而能用更短的时间获得第一手信息。也正是因为系统专注于灾害信息、防灾、减灾信息，所以能得到更全面的相关信息。在汶川地震发生后第一时间内，使用通用搜索引擎检索到5万～6万条数据，而领域垂直搜索引擎针对地震灾害就有接近10万条信息。

第二，领域搜索引擎更有针对性和灵活性，也就是针对灾害的不同种类或者每个地区，具有信息分类的功能。例如，基于后台减灾专家提供的领域知识和指导，对冰雪、火灾、沙尘等做相应的、不同的信息整理、整合后，可以更灵活地进行针对性信息处理。

举例来说，我们的系统在雪灾发生以后及时提供了对预警、雪灾报道、损失情况以及重建信息的处理。这些处理看似是信息的简单整合，实际上后台进行了很多对数据的智能

化分析，并完成了一些减灾专家要求的信息分析。针对雪灾提供的基础数据，从预警、大概信息发布情况、雪灾报告的情况、灾后重建多个方面进行汇总，能够在一定程度上反映灾害发生的时候，对不同作物、不同部门的影响。在汶川地震之后，系统及时获取了灾害现场、灾害搜救，包括政府应急转移、响应等信息，并完成了这些方面的总体情况分析。当时的数据分布情况为减灾专家提供了很有效的信息依据，相关的数据和图片是地震发生以后的一些信息结果汇总和数据分布列表。在实践中，利用综合风险防范领域的领域搜索引擎信息，可以大致判断在减灾救灾中某一部分工作的进展，或者说大概的直观状况。

更进一步的，构建综合风险领域搜索引擎，面对各种综合风险防范应用，希望能够在灾害发生以后或者在发生之前，将各种公布的气象信息、一些预警信息、灾害报道信息通过一个平台整合在一起，进行预警，并提供给决策部门或者说专业的人员使用。此外，综合风险领域搜索引擎的应用目的还包括在灾害发生之后，能够实现信息的实时监控，实现社会救灾、民众的反映等各方面渠道信息的汇总。当然，目前我们这个专业搜索引擎仅仅能够实现基本的信息采集、处理、分析和报告，要达到上述目标，还有很多需要进一步研究和实现的部分。

综合风险领域搜索引擎上的信息采集主要针对与灾害预警、防范等相关的专业网站。信息采集主要面向灾害、减灾、救灾有关的网站，大量的官方、非官方的信息汇聚起来，形成第一手资料。系统的反应速度快，可以在信息发布后几分钟之内把相关的信息汇集起来。系统平台的主体信息通过对互联网信息的直接监控得到。对信息做后台的一些技术性处理，并补充一部分百度、谷歌等搜索结果，可得到一个综合分析的结果。

1.2 综合风险网络信息分布模型

1.2.1 Web 网状模型

互联网的综合风险相关信息是通过网页进行组织的，网站中的网页通常通过内部链接和外部链接与其他网页和信息关联，外部链接是那些链接目的地址为其他网站或其网页的超链接。

网站的内部链接是指网站内不同网页之间的链接关系。网站的内部链接结构可以表示成一种树或图的结构。网页链接实际上是一种带有树状层次结构的有向图，最终形成一种非均衡的网状结构。网页内部包括更小的文本块、表格与图像等。这些基础元素组成的信息块之间的链接也是网状结构。因此，可以使用有向图来表示互联网资源之间的超链接。

从网站拓扑结构模型的角度来看，Web 网站可以理解为一个并不规范的实体关系图：Web 文档页面可以表示为一个实体，相互之间的超链接形成实体间的相互关联，文档的元数据、（标签）Tag 及描述信息成为实体的属性；网页的文本正文是实体对应的内部成员，而网页之间的关联信息成为对象链接。

同一网站内的网页往往属于一个或多个主题，以 2 ~ 3 层的目录形式组织。网站内拥有的文档型和网页型的资源内容，通过网页之间的链接信息互相关联（图 1-1）。

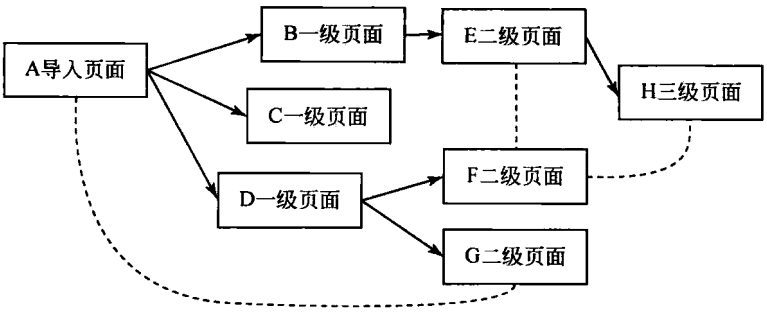


图 1-1　网站内部页面间链接结构示意图

注：实线表示直接上下层网页间的关联；虚线表示跨层或同层网页间的关联

同一网页内部往往也包含页内链接关系，页内链接虽然信息相关度不如页面之间的链接强，但是页内链接关系表明了信息块的边界以及相互间的层次关系，在后续对综合风险信息的语义分析中具有重要作用。这是因为根据综合风险信息特点，同一个文档内往往包含了多个独立的信息内容，利用页内链接进行信息块划分，可配合 HTML 标识符有效地分隔信息块，建立小粒度的信息对象。页内链接关系如图 1-2 所示。

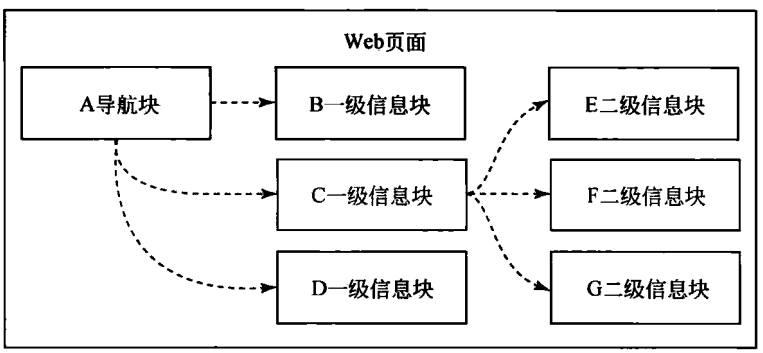


图 1-2　网页内部信息块间链接结构示意图

对于综合风险信息，资源内容分布更为集中在一个网站上，往往对于一个网站，要么具有大量有价值的灾害预防、新闻信息，要么只是具有部分简单报道。分析主题权威网站的结构，能够帮助搜索引擎快速获得大量的有效资料，并且借助于资源之间的链接关系，根据站点内部的物理链接和逻辑链接，将多个同一主题的网页合并为高粒度的信息资源。内部链接往往表示网页之间在内容上具有很好的相似性或相关性，采用内部链接可以分析网站的主题是否与综合风险防范相关，站内是否存在较多的相关资料和信息。

规则 1-1 内部链接的内容相关性规则

如果网站内部两个网页 A、B 之间存在 A 指向 B 的内容相关链接，则表明网页 A 中包含了网页 B 的相关信息或者两者具有相似性。

网站内网页之间也存在用来完成网站组织的结构相关链接和特定用途的特殊相关链接。这些链接虽然会给链接的内容分析造成困难，但是可以作为中介和桥梁，将网站变成

可连通的有向图，便于信息采集的最大化。

对于网站内部的网页，如果该网页所指向的一系列目标网页具有相同的目录层次和内容结构，则该网页可以成为目录页面。网页链接中指向目录页面的网站成为目录链接。目录页面是网站内部的特殊页面，其内容主要是相关资源的列表和链接地址，是具有一定数量节点链接的集合。目录页面区别于网站普通页面，具有独特的特征。

规则 1-2 目录页面下级兄弟节点相似性规则

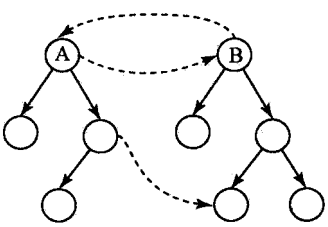
目录页面的下级兄弟节点链接在 URL 结构形式和相关内容上具有相似性。相似性主要指链接对应的锚文本、链接地址格式、参数名称、参数个数等方面的相似。锚文本包含的链接描述信息具有更为完整的语义，通常是资源页面内容的概括，区别于其他普通页面链接。

1.2.2 Web 链接模型

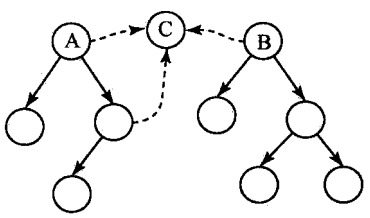
搜索引擎对互联网上海量网站的访问主要利用随机冲浪模型，即认为在某一时刻随机访问某起始网站，但对于下一次访问的网站，是同等转移概率的随机访问。这种跳转过程需要利用网站之间的外部链接来完成，没有网站的外部链接，学习资源网络爬虫就难以获得大量的学习材料和资源。

研究表明，由于专业综合风险相关网站相互之间的协作与竞争关系，彼此之间的链接关系并不是很完备，但往往会共同关联到另一个比较重要的综合风险防范权威机构，也存在一些网站目录和导航网站同时指向这些网站。

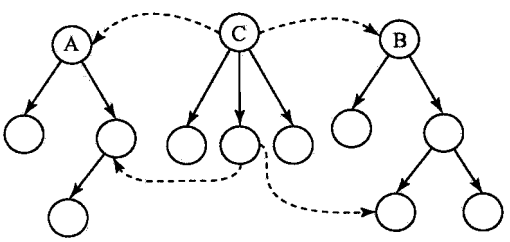
总体来看，综合风险防范信息相关网站之间的链接关系主要包括三类：第一类是彼此之间存在相互链接关系或者单项链接关系，如图 1-3（a）所示；第二类是综合风险防范网站指向共同的目的链接，如图 1-3（b）所示；第三类是一个权威的资源导航网站所指向的网站链接，如图 1-3（c）所示。



(a) 网站间互相链接关系



(b) 网站指向共同目标链接



(c) 权威网站指向的网站链接关系

图 1-3 网站之间链接关系示意图

1.3　综合风险网络信息搜索与服务

综合风险网络信息搜索引擎通过收集整合各种综合风险相关信息，根据不同用户对综合风险信息的需求，基于数字地球模型，建设综合风险网络服务系统，提供准确、权威、方便、快捷、全面的综合风险搜索功能，实现政府、企业、普通公众的定制搜索服务。

1.3.1　研究内容与目标

综合风险网络信息搜索与服务面向的行业、信息的内容和针对的用户群体都有特殊性，相关研究也要在通用技术的基础上作针对性探索。综合风险搜索针对三方面内容展开研究、解决各方面的问题。

1. 综合风险信息空间与行业搜索

根据用户发出的经纬度范围、行政区范围、地名、交通线路、重要生命线或者生产线的分布等空间与行业分布等属性查询条件，浏览器向搜索入口平台发送搜索请求，然后将搜索信息转向网络地理信息系统（WebGIS）服务器，再对相关的综合风险数据库进行查询和搜索，并返回与用户查询条件相符的包含综合风险信息的空间数据图层、不同行业和综合风险的属性信息，实现通过浏览器进行实时交互式的信息查询与空间制图。

2. 综合风险信息的分类整理

综合风险数据搜索与网络信息服务首先根据搜索结果的来源、准确性、标准化程度、现实有效性、更新频率等进行综合评价；在此基础上对搜索结果进行筛选，并根据以上指标进行排序调整和进一步评价，按照综合风险信息的可靠度、权威程度、标准化程度进行排序，做到向用户提供及时、准确的信息；此外，综合风险数据搜索与网络信息服务实现对搜索结果的自动整理功能，即根据主题、文件类型、更新时间等对搜索结果进行归类和整理。

3. 综合风险信息的用户分类

综合风险数据搜索与网络信息服务将具有向不同用户提供不同服务的功能。该服务将根据不同用户的需求确定搜索范围及搜索内容，针对政府风险与防范部门、研究机构、普通公众进行定制搜索。由于部分综合风险信息有可能会涉及保密问题，该系统还将对不同用户的访问权限进行限制。

1.3.2　技术路线

综合风险数据搜索和网络信息服务的技术路线（图1-4）：用户向网络服务器发出空间或者行业搜索信息，浏览器向搜索入口平台发送搜索请求，将搜索信息转向 WebGIS 服务器，经过用户权限认证后，再对相关的综合风险信息数据库进行搜索和分类整理，以基

于数字地球模型的综合风险信息表达方式返回给用户。

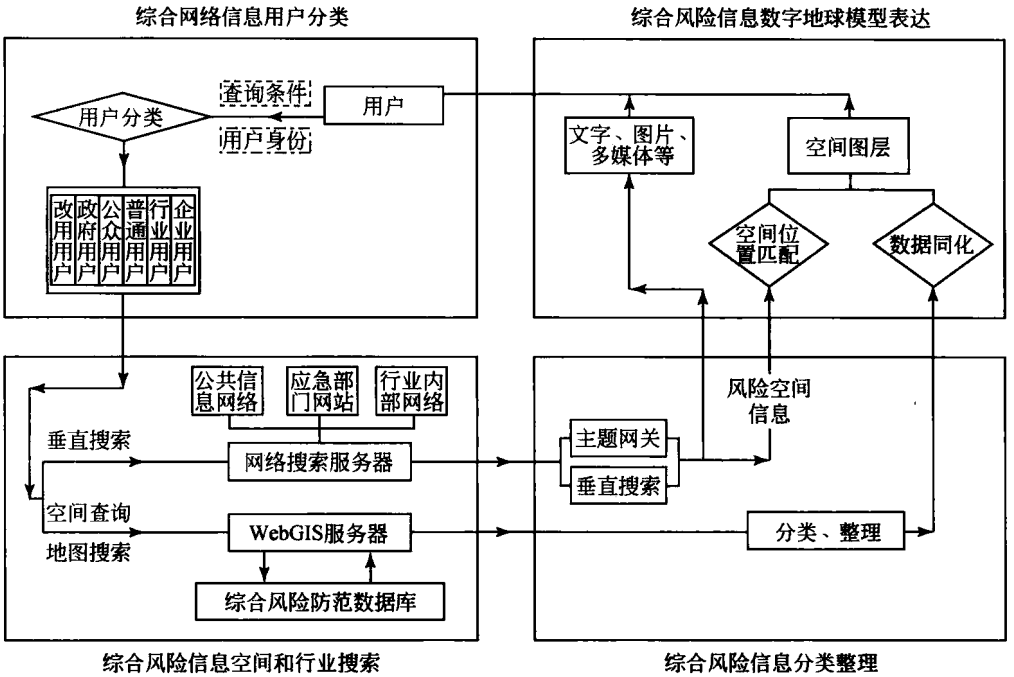


图 1-4 综合风险数据搜索和网络信息服务的技术路线

1. 综合风险信息的空间搜索服务

综合灾害风险防范技术集成平台课题将建立地理信息处理软件和大型数据库软件支持下的分布式综合风险信息空间及其属性数据库。首先，以该数据平台为基础，使用跨平台和适于分布式计算的语言（如 Java 等）、分布式计算的体系结构等，开发基于浏览器/服务器（B/S）的 WebGIS 系统，实现综合风险信息的空间搜索。然后，根据用户分类控制用户的访问权限，以保证系统和信息的安全性。其次，在通过权限认证后，以文字输入或者圈选位置等方式向网络服务器发出经纬度范围、行政区范围、地名、交通线路、自然区域等空间信息，浏览器向搜索入口平台发送搜索请求，将搜索信息转向 WebGIS 服务器，再对相关的综合风险信息数据库进行查询和搜索，并返回与用户查询条件相符的包含综合风险信息的空间数据图层、不同行业和综合风险的属性信息等。最后，信息以统一的格式返回给用户，既包含信息的基本信息（摘要、出处、作者、时间等），又包含对该信息的分类、评价（以项目制定的标准为依据）等信息，用户可以很容易地找到自己所需要的信息。同时，用户还可以通过事先设定偏好，如设定分类、信息源发布时间等条件来实现有针对性的查询。

2. 综合风险信息的行业搜索服务

本书通过对各个行业信息的整合，提供针对主要行业的专业化服务，主要服务对象为

国家灾害救助相关部门、保险和再保险公司等。综合风险信息的行业搜索服务包括：①综合风险信息垂直搜索，过滤掉无关以及相关度低、权威性和时效性差的信息，通过结构化抽取等方式，以统一、直观的方式全面、准确地将数据返回给用户；②主题网关，通过专家加工方式对已有的和垂直搜索所得的信息进行整合，以专题的方式提供信息服务。主题网关将通过站内查询和目录的方式，方便用户对所需信息的检索。对应于系统的主要服务对象，拟建立国家灾害救助和保险两个主题网关。

搜索结果排序主要依据词频和词位置信息、网页之间的链接信息、用户行为信息、主题相关度和查询语文档的相似度。首先，综合运用词频和位置加权排序算法、用户行为信息算法、网站排名算法、网页排序算法、相似度算法等计算出各影响因素的值；其次，根据专家经验等方法确定各影响因素的权重；最后，根据影响因素的值和权重求出贡献值，将各影响因素的贡献值相加得出网页的排序值。

3. 综合风险信息的分类整理

综合风险信息的分类整理采用以主题网关为主，垂直搜索为辅的方式进行，即主要通过专家参与和自动的方式对主要的、权威性强的网站上的信息进行评价、筛选、结构化处理和再加工，在分类整理后存入本地数据库，使用户可以很容易地获得全面、准确的信息。手动分类整理以整理现有数据和对自动收集的数据进行深加工为主；自动分类整理则以垂直搜索为主，即定向性地采集和垂直搜索范围相关的网页，忽略不相关的网页和不必要的网页，选择内容相关的以及适合做进一步处理的网页深度优先采集、对页面有选择地调整更新频率。自动采集以人工设定网址和网页分析方式共同进行，并根据相关信息的重要性和时效性设定信息采集周期。

4. 综合风险信息的用户分类

综合风险信息数据库系统为异质数据库系统，包含大量的异构分布式数据资源。综合风险信息服务的对象主要为政府、企业以及普通用户三大类，他们对综合灾害风险防范数据库的访问权限各不相同。本研究将通过运用基于角色的访问控制（role based access control，RBAC）来进行用户访问权限控制，在用户和权限之间引入角色，访问权限和角色相关联，角色再与用户关联，将用户与访问权限逻辑分离。安全管理员根据实际需要（如数据共享程度变化等）改变与角色对应的访问权限，普通管理员通过角色指定来设定或者改变用户角色。

在用户权限设定后，用户从应用层中的任意一个应用系统登录，首先要通过用户认证，生成安全证书。证书中包括主体的特权属性、控制属性、操作和对象引用等访问允许函数所需要的参数。对象请求代理负责对象在分布环境中透明地收发请求和响应。通过应用安全控制把证书传到访问运行对象，从而实现用户对目标对象的安全访问控制。

1.3.3　主要技术难点和问题

网络搜索技术涉及海量信息获取和分析，需要解决大数据量环境下数据采集、存储效

率和实时交互检索的速度问题，在技术实现上有较大难度。

综合风险网络搜索的信息评价和排序算法需要针对风险数据特点定制，实现语义信息识别、信息自动分类、地图数据的搜索等功能，在算法和原理上有较大难度。

1.4 领域搜索引擎研究进展与现状

1.4.1 搜索引擎研究进展

随着互联网普及程度的不断提高，人们对信息的需求更多地趋向于专业化和个性化，导致搜索引擎的行业化和细分化趋势日益显著。据赛迪网调查，有六成的网民认为面向某一领域的搜索引擎对其而言非常或比较重要。

通用搜索引擎一方面缺乏深加工及个性化的服务信息，已经无法满足人们对个性化信息检索服务日益增长的需求；另一方面，其缺陷与不足也在不断凸现，如信息大量重复赘积、垃圾数据混杂、有效信息漏失、更新不及时等。专业化的搜索引擎在此背景下应运而生，成为搜索引擎发展的必然趋势之一。所谓专业化，就是该搜索引擎专注于某种专业的信息检索，比起常规的搜索引擎，它可能在信息的广度上略有不足，但就某一专业而言，它的检索深度和分类细化程度远远优于常规的搜索引擎。

领域搜索引擎是一种全新的搜索引擎服务模式，它是通用搜索引擎的细化和延伸。领域搜索引擎是针对通用搜索引擎信息量大、查询不准确、深度不够等问题提出的搜索服务模式，面向某特定领域、某特定人群或某一特定需求，提供更精确的信息和更准确的信息查询结果，其特点是“专、精、深”，且具有行业色彩。领域搜索引擎为用户提供的并不是上百甚至上千万相关网页，而是在限定范围内、极具针对性的具体信息。因此，特定行业的用户更加青睐领域搜索引擎。

领域搜索引擎也被称为垂直搜索引擎或主题搜索引擎，它专注于具体、深入的纵向服务，专门收录某一方面、某一行业或某一主题内的信息，专为查询某一个学科或某一主题的信息提供检索服务。通用搜索引擎对于信息需求相对集中、分类更加详细的行业客户缺乏导向性，因此在解决某些实际查询问题的时候，领域搜索引擎比通用搜索引擎更有效。TRS 公司对领域搜索引擎的定义是：领域搜索引擎是针对某一个行业或组织，满足行业专业需求或者组织某项业务需求的专业搜索引擎，是通用搜索引擎的细分和延伸，是对某类网页资源和结构化资源的深度整合，并为用户提供符合专业用户操作行为的信息服务方式。具体而言，领域搜索引擎就是把网页库中的某类专门信息进行整合，定向、分字段地抽取出需要的数据，然后进行深度加工处理，如去重、分类、分词、索引等，最后再以某种特定的形式返回给用户。它能为用户提供针对性更强、精确度更高的信息检索服务。

领域搜索引擎的应用方向很多，如地图搜索、音乐搜索、图片搜索、文献搜索、企业信息搜索、求职信息搜索、供求信息搜索引擎、购物搜索、房产搜索、人才搜索等，几乎各行各业、各类信息都可被细化成相应的领域搜索对象。要构建领域搜索，不仅需要掌握搜索的关键技术，更重要的是需熟悉其针对的具体行业，掌握行业特点，深入了解行业的需求。

1.4.2　领域搜索引擎通用架构

领域搜索引擎的体系结构与通用搜索引擎基本类似，它们之间最大的区别是前者引入了领域搜索的概念，即在网页搜索过程中进行主题评价和相关度计算，以确定是否采集网页，因此其网络机器人较通用搜索引擎要复杂。领域搜索引擎通常由三大部分组成：抓取系统、索引系统和搜索系统（图 1-5）。

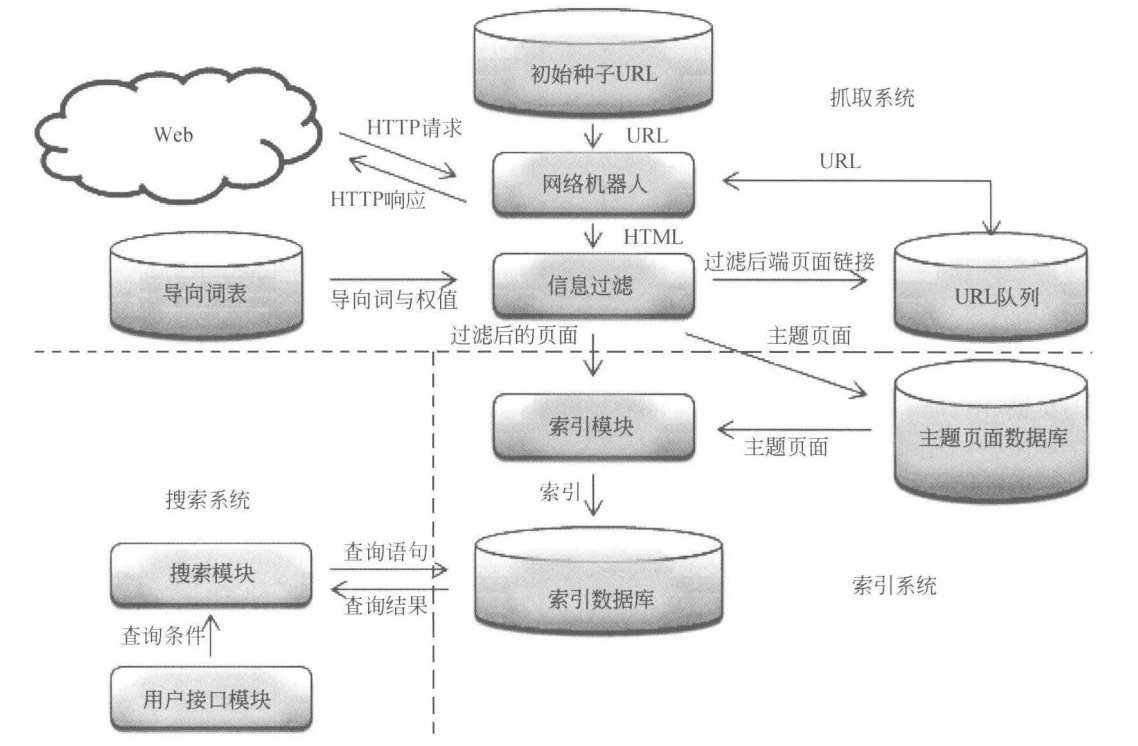


图 1-5　领域搜索引擎系统结构图

1. 信息采集抓取系统

信息采集抓取系统也就是蜘蛛（spider）程序，负责从信息源抓取数据。蜘蛛程序通常基于预先构造的模板工作，无模板的蜘蛛程序只能处理结构相对简单的信息。抓取系统涉及的关键技术点有爬行路径分析、增量抓取与全抓取、信息构造完整性、信息唯一性识别、多网页信息整合、结构化信息提取等。

领域搜索引擎的抓取程序和通用搜索引擎的抓取程序相比应该更加专业，并且是可定制的。抓取程序定向地采集与专业领域相关的网页，忽略不相关的网页和不必要的网页，并选择内容相关的以及适合做进一步处理的网页进行采集，采集工作可通过人工设定网址或网页 URL 分析的方式进行。领域搜索依据网站内容变化的频率、网站内容的重要性、用户点击频率及网站稳定性等几个方面指标，确定对网站内容更新采集的频率。