支持内容智能治理的双结构互联网

杨鹏[1,2,3]，李幼平[1,2,3]

（1. 计算机网络和信息集成教育部重点实验室（东南大学），江苏 南京 211189；
2. 东南大学计算机科学与工程学院，江苏 南京 211189；3. 东南大学网络空间安全学院，江苏 南京 211189）

摘 要：针对互联网面临的内容大数据趋势显著、内容语义标识缺乏和内容安全态势严峻等内容治理挑战，提出了一种支持内容大数据智能治理的双结构互联网，以现有互联网体系结构作为主结构，以基于辐射−复制范型的播存网络作为次结构。介绍了双结构互联网的体系结构设计原则、核心理念和内容智能治理实现机制，尤其对统一内容标签 UCL 国家标准与富语义矢量编码、热门内容汇聚与 UCL 安全能级模型、UCL 知识空间与内容汇聚研讨厅等内容智能治理关键技术进行了详细阐述。最后，通过研发双结构互联网内容智能治理原型系统，对双结构互联网及其内容智能治理能力进行了验证。双结构互联网为破解内容大数据治理难题提供了网络体系结构层面的创新解决思路。

关键词：双结构互联网；内容智能治理；统一内容标签；内容大数据；综合集成方法
中图分类号：TP393
文献标识码：A
doi: 10.11959/j.issn.1000−436x.2019197

Dual-architecture Internet supporting intelligent governance of cyber content

YANG Peng[1,2,3], LI Youping[1,2,3]

1. Key Laboratory of Computer Network and Information Integration (Southeast University), Ministry of Education, Nanjing 211189, China
2. School of Computer Science and Engineering, Southeast University, Nanjing 211189, China
3. School of Cyber Science and Engineering, Southeast University, Nanjing 211189, China

Abstract: For the current Internet architecture is suffering many challenges of cyber content governance, including big data trends of cyber contents, lack of semantic identification, and the severe situation of content security. In seeking to address these challenges, a dual-architecture Internet which integrated the current Internet architecture with a complementary secondary broadcast-storage network architecture characterized as the combination of broadcast transmission and ubiquitous storage was proposed. The design principles and core concepts of the dual-architecture Internet and its intelligent realization mechanism of cyber content governance were introduced, and especially, the national standard of uniform content label (UCL) and UCL-based rich semantic vector coding of contents, the gathering of popular contents the UCL-oriented security energy level model, the UCL knowledge space and content clustering hall for workshop of meta-synthetic engineering were detailed. Consequently, a prototype of dual-architecture Internet supporting intelligent governance of cyber contents was developed to verify the design principles and core concepts of the dual-architecture Internet and its capability for intelligent governance of cyber contents. The dual-architecture Internet proposed can provide a brand-new approach to solve the problem of cyber content governance from the high-level perspective of network architecture.

Key words: dual-architecture Internet, intelligent governance of cyber content, uniform content label, content big data, meta-synthesis

收稿日期：2019−02−14；**修回日期**：2019−08−05
基金项目：国家自然科学基金资助项目（No.61472080, No.61672155）；中国工程院咨询研究基金资助项目（No.2018-XY-07）；软件新技术与产业化协同创新中心基金资助项目
Foundation Items: The National Natural Science Foundation of China (No.61472080, No.61672155), Consulting Project of Chinese Academy of Engineering (No.2018-XY-07), Collaborative Innovation Center of Novel Software Technology and Industrialization

1 引言

互联网是网络空间的主要载体，与人们生产生活、社会发展、国家安全休戚相关，已经对全球政治、经济、文化等产生深远影响。但是，由于互联网的开放性和便捷性，今天的互联网已经成为内容大数据的集散地，各种海量化、碎片化的内容不断涌现，日益呈现出异构驳杂和混乱失序等特征。互联网不是法外之地，它理应成为人类共同的精神家园，肩负着传播人类优秀文化的重要使命。但是，在今天的互联网中，因为内容不能得到有效治理而导致的安全问题正变得越来越突出。

内容治理是互联网治理体系变革的核心目标和关键环节。但是，由于当前互联网在体系结构和治理机制方面存在欠缺，因而难以对不断涌现的海量化、异构化、碎片化和混乱失序的内容大数据进行有效治理，互联网内容乱象愈演愈烈，因此如何通过变革现有互联网体系结构使之支持互联网内容大数据的高效治理，已成为当前互联网体系结构研究的燃眉之急。

为此，本文提出了一种支持内容智能治理的双结构互联网，它以现有互联网体系结构作为主结构，以基于辐射−复制范型的播存网络作为次结构，在确保互联网平滑演进的基础上，以较小的网络体系结构变革代价，实现互联网内容治理能力的显著提升。双结构互联网遵循新型互联网体系结构 3 条设计原则，从总体结构、核心基元、治理方法学 3 个方面进行创新，以统一内容标签（UCL, uniform content label）[1]内容驱动基元，对互联网内容进行富语义矢量编码，建立 UCL 多标识维度语义关联模型，引入数据与知识联合驱动的安全能级模型，借助知识图谱刻画基于语义的内容关联，建立内容大数据 UCL 知识空间，按照钱学森先生提出的综合集成方法构建内容汇聚研讨厅，实现对互联网内容大数据的智能治理，为消除互联网内容混乱失序顽疾提供了网络体系结构层面的创新解决思路。

2 互联网体系结构面临的内容治理挑战

互联网的设计初衷和基本运作理念是为了支持端到端通信，因此传统互联网采用的是基于对流传输模型的 TCP/IP 结构，它虽然对端到端交互型应用存在优势，但是现今互联网的主流应用范型已经发生根本改变，从端到端通信转变为向海量用户提供海量内容的内容共享服务[2]。但是，由于网站、论坛、微博、微信、社交网络以及各种自媒体渠道的便捷畅通，互联网中的各种内容正在快速无序化增长，这些内容中包含大量虚假信息、片面信息、甚至恶意谣言，造成互联网内容良莠不齐和混乱失序。网络空间是亿万民众共同的精神家园，只有网络空间生态良好，才符合人民利益。因此，如何从根本上解决互联网内容治理难题，成为当前互联网体系结构研究领域亟待解决的重要课题之一。概括起来，当前互联网体系结构研究所面临的内容治理挑战主要体现在内容大数据趋势显著、内容语义标识缺乏和内容安全态势严峻 3 个方面。

首先，以富媒体化和海量化等为特征的内容分发与共享，已经成为互联网发展的主旋律，互联网中的新闻资讯、音视频、流媒体、自媒体等内容正呈现出爆炸性增长趋势。根据最新的 Cisco VNI 预测报告，全球固网和移动网络的互联网 IP 流量中 90%以上的流量与内容共享应用有关，预计 2022 年这部分流量将高达 4.8 ZB[3]。此外，按照互联网数据中心（IDC, Internet data center）的报告[4]，预计到 2020 年全球的数据总量将达到 44 ZB，远远超过人类有史以来所有印刷材料的数据总量（200 PB）。在大数据和泛媒体环境下，不断涌现的互联网内容大数据（content big data），由于治理机制的缺位，正表现出复杂异构、良莠不齐和混乱失序等特征。作为一类以内容为主体的特殊大数据[5]，互联网中内容大数据的特征同样可以用描述一般大数据的多个“V”来进行刻画，包括体量大（volume）、快速化（velocity）、类型杂（variety）、有价值（value）、待辨识（veracity）和强关联（viscosity），如图 1 所示。治理互联网中复杂异构、良莠不齐和混乱失序的内容大数据，比处理特定领域中的一般大数据更加复杂，必须在互联网体系结构和关键治理机制等方面进行创新。

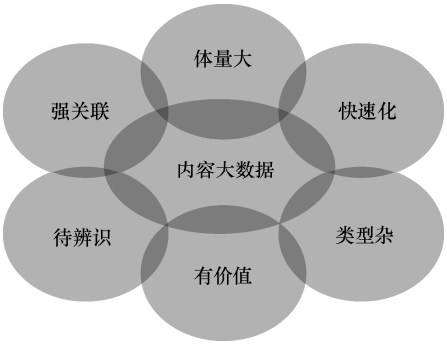

图 1 互联网内容大数据的多“V”特征

其次，当前互联网体系结构难以满足内容大数据的治理需求，还体现在缺乏“以内容为中心”的内容语义标识。传统互联网本质上是以地址为中心的网络，互联网体系结构中的面向地址特征，不但体现在以 IP 为代表的 TCP/IP 中，而且体现在 Web 中广泛采用的统一资源定位符（URL, uniform resource locator）中。Web 中所有内容均按照统一资源定位符 URL 来进行组织，这虽然极大推动了互联网主流应用范型向内容共享应用的跃迁，但正如 URL 名字“Locator”所强调的那样，它只能表示内容在 Web 中的位置，无法描述内容资源的丰富语义，因此难以支持基于语义的内容大数据描述、关联和管理等[6]。内容大数据的治理需求，本质上反映的是一种以内容为中心（而非以地址为中心）的需求。近年来，学术界注意到网络体系结构关注重心向面向内容的转变，提出了以结合广播与基于内容的路由（CBCB, combined broadcast and content-based）、发布订阅互联网路由范型（PSIRP, publish subscribe Internet routing paradigm）、信息网络（NetInf, network information）、内容中心网络（CCN, content-centric networking）和命名数据网络（NDN, named data networking）[7]等为代表的信息中心网络（ICN, information-centric networking）[8-9]。在这些 ICN 研究方案中，体现以内容为中心设计理念的是各种内容标识[10]，主要包括 CBCB 所采用的基于属性的标识、PSIRP/NetInf 所采用的扁平化内容标识、CCN 和 NDN 所采用的层次化内容标识等，但总体来讲，这些内容标识大多没有摆脱“重路由、轻语义”的传统设计思路，无法从体系结构层面提供对内容丰富语义的感知能力，因此难以从根本上解决网络空间中内容混乱失序的顽疾。

最后，互联网的功用本质上由它所承载的内容体现，互联网的内容安全与国家安全休戚相关。当前，互联网中不断发生的各种网络攻击事件正在威胁社会稳定和国家安全。由于互联网体系结构在安全方面（尤其内容安全与可信方面）的先天缺陷，互联网的安全态势变得越来越严峻[11]，而且互联网中日益频发的安全事件都或多或少与内容安全有关。近年来，美国国家科学基金会（NSF, National Science Foundation）、美国国防高级研究计划局（DARPA, Defense Advanced Research Projects Agency）、欧盟“地平线 2020”计划以及我国国家自然科学基金委员会等，都对网络安全给予了重点资助[12]。一些重要的网络安全研究方案包括美国的移动目标防御（MTD, moving target defense）[13]和定制可信赖空间（TTS, tailored trustworthy space）、信息工程大学邬江兴院士的拟态防御[14]、北京邮电大学方滨兴院士的使命确保技术、中国科学院信息工程研究所的自重构可信赖，以及各种事件跟踪和舆情监测研究等。这些研究方案或者并不针对内容治理这一难题而提出，或者难以突破传统互联网体系结构在内容安全与可信等方面的固有局限，因此未能改变当前互联网内容安全态势日益严峻的棘手现状。

与此同时，学术界对互联网中内容大数据及其影响的重视已现端倪。一方面，人们对 Twitter、微博等社交媒体上事件（event）的关注由来已久，先后提出了 Twitinfo、Twevent、MABED[15]等事件检测方法。另一方面，近年来互联网媒体领域正在发生深刻变革，Facebook 和 Apple 紧跟媒体融合与转型趋势，相继推出了 Instant Articles 和 Apple News，意在改变媒体内容的生产、组织和呈现形式；国际著名媒体纽约时报（New York Times）则创新性地提出了“新闻编码（particles code）”[16]，通过编码标识支持对新闻以时间轴和知识点进行组织，从而把意义上相关的多个内容有机关联。此外，W3C 还研发了基于标签元数据的互联网内容访问管理系统（PICS, platform for Internet content selection）[17]。2019 年初，美国 DARPA 宣布开展 KAIROS（knowledge-directed artificial intelligence reasoning over schema）研究[18]，凸显了美国对内容大数据智能处理的高度重视。KAIROS 项目的实现框架如图 2 所示，旨在通过人工智能、知识图谱和机器学习技术，在日益复杂的全球环境中更好地追踪、分析世界各地每天产生的无数事件和媒体片段，自动识别其中的关联性或线索，理解和预测导致世界混乱与动荡的因素。

综上所述，以地址为中心的现行互联网体系结构难以满足内容治理需求，正在面临内容大数据趋势显著、内容语义标识缺乏和内容安全态势严峻等多方面挑战。学术界围绕 Twitter、微博等社交媒体的分析研究，以及 Facebook、Apple、纽约时报、W3C 等关于内容组织、管理与访问的应用实践，虽然并不直接针对互联网内容治理，但从侧面反映出人们对互联网内容大数据的重视。DARPA 站在从混乱与动荡中建立秩序的角度，高调资助 KAIROS 项目研究，表明美国已经开始直面这一问题。然而，

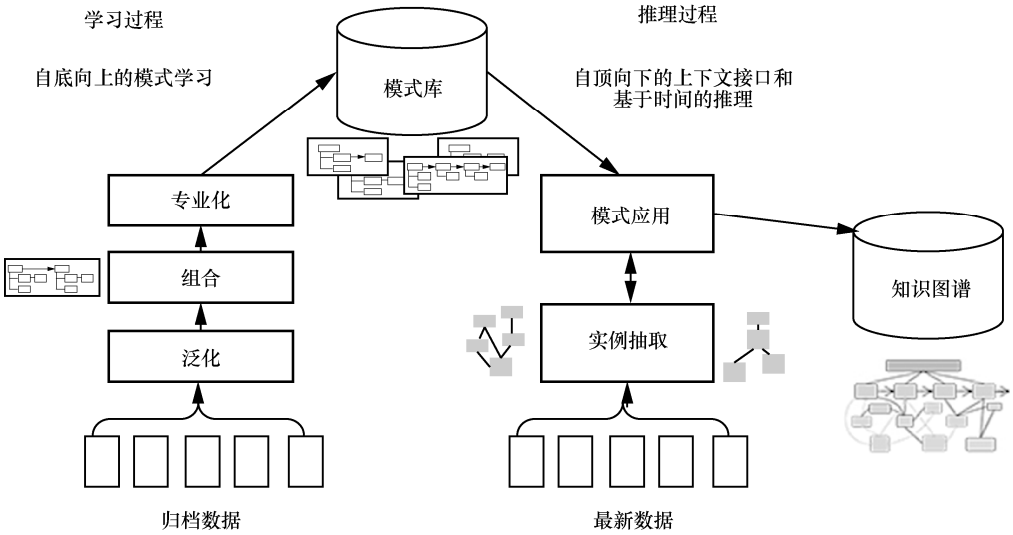

图 2　DARPA 资助的 KAIROS 项目实现框架

现有各种解决思路（包括 KAIROS）鲜有从网络体系结构的全局高度着眼，从变革传统互联网体系结构的角度给出的创新方案。为此，本文提出一种支持内容智能治理的双结构互联网，它能在维持互联网体系结构演进性的基础上，以较小代价换取互联网内容治理能力的显著提升。

3　新型互联网体系结构设计原则与模型

网络体系结构（network architecture）是设计和构造网络系统的科学，是对网络系统的总体结构规约。网络体系结构设计原则是指针对网络系统设计目标而提出的一系列具有指导意义的抽象设计原则。网络体系结构设计原则是计算机网络研究的“第一性问题”，有何种网络体系结构设计原则，才会有与之对应的网络体系结构。网络体系结构设计原则决定了网络系统的全局组织、总体结构和技术选择标准，各种具体实现技术都是在网络体系结构设计原则的指导下派生而得。针对当前互联网所面临的内容治理挑战，借鉴信息中心网络“以内容为中心”的合理研究思路，本文提出 3 条新型互联网体系结构设计原则。

1) 双驱动二元结构原则（P1）

基于“边缘论（end-to-end argument，也称端到端原则）”构建的互联网体系结构，是一种地址驱动的、以数据传输可达性为目标的网络体系结构。面向地址的设计理念贯穿整个互联网体系结构，体现在链路层地址、IP 地址、URL 地址等实现技术中。纯地址驱动的单一体系结构与“以内容为中心”主流应用泛型的不匹配，是互联网面临诸多挑战（包括内容治理）的本质根源。破解互联网发展困局的可行出路，是设计一种适配内容中心主流应用泛型的内容驱动次结构（secondary structure），用以辅助单一地址驱动的现行互联网体系结构，形成同时包含地址驱动主结构（primary structure）与内容驱动次结构的双驱动二元结构新型互联网，即双结构互联网（dual-architecture Internet）。

2) 富语义内容基元原则（P2）

网络体系结构的基元（building-block）是最能代表网络系统设计思想与核心理念的基础性构件，它体现特定网络系统的设计特色，又作为网络体系结构的基本单元，支撑网络系统的各种派生功能和上层应用。IP（或 IP 分组）是传统互联网体系结构的地址驱动基元，它是互联网面向地址设计理念的集中体现。应对传统互联网体系结构缺乏内容语义标识的挑战，必须以内容大数据的规范标引、高效共享与依法治理为目标，引入“以内容为中心”的新型内容标识作为双结构互联网的内容驱动基元，确保为海量、无序的无结构或半结构模量化内容大数据提供统一格式富语义矢量化内容标识。这种内容基元既是双结构互联网实现内容智能治理的抓手，又是沟通主结构和次结构的桥梁。

3) 定性定量综合集成原则（P3）

互联网海量内容大数据难以有效治理的主要原因，在于传统互联网信息处理领域一直缺乏有效的理论和方法。在系统工程领域，我国著名科学家钱学森先生提出从定性到定量的综合集成（meta-synthesis）方法[19]，利用现代信息理论、人工智能、知识工程等技术构建智能化综合集成研讨厅，实现定性的和定量的知识综合集成与复杂系统深层认知。在大数据和泛媒体环境下，一方面各种媒体信息（包括自媒体）在快速无序化“野蛮生长”，另一方面网络空间中充满大量虚假信息、片面信息，甚至恶意谣言，这大大增加了互联网内容治理的难度。因此，应该吸纳综合集成方法在开放复杂巨系统理论与实践方面的成功经验，采用定性定量综合集成原则指导互联网内容智能治理关键技术。

支持内容大数据智能治理的双结构互联网，完全遵循上述 3 条网络体系结构设计原则进行设计。双驱动二元结构原则（P1）为变革互联网体系结构提供了“设计原理（design philosophy）”创新，按照双驱动二元结构原则设计的双结构互联网，完全摒弃“非此（互联网）即彼（非互联网）”的网络体系结构一元论思维，在不改变现行地址驱动互联网体系结构的主体地位的基础上，借助多种网络（互联网、电信网和广播网等）优势互补的协同变革思路，采取“双重驱动、结构共轭”的二元结构创意建立起具有双体系结构的新型互联网。

在此基础上，按照富语义内容基元原则（P2）设计统一内容标签 UCL[1]，充当双结构互联网的“以内容为中心”新型内容标识，为繁杂异构内容大数据提供格式统一、语义丰富的内容驱动基元，直接支撑并简化了复杂的互联网内容大数据治理需求。进一步遵循定性定量综合集成原则（P3），设计内容大数据智能治理关键技术和实现机制，将常规技术难以解决的复杂巨系统问题（由无限用户、无限内容构成的单一地址驱动网络中的混乱无序内容大数据治理问题），转换成钱学森先生的综合集成方法可以求解的系统科学问题，运用人工智能、知识图谱、网络空间安全等技术，对异构、碎片化内容进行 UCL 自动标引，建立 UCL 多标识维度语义关联模型，引入数据与知识联合驱动的安全能级模型，借助基于知识图谱的内容大数据 UCL 知识空间，构建综合集成内容汇聚研讨厅并智能治理互联网内容大数据。

双结构互联网的体系结构参考模型如图 3 所示，它以地址驱动的互联网 TCP/IP 结构作为主结构，以内容驱动的“辐射−复制范型”播存网络[20]作为次结构。这种双驱共轭二元体系结构思路，显著区别于单纯的“打补丁”式演进路线或“推倒重建”式重构路线，既有利于继续发挥互联网 TCP/IP 主结构在端到端通信方面的既有优势，又能将单一的地址驱动网络迅速升级为“以内容为中心”的复合网络，不但能显著提升互联网的内容共享能力[21-22]，而且在应对互联网内容治理这一全球性难题方面有突出优势。

4　双结构互联网内容智能治理关键技术

4.1　UCL 国家标准与富语义矢量编码

弥合“以地址为中心”的传统互联网体系结构与“以内容为中心”的内容大数据治理需求之间的巨大沟壑，必须对网络体系结构的基元进行创新。互联网中的内容资源普遍采用 URL 进行组织，URL 既描述内容资源的地址，又充当内容资源的标识。作为内容标识，URL 的内容语义描述功能非常弱，由此带来互联网内容资源难找、难管、失序等弊端。为此，Tim Berners-Lee 提出了语义网（semantic web）概念[23]，试图使 Web 变成能够自动理解词语和概念，以及它们之间逻辑关系的智能网络，实现更加人性化和主动化的内容服务。但是，语义网要求机器能够“读懂自然语言”，实现起来非常困难。互联网中的内容大数据来源广泛且更新频繁，并具有非结构化（或半结构化）和高度异构等特点，因此治理互联网内容大数据的关键在于网络体系结构语义基元创新。

双结构互联网按照富语义内容基元原则，从全方位支持互联网内容大数据智能治理的角度，提出以统一内容标签 UCL 作为新型互联网体系结构的内容驱动富语义基元。UCL 本质上是一种面向内容的元数据（Metadata），它从互联网中海量内容资源难找、难管和失序等问题的根本症结入手，兼顾内容的生产者、消费者和管理者 3 个重要角色，能够有效弥补 URL 的语义缺失和管理缺失，成为双结构互联网中描述、引领和治理内容大数据的基石。图 4 是按照富语义内容基元原则（P2）进行全新设计后的 UCL，已经发布成为中华人民共和国国家标准 GB/T 35304-2017[1]，从 2018 年 4 月起在全国正式实施。UCL 国家标准能够有效支持内容大数

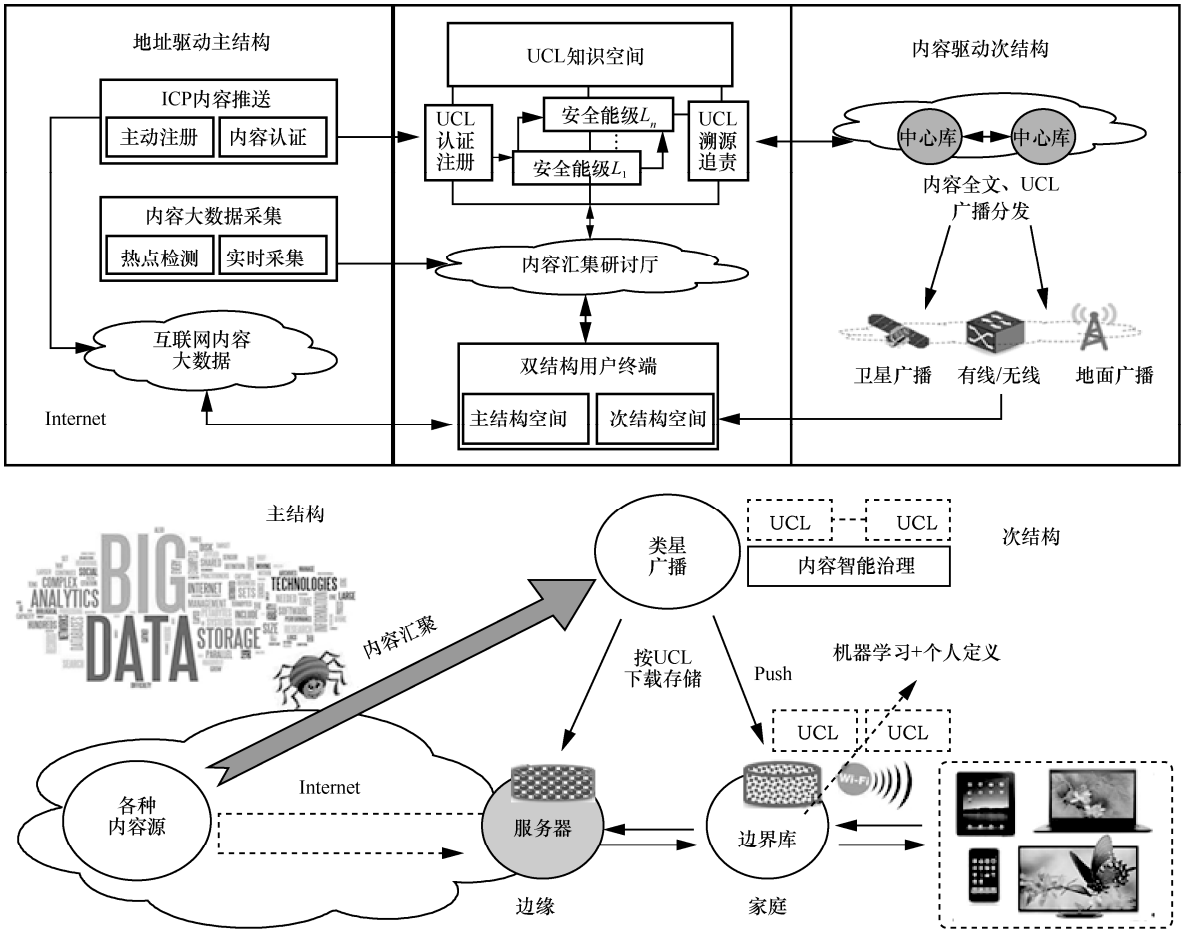

图 3　双结构互联网体系结构参考模型

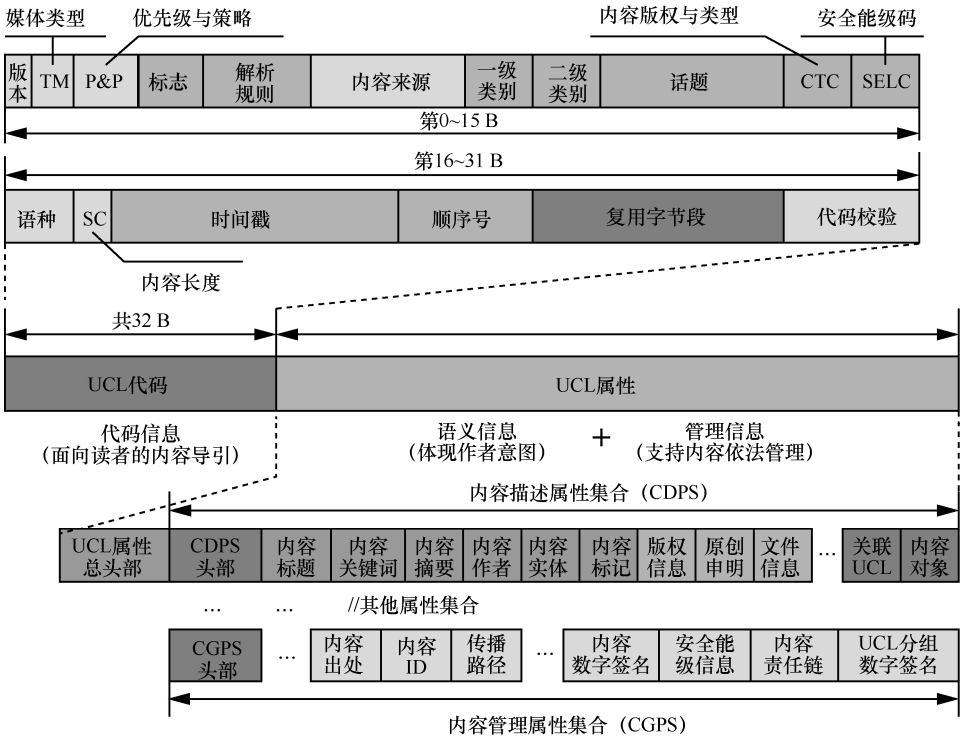

图 4　国家标准 GB/T 35304-2017 中的 UCL 格式

据的高效聚合与泛在分发、个性化主动服务、语义分析与知识萃取、认证注册物证链管理、依法治理与溯源追责等，为双结构互联网提供了标准化的富语义内容基元支持。

互联网中海量、繁杂、无序的内容大数据，本质上是一种模量形态的数据，治理互联网内容大数据的关键在于对模量化数据进行矢量化。UCL 国家标准是一种单位矢量性质的内容元数据，它可以在多个语义抽象层次上全方位描述内容资源的丰富语义信息。UCL 中的标题、摘要、话题、关键词和实体（Entity）等内容语义关键标识域，彼此关联又各有侧重（如图 5 所示），分别表征了内容的部分语义信息。如果把内容全文视为语义的零阶表述，摘要（即“有关内容的内容”）是语义的一阶表述，标题（即“摘要基础上的内容抽象”）是语义的二阶表述，而话题则是语义的高阶表述。UCL 基于五要素（5W）方法进行内容实体编码，描述何时（when）、何地（where）、何人（who）、何事（what）、何因（why）5 个方面基本要素。进而从多个语义关键标识域之间的联系出发，借助语义分析、知识库和实体链接等技术，建立 UCL 多关键标识维度间的语义关联模型，实现基于 UCL 国家标准的富语义矢量自动编码。

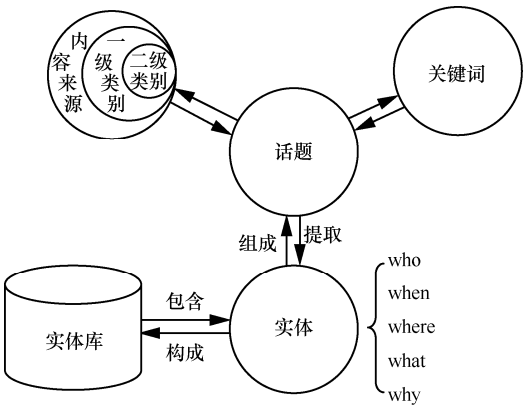

图 5　UCL 多关键标识维度间的语义关联

4.2　热门内容汇聚与 UCL 安全能级模型

进行互联网内容治理的首要问题，是如何有效地获取被治理的内容源，因为互联网的内容浩如烟海、层出不穷，无一遗漏地采集所有内容显然不太现实。所幸由复杂网络的研究揭示，在今天的互联网中，虽然一方面内容发布渠道的便捷性使碎片化内容大数据不断涌现，但另一方面用户对内容的访问又表现出明显的聚集性和无尺度（scale-free）性：全球有近 20 亿个网站[24]，其中绝大多数乏人问津，只有少数热门网站能吸引大多数访客，而一些热门内容更是被成千上万次频繁地访问。根据互联网内容访问的无尺度与幂律特征，互联网中的内容虽然浩如烟海，但往往其中起关键影响的只是少数热门内容，因此只需在确保尽可能高的内容覆盖度的同时，对热门内容进行重点关注。基于这一理论，双结构互联网在主结构和次结构之间安置内容大数据汇聚中心，它实时采集互联网中的热门内容并进行 UCL 矢量编码。

按照定性定量综合集成原则（P3），双结构互联网对热门内容进行多视角、多维度的话题汇聚和分析，运用知识发现与数据聚焦搜索技术，快速采集、汇聚特定话题相关的网站和社会媒体内容，对海量热门内容按照事件进行定性归类和定量关联。热点事件是具有重大影响的高热度事件[25]，热点事件及其关联的评论具有很强的社会舆论导向性，需要尽早发现并进行跟踪监测。双结构互联网对热点事件的挖掘分析流程如图 6 所示。利用“词袋模型”中词共现理论和 UCL 中的内容摘要和关键词等属性，通过大数据处理框架实现从定性到定量的迭代过程，通过动态调整关联规则挖掘算法的参数，智能挖掘热点事件并对关联评论的情感导向[26]进行挖掘分类。再根据联想型认知模式和知识图谱相关理论，实现基于事件评论情感极性的热点事件分类和聚类，跟踪热点事件的演化脉络（发生、发展、高峰、回落、平息），为网络舆情预警、舆情分析和应急响应等提供支持。

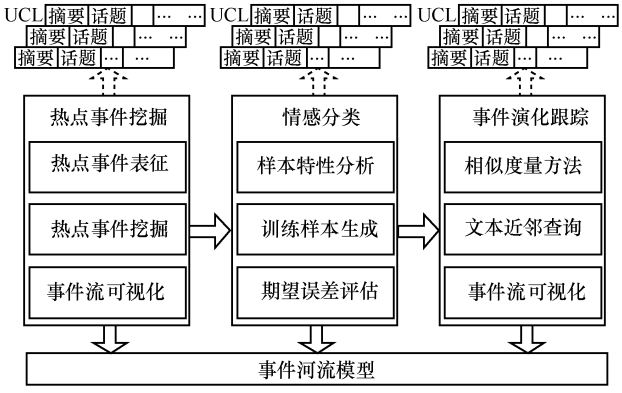

图 6　双结构互联网热点事件挖掘分析流程

在第 3 节提出的 3 条网络体系结构设计原则的指导下，双结构互联网对基于 UCL 的内容治理方法以及网络空间安全确保技术[11]进行了创新和突破。UCL 国家标准是支持内容大数据智能治理的利

器，它采取内容驱动理念对双结构互联网面向内容的基元进行了全新设计，形成生产、消费和管理三位一体的内容大数据创新标识体系。并且在 UCL 国家标准中，对双结构互联网安全能级模型（SELM, security energy-level model）给予了内嵌（built-in）支持[1]。安全能级模型将主结构模量内容向次结构空间的汇聚，细化为多个分离的安全能级（类似电子绕核运动的轨道能级），如图 7 所示。安全能级不但含有对内容的安全等级进行定级的概念，同时还有对内容安全等级进行动态调整的“能级跃迁”概念。依据来源路径的安全性、内容的质量和可信度等，设定内容的初始安全能级，然后借助知识萃取技术、深度学习神经网络和 UCL 知识空间等，对进入次结构的内容进行逐级趋严的智能化能级跃迁检查。内容安全能级信息记录在 UCL 代码部分和属性部分，再结合基于 UCL 的富语义矢量编码技术、多维度语义关联模型与 UCL 知识空间等，形成一种“以疏代控”的内容“依法治理”体系，实现网络空间安全从处理数据向治理内容的跃升，形成基于安全能级模型的数据与知识联合驱动智能化治理体系。

4.3 UCL 知识空间与内容汇聚研讨厅

由于互联网内容大数据具有碎片化和缺少关联等特点，实现内容智能治理还必须将这些碎片化内容按照语义进行有机关联。双结构互联网针对此问题的解决办法是，基于实体链接技术构建 UCL 知识空间，如图 8 所示。构建 UCL 知识空间首先需要一个基础 UCL 知识空间，然后将采集到的内容及其对应的 UCL 不断与该知识空间进行链接。基础 UCL 知识空间有多种构建方法，可基于维基百科（Wikipedia）、百度百科、ACE 中文语料库和 KBP 语料库等多种语料源来进行构建。首先根据抽取的词条信息，结合综合词频和位置图等实体语义权值计算方法，建立实体名称映射词典和关系映射词典，得到基本知识实体的逻辑关联知识图谱[27]。然后对从互联网采集到的每一份内容，抽取该内容对应 UCL 中的命名实体及其语义权重信息，并通过基于语境相似的实体消歧进行实体链接[28]。最后根据对应 UCL 实体之间的关联关系链接生成 UCL 知识空间。UCL 知识空间是互联网内容深度治理的基础，既可以根据一个 UCL 直接获得对应内容（也包括话题、事件等）的实体以及实体间的关联链接关系，又可以通过基于语义关联度排序的查询获得与内容关联的 UCL 集合，还可以支持内容的实体消歧、隐含知识萃取和 UCL 能级跃迁。

互联网中各种海量化、异构化、碎片化内容正在快速无序增长，其中充满大量虚假信息、片面信息，甚至恶意谣言，大大增加了互联网内容治理的难度。治理内容大数据的目的是把大量有待辨识（veracity）的数据，转换成有价值的、彼此关联的“知识”[18]。双结构互联网遵循定性定量综合集成原则（P3），在 UCL 富语义矢量编码技术、UCL 多维度语义关联模型、UCL 安全能级模型与 UCL 知识空间等的支持下，构建互联网内容大数据汇聚

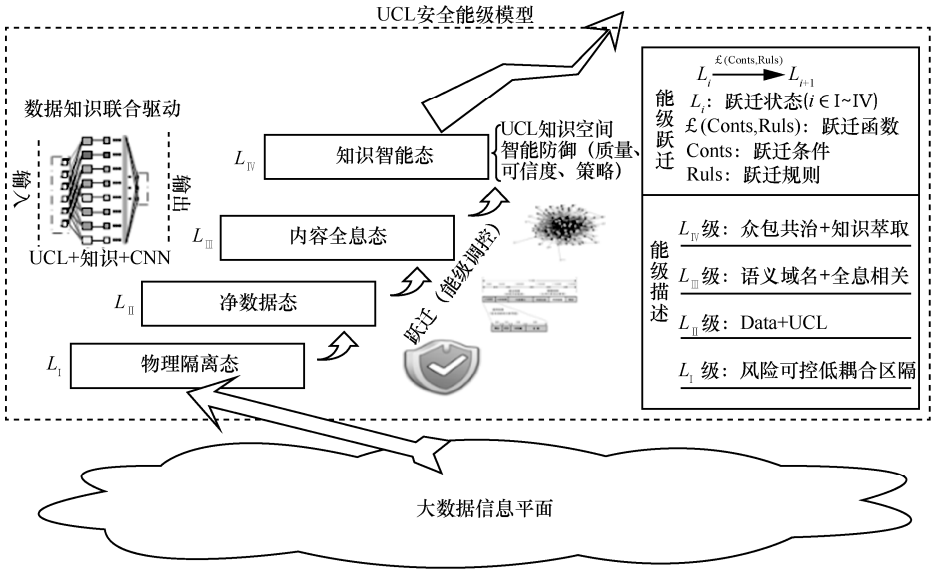

图 7 双结构互联网 UCL 安全能级模型

研讨厅（简称内容汇聚研讨厅），如图 9 所示，通过内容汇聚研讨厅实现对互联网内容大数据的深度治理。内容汇聚研讨厅的工作机理介绍如下。

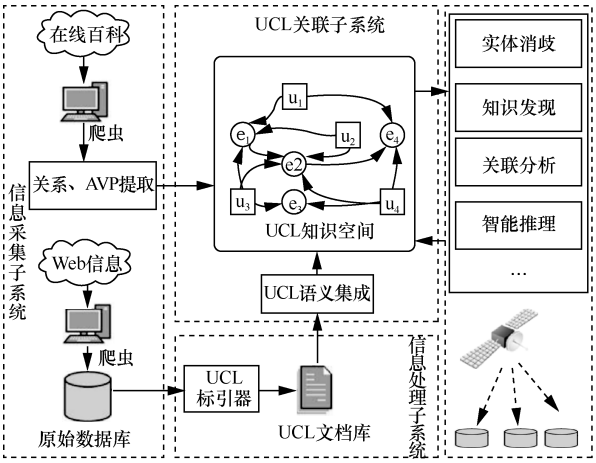

图 8　双结构互联网 UCL 知识空间

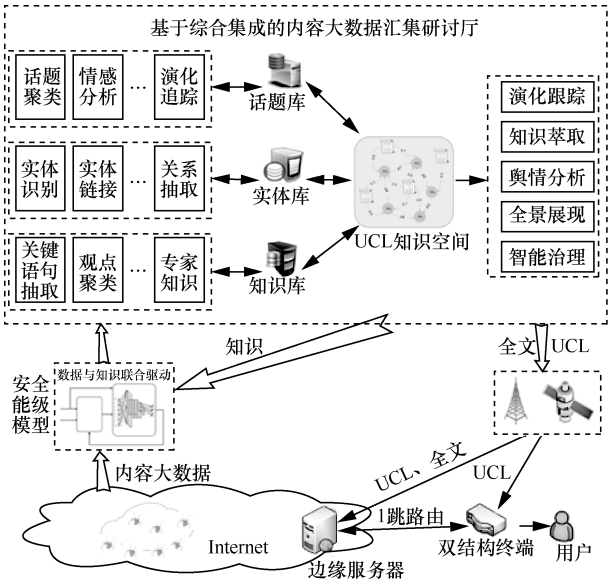

图 9　基于综合集成的内容大数据汇聚研讨厅

1) 从互联网实时采集的内容大数据在数据与知识联合驱动下进行 UCL 初级编码，并携带安全能级信息穿越 UCL 安全能级模型。新的知识同时被记录和关联到系统的知识库中。

2) 在 UCL 知识空间的支持下，进入内容汇聚研讨厅的内容进一步从话题、事件线（含事件）、实体、观点、知识等层面进一步解析和关联。

3) 碎片化内容不断卷积联合过往内容和知识，逐渐形成“各态历经（ergodic）”的时间链和知识链，孤立内容被自动置于话题和事件的全貌中。

4) 研讨厅展示细节、曝露联系、呈现全貌、跟踪演化，观点趋同的给予归并，观点趋反的则通过存储给予尊重，信息不确定性（熵）得以消除，达到去伪存真、以疏代控、由乱入治的目的。

5　双结构互联网内容智能治理原型系统

双结构互联网的 3 条网络体系结构设计原则均以内容大数据智能治理作为核心目标。双驱动二元结构原则（P1）强调用内容驱动的次结构播存网络辅助和改造单一地址驱动的互联网，形成兼含主、次二元结构的双驱共轭新型互联网，既维持互联网主结构的/平滑演进路线，又为治理互联网内容大数据提供网络总体结构支持。富语义内容基元原则（P2）直接聚焦新型互联网体系结构的基础性构件，将双驱动二元结构原则（P1）贯彻至新型互联网体系结构支持内容治理的核心基元，指导形成格式统一的富语义矢量化内容标签，并制定统一内容标签 UCL 国家标准。基于总体结构与核心基元的创新，定性定量综合集成原则（P3）进一步对治理内容大数据的方法学（methodology）进行创新，运用钱学森先生的综合集成方法的系统科学思维，求解异构、碎片化、混乱无序内容大数据的治理难题。在上述 3 条网络体系结构设计原则的指导下，本文研发了双结构互联网内容智能治理原型系统，对双结构互联网及其内容智能治理机制的可行性和有效性进行验证，原型系统的实现框架如图 10 所示。

双结构互联网内容智能治理原型系统主要包括热门内容汇聚子系统、UCL 知识空间子系统和内容汇聚研讨厅子系统。热门内容汇聚子系统首先实时采集互联网中的热门内容，然后利用自然语言处理技术（分词、去停用词、自动摘要等）和 UCL 多维语义关联模型，生成内容对应的 UCL 富语义矢量编码，并借助大数据处理平台（Hadoop 和 Spark）利用知识萃取技术和深度学习神经网络，实现热门内容聚类和热点事件发掘，将这些信息与系统中的既有知识不断卷积联合，在 UCL 安全能级模型与 UCL 知识空间子系统支持下，对进入次结构的内容进行认证注册与智能化能级跃迁检查。UCL 知识空间子系统首先利用维基百科和百度百科等构建基础 UCL 知识空间，然后对热门内容提取 UCL 命名实体，经过实体消歧等处理后将 UCL 链接到 UCL 知识空间，实现内容之间基于语义的深度关联，为 UCL 安全能级跃迁

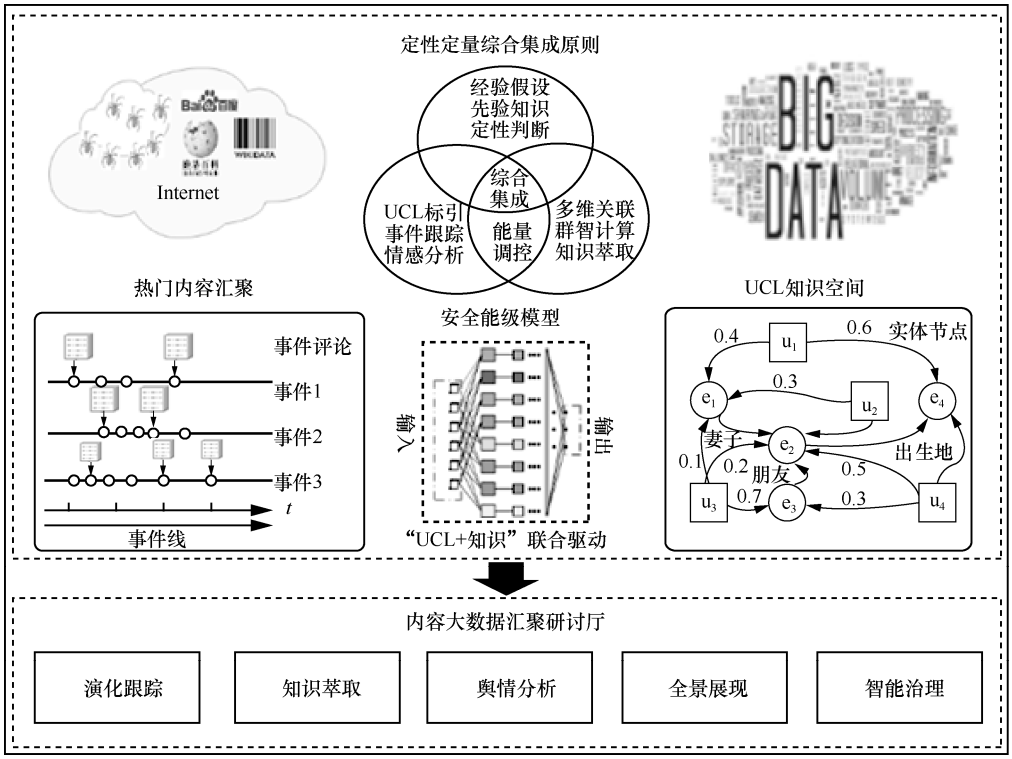

图 10　双结构互联网内容智能治理原型实现框架

与内容汇聚研讨厅子系统奠定实现基础。

通过内容汇聚研讨厅支持海量信息的汇集聚类，它并不立即为用户提供可信答案，而是提供一种信息汇聚场所，经历长时间的信息累积和开放式研讨，渐次获得大众信服的客观认知。信息集成研讨厅把信息按照话题进行定性归纳，“定性”指借助专业人士的智慧，把有争议话题归纳为少数几种观点（例如正、中、反等）。“定量”指计算机对同类观点进行时间与空间的定量关联。随着信息的汇聚累积，反映事物内在本质的内容将随时间浮现出来（emergence），而反映事物表层的非本质内容将随时间逐步湮灭，形成一种“以疏代控、和谐民主”的互联网内容治理环境。

内容汇聚研讨厅子系统集中体现双结构互联网的内容大数据智能治理效果，它遵循定性定量综合集成原则（P3）进行设计，其目的在于提供一种以话题和事件作为线索来组织内容大数据的汇聚场所（研讨厅），而不是立即为用户提供可信答案。内容汇聚研讨厅子系统中话题观点聚类与观点强度计算的实现框架如图 11 所示。用户可以借助内容汇聚研讨厅了解各方观点，并通过浏览观点语句及其来源了解每一个观点类的论点、论据和论证过程，获取最具有价值的支持观点的材料。用户通过以时间轴组织的内容汇聚研讨厅，可以了解观

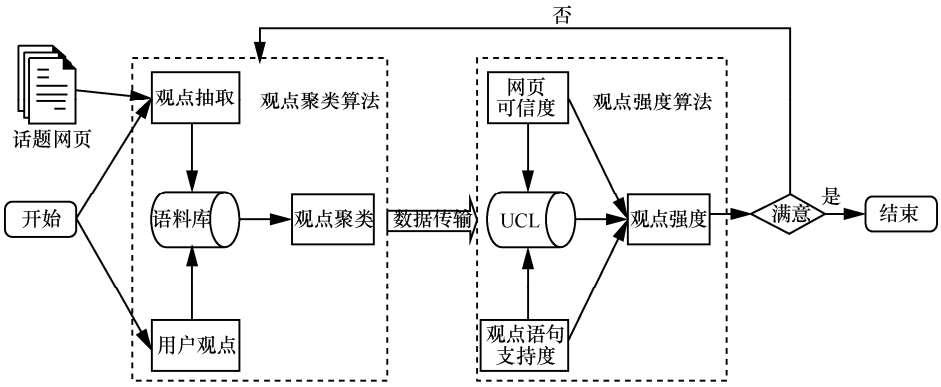

图 11　话题观点聚类与观点强度计算实现框架

点的变化过程，进而在全面了解各方观点的基础上对内容是否可信做出理性判断。

随着内容大数据的不断汇聚累积，内容汇聚研讨厅中反映事物内在本质的内容将随时间浮现出来（emergence），而反映事物表层的非本质内容将随时间逐步湮灭。借助内容汇聚研讨厅，经过长时间信息累积和开放式研讨之后，渐次逼近令大众信服的客观认知。内容汇聚研讨厅子系统通过可视化界面展示内容出处、发布时间、热门程度等基本属性，并借助内容汇聚子系统和 UCL 知识空间子系统提取内容的观点和计算内容的可信度，展示热点事件的挖掘结果和演化轨迹，全景呈现事件的各方观点及观点的可信度。

基于双结构互联网内容智能治理原型系统，本文以 2019 年 1 月 1 日—2 月 13 日这段时间为例，说明原型系统对互联网新闻的治理效果。采取去重策略后，该时间段原型系统自动从国内主要新闻门户网站共采集了 99 581 条新闻，原型系统从这些新闻中自动提取出 8 510 条事件线数和 67 106 个事件。如果用户关注其中的烟花爆竹安全燃放事件，则系统自动生成如图 12 所示的事件实体关联图，其中位于最内层的实心圆表示该事件线实体，其外第二层排布的实心圆表示命名实体，其外第三层排布的实心圆表示事件实体（用事件在系统中的存储 id 表示），最外层排布的实心圆表示新闻实体。并且，系统还能根据 UCL 知识空间和内容汇聚研讨厅中实时采集的新闻，展示烟花爆竹相关命名实体关系及与烟花爆竹安全燃放事件有关的命名实体词云，如图 13 所示。

双结构互联网内容智能治理原型系统也能进一步展示热点事件线的发展脉络。在上面的测试场景中，系统自动进行事件线提取和对应热度计算，并根据计算结果调用 CanvasJS 插件绘制出的 2019 年 1 月 1 日—2 月 13 日最热 15 条事件线如图 14 所示。其中事件线上每一个点代表一个事件，该点的纵坐标表示涉及该事件的新闻篇数，如其中事件线 5，在此期间的最热事件共出现了 19 篇报道，并在图中对应方框给出了从中随机选

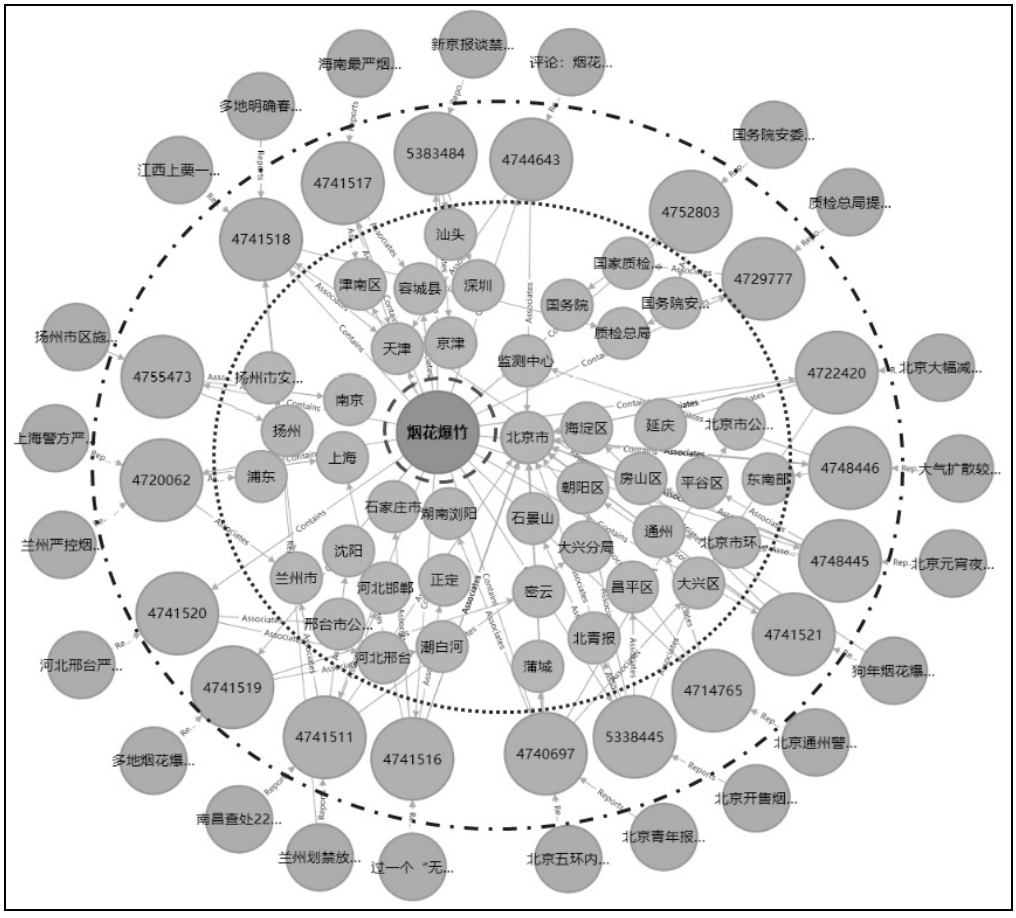

图 12　烟花爆竹安全燃放事件实体关联图

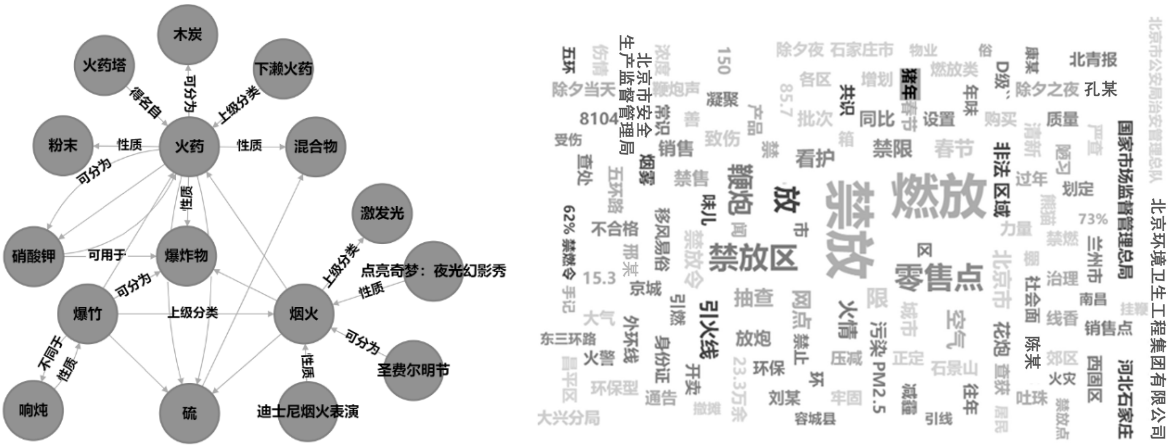

图 13　烟花爆竹命名实体关系和烟花爆竹安全燃放事件词云

取的一篇来自环球网的新闻题目。

原型系统还提供了对单条事件线演化发展轨迹的细节展示功能，如针对图 14 中事件线 5（特征词为：爆竹/燃放烟花/禁放/区域），其演化发展轨迹细节展示如图 15 所示。需要说明的是，系统中的事件线、事件、话题、观点等的提取及特征表

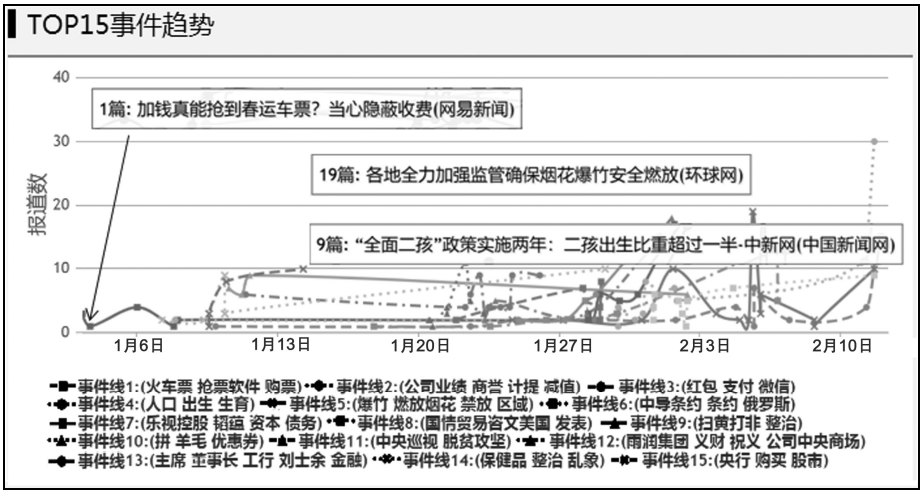

图 14　最热 15 条事件线（2019 年 1 月 1 日—2 月 13 日）

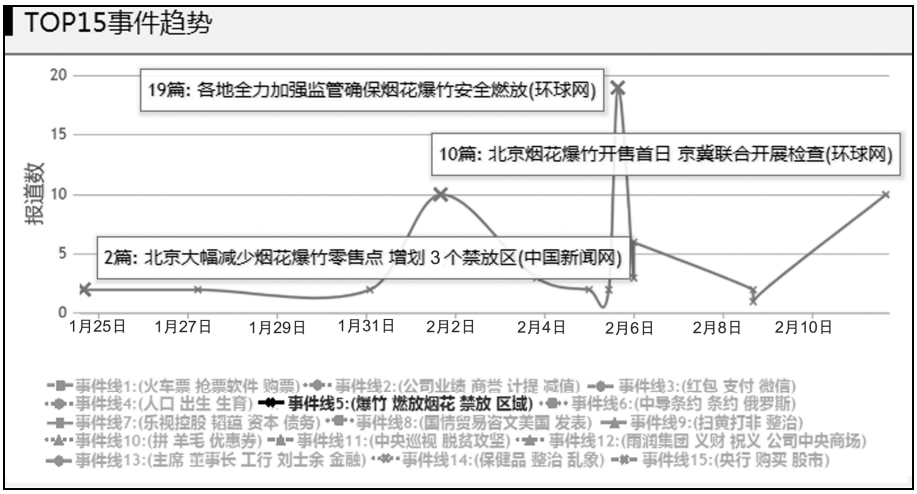

图 15　事件线 5 的演化发展轨迹

示，以及它们之间的从属关系和实体关联等，均是由系统根据内容的语义（UCL 富语义矢量编码、UCL 多维语义关联、UCL 安全能级模型与 UCL 知识空间等）自动处理得到，不需要人工干预。

双结构互联网内容智能治理原型系统不但验证了主、次二元结构共轭协同的可行性和有效性，而且依据双驱动二元结构原则（P1）、富语义内容基元原则（P2）、定性定量综合集成原则（P3）研发的内容汇聚子系统、UCL 知识空间子系统和内容汇聚研讨厅子系统，在基本功能和性能方面符合设计预期。对互联网中不断涌现的海量化、异构化、碎片化和混乱失序的内容大数据进行智能治理，是一项极具挑战性的课题，目前原型系统尚在进一步完善中。原型系统较好地体现出本文关于互联网内容治理的愿景：从总体结构、核心基元、治理方法学 3 个方面，对互联网进行网络体系结构层面的创新，营造让互联网内容大数据自动自发地由片面到全面、由无序到有序的智能化环境，形成一种“以疏代控、和谐民主”的内容“有序组织、依法治理”智能治理综合体系，借助格式统一、语义丰富的 UCL 国家标准，弥平无结构或半结构内容大数据处理的浅层、冗余和低效，并能通过内容汇聚研讨厅建立内容间深层语义关联，洞悉和发掘看似无关内容或事件碎片之间的相关性，全景展示事件的演化趋势和话题观点的客观可信度，实现网络空间安全从处理模量化内容大数据向治理结构化富语义内容元数据的巨大跃升。

6　结束语

由于在网络体系结构和治理方法等方面存在欠缺，互联网已经成为海量化、异构化、碎片化和混乱失序内容大数据不断涌现的集散地。然而，以地址为中心的现行互联网体系结构难以满足内容治理需求，正在面临内容大数据趋势显著、内容语义标识缺乏和内容安全态势严峻等诸方面挑战，如何高效治理内容大数据已经成为当前互联网体系结构研究的燃眉之急。

本文聚焦互联网内容治理这一棘手难题，深入分析互联网体系结构所面临的挑战，从总体结构、核心基元、治理方法学 3 个方面入手，提出支持内容智能治理的新型互联网体系结构的 3 条设计原则，即双驱动二元结构原则（P1）、富语义内容基元原则（P2）、定性定量综合集成原则（P3）。遵循这些设计原则，本文介绍了双结构互联网的体系结构核心理念和内容智能治理实现机制，尤其对 UCL 国家标准与富语义矢量编码、热门内容汇聚与 UCL 安全能级模型、UCL 知识空间与内容汇聚研讨厅等内容智能治理关键技术进行了详细阐述。最后，通过研发双结构互联网内容智能治理原型系统，对双结构互联网及其内容智能治理能力进行了验证。双结构互联网实现了网络空间安全从处理数据向治理内容的巨大跃升，为破解互联网内容大数据治理难题提供了网络体系结构层面的创新解决思路。

参考文献：

[1] 全国中文新闻信息标准化技术委员会. 统一内容标签格式规范:GB/T 35304-2017[S]. 北京: 中国标准出版社, 2017-12-29.
National Chinese News Information Standardization Technical Committee. Uniform content label format specification: GB/T 35304-2017[S]. Beijing: Standards Press of China, 2017-12-29.

[2] CHEN F, SITARAMAN R K, TORRES M. End-user mapping: next generation request routing for content delivery[J]. ACM SIGCOMM Computer Communication Review, 2015, 45(4): 167-181.

[3] CISCO. Cisco visual networking index: forecast and trends, 2017-2022 white paper[R]. (2019-02-27)[2019-08-05].

[4] IDC. The digital universe of opportunities: rich data and the increasing value of the Internet of things[R]. (2014-04)[2019-08-05].

[5] YU S, LIU M, DOU W, et al. Networking for big data: a survey[J]. IEEE Communications Surveys & Tutorials, 2017, 19(1): 531-549.

[6] ZOIDI O, FOTIADOU E, NIKOLAIDIS N, et al. Graph-based label propagation in digital media: a review[J]. ACM Computing Surveys, 2015, 47(3): 1-35.

[7] BOUK S H, AHMED S H, KIM D, et al. Named-data- networking-based ITS for smart cities[J]. IEEE Communications Magazine, 2017, 55(1):105-111.

[8] DIN I U, HASSAN S, KHAN M K, et al. Caching in information-centric networking: strategies, challenges, and future research directions[J]. IEEE Communications Surveys & Tutorials, 2018, 20(2): 1443-1474.

[9] AL-NADAY M F, THOMOS N, REED M J. Information-centric multilayer networking: improving performance through an ICN/WDM architecture[J]. IEEE/ACM Transactions on Networking (TON), 2017, 25(1): 83-97.

[10] VASILAKOS A V, LI Z, SIMON G, YOU W. Information centric network: research challenges and opportunities[J]. Journal of Network and Computer Applications, 2015, 25: 1-10.

[11] 尹浩, 郭东超, 吕勇强, 等. 主动防御的双结构网络[J]. 中国科学: 信息科学, 2018, 48(12): 1651-1669.
YIN H, GUO D C, LYU Y Q, et al. Dual-structural network of active defense[J]. SCIENTIA SINICA Informations, 2018, 48(12): 1651-1669.

[12] AMBROSIN M, COMPAGNO A, CONTI M, et al. Security and

privacy analysis of national science foundation future Internet architectures[J]. IEEE Communications Surveys & Tutorials, 2018, 20(2): 1418-1442.

[13] 蔡桂林, 王宝生, 王天佐, 等. 移动目标防御技术研究进展[J]. 计算机研究与发展, 2016, 53(5): 968-987.
CAI G L, WANG B S, WANG T Z, et al. Research and development of moving target defense technology[J]. Journal of Computer Research and Development, 2016, 53(5): 968-987.

[14] 仝青, 张铮, 张为华, 等. 拟态防御 Web 服务器设计与实现[J]. 软件学报, 2017, 48(4): 883-897.
TONG Q, ZHANG Z, ZHANG W H, et al. Design and implementation of mimic defense Web server[J]. Journal of Software, 2017, 48(4): 883-897.

[15] GUILLE A, FAVRE C. Event detection, tracking, and visualization in twitter: a mention-anomaly-based approach[J]. Social Network Analysis and Mining, 2015, 5(1): 1-18.

[16] LLOYD A. The future of news is not an article[R]. (2015-10-20)[2019-08-05].

[17] W3C. Platform for Internet content selection[R]. (2009-05-13)[2019-08-05].

[18] DARPA. Knowledge-directed artificial intelligence reasoning over schemas (KAIROS)[R]. (2019-01-23)[2019-08-05].

[19] 钱学森, 于景元, 戴汝为. 一个科学新领域——开放的复杂巨系统及其方法论[J]. 自然杂志, 1990, 13(1): 3-10.
QIAN X S, YU J Y, DAI R W. An ewdiscip line of science – the study of open complex giant system andit smethodology[J]. Chinese Journal of Nature, 1990, 13(1): 3-10.

[20] 杨鹏, 李幼平. 播存网络体系结构普适模型及实现模式[J]. 电子学报, 2015, 43(5): 974-979.
YANG P, LI Y P. General architecture model of broadcast-storage network and its realization patterns[J]. Acta Electronica Sinica, 2015, 43(5): 974-979.

[21] 刘旋, 杨鹏, 董永强. 双结构互联网内容共享能力研究[J]. 电子学报, 2018, 46(4): 849-855.
LIU X, YANG P, DONG Y Q. Research on content sharing capability for dual-architecture network[J]. Acta Electronica Sinica, 2018, 46(4): 849-855.

[22] XUAN L, PENG Y, YONGQIANG D, SYED H A. An analysis of content sharing hops for dual-structural network based on general random graph[C]//IEEE Global Communications Conference 2018 (Globecom 2018). 2018.

[23] W3C. Semantic web[R]. (2015-05)[2019-08-05].

[24] Internet Live Stats. Total number of Websites[R]. (2018-06)[2019-08-05].

[25] ADEDOYIN-OLOWE M, GABER M M, DANCAUSA C M, et al. A rule dynamics approach to event detection in twitter with its application to sports and politics[J]. Expert Systems with Applications, 2016, 55: 351-360.

[26] TAN S, LI Y, SUN H, et al. Interpreting the public sentiment variations on twitter[J]. IEEE transactions on knowledge and data engineering, 2014, 26(5): 1158-1170.

[27] WANG C, SONG Y, EL-KISHKY A, et al. Incorporating world knowledge to document clustering via heterogeneous information networks[C]//The 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. ACM, 2015: 1215-1224.

[28] LIN Y, LIU Z, SUN M, et al. Learning entity and relation embeddings for knowledge graph completion[C]//Twenty-ninth Aaai Conference on Artificial Intelligence. 2015: 2181-2187.

[作者简介]

杨鹏（1975− ），男，四川南充人，博士，东南大学副教授，主要研究方向为双结构互联网、统一内容标签、互联网内容治理、网络空间安全等。

李幼平（1935− ），男，福建泉州人，中国工程院院士，东南大学教授，主要研究方向为播存网络与 UCL、新型互联网体系结构、互联网内容治理等。