DOI:10.13530/j.cnki.jlis.140012

网络科技信息结构化监测的思路和技术方法实现*

张智雄　张晓林　刘建华　邹益民　谢　靖　钱　力　王　颖

摘　要　网络科技信息具有开源、发布及时等特点,目前已成为战略情报监测的重要资源。但这类资源又具有非结构化、无语义描述等特点,如何将 Web 信息从非结构的自由信息转为可分析的结构化、语义化信息成为一个亟需解决的问题。针对这一问题,笔者提出了网络科技信息结构化监测的思路方法。这一方法通过知识抽取技术,从网络信息资源中抽取出嵌在其中的知识对象以及对象间的相互关系,将自由文本转换为结构化的可计算的知识单元,在此基础上构建各类监测模型,进而实现对研究领域的态势监测。基于这一思路,笔者开发了"网络科技信息自动监测系统",并基于监测数据所形成的语义资源,进行了监测态势分析实验。图 6。表 1。参考文献 16。

关键词　网络科技信息　结构化监测　内容监测对象　知识抽取　自动监测系统　领域监测

分类号　G350

The Ideas and Methods of Structural Monitoring of the Scientific and Technical Information Resources on the Web

Zhang Zhixiong, Zhang Xiaolin, Liu Jianhua, Zou Yimin, Xie Jing, Qian Li & Wang Ying

ABSTRACT　Due to its openness and timeliness, S&T web information has become one of the most important resources for strategic intelligence monitoring. However, since S&T web information is unstructured and lack of semantic description, it is a challenge to transfer the unstructured web information into structured semantic knowledge. To solve this problem, the authors propose a method for structural monitoring of the S&T web information resources. By using the knowledge extraction technology, the authors firstly extract the knowledge objects as well as the relationship between objects from the web resources and convert the free text into calculable structured knowledge unit. Based on those extracted structured information, the authors build various kinds of monitor models to realize research profiling for specific research field. Based on those ideas, the authors implemente the automated web information monitoring system suitable for research field monitoring. A research profiling experiment also is carried out based on the semantic resources which are converted from the monitored web data. 6 figs. 1 tab. 16 refs.

KEY WORDS　S&T web information. Structural monitoring. Monitored content object. Knowledge extraction. Automatic monitoring system. Field monitoring.

1　引言

在网络日益成为科学交流和科学传播最重要渠道的今天,很多与科技创新相关的重要科研信息,如科技战略、科研项目计划、科研投入、科技合作、科研成果、科技指标等都是首先通过网络渠道对外发布。美国奥巴马政府 2011 年

* 本文系国家自然科学基金项目"基于语言网络的文本主题中心度计算方法研究"(批准号:61075047)及中国科学院文献情报能力建设专项项目"网络科技信息自动监测系统二期建设"(编号:院 1306)的研究成果之一。

通讯作者:张智雄,Email:zhangzhx@mail.las.ac.cn

“国家创新战略”[1]和2012年的“大数据计划”[2]，欧盟的“创新积分榜”[3]，世界经合组织的主要科学和技术指标[4]等与科技创新态势密切相关的信息都可以直接通过网络获取。

网络信息资源具有公开发布、及时(随时)获取的特点，目前已成为国内外情报机构的重要情报源，是开源情报(Open-source intelligence, OSINT)[5-6]的重要组成部分。一些与计算机科学相关的研究领域，如Web智能(Web Intelligence)[7]、话题追踪和探测(Topic Detection and Tracking)[8]、Web数据挖掘(Web Data Mining)[9-11]、舆情监测(Opinion Mining)[12-13]等也以其为主要研究对象，试图从中挖掘出有用的知识。

网络科技信息是一种非结构化、无语义描述的信息，各个网站上各条网络科技信息的结构不同、内容布局不一，没有足够的元数据对信息进行语义描述，这使得网络科技信息在自动计算分析上的可用性差。如何实现网络科技信息从非结构的自由信息向可分析的结构化、语义化信息转化，成为网络科技信息开发利用中的主要问题。

针对上述问题，笔者提出了一个进行网络科技信息结构化监测的思路，并基于这一思路实现了网络科技信息的结构化监测，设计和开发了适用于领域监测的“网络科技信息自动监测系统”。本文主要针对网络科技信息结构化监测的主要思路和框架、关键技术方法实现和实际应用效果进行论述。

2 结构化监测的思路和技术框架

科研机构在网页上发布的科技信息中，常常嵌有情报人员所关注的各种重点内容，如战略计划、科研项目、重要研究报告、科研创新投入、各项科技指标等。这些重点内容揭示了网页所表述的主体内容，而重点内容之间的相互关系揭示了这一网页中各项关系的骨干架构，反映着这一网页对于情报人员的价值。我们将这些反映网页信息的重点内容，称之为内容监测对象，将重点内容之间的各种关系(如语法、共现、语义)称之为对象关系。

结构化监测的主要思路就是从采集到的特定科研领域的信息资源中，抽取出嵌入其中的内容监测对象，如科研机构、科研人员、重要战略、重大项目计划、重要研究报告、积分榜、R&D投入等，并通过语法分析、共现分析、语义计算等方法，构建监测对象关系，将自由文本转换为结构化的可计算的对象网络，再基于此，构建各类监测模型(如重要内容判断、热点监测、重要对象跟踪等)，实现对研究领域的态势监测。

具体而言，对于每一条从网络上采集到的科技信息资源(如HTML页面、PDF文件、WORD文档等)，网络科技信息自动监测系统首先通过知识抽取技术，从这些网络信息资源中抽取出嵌在其中的知识对象以及对象间的相互关系。例如，对于“July 13, 2010, White House announces National HIV/AIDS Strategy”这一句子，通过内容监测对象的抽取，系统将识别出“National/HIV/AIDS Strategy”是一项重要战略(Strategy)，形成了“对象，类型，时间戳”的结构，如“National HIV/AIDS Strategy, Strategy, July 13, 2010”。同时，系统还通过语法分析，进一步分析出“White House”发布了“National HIV/AIDS Strategy”，进一步形成了“对象，对象，关系，时间戳”的结构，如“White House, National HIV/AIDS Strategy, Announces, July 13, 2010”。

通过对内容监测对象及对象关系的抽取，可以实现信息从自由文本向可供计算的结构化数据的转换。结合实际科研领域监测的需要，我们可基于这些结构化数据，实现重要目标对象的识别、重要目标对象的跟踪、热点内容的监测、特定情报内容的价值判断等功能，从而帮助战略情报人员实现相关科技领域的态势捕捉、态势跟踪、态势分析和态势的可视化表述。

上述网络科技信息结构化监测的思路，可进一步细化为图1所示的结构化监测的技术框架。这一框架重点完成结构化监测的四项重要

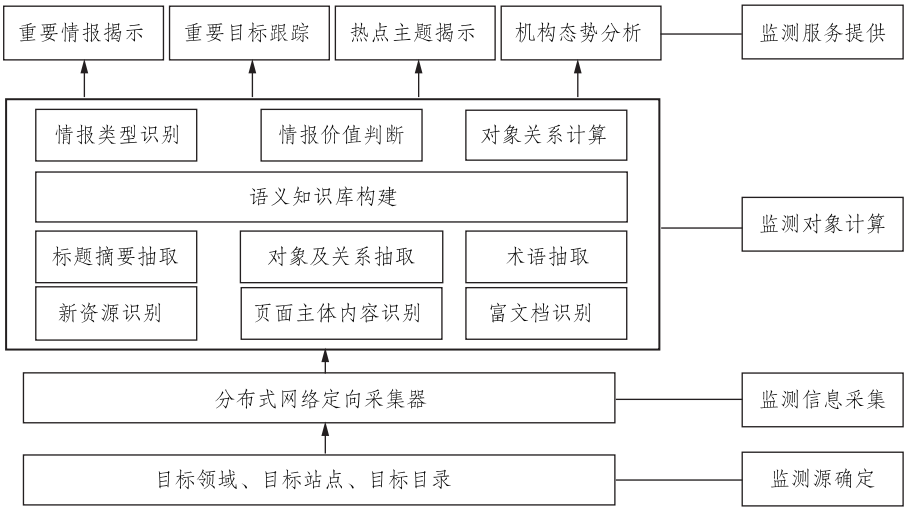

图 1　结构化监测的技术框架

逻辑任务:监测源确定、监测信息采集、监测对象计算和监测服务提供。

“监测源确定”是指确定需要监测的目标领域、目标站点以及站点之下的目标目录。对于战略情报监测来讲,需要监测的目标资源往往是一些机构网站上的资源,如国家科技政策机构、国家研究理事会、科技主管部门、科技资助机构、重要国际科技组织、科技咨询机构等。

“监测信息采集”是指定期对目标资源进行采集和收割。可构建一系列分布部署的网络定向采集器来实现对目标资源的精准采集。

“监测对象计算”完成结构化监测的主体工作。对于每一条采集到的信息资源,需要进行新资源识别,识别出原先系统中没有的资源。对于被识别出的新 HTML 页面,进行结构分析,抽取出页面中的主体文本内容。如果这一资源是 PDF、WORD、PPT 等富文档资源,还需要自动实现对富文档资源的解析,抽取出相应的文本内容。接下来,需要实施对资源的标题和摘要抽取、监测内容对象及对象关系抽取、领域术语抽取,将原先的自由文本转化为具有一定语义支撑的结构化数据,存储到相应的语义知识库中。在此之后,对于这些已经由结构化的监测对象和对象关系表达的资源,可以采用相应的数据挖掘方法,实现对这些资源的情报价值判断;通过自动分类工具,实现情报类型的识别;通过语义相似性计算,揭示情报之间的聚类和分布关系。

“监测服务提供”提供面向战略情报人员的自动监测服务。通过上述一系列工作,可提供重要情报内容揭示、重点监测目标跟踪、热点主题和热点对象揭示、机构整体态势揭示等服务。

3　结构化监测关键技术方法实现

进行领域战略情报监测,通常需要回答四个方面的主要问题:为了进行这一领域的战略情报监测,当前需要重点关注的这一领域的重要组织、人物、计划、战略等都有哪些;所获取的信息资源中是否有上述关注的内容;哪些资源有重要的价值;如何从众多信息资源中理出与当前关注点相关的情报,并发现一些原来没有关注而今后需要引起关注的重要内容。

围绕上述四个关键问题,笔者基于结构化监测的思路,按照结构化、语义化表示,对象化、自动计算分析的要求,提出了上述四个问题的

关键技术解决方案：构建监测本体指导结构化的目标内容监测，基于对象及对象关系抽取实现网页内容的结构化表示，基于对象指标实现网页内容的情报价值计算，基于对象计算实现监测目标的态势分析。下文对结构化监测中这四个关键技术方法实现进行论述。

3.1 构建监测本体指导结构化的目标内容监测

目标内容是指战略情报研究团队希望监测到的与本领域重大科研活动相关的内容。通过调研，笔者发现这种目标内容是可以通过结构化的方式来表达的。

战略情报研究团队需要随时了解目标科研机构在使命、愿景、战略定位、研究布局、主要科研活动、绩效指标、年报等方面的情况。国家领导人对相关领域的重要讲话、重大科技战略规划的出台、重要组织结构的调整、预算分配的变化、重要报告的发布、领域科技的重要进展等都是战略情报研究团队高度关注的目标内容。基于战略情报研究团队的监测内容需求，笔者提出了科研领域监测本体的概念，将科研领域的监测，转化为对一系列重点目标对象的监测，形成科研领域的监测本体，指导战略情报团队的内容监测。

科研领域监测本体从战略情报研究团队的需要出发，将科研领域监测的目标内容划分为四个大的概念范畴，即：被监测的目标主体、目标主题、目标活动和目标国家地区，并以此为基础进一步细化。具体的监测本体设计如图 2 所示。

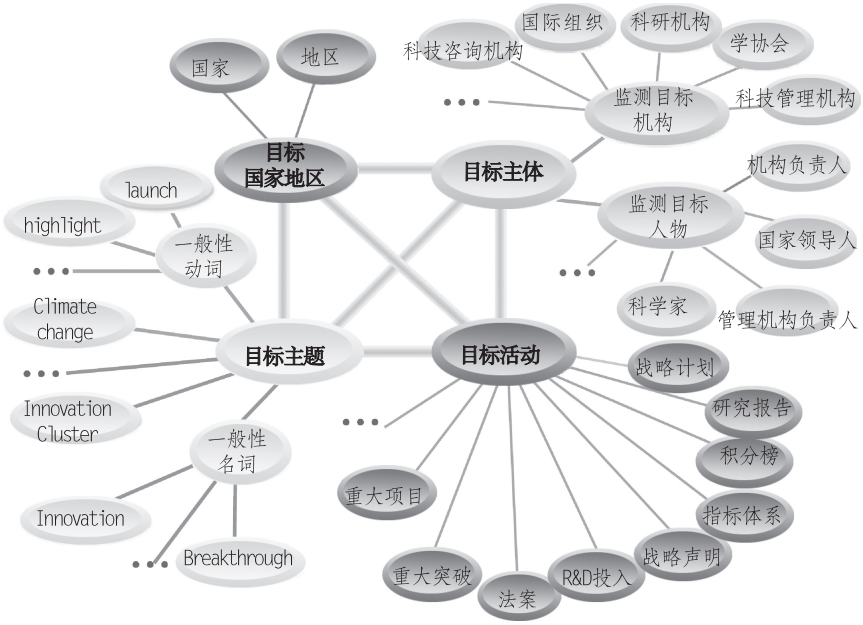

图 2 科研领域监测本体结构

目标主体包括被监测的目标机构和目标人物。目标机构又可细分为科技管理机构、政府机构、科技咨询机构、研究机构、国际组织、科技企业、学会协会等更能够反映机构性质的核心概念。而目标人物也可进一步划分为国家领导人、科技管理机构负责人、科研机构负责人、科学家等几种不同的类型。

目标主题由具体的领域主题、一般性监测动词、一般性监测名词组成。领域主题是被监测的科研领域的主题，它可以通过一系列本领

域的主题词来表示。而一般性监测动词和一般性监测名词由一系列表征重要监测内容的动词和名词组成。笔者在实际工作中发现,一些表征“宣布”、“发起”等行为的动词,如“announce”、“award”、“advance”、“spur”、“launch”等,被战略情报人员高度关注,是重要的监测动词。而一些反映重要科研进展的名词,如“breakthrough”、“innovation”等,也是战略情报人员关注的重点。

目标活动是战略情报人员关注的一些具体的科学研究、创新开发、科技管理等活动。通过调研,笔者发现在战略情报监测中,一些重要战略、重大项目计划、重要研究报告、重要科技指标体系(积分榜)、R&D 投入等是情报人员关注的重点。结构化监测的系统实现,需要收集和整理出相关科研领域下的重要战略、重大项目计划等的具体实例,指导自动监测。

在战略情报监测中,目标国家和目标地区各自的重要程度是不同的。科技大国、发达国家、金砖五国、新兴经济体的科研活动是目前战略情报监测的重点。

通过科技领域监测本体的构建,实现了战略情报人员希望监测到的目标内容的结构化表达,为网络科技信息的结构化监测提供了指导框架。

3.2 基于对象及对象关系抽取实现网页内容的结构化表示

上文中的领域监测本体已经明确了科研领域监测的内容,通过知识抽取的方法自动识别嵌于网页内容中的监测对象及对象关系,将有效地实现网页内容的结构化表示。笔者基于网络科技信息的特点,结合句法深度解析、语言分析词典和模式规则,提出了多种方法混合的内容监测对象抽取方法。具体流程如图 3 所示。

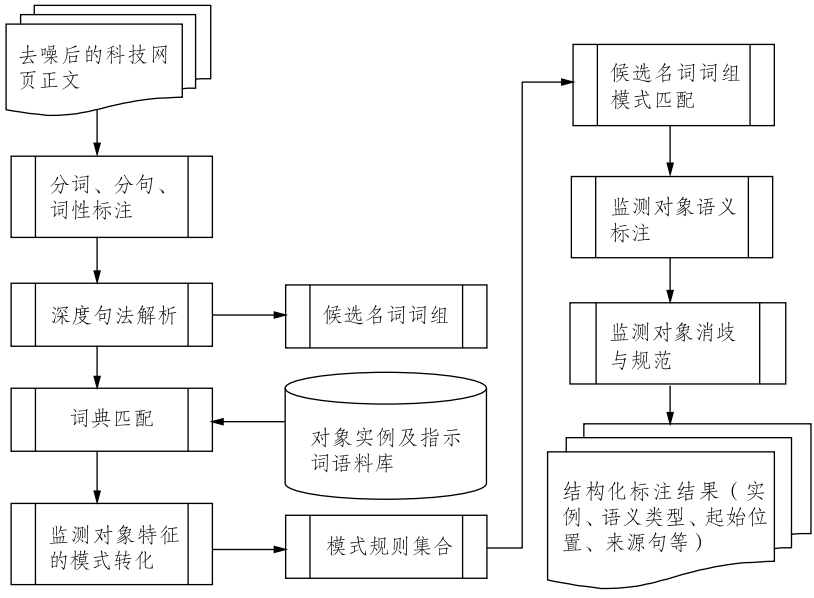

图 3 监测对象及对象关系的抽取流程

这一方法通过五个步骤实现内容监测对象及关系抽取。

(1)通过深度的句法解析获取候选名词词组。本文采用了 GATE (General Architecture for Text Engineering)[14]、Stanford Parser[15]等开源工具来进行分词、分句、词性标注等自然语言处理,对句子进行语法树解析,在深度句法标注的基础上获取其中谓语部分之外其它句法成分中的各个名词词块,形成候选名词词组。这些候选名词词组构成基于指示词典和基于模式特征

匹配的内容监测对象实例识别的基础。

(2)通过监测对象的指示词典和实例词典实现对象语义类型的初步判断。在候选的名词词组中,通常会存在语义指示性很强的指示词(如 University、Conference、Project 等)揭示这一名词词组的语义类型(如中心词为“Project”的往往是科研项目)。笔者根据领域监测本体,构建了各类监测对象的指示词典,并收集整理了一些重要监测对象的实例词典(包括各个实例的规范表达、缩写、变体表达)。通过实例词典,可以较为精确地匹配出候选名词词组的语义类型;通过指示词典,在进一步精选候选名词词组的同时,初步实现对这一候选的名词词组的语义类型判断。

(3)通过对象特征模式实现监测对象实例的识别。笔者根据一些监测内容对象实例在词形(首字母大小写、单复数等)、组成结构、指示词类型、指示词位置、上下文环境等方面的特点,构建了各类对象实例的识别模式和规则。对于上述经过语义初判的候选名词词组,进一步基于这一词组模式和规则的匹配,从中识别出需要的对象实例。例如,基于下一识别模式,可识别 University of New South Wales Australia 为一个大学的对象实例。

{(Token. string = = University)(Token. string = = of)(Token. orth = = upperInitial, Token. category = = NNP) * 4}

其中,Token 代表这一模式中的各个组成分词,Token. orth 表示其中一个分词的拼字法(大写、小写、混合拼写等),Token. category 表示分词的词性构成,NNP 表示专有名词单数。

(4)进行共指的内容监测对象的消解并为其确定规范的表达。如一个网页中会同时出现“Barack Obama”和“President Obama”,两词共指向同一个对象,需要合而为一,并为其选择一个规范的表达方式。在本研究中,笔者主要通过词典、共指模式规则、共现对象实例语义相似度三种方式实现消解。

(5)对象关系抽取。在识别出监测对象之后,基于深度句法解析过程中提供的主谓宾句法角色标注,进一步在同一个句子中,识别出两对象之间存在的主谓宾关系和基于共现的关联关系,以实现网页内容的结构化表示。

通过上述抽取方法,系统自动抽取出网页内容中内嵌的监测对象,将网页转化为一系列<监测对象,对象类型,来源网页,发布时间>和<监测对象,监测对象,相关关系,来源网页,发布时间>的结构化模式,利用这些结构化的对象及对象关系揭示网页的主要内容,实现网页内容的结构化表达。

3.3 基于对象指标实现网页内容的情报价值计算

网络科技信息自动监测系统每天会采集到大量的网页信息。如何从这些网页信息中,准确发现并有效揭示有重要情报价值的信息是结构化信息监测需要解决的一个重要问题。笔者针对这一问题,提出了基于监测对象指标实现网页内容情报价值计算的方法[16],实现对所采集网络信息资源的情报价值判断,以揭示重要情报资源。

具体而言,这一方法基于情报人员对内容监测对象的重要程度判断,构建了相关领域的监测内容对象重要度指标体系。对于采集到的每条网络科技信息,分别从情报来源的权威性、情报的类型、情报中内容监测对象的重要程度、情报的科技相关度和情报的主题相关度五个维度进行情报价值的判断,在此基础之上,确定这条网络信息对于特定领域的情报价值。

情报来源的权威性从发布这一内容的机构性质、来源目录等几个方面进行判断。发布科技信息内容的机构按性质可以划分为科技管理机构、科技咨询机构、政府部门、研究机构、国际组织、科技企业、新闻网站等类型。不同类型机构发布的信息重要度各不相同,对于科技战略和科技政策领域的情报监测而言,来源于科研管理机构的信息的重要程度往往高于来源于研

究机构的信息。在同一机构的网站上,不同目录下信息的重要程度也不一样,对于战略情报研究而言,发布在研究报告(或出版物)目录下的信息的重要性可能会远远高于新闻目录或事件目录下的信息。

情报类型是指情报外在的发布类型。笔者将情报类型划分为新闻报道、专家观点、深度分析报告、官方重要信息发布、研究成果等类型,并对不同类型的情报赋予不同的重要程度。监测系统通过识别信息的情报类型来判断其情报价值。在实际计算中,该维度包括信息的载体类型(HTML,还是 PDF、DOC)、标题中情报类型敏感词(如 Report、Press Release、Announcement 等)的数量和比例、正文中情报类型敏感词数量和比例、来源目录类型、正文长度、正文占全文内容的比例等多个指标。

情报中内容监测对象的重要程度是指在网页内容中包含的监测本体中定义的各类型重要对象的情况。在战略情报监测中,科技大国、发达国家、金砖五国、新兴经济体的科研活动,如重要战略、重大项目计划、重要研究报告、重要科技指标体系(积分榜)、R&D 投入等都是情报人员关注的重点,依据这些具体的目标活动、目标地区在实际情报工作中的不同重要度,可以进一步计算出情报中内容监测对象的综合重要度。在计算网页资源中涉及的内容监测对象重要度时,本文主要考虑四个方面的因素,包括某个对象在网页资源中出现的频次 F(O)、对象的重要度分值 IS(O)、对象的长度 L(O)、网页主体内容的文本长度 L(D)。网页资源中包含的多个不同语义类型的对象实例通过累加最终得到网页资源的内容监测对象重要度。具体计算如公式(1)所示:

$$\text{ObjectRelevancy}(D) = \sum_{i=1}^{n} F_i(O) * L_i(O) * IS_i(O)/L(D) \quad \text{公式(1)}$$

该公式中,ObjectRelevancy(D) 为一个网页的对象重要度,$F_i(O)$为某个对象实例 O 在文中出现的频次,$L_i(O)$为该对象实例的长度,$IS_i(O)$为对象 O 的重要度分值,L(D)代表了网页主体内容的长度,i 为网页资源中具有不同语义类型和值的监测对象实例数量。

情报的科技相关度指的是网页资源内容与科技内容的相关程度。富含科技主题词的网页资源才有科技战略情报价值。在计算该维度时,同样需要考虑某个科技主题词在网页资源中出现的频次、重要度分值、长度、网页主体内容的文本长度四个方面,其计算公式与公式(1)类似。

情报内容的领域主题相关度主要以网页内容中出现的领域监测本体、领域主题词和领域热点词来计算。与情报中内容监测对象的重要程度和情报科技相关度相似,利用从网页内容中识别出的领域监测本体、领域主题词和领域热点词,综合计算这些主题词在网页资源中出现的频次、重要度分值、长度、网页主体内容的文本长度,可以得到某个网页的领域主题相关度。

在计算上述五个维度的基础上,笔者还基于情报分析人员的经验知识,从一个维度或不同维度中抽取出多个指标,综合组成了一些计算规则,辅助情报价值的判定。以科技政策领域为例,重要关注人物与重大科技创新关键词(如资源标题中 Barack Obama 与 Sci&Tech Innovation 共现)共同出现的网页资源比某个重大科技创新关键词单一出现的资源具有更高的战略情报价值,而仅仅报道非重要国家的科技创新活动的网页资源(如资源标题中出现 Nepal 和 Sci&Tech Innovation)则没有太大的战略情报价值。

基于上述各指标与规则的计算值,可以自动实现某个网页资源情报价值的判定。

3.4 基于对象计算实现监测目标的态势分析

通过上述工作,笔者将一系列的网页文本资源变成了可供分析和计算的语义资源。综合内容监测对象、监测对象关系、网页内容、来源机构、网页发布时间、发布机构的空间分布等多

个分析维度，笔者以内容监测对象为主线，通过监测对象间的关系、监测对象和相关网页、相关机构、相关科研活动的各种关系揭示，提出了基于对象计算的监测目标态势分析方法。该方法主要包括以下几个方面的内容：

基于新对象的识别实现新趋势的识别。笔者将抽取到的资源内容中包含的显著对象同语义知识库中已有的对象进行比较，形成新对象列表，并将其同后续监测到的包含该对象的资源相关联，如果资源数达到了一定阈值，则作为新对象引起的新趋势加以推荐。

基于监测对象的频次分析实现监测目标的重要对象发现。通过对高频出现内容监测对象的统计分析，去除其中的噪音数据，供情报人员发现和识别监测目标下的重要对象。

基于监测对象的频次变化实现特定时间周期内的热点分析。通过对监测对象在特定时间窗口出现的频次进行统计，可以对时间周期内的热点对象进行发现，包括：热点科研活动、热点人物、热点主题等，有利于战略情报人员系统把握特定对象的动态变化情况。

基于监测对象的共现分布实现监测目标的关联描绘。对资源的类别、来源国家或组织、来源机构、来源机构性质、发布时间、监测对象、监测对象关系等建立多维索引，实现基于监测对象共现分布的关联描绘。

基于监测对象的网络资源聚类实现科研活动事件关联揭示。根据网络资源中包含相同对象的多少，以及对象在相应资源中的权重值，将采集到的网络资源集自动聚类为多个不同的科研活动事件，实现科研活动事件的关联揭示。

基于对象计算实现监测目标的态势分析。帮助战略情报人员对目标领域内的科技创新活动进行全面的了解，提升战略情报人员对目标机构战略定位、科技布局、政策调整、资源配置、科研产出、科研活动等各方面情况的洞察能力、监测能力和分析能力。

4 结构化监测的实际应用效果

基于上述网络科技信息结构化监测的思路和方法，笔者设计和实现了适用于领域监测的“网络科技信息自动监测系统”。这一系统能够全面自动监测特定领域内一些重要科研机构发布的网络信息资源，及时跟踪相关领域的科技动态，自动辨别重要情报资源，自动汇集重要富文档资源，自动揭示重要对象和主题，自动揭示热点对象和主题，并且结合战略情报监测的实际需要，实现了重要信息推送、快报辅助编辑加工等功能。

目前，中国科学院“科技战略与政策”、“空天科技”、“资源与环境”、“能源科技”和“信息科技”等五个战略情报研究团队和中国科学院青岛生物与能源所、上海光机所、上海药物研究所、大连化物所和生物物理所都使用了这一网络科技信息自动监测服务平台。

在将一系列的网页文本资源变成可供分析和计算的语义资源之后，笔者尝试利用美国白宫科技政策办公室（OSTP）网站上 15 个重要发展计划（或项目）进行基于对象计算的监测目标态势分析实验。

基于 OSTP 网站的资源，笔者根据上述方法，基于自动监测系统抽取的重要计划及其在 OSTP 网站上出现的频次，识别出近四年美国白宫科技政策办公室（OSTP）网站上最重要的 15 个发展计划（或项目）（见表 1）。

笔者进一步基于 15 个发展计划（或项目）中抽取的术语和监测对象，对这 15 个计划进行重要主题揭示、计划内的重要对象揭示和 15 个计划之间的关联关系描绘。

15 个 OSTP 重要计划的主题揭示如图 4 所示。从中可以看到 advanced manufacturing、economic growth、tax credit、national security、climate change、renewable energy、information technology、advanced material、business innovation 等是这些计划的重要主题。

表 1　美国白宫科技政策办公室(OSTP)网站上的重要计划和项目

重要计划和项目	提及频次数	提及文件数	首次出现时间
Smart Grid	185	20	2009-9-20
National Nanotechnology Initiative	120	28	2009-5-13
National Information Technology Research and Development Program	81	13	2009-5-13
Methane Opportunities for Vehicular Energy	65	45	2009-9-20
Materials Genome Initiative	51	14	2011-6-24
Federal Energy Management Program	31	8	2012-5-15
Startup America initiative	20	11	2011-3-2
US Global Change Research Program	18	13	2010-3-24
National Acid Precipitation Assessment Program	18	5	2012-1-20
National Robotics Initiative	16	11	2011-6-24
Open Government Initiative	12	10	2009-9-20
Wireless Innovation and Infrastructure Initiative	10	10	2011-2-15
Rural Energy for America Program	10	9	2009-9-20
Energy Data Initiative	10	4	2012-5-24
Advanced Manufacturing Initiative	8	2	2011-6-23

注:表中“提及频次数”指这个计划(或项目)在 OSTP 网站上被提及的次数,“提及文件数”指该网站上提及这个计划的文件数,而“首次出现时间”表示该网站上的相关计划(或项目)首次在自动监测系统中出现的时间。

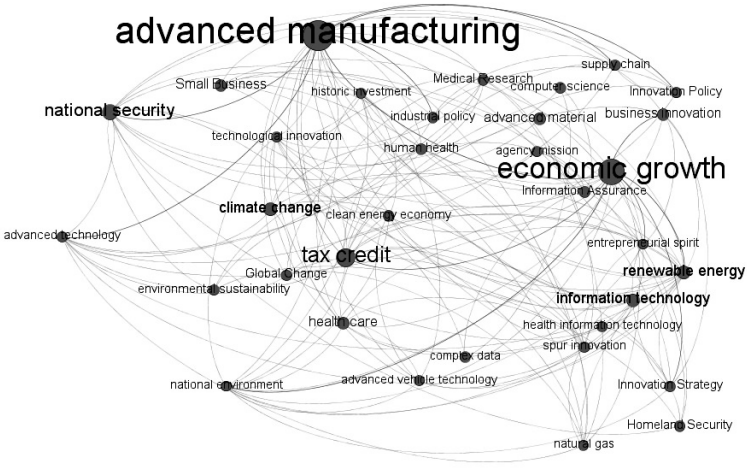

图 4　15 个 OSTP 重要计划的主题揭示

15 个 OSTP 重要计划的重要对象识别和揭示如图 5 所示。从中可以了解到与这 15 个 OSTP 重要计划相关的重要机构有 National Science Foundation、National Institute of Standards

and Technology、Department of Energy、National Institute of Health、U.S Department of Agriculture 等,相关的重要大学有 Massachusetts Institute of Technology、Harvard University、University of Illinois、University of Maryland 等,相关的重要人物有 Barack Obama、Jeo Biden、John P. Holdren 等,相关的重要实验室有 Oak Ridge National Laboratory、Argonne National Laboratory、Lawrence Berkeley National Laboratory 等。

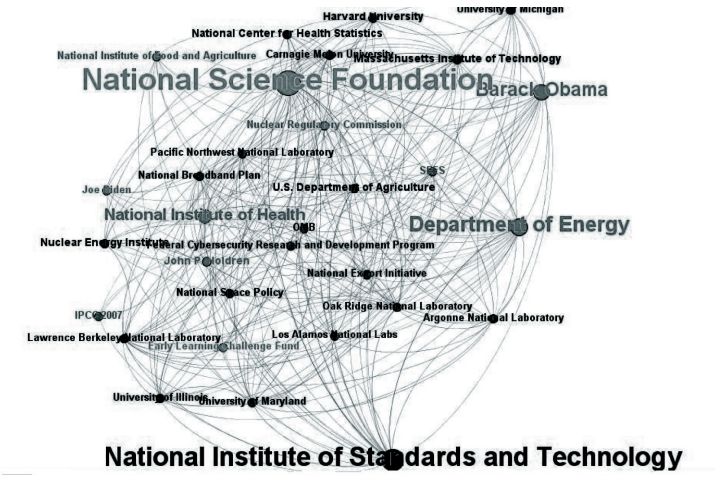

图 5　15 个 OSTP 重要计划中的重要对象及相关关系揭示

基于 15 个计划中出现的术语和监测对象的共现情况,笔者对这 15 个计划进行关联关系描绘(见图 6)。从图中可以看到这些计划之间存在着的相关关系。以 Smart Grid 计划为例,它与 National Nanotechnology Initiative、National Information Technology Research and Development Program、National Robotics Initiative、Methane Opportunities for Vehicular Energy 等都有相关性。

上述实验表明,“网络科技信息自动监测系统”及其产生的语义资源,可以支持基于对象计

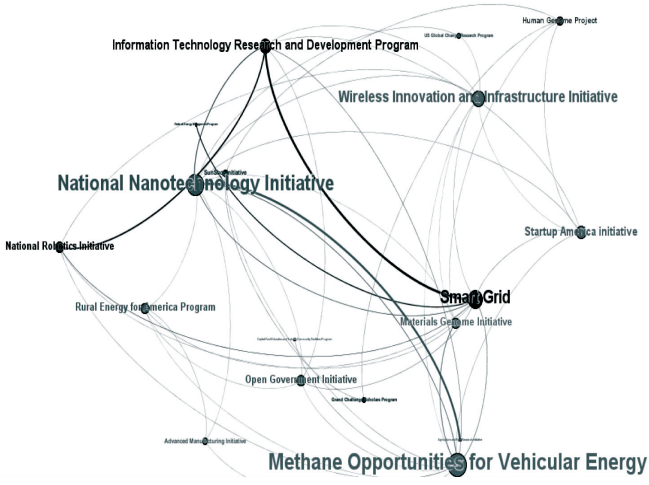

图 6　15 个 OSTP 重要计划的关联揭示

算的监测目标态势分析,能在一定程度上揭示监测目标内部的科研活动态势,以及监测目标之间存在的相关关系。

5 结论

网络科技信息具有发布及时、开放以及情报价值高等特点,已成为战略情报监测的重要资源。笔者提出的网络科技信息结构化监测的思路和方法,能够将网络科技信息从自由文本信息转为结构化、语义化的信息资源,并且可以利用这些资源实现科技战略情报的监测和跟踪。在本文中,笔者重点对结构化监测的思路方法、技术框架、方法实现进行了阐述。基于这一思路,设计和开发了适用于领域监测的"网络科技信息自动监测系统",并基于监测数据所形成的语义资源,进行基于对象计算的监测目标态势分析实验,验证了网络科技信息结构化监测这一思路的可行性和有效性。

当然,基于这一思路开发的"网络科技信息自动监测系统"还有需要改进和提高之处。例如:在对象及对象关系抽取方面,可以利用外部知识库提高抽取的准确率和覆盖率;在情报价值计算方面,可以考虑利用具有明显情报价值指示作用的动词和特征名词,发挥它们在价值判断中的积极作用;在监测目标的态势分析方面,还没有很好地以科研活动事件为主线对资源进行组织。这也是我们未来工作的重点所在。

致谢

"网络科技信息自动监测系统"的开发和应用得到了国家科学图书馆战略情报研究团队和中国科学院相关合作研究所的大力支持,在此一并致谢!

参考文献

[1] A strategy for american innovation: securing our economic growth and prosperity[EB/OL]. [2013-06-20]. http://www.whitehouse.gov/sites/default/files/uploads/InnovationStrategy.pdf.

[2] Obama administration unveils "big data" initiative: announces $200 million in new R&D investments[EB/OL]. [2013-06-20]. http://www.whitehouse.gov/sites/default/files/microsites/ostp/big_data_press_release_final_2.pdf.

[3] Innovation union scoreboard[EB/OL]. [2013-06-20]. http://ec.europa.eu/enterprise/policies/innovation/facts-figures-analysis/innovation-scoreboard/index_en.htm.

[4] Main science and technology indicators, Volume 2011 Issue 2[EB/OL]. [2013-06-20]. http://www.oecd-ilibrary.org/science-and-technology/main-science-and-technology-indicators_16097327.

[5] Open-source intelligence[EB/OL]. [2013-06-20]. http://en.wikipedia.org/wiki/Open-source_intelligence.

[6] Central Intelligence Agency. Establishment of the DNI open source center[EB/OL]. [2005-11-08][2013-06-20]. https://www.cia.gov/news-information/press-releases-statements/press-release-archive-2005/pr11082005.html.

[7] Web intelligence consortium (WIC)[EB/OL]. [2013-06-20]. http://wi-consortium.org.

[8] Topic detection and tracking evaluation[EB/OL]. [2013-06-20]. http://www.itl.nist.gov/iad/mig/tests/tdt.

[9] Liu B. Webdata mining, exploring hyperlinks, contents, and usage data[M]. Heidelberg: Springer, 2011.

[10] Liu M, Liu Y, Xiang L, et al. Extracting key entities and significant events from online daily news[C]//Intelligent Data Engineering and Automated Learning-IDEAL 2008, 2008(5326): 201-209.

[11] Ploeger T, Armenta B, Aroyo L, et al. Making sense of the arab revolution and occupy: visual analytics to under-

stand events[C]//Detection, Representation, and Exploitation of Events in the Semantic Web, Boston, USA, 2012:61-70.

[12] Pang B, Lee L. Opinion mining and sentiment analysis[EB/OL].Foundations and Trends in Information Retrieval,Vol. 2, No 1-2 (2008) ,1 - 135. [2013-06-20].http://www. cs. cornell. edu/home/llee/omsa/omsa. pdf.

[13] Risse T, Dietze S, Maynard D, et al. Using events for content appraisal and selection in web archives[C]//Proceedings of the 11th Interational Semantic Web Conference ISWC2011, Bonn, Germany, 2011:1-10.

[14] GATE (General Architecture for Text Engineering)[EB/OL]. [2013-06-20]. http://gate. ac. uk.

[15] The Stanford NLP (Natural Language Processing) Group. The Stanford parser: a statistical parser[EB/OL]. [2013-06-20]. http://nlp. stanford. edu/software/lex-parser. shtml.

[16] 刘建华, 张智雄. 情报重要度的指标体系和计算方法[R]. 北京: 中国科学院文献情报中心, 2011. (Liu Jianhua,Zhang Zhixiong.The index system and calculation method of information importance[R].Beijing:National Science Library,Chinese Academy of Sciences,2011.)

张智雄 中国科学院文献情报中心研究馆员,博士生导师。
通讯地址:北京中关村北四环西路33号中国科学院文献情报中心。邮编:100190。
张晓林 中国科学院文献情报中心教授,博士生导师。通讯地址同上。
刘建华 中国科学院文献情报中心博士研究生,馆员。通讯地址同上。
邹益民 中国科学院文献情报中心博士研究生,助教。通讯地址同上。
谢 靖 中国科学院文献情报中心馆员。通讯地址同上。
钱 力 中国科学院文献情报中心博士研究生,馆员。通讯地址同上。
王 颖 中国科学院文献情报中心馆员。通讯地址同上。

(收稿日期:2013-09-28;修回日期:2013-12-14)