文章编号: 1673-3819(2018)06-0064-04

互联网大数据在指挥决策中的应用研究

高晨旭[1]，张鹏乐[1]，邢 萌[2]，李海龙[1]
(1. 军事科学院，北京 100141; 2. 陆军装甲兵学院，北京 100072)

摘 要: 针对作战指挥决策领域对互联网大数据应用匮乏、需求结合不紧密、应用模式不系统等问题，提出了互联网大数据在指挥决策领域应用的方法。首先，基于指挥决策领域本体进行需求建模，然后从需求出发，对全互联网数据进行主题爬虫，对获取的数据进行命名实体识别、信息抽取、事件聚类等处理计算，最后以定制化的服务模式支撑作战指挥决策。该方法创新性的将作战需求模型和互联网大数据挖掘分析过程有机融合，提升指挥决策的效率和准确性。

关键词: 需求建模; 大数据; 挖掘分析; 微服务; 主题爬虫
中图分类号: E27 **文献标志码:** A **DOI:** 10.3969/j.issn.1673-3819.2018.06.014

Research on the Application of Internet Big Data in Command Decision Making

GAO Chen-xu[1], ZHANG Peng-le[1], XING Meng[2], LI Hai-long[1]
(1. Academy of Nilitary Sciences, Beijing 100141; 2. Army Armored Military Academy, Beijing 100072, China)

Abstract: In the field of traditional Integrated Combined Operation Command, Internet data is used inefficiently and is not closely connected with requirement. And also, the application model of internet data is not systematic. Therefore, this paper puts forward the application of Internet big data in the field of Integrated Combined Operation Command. First of all, the command decision requirement modeling based on domain ontology. And then, starting from the requirement, we crawl the whole Internet data topically. Next, we process the acquired data by named entity recognition clustering, information extraction and event clustering. Finally, it support the operation command and decision by customized service model. This method can promote efficiency and accuracy of command decision by fuse the operational requirement model and Internet data mining process innovatively.

Key words: requirement modeling; big data; mining analysis; micro service; topical crawler

联合作战指挥是目前世界主要国家的现代作战思想，在现有的战争模式下，快速获取战场信息，准确还原战场态势，保证所有参战单元对战场态势有一个共同理解，在高度协同的指挥控制下实现战争胜利。其中，在联合作战指挥决策过程中，制信息权是关键，成为继制海权和制空权之后新的战场制高点，而大数据技术是夺得制信息权的重要利器。

随着互联网技术的飞速发展，数据成了信息时代的核心资源[1]，成为战斗力生成的核心要素。然而，由于互联网大数据具备体量大、种类繁多、价值密度低等特点，加剧了信息过载问题，无法直接形成战斗力[2]。为有效提升数据价值，挖掘有用信息，金融、网购等应用领域借助大数据挖掘分析技术，获得了显著的成功。我军在数据的挖掘分析方面也做了很多尝试，通过大数据采集、处理、存储、分析、挖掘等过程，并借助 Hadoop 等分布式技术，加速数据分析计算，实现实时动态的大数据态势展现[3-4]，取得了一定效果，但也存在诸多问题: 1) 大数据应用与作战需求联系不紧密，没有形成科学的需求建模体系，导致挖掘出的有用信息依然无用; 2) 作战领域欠缺对互联网数据的使用，虽然互联网上的大多数据与军事无关，但在舆情分析、情报获取等方面具有重要应用价值，能有效支撑辅助决策; 3) 目前的大数据挖掘分析工具大多以简单的工具包交付，缺乏系统的应用模式，不能有效形成使用反馈、迭代更新、系统完善的良性循环。

本文针对目前大数据挖掘分析在指挥决策领域遇到的一些问题，提出了一种基于本体(Ontology)的需求建模方法，并从需求出发，对全互联网数据进行主题爬虫，对获取的数据进行命名实体识别、信息抽取、事件聚类等操作，最后以定制化的服务模式支撑作战指挥决策。

1 指控领域大数据使用需求分析建模

本文主要基于指控领域本体[5]进行需求建模，实现对指控领域大数据使用需求分析。本体是对特定领域之中某套概念及其相互之间关系的形式化表达，是一种特殊类型的术语集，具有结构化的特点。对于本

收稿日期: 2018-02-27 修回日期: 2018-04-01
作者简介: 高晨旭(1990-)，男，河北廊坊人，工程师，研究方向为人工智能。
张鹏乐(1990-)，男，工程师。

体的构建，国内外也已经提出了一些有效的解决方案，如 Skeletal Methodology（骨架法）[6]、Methontology 方法[7]、On-To-Knowledge 方法[8]、UPON（the Unified Process for Ontology）Methodology[9] 等，目前，这些方案仍处于基础理论研究解决，尚未广泛应用于实际案例，也没有形成被广泛认可的标准。

本文结合上述基本方法，根据指控领域特点，通过计算机技术、本体技术来模拟标引指控领域，构建一个统一的、规范的本体库来描述专业领域的知识，并通过本体库进行需求建模。方法流程如图 1 所示。

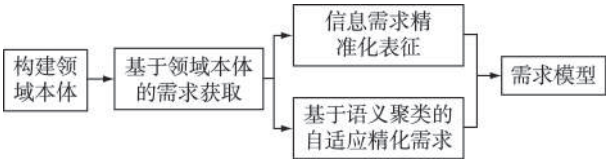

图 1　基于本体的需求建模流程图

1）构建领域本体。首先，确定本体库的专业领域，搜集领域内相关知识以及已有的本体库，在复用已有本体库的同时，增加专业领域中的术语、概念，然后针对上述内容建立专业领域的知识分类，并进行编码，形式化本体库内容，便于计算机处理，最后对输出结果进行评价和检验。

2）基于领域本体的需求获取。针对对口决策部门对互联网大数据的需求难以准确表示的难题，面向战区级、军种级、集团军级等各级决策部门用户，以指挥决策为背景，在迭代整合专家和用户意见的基础上，基于领域本体的业务信息需求分析技术，以各部门业务本体和领域本体作为需求获取过程的基本线索，引导各决策部门用户以规范化方式描述其信息需求，并通过可复用领域模型，构造军政各层级用户的信息需求文档，达到系统、高效和规范获取对口决策信息需求的目的。

3）信息需求精准化表征。针对指挥决策过程中数据需求往往不确定和变更频率高等难题，根据用户行为数据抽象出一个标签化的用户模型，即用户画像，然后基于用户画像进行互联网信息需求精准表征。该方法采取在系统中缩短用户使用路径并持续更新完善用户需求模型的方式，通过非结构化数据抽取能够自动识别的指标信息，通过需求刻画模板描述用户信息需求的要素，包括描述业务领域的重要关键词和语义实体，以及与自身业务相关的重要门户网站和社交媒体账号等信息源。挖掘用户历史行为、位置、时间等信息，通过不断缩短用户的使用路径，提高检索效率，满足用户主动隐性的需求或场景的即时需求，通过迭代优化的方式不断细分和描绘用户画像，达到超出用户体验预期的效果，指导后续数据挖掘。

4）基于语义聚类的自适应精化需求。针对互联网大数据的语义理解与分析和多模态关联与融合等难题，基于语义聚类技术，以各级业务主管部门用户提供的初始信息需求要素为基础，进一步理清众多业务门类、不同领域、不同层级主管部门的信息需求偏好，充分体现覆盖面广、信息源复杂、综合程度高、描述结构抽象等信息特点。基于热点实体发现技术，通过分析有关互联网信息形成反馈，经过用户确认后完善信息需求档案，进一步刻画用户需求并支持不断更新完善，用以指导系统的信息获取和增值分析。

2　互联网数据获取与挖掘分析

互联网大数据挖掘分析系统架构如图 2 所示，主要由数据基础层、预处理层、数据挖掘分析层、业务应用层组成。其中，数据基础层主要通过主题爬虫技术获取互联网文本数据，通过中文分词、命名实体识别、词性标注、句法分析等技术对文本数据预处理，为信息抽取、事件聚类等挖掘分析提供支撑，以实现业务层的应用。

2.1　基于主题爬虫的数据获取

为满足多业务主管部门个性化服务需求，本文提出了一种基于自适应主题的数据获取策略，只需要根据需求主题，提供相应主题的一组链接，数据获取组件即可根据链接地址对应的网页完成主题建模，并基于此进行主题爬行，取得与使用训练集的主题爬虫相当的数据获取效率[10-11]。

主题爬虫可以分划分为 4 个部分，网页爬行器、页面分析器、相关度评价器、主题表达器。

1）网页爬行器。网络爬虫的爬行是请求服务器、下载网页、抽取链接过程的不断循环迭代，爬虫在请求服务器阶段是相当耗时的，需要与服务器建立 HTTP 链接，然后等待对方反应，这也成了爬虫性能的一个瓶颈，正因为如此，需要应用多线程技术解决这个问题。

2）页面内容分析。在爬虫获取页面之后，需要分析页面信息，如标题信息、关键词信息、正文信息、超链接信息等。这些信息可用于计算该网页与主题的相似度，并获取更多的 URLs。

3）相关度评价器。相关度评价器由两部分组成，分别为页面相关度评价和 URL 相关度评价。页面相关度评价主要是分析从页面分析器出来的页面与主题的相似度，URL 相关度评价主要是用来判断页面中的链接与当前主题的相似程度。

4）主题表达器。基于层次目录树、基于本体、基于关键词等方法都是主题的有效表达方式，本文所采用

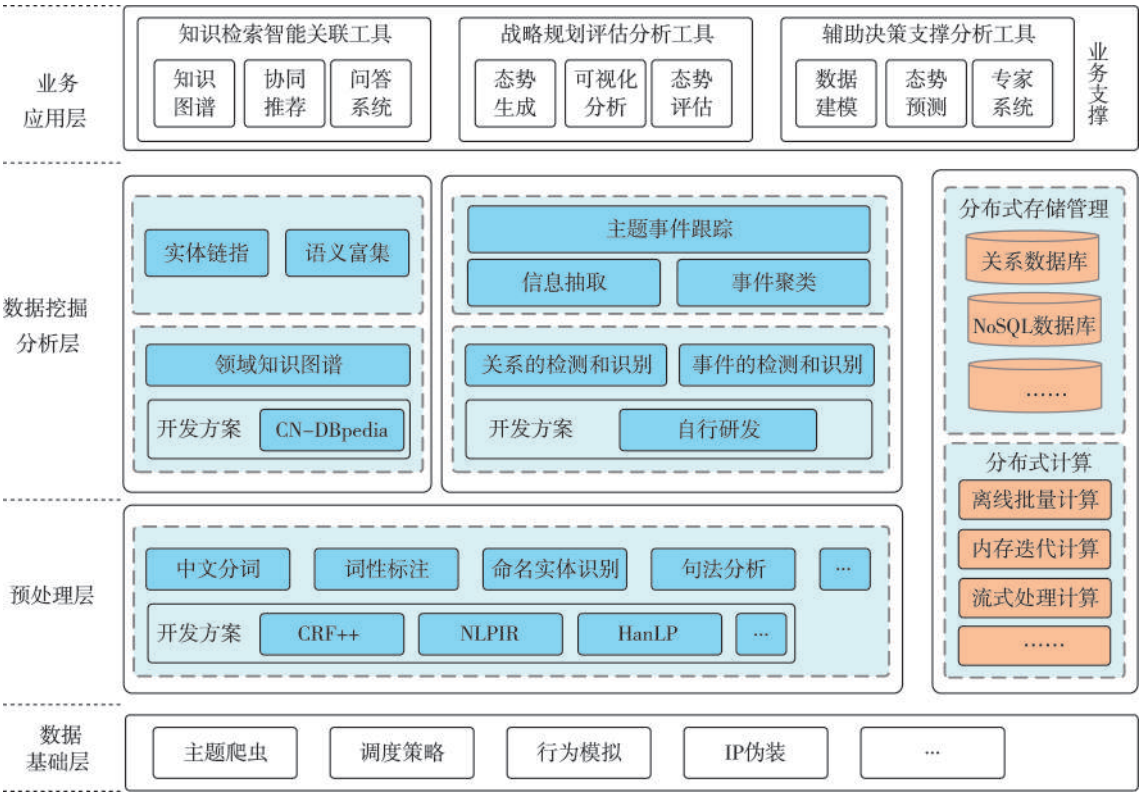

图 2 大数据挖掘分析系统架构图

的基于关键词的主题表示是最基本的表示方法之一。基于关键词的主题表示是指用一组代表主题特征的关键词集合来表示主题内容。本文是在爬虫爬行的过程中自动的扩充主题词,用以扩充主题库,完整的表达用户的需求,以求得更多的主题相关页面。

2.2 大数据挖掘分析

首先,针对 2.1 中已经获取的互联网文本数据,进行数据清洗,过滤垃圾信息。然后,针对过滤后的文本,进行数据预处理,主要包括命名实体识别、语义抽取、语义富集等操作,实现对文本数据的结构化预处理,丰富实体语义,为指挥决策提供领域知识支撑[13]。根据结构化的文本数据,构建完善领域知识库,并基于此实现更粗粒度的信息抽取,如事件检测、关系抽取等,为挖掘分析高层语义信息提供支撑。

为了跟踪研究互联网上与辅助指挥决策有关的信息,同时降低业务用户的信息过载压力,应当识别在整个网络上传播的基本单位,此单位的粒度应当介于一篇新闻报道与术语或主题词之间。此外,当确定了这个跟踪粒度之后,还要考虑到它在互联网传播的过程中动态变化的因素。因此,本文需要研究一种发现和追踪对同一事件不同表述形式的方法。

本文拟将互联网信息中的原文语句引用作为发现和跟踪事件的基本单位,并支持对其在互联网传播过程中的变异形式进行聚类识别[12]。具体做法是将原

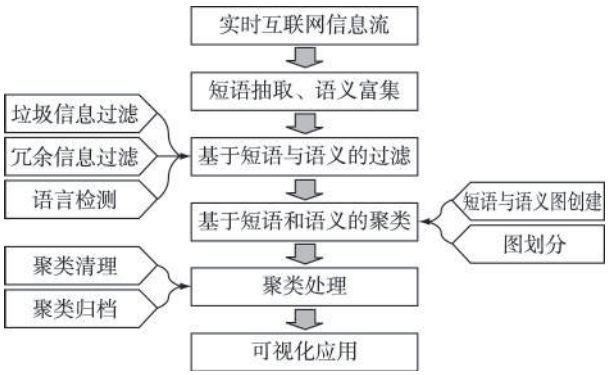

图 3 互联网文本大数据挖掘分析流程图

文中的短语或语义元素分割成若干个簇作为模块进行发现和跟踪,其中每个簇表示一个事件,并且簇中的短语或者语义可以是单个短语或者语义元素的突变变体,短语的变体形式可以包含部分词法变化或者部分增量信息,语义的元素的变体形式可以关联语义元素的形式出现。

为了支持以事件为导向的挖掘分析,系统通过图 3 所示步骤实现从互联网信息中检测事件并进行聚类跟踪。从短语抽取与语义富集开始,而后进行基于短语与语义的过滤,从而消除垃圾信息和冗余内容。增加过滤步骤是考虑到互联网开源情报常常包含广告、低质量信息的情况,一般系统常常忽视这个处理环节,但对于实现高质量产出来说这个处理至关重要。随后,

系统将实现基于短语与语义的聚类。由于在分析处理的过程中,描述同一事件的短语和语义可能经历了演进变化,而现有方法主要基于文本距离度量和传统聚类方法实现,因此常常效果不佳。解决方法是将描述同一事件的短语、语义信息以及他们的变体聚类在一个簇里,作为聚类分析的输出结果,可在简单处理后供信息检索或可视化分析使用。

聚类处理步骤执行的是输出最终质量检查,清理旧的聚类,将已经结束的聚类进行归档,并增量式地为已有聚类提供信息更新。最终,可视化应用使用最终的聚类存档通过可视化组件来呈现聚类。

3 定制化服务模式

微服务应用框架是实现轻量级的操作系统虚拟化解决方案,主要是以 Linux 容器(LXC) 等技术为基础,并在此基础上进行了封装,针对用户屏蔽掉容器的相关管理,使得操作更为简便方便,为用户提供一种类似于操作快速轻量级虚拟机似的体验。

微服务应用框架具备隔离性、资源可度量性等 2 个特性:

1) 隔离性。对于不同业务实例之间相互隔离,采用基于 LXC 的 Container 方式进行隔离。主要通过内核的 namespace 将进程、网络、文件系统等隔离开。

2) 资源可度量。cgroups(controlgroups) 是 Linux 内核提供的一种机制,该机制可以隔离、记录、限制进程组所使用物理资源,它提供了一种类似于文件形式的接口,通过将数据内容导入文件的形式实现资源控制度量。cgroups 可以实现对 bikio、CPU、cpuacct、cpuset、devices、freezer、memory、net_cls、ns 九大子系统的限制。

面向指控领域辅助决策的复杂使用需求以及互联网大数据的应用特点,通过容器的方式灵活配置需求、采集、处理、组织、计算、可视化等各类技术组件,进行面向业务的微服务组织与封装,并将生成的各类服务录入容器进行注册、调度与管理。通过容器在已有的服务组件中动态适配符合业务场景要求的微服务单元,形成细粒度、松耦合、可灵活组合的自治单元,为用户提供定制化的军政协同业务信息和应用服务模式,具体包括可视化组件重组与信息动态补偿等服务模式。

4 结束语

本文针对互联网大数据在指控领域辅助决策的应用,研究了使用需求分析建模、大规模互联网数据收集获取与挖掘分析,并基于定制化的服务模式为辅助决策应用提供支撑。下一步,我们将继续在数据挖掘分析方面进行应用拓展,增加信息提取样式,丰富态势信息展现,不断拓展其在作战指挥领域的应用。

参考文献:

[1] 孟小峰, 慈祥. 大数据管理: 概念、技术与挑战 [J]. 计算机研究与发展, 2013, 50(1) : 146-169.

[2] 覃春莲, 刘东波, 张红亮. 我军信息化建设中的软件工程与数据工程 [J]. 军队指挥自动化, 2004: 43-44.

[3] 程龙军. 面向大数据的指挥决策系统模型研究 [J]. 山西电子技术, 2015(1) : 85-87.

[4] 雷锟, 王劲松, 阳名喜. 大数据在信息作战指挥决策中的运用 [J]. 指挥控制与仿真, 2016, 38(3) : 24-27.

[5] 金芝. 基于本体的需求自动获取 [J]. 计算机学报, 2000, 23(5) : 486-492.

[6] Gruninger M, Fox M S. Methodology for the Design and Evaluation of Ontologies [J]. Workshop on Basic Ontological Issues in Knowledge Sharing, 1995.

[7] Fernandez M, Gomez-Perez A, Juristo N. METHONTOLOGY: From Ontological Arts Towards Ontological Engineering [J]. Proc. AAAI-97, 1997.

[8] Fensel D, Harmelen F V, Klein M, et al. On-To-Knowledge: Ontology-based Tools for Knowledge Management [J]. Ebusiness & Ework, 2000.

[9] Nicola A D, Missikoff M, Navigli R. A Proposal for a Unified Process for Ontology Building: UPON [C] //International Conference on Database and Expert Systems Applications. Springer-Verlag, 2005: 655-664.

[10] 汪涛, 樊孝忠. 主题爬虫的设计与实现 [J]. 计算机应用, 2004, 24(s1) : 270-272.

[11] 张航. 主题爬虫的实现及其关键技术研究 [D]. 武汉理工大学, 2010.

[12] 谭红叶. 中文事件抽取关键技术研究 [D]. 哈尔滨工业大学, 2008.

[13] 宋瑞亮. 面向军事领域的命名实体识别及相关信息提取关键技术研究 [D]. 哈尔滨工业大学, 2016.