● 由　凯，钟慧娟，游宏梁（中国国防科技信息中心，北京　100142）

面向情报工作的可视化分析系统研究

摘　要：当前情报研究工作所要处理的信息规模变得越来越庞大，信息结构变得越来越复杂，为让情报人员更好地对海量复杂信息进行有效分析，文章研究并设计了科技信息可视化分析系统框架。该系统采用基于组件的体系架构，使用统一的元数据标准和接口标准，融合了数据挖掘工具与可视化工具，具有较为完善的可视化分析功能。该系统能够有效提升信息挖掘的效能，帮助用户更好地理解数据，丰富情报产品的表现形式。

关键词：情报研究工作；可视化分析；框架系统

Abstract: In the field of intelligence research work, the information scale is becoming larger and larger and the structure is becoming more and more complex. In order to analyze the massive and complex information effectively, this paper researches and designs a scientific and technical information visualized analysis frame structure. This system adopts component-based architecture (CBA), uses unified metadata standard and interface standard, integrates the data mining tools and visualized tools, and contains perfect visualized analysis function. The system can effectively improve the efficiency of information mining, help users to better understand data, and enrich the forms of information products.

Keywords: intelligence research work; visualized analysis; frame structure

当前，科技情报工作面临着严重挑战，一方面，随着互联网、传感器等技术的飞速发展，信息的规模急剧扩增，由于缺乏有效分析手段，无法从大量的数据中提取出对研究人员有用的信息；另一方面，情报研究所要处理的信息结构也越来越复杂，高维数据与复杂关联信息层出不穷，远远超出了人脑分析解释这些数据的能力。在研究对象日益复杂的同时，社会各界对情报工作的要求却在逐渐提高。在企业界，为了抢占先机，需要情报人员必须在短时间内对某一技术作出准确评估。如何快速有效地对各种复杂数据进行有效分析，已成为情报工作亟待解决的问题。可视化技术是一种十分有效的数据理解手段，最早被应用于科学与工程计算领域，现已发展为一个十分热门的研究领域。简单地说，可视化就是将复杂的信息或规律以图形符号的形式表达出来，使人们能快速获取数据中所蕴含的关键信息[1]。可视化分析技术是可视化与数据挖掘、自然语言处理等分析技术的融合，将可视化分析技术应用在情报工作中，可以弥补传统方法中成果展示方式单一、研究手段固化不足等缺陷，对信息从一个全新的角度进行观察分析，发现以往的方法所不能发现的隐藏情报，并对其进行分析解释，得出有价值的结论，为决策提供有力支撑，从而大大提高情报分析的效率和效果。中国科学技术信息研究所梁战平先生更是认为可视化原理是情报研究的十大基本理论之一[2]。

1　可视化分析技术概述

可视化分析技术是建立在可视化与分析过程的基础上，以深入刻画数据特征和人类感知模式的能力为基础，可加强数据挖掘分析的能力和效果[3]。可视化分析技术是一个多学科交叉的领域，包含可视化、数据挖掘（文本挖掘、图挖掘等）、自然语言处理、机器学习等多个学科。对于情报研究人员而言，在使用各种挖掘分析技术时关注的是数据本身，而不是各种挖掘分析算法，但现有的一些挖掘技术和算法往往过于复杂，难以理解和使用。可视化使分析过程和结果更加直观，让人更容易理解，为情报人员更好地使用各种挖掘分析技术提供了便利。

当前，可视化分析作为一个新的研究领域，吸引了越来越多的学者对其进行研究。从谷歌趋势（见图1）中可以发现可视化分析最开始引起人们注意是在2005年，2012年至今对其关注度逐渐提高。IEEE从2006年开始，每年都要举办“可视化分析科学与技术”（IEEE Visual Analytics Science and Technology，VAST）会议，从历年VAST会议上的论文关键词统计来看，可视化分析的研究热点主要有：图数据（Graph）可视化分析、文本（Text）可视化分析、高维数据可视化分析、地理及时空信息分析，以及可视化分析过程等。

在情报研究领域，也有不少学者开始用可视化分析技术来解决研究过程中遇到的问题，比较有代表性有社会网

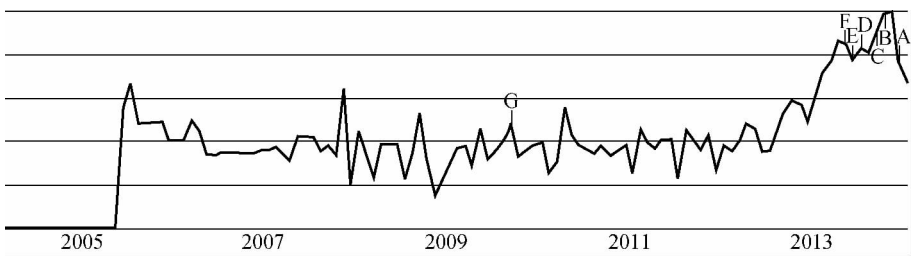

图 1　谷歌搜索中“可视化分析”的变化趋势

络分析[4]和信息检索可视化[5]。但可视化分析技术所解决的问题还比较局限，可视化与信息分析的结合也比较松散。这是因为现有的可视化分析工具还不足以为情报研究提供有力支持，常用的几种工具如 Pajek，CiteSpace，SPSS 等各有其局限性，如 Pajek 只支持对图数据的分析，CiteSpace 面向的是科学引文数据分析，SPSS 中虽然集成了多种数据挖掘工具，但其可视化展示手段单一。

2　可视化分析系统架构

情报研究中的可视化分析系统主要是为了支持情报人员更好、更快地分析理解各种数据和信息。一方面，要有信息分析挖掘功能；另一方面，要实现对信息及挖掘结果和挖掘过程的可视化展示，增强对研究成果的理解和研究过程的可控性。根据以上需求，采用自底向上的设计思路，系统应当主要包括可视化工具和信息挖掘分析工具。系统的整体框架见图 2。

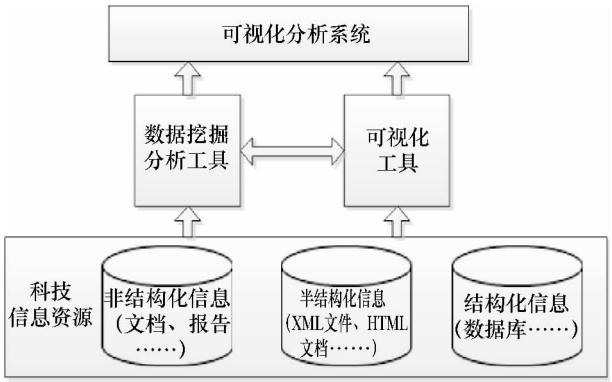

图 2　情报研究中的可视化分析系统框架

系统采用基于组件的体系架构，数据挖掘工具与可视化工具都是由众多的功能插件构成。可视化分析系统本身作为软件平台，定义数据标准及交互标准等，不同的组件采用平台统一标准，确保程序的兼容性。用户根据需要在系统中选用相应的功能组件来辅助自己进行情报研究。下面来分析一下系统实现中所需解决的关键问题。

2.1　技术标准

现今，可视化技术与数据挖掘技术都是非常活跃的研究领域，新的技术和算法层出不穷，为确保系统可以较快地融合新的研究成果，平台必须具有良好的可扩展性。同时，也要保证不同的插件之间具有可交互性，以使多个插件可以协同使用。这些都需要一个统一的技术标准才能实现，本文在已有的数据挖掘技术标准[6]基础上，分析了可视化分析系统中所应具有的技术标准。

可视化分析系统中集成了多种不同的数据集，这些数据集最后又汇总到数据仓库中，为分析决策提供支撑，在此过程中要有一个数据传输标准对数据传输格式等进行规范。为方便平台对数据进行统一管理与调用，需要采用同一个元数据标准。数据挖掘工具与可视化工具通过统一的数据库驱动程序来存取仓库中的数据，用户可以通过相应的接口调用这些服务，接口由接口标准进行规范（见图 3）。

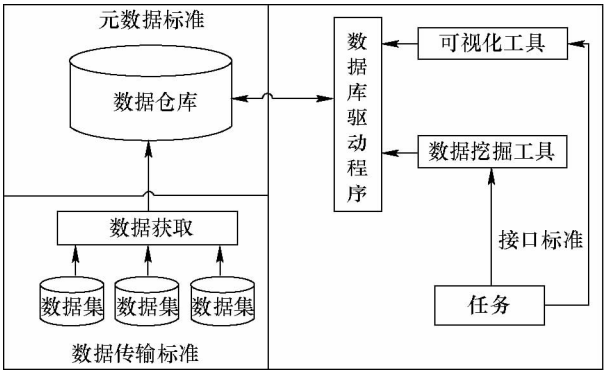

图 3　平台标准示意图

2.2　可视化工具

可视化可以看作是从数据到可视化形式再到人感知系统的可调节的映射。如图 4 所示是这些映射的简单描述。在该模型中，从原始数据到人的感知要经过如下变换：数据提取、可视化映射和视图转换[7]。数据提取是对原始数据进行提炼，挖掘出其中需要可视化的内容；可视化映射是对数据结构进行变换，将原有数据的结构转变为可视化结构（结合了空间基、标记和图形属性的结构）；视图转换是将转变后的数据通过定义位置、缩放比例、裁剪等图形参数创建可视化结构的视图。可视化的 3 个步骤都是紧紧围绕任务和数据特征开展的，需要人的控制和指导（见图 4）。

情报研究所要面临的数据类型是多种多样的，有非结构化的文本信息、半结构化的网页信息以及数据库中的结

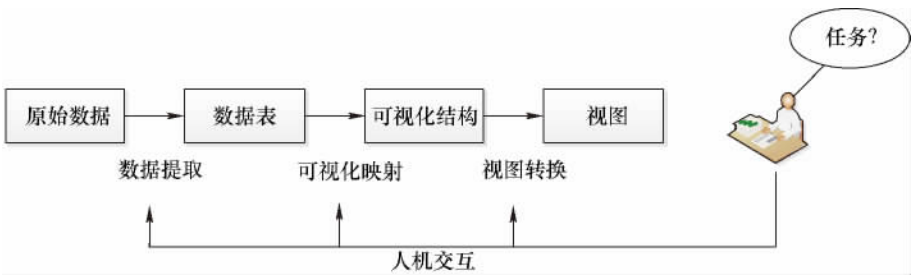

图4 可视化概念图

构化信息等。可视化工具要尽可能地涵盖这些不同类型的信息，也要较好地融合新的可视化技术。可根据当前的研究热点将可视化工具分为如下几类：基础可视化工具、文本可视化工具、高维数据可视化工具、时序数据可视化工具及层次数据可视化工具等。

1) 基础可视化工具针对的是一些简单的低维数据的可视化，一般通过概率统计等方法来总结出数据分布规律及变化规律等。其具体的展现形式一般为折线图、柱状图及圆饼图等。

2) 文本可视化是一种对文本内容、文本内部结构、文本间关系等信息进行可视化的技术[9]，以下是几种常见的文本可视化方法。标签云（Tag Cloud）将文章的关键词按一定顺序和规律排列，如频度递减等，并用文字的大小来表示词的重要程度，通过标签云可以快速直观地了解文章的主要内容[8]；Phrase Net 以网络图的形式展示了文章中命名实体的多种关系，如从属关系、并列关系等，它深入描述了文本的语义层次结构[8]；也有用有向网络图来表示文本之间的应用关系，其中每个节点表示文章，有向线段表示文本之间的引用关系[4,10]。

3) 高维信息可视化是要设计合适的可视化展现形式，将高维空间信息映射到三维或二维可视空间中，从而实现对高维信息的可视化[11]。目前针对高维信息的可视化方法主要分为如下几类基于几何的技术、基于图标的技术和基于降维映射的技术等。比较有代表性的方法有平行坐标系法[12]、等距映射[11]及散点图矩阵[13]等。

4) 时序数据可视化主要是指对具有较好时间连续性的数据进行可视化展现的技术，针对时序数据的展现形式主要解决如下3个关键问题[14]：数据是按线性时间还是周期时间排列；关注的是时间点还是时间间隔；数据序列是只有一条时间主线（序列时间）还是多条主线（多分支时间）。针对此时序数据的具体表现形式有螺旋图、TimeWheel 以及 ThemeRiver 等。

5) 层次数据主要包含两类信息，一是结构信息，二是内容信息[15]，结构信息主要由层次结构的上下级关系确定，内容信息主要存储在图形的节点中。针对可视化图形的空间特征，可以分为2D图、3D图和混合图，具体的表现形式有双曲树、径向树、圆锥树等。

2.3 数据挖掘工具

数据挖掘是指通过分析大量的数据，来揭示数据内在的规律和联系[16]。一般来说，数据挖掘主要包括如下过程：数据选择、数据预处理、数据转换、数据挖掘和解释分析。前面3个步骤可以看作是数据准备阶段，为挖掘任务和挖掘算法提供适合有效的数据。数据挖掘是选择正确的模型算法，对待处理数据进行分析的过程。相应的数据挖掘工具中主要包含这些功能。

由于情报研究所需处理的数据有多种类型，为尽可能地满足研究需求，挖掘工具中要集成多种分析工具。如文本分析、数据挖掘等。文本挖掘当中主要包括文本词法、句法的分析工具、信息抽取工具和文本检索工具等。数据挖掘技术主要包括信息分类、聚类、回归分析、关联发现、序列预测及奇异值探测等工具。这里每种工具的实现都可以采用多种算法，例如分类工具中有贝叶斯算法、人工神经网络、最大熵等。当前新的算法还在不断出现，每个工具也都可以不断地扩充。但在使用时要注意，每种算法都有自己所适合的使用范围和领域，针对具体问题要经过仔细分析，选用合适的算法工具来实现。当然，也可以同时使用多种工具，选取其中评估得分最高的结果。另外，根据实际问题，有时需要多种工具协同使用，如对中文文本关键词提取的任务中，就需要分词、词法分析、TFIDF 值等多个功能组件的共同使用。

3 可视化分析流程

可视化分析的流程一般为数据准备、数据挖掘、结果分析和可视化展现这4个部分，目的是通过大量的数据分析获取未知的、有效的、实用的信息，并通过这些信息获得所需要的知识或作出决策。

数据准备包括数据选择、预处理和数据变换，其根本目的是根据分析任务筛选出可用的数据；数据挖掘阶段首先要明确分析的目的，如分类、趋势分析等。确定了挖掘目的后，就需要决定本次任务的技术和算法，同样的任务可以用不同的技术和算法来实现，有时也要多种技术综合使用；数据挖掘阶段发现出来的信息和模式，经过人的分析和归纳后就形成有用的知识；在整个挖掘任务中，用户都可以采用可视化工具动态地监测到数据的变化（见图5）。过程中各步骤的大体内容如下。

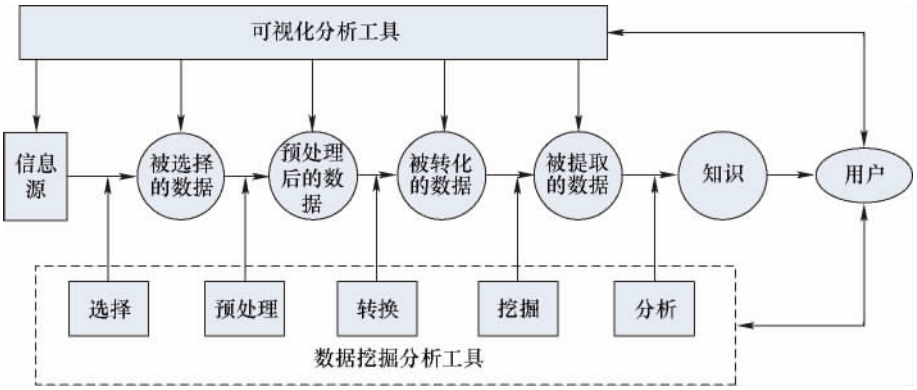

图5　可视化分析流程

1) 确定研究对象、明确研究任务: 分析实际需要，科学定义研究问题，明确任务。进行可视化分析之前要理清所要探索的问题，只有问题明确，才能进行有效分析。虽然最后结果是不可预测的，但要探索的问题应是有预见的。

2) 数据选择: 从数据源中搜索所有与本次任务相关的数据信息，并从中选择出适用于数据挖掘应用的数据，这里主要用到了文本检索、数据库查询等工具。

3) 数据预处理: 研究数据的质量，过滤噪音，为进一步分析作准备。确定将要进行的挖掘操作的类型，以及所要采用的挖掘技术和算法。

4) 数据转换: 将数据转换成一个分析模型。这个分析模型是针对挖掘算法建立的。

5) 数据挖掘: 对所得到的经过转换的数据进行挖掘。这一步在确定挖掘工具和输入输出文件后，所有工作都能自动地完成。

6) 结果分析: 解释并分析结果。使用的分析方法一般应根据数据挖掘操作而定。

7) 可视化交互: 分析对象特征，采用适当的可视化工具。对挖掘过程中产生的数据进行可视化，用户据此来判断下一步工作。不同的信息类型，要采用不同的可视化方法，只有方法正确才能有效地揭示信息内在的特征和规律。该步骤应该贯穿整个挖掘流程中，时时地纠正挖掘中产生的偏差。

可视化分析是数据挖掘技术和可视化技术的有机结合。这种结合强调的是以人为中心，一方面充分利用人类的知识领域和模式感知能力，另一方面也促进了用户对挖掘结果的理解和利用。可视化的方法使数据挖掘技术的应用更具形象性和直观性，挖掘的过程加入更多人的参与和指导，可以提高数据挖掘的效率和可靠性，也保证了数据挖掘结果的可信度、可理解性和可用性。

4　可视化分析实例

采用可视化分析系统中的工具按上面的分析流程，对简氏（JANE'S）年鉴中的装甲与火炮系统年鉴中的数据进行分析。该年鉴提供了全球目前在用或在研的装甲与火炮系统的重要信息，包括装备发展动态、生产商、研发国家等。

本次主要对装备的生产商和研发国家之间的合作信息来分析不同国家在装甲与火炮系统方面的发展情况。简氏年鉴中，每种型号的装备信息都由一个 HTML 文件存储。采用 Web 信息抽取工具，提取每个文件中“Country”与“Company”标签中的数据，初步统计有 1013 条数据，去除其中有缺漏的信息后得到 848 条。分别对“国家”和“公司”数据进行统计，发现共涉及 54 个国家和 236 个公司。以“国家”信息为例，采用标签云对其进行可视化展示（见图 6），图中字体大小表示了该国家装甲与火炮系统型号的数量多少。

图6　“国家”信息

从图 6 中可以发现，中国、美国、英国、法国、俄罗斯、南非、德国等国家的出现频次最高。这些国家中前 5 个都是典型的军事强国，德国与南非的总体军事实力相比这 5 个国家较弱，但是其装甲与火炮系统在世界上占有重要地位，其典型型号综合水平较为突出。比这些国家稍逊色的有以色列、瑞典、土耳其、瑞士、意大利等国家。

针对国家与公司之间的合同关系，绘制二者的共现网络图（见图 7）。图中深色的点表示国家，浅色的点表示公司，曲线代表国家与公司之间的合同关系。从中可以发现相比于中、俄两国的封闭式发展，美、英、南非、意大利和德国等选择的公司都比较国际化，很多公司都曾为不同的国家研发过武器，特别是英国 BAE 系统公司，它是世界第三大军火公司，在美国及欧洲国家的地面武器系统方面占有很大的份额。

从该实例可发现，可视化分析系统在实际应用中具有较好的实用性，结果更为直观，用户可以方便地从分析图中得出结论。

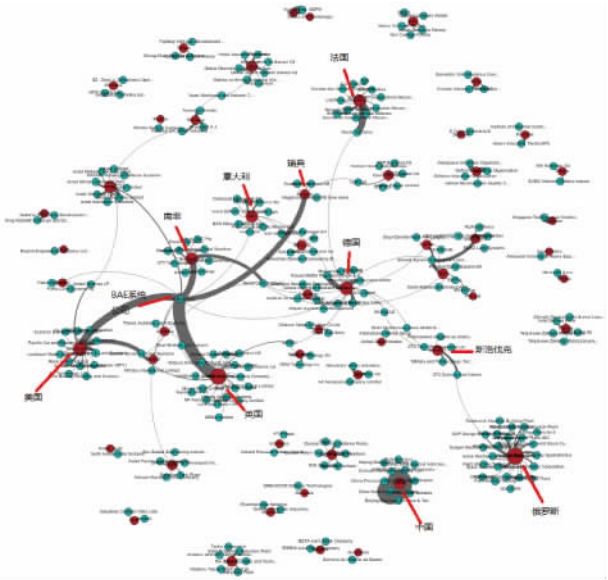

图7　国家与公司之间合同关系网络图

5　结束语

情报研究的可视化分析系统综合集成了可视化技术与数据挖掘分析技术，并采用基于组件的平台架构，具有通用的数据存取和工具使用接口，可以方便地对功能进行增删和更改，也便于用户使用。可视化分析系统对情报研究工作将有如下三方面促进作用。

1）提升科技信息挖掘效能。信息挖掘是解决当下所面临的信息爆炸而知识贫乏问题的一种有效方法，但是其算法的复杂性往往令情报研究人员无所适从。把可视化技术引入到信息挖掘中，充分利用可视化技术的直观性，使科研人员更加容易地参与到信息挖掘的过程中，从而提高信息挖掘的效能。

2）增强用户对科技信息的理解。可视化技术的研究与心理学的认知理论密切相关，可以有效增强人类对于信息的认知能力，更好地理解海量信息之间的潜在联系，发现深层次的具有重要意义的知识。

3）提升科技信息分析服务手段。科技信息研究成果通常以研究报告的形式呈现，难以给科研人员直观、生动地展示，可视化分析技术作为科技信息研究成果的表达手段，可以根据研究成果对其内在推理过程、核心结论及关键数据进行展示，从而提高科技信息分析服务手段。□

参考文献

[1] 唐家渝，刘知远，孙茂松．文本可视化研究综述 [J]．计算机辅助设计与图形学学报，2013，25（3）：273-285.

[2] 梁战平．我国科技情报研究的探索与发展 [J]．情报探索，2007（7）：3-7.

[3] 耿学华，傅德胜．可视化数据挖掘技术研究 [J]．计算机应用与软件，2006，23（2）：85-87.

[4] 周金侠．基于 Citespace Ⅱ 的信息可视化文献的量化分析 [J]．情报科学，2011，29（1）：98-101.

[5] 陈艳．信息检索可视化技术 [J]．情报理论与实践，2006（5）：618-621.

[6] 刘明亮，李雄飞，孙涛，等．数据挖掘技术标准综述 [J]．计算机科学，2008，35（6）：5-10.

[7] CARD S，MACKINLAY J，SHNEIDERMAN B. Readings in information visualization：using vision to think [M]．San Francisco：Morgan Kaufmann，1999.

[8] VIEGAS F B，WATTENBERG M. TIMELINES：tag clouds and the case for vernacular visualization [J]．Interactions，2008，15（4）：49-52.

[9] VAN HAM F，WATTENBERG M，VIEGAS F B. Mapping text with phrase nets [J]．IEEE Transaction on Visualization and Computer Graphics，2009，15（6）：1169-1176.

[10] CHEN C. CiteSpace II：detecting and visualizing emerging trends and transient patterns in scientific literature [J]．Journal of the American Society for Information Science and Technology，2006，57（3）：359-377.

[11] 孙扬，封孝生，唐九阳，等．多维可视化技术综述 [J]．计算机科学，2008，35（11）：1-7.

[12] KEIM D A，ANKERST M. Visual data mining and exploration of large databases [C]．PKDD. Freiburg，Germany，2001.

[13] NIST/SEMATECH e-handbook of statistical methods [EB/OL]．[2012-04-06]．http：//www.itl.nist.Gov/div898/handbook.

[14] AIGNER W，MIKSCH S，MULLER W，et al. Visual methods for analyzing time-oriented data [J]．Visualization and Computer Graphics，IEEE Transactions on，2008，14（1）：47-60.

[15] 肖卫东，孙扬，赵翔，等．层次信息可视化技术研究综述 [J]．小型微型计算机系统，2011，32（1）：137-146.

[16] 王光宏，蒋平．数据挖掘综述 [J]．同济大学学报：自然科学版，2004（2）：32.

作者简介：由凯，男，1989 年生，硕士生。
钟慧娟，女，1981 年生，助理研究员。
游宏梁，男，1971 年生，高级工程师。

收稿日期：2014－05－19