CORE
Metadata, citation and similar papers at core.ac.uk
Provided by National Science Library,Chinese Academy of Sciences

第26卷第5期
2008 年 5月

情 报 科 学

Vol. 26 ,No. 5
May ,2008

情报分析中五项新技术的应用解析

李 娜[1,3]，吴清强[1,2]，侯 丽[1,2]

(1. 中国科学院 研究生院，北京 100049；2. 中国科学院 国家科学图书馆，北京 100080；
3. 中国科学院 国家科学图书馆成都分馆，四川 成都 610041)

摘 要：随着全球信息化的进程，情报研究的任务和方式都发生了重大的变革。情报研究当中的重要环节情报分析，也在计算机技术的日益发展和应用的不断深入当中不断变化着。本文主要从五个方向的技术：基于网络的技术、信息抽取技术、语义网络技术、数据挖掘技术和信息可视化技术，应用到情报分析当中的情况来探讨情报分析工作的变化和现状，从中发掘出需要进一步解决的问题。

关键词：情报分析；基于网络的技术；信息抽取；语义网络；数据挖掘；信息可视化
中图分类号：G350.7 **文献标识码**：A **文章编号**：1007 - 7634(2008)05 - 0683 - 05

The Application of Five New Technologies in Intelligence Analysis

LI Na[1,3]，*WU Qing - qiang*[1,2]，*HOU Li*[1,2]

(1. *Graduate School of Chinese Academy of Sciences*, *Beijing* 100049, *China*;
2. *National Science Library of Chinese Academy of Sciences*, *Beijing* 100080, *China*;
3. *Chengdu Departments*, *National Science Library of Chinese Academy of Science*, *Chengdu* 610041, *China*)

Abstract: Along with the tenor of global informationization, the task and method of intelligence investigation has transformed gravely. Intelligence analysis, which is a crucial tache in the process of intelligence investigation, has been moving with the development of computer technologies and its adhibition. This paper talks about five aspects of technologies: technology based on the web, information extraction, semantic web, data mining, and information visualization. Lay hands on the application of these techniques in the course of intelligence analyses and find some issues that could be raveled out.

Key words: intelligence analysis; technology based on the web; information extraction; semantic web; data mining; information visualization

1 引 言

计算机和网络技术的不断发展，推动着全球信息化的进程，当今社会，信息资源早已不仅限于传统纸质的图书、期刊、会议记录和专利等，网络信息和各种数字化资源已经成为信息资源最重要的组成部分之一。同时，信息技术的快速发展，为获取和处理各类信息提供了新的方法和工具。

情报分析是情报研究当中的一个重要环节，传统的分析方法，大多是采取人工方式，然而，网络时代的到来，可以获取的信息资源越来越多，信息

收稿日期：2007 - 11 - 06

作者简介：李 娜（1984 - ），女，湖北武汉人，硕士研究生，从事情报分析技术与方法、信息服务与用户研究；吴清强（1974 - ），男，福建福清人，博士研究生，从事数据挖掘与情报分析技术研究；侯 丽（1978 - ），女，湖北荆门人，博士研究生，从事情报研究和竞争情报研究.

© 1994-2008 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

量几何级数增加，怎样从海量信息当中获取有价值的情报，对获取到的巨量信息进行分析，以及分析结果的明晰表达，都是急迫需要解决的问题。面对数量庞大和形式复杂的信息资源，已经有越来越多的研究，将新的计算机技术应用到情报分析当中，具体来说，主要有以下五个方向：基于网络的技术、信息抽取技术、语义网络技术、数据挖掘技术和信息可视化技术。

2 新技术的应用

信息技术的发展，是现代社会变革的主要驱动力，对社会和经济产生了根本性的影响，同时，这也是情报分析发展的推力。下面就对应用到情报分析当中的五个方面的计算机技术进行具体的说明。

2.1 基于网络的技术

环境监控和对手分析是情报分析工作当中的重要部分。随着网络的发展，丰富的网络资源已经成为最主要信息源之一。因特网的信息量大，而且信息的传播和更新快，许多信息都可以通过网站获得[1]。如果采用传统的人工方式搜集网络信息，工作效率低下而且具有很大的随意性。现在可以利用一些联机服务系统跟踪和监测特定的目标网站，通过对网站信息的挖掘来获得有用的情报[2]。

目前，因特网上的信息大部分是以超文本的形式存储，通过超链接提供服务，因此，除了跟踪监测网页内容的变化，还可以使用 Robot 程序，沿着网页中的链接自动漫游和下载，从而完成自动获取网络信息。获得相关信息以后，可以运用 WEB 文本分类技术对采集结果进行处理[3]。从网络上获取的信息不仅有文字，一般还有声音、图片和视频等多媒体信息，其中，最便于利用的是文本信息，因此，在进行 WEB 文本分类之前一般剔除非文本信息，并且对照停用词表，去除一些虚词和介词等没有具体含义的词。然后，对 WEB 文档进行词法分析和词条分割，如果是英文文档还需要进行词干抽取。接下来，通常采用向量空间模型（VSM，Vector Space Model），进行 WEB 文档的特征提取，在这个模型中，特征项由字、词或短语构成，通过相似度计算找到所有特征项，由它们组成特征项集[4]。在此基础之上，通过文本分类的算法，比如 K 最近邻分类算法和贝叶斯分类算法，将上面的处理结果归到一个或多个主题类别，加工整理后归档，将符合用户需求的结果提交给用户，其他内容可以保存下来，作为长期跟踪的材料[5]。

现在，国内外已经有一些基于网络技术的情报分析工具，比如 ChangDetect（http：//www.changedetect.com/）提供网页内容监视的服务。注册用户可以监测一个网页，也可以监测一组网页，监测的网页内容发生变化时，系统就会自动发送电子邮件通知用户。中国网络情报中心（http：//chinawi.tixa.com/index.html）提供基于网络技术的多种情报分析工具。天下通企业情报门户系统，根据企业用户的情报定制，提供情报监测和情报分析服务。它可以实时全面监测指定网络信息源，并且定向发送至制定人员[6]。天下通专业网媒监测，集成情报监测、管理、分析、统计和通知为一体，可持续获取全面及时的目标信息，并且将监测到的信息进行系统过滤筛选、实时匹配、编辑排重等处理，然后按照用户定制的时间和电子邮件发送给用户[7]。

2.2 信息抽取技术

面对海量信息，情报分析人员需要一些自动化的工具来获取需要的情报，在这样的背景下，出现了信息抽取技术的研究。信息抽取是面向结构化、半结构化和非结构化文本所进行的文本理解技术，其定义为从一段文本中抽取指定的一类信息并将其形成结构化的数据填入到特定数据库中供用户查询使用的过程。它从文本中抽取用户感兴趣的事件、实体和关系，然后存入数据库中进行分析，给出文摘或提供在线服务。也有学者认为，信息抽取可以看作信息检索的进一步深化，研究指定信息的查找、理解和抽取，并将指定信息以适当的方式输出[8]。

情报分析需要的是智能化信息处理技术，来解决信息过载的问题。信息抽取作为智能化信息处理的前沿技术，可以嵌入到情报分析系统当中来发挥作用。基于信息获取的情报分析系统中，信息抽取技术可以进行概念描述、关联分析、分类和聚类等功能，从而实现信息的智能化分析。具体来说，在对获取到的各种信息进行预处理以后，通过对文本信息进行语义分析，可以获得预定主题的相关信息，然后从中抽取出相关的特征项，并将处理后的结构化文本信息存入数据库中，再进行下一步的各种分析，最终得到用户需要的结果[9]。

国内外都有研究将网页作为信息抽取的对象，

© 1994-2008 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

进行情报分析。网页信息抽取的工作原理有以下几类：利用网站查询表格；基于归纳学习；基于网页结构分析、基于隐式马尔科夫模型和基于模式匹配[10]。从网页当中提取信息的技术原理和流程如下：首先，采集大量类似于指定目标的网页信息，进行过滤处理，以文本形式保存相关信息以待抽取；同时，依据特定目标的特点结合样本网页，提炼出抽取模型和算法，确定需要抽取的特征信息；最后，在处理过的采集信息当中抽取出相关特征信息项，并且以统一的格式保存于数据库中[11]。

2.3　语义网络技术

在越来越复杂的信息环境当中，用计算机技术实现数量化方法进行信息的分析和预测已经得到广泛应用。其中，语义网络技术是基于网络结构的一种知识的图解表示，以网络形式实现知识语义结构使之能够通过多种机制来表达概念、规则及其之间的关联知识[12]。语义网络作为一个带有标识的有向图，其节点表示各种事物、概念、属性和知识实体，链则表示所连接节点之间的各种语义关联[13]。

国外对语义网络技术的研究很多，其中有学者设计出算法能够从语义网络中半自动地获取知识，并且开发了一组基于JAVA的API来提供一些基础的服务。通过实验验证了这组工具能够作为一个通用的工具箱应用到自然语言处理当中[14]。

国内有学者受到狄克问题求解的启发，设计出定量化情报分析的数学模型——三因素（三方）二分（两种情况）网络，并且推广到m因素n分网络及k值模型。编制软件实现了三因素和四因素二分模型进行信息分析，通过应用“某空军夜行团训练方式”的实例验证了语义网络模型进行情报分析的可行性。实例说明依靠计算机软件来辅助定量化情报分析，比传统方法的效率和可靠性要高出很多[15]。

2.4　数据挖掘技术

数字时代，情报分析的新使命是利用基于知识的新方法和新技术，为用户服务。数据挖掘技术，是基于数据仓库发展起来的一种知识发现技术，它是一个从大量的、不完全的、有噪声的、模糊的、随机的实际应用数据中，提取隐含在其中的、人们事先不知道的但又是潜在有用的信息和知识的过程[16]。经过几年的发展，自然语言处理、语义关联分析、词频分布统计和语料学研究等可以用于情报分析的技术方法，已经成为数据挖掘的重要研究方向，并且有一些成型的软件工具[17]。一般情报分析系统的功能模块包括收集、分析和发布等几个部分。而常用的数据挖掘工具是由数据收集、处理和结果输出等几个功能模块组成。因此数据挖掘技术完全可以应用到情报分析工作当中。

国外有研究将文本数据挖掘技术应用到情报分析当中，开发出技术情报的问答系统，其中包含一个应用于期刊数据库的文本挖掘工具。这个系统可以按照问卷表达的信息需求做出趋势分析。另外，还利用多种不同的聚类技术识别出对于决策者来说重要的信息[18]。

国内已经有学者将数据挖掘应用到专利情报分析当中。专利情报分析的研究对象是专利数据，将专利数据的技术内容进行数据化和集成化之后，就可以运用数据挖掘算法对其进行分析并且识别出有用的知识。具体的分析流程包括：专利数据的获取、数据预处理以及数据分析和报告。数据挖掘技术能够分析特定的情报，并且各种单一的情报综合起来。然后采取统计分析、技术群组、文本挖掘、组合理论、专利地图等技术，对其进行分析，然后以统计图谱、关联图谱和报告等形式展现出来[19]。

2.5　信息可视化技术

传统的情报分析系统提供给用户的大多是文本或者数值数据，千篇一律的文字或者数字让人很难从中一眼识别出有用的信息。除了需要前面几个技术对信息进行分析以外，情报分析人员还需要通过一些形象化的方法来处理和发布信息，信息可视化技术就是这样一种技术。它首先对信息进行分析和提取，然后以各种图像形式展现出原始数据之间的关系和发展趋势[20]。这种形象化的表达方式可以让情报分析人员更好的发掘和利用信息资源。

信息可视化技术的显示对象一般是多维的标量数据，其本质是将抽象的数据转化成为形象化的可视结构。具体的实现方式可以通过OpenGL或者Java3D等从底层编程实现，也可以使用可视化工具进行实现。现有的数据管理工具中，越来越多的软件都集成了可视化功能，比如常用的Microsoft Excel就可以将表格当中的数据制图显示出来。还有包含可以直接使用的可视化工具，或者可以进行二次开发的可视化环境，比如AVS Express和OpenDX，以及一些可以用来开发可视化工具的组件，比如VTK和OpenViz等[21]。

© 1994-2008 China Academic Journal Electronic Publishing House. All rights reserved.　http://www.cnki.net

互联网上的信息量巨大，用户可以通过浏览器浏览或者关键词搜索来找寻信息。然而，这两种方式都费时费力，而且，不一定能够得到满意的结果。国外有研究将信息可视化技术和信息检索技术融合应用到万维网上的知识发现。在这个研究当中，网站的结构以三维的双曲线树来表示，每一个网页的节点高度由计算过的与用户兴趣的相关度决定。这些功能是嵌套在浏览器中实现的。可以帮助用户在大型网站当中抓取出最相关的网页信息[22]。

国内也有研究将信息可视化技术应用到情报分析当中，针对两个时段相关媒体的报道，利用河流模型与关联分析模型进行对比分析。实验得出的结论是，利用可视化的分析方法，不仅可以帮助研究人员从大量的文档集合当中提取出主要因素，而且还能够利用这些主要因素研究事件的发展过程以及可能的趋势，同时为文档的整理和分类提供帮助[23]。

3 新技术应用的优势

以上5个方向的计算机技术使得以下情报分析问题得以有效解决。

(1) 海量网络信息的监测和挖掘。网络的发展让情报分析所需要的信息资源发生了巨大的变化，在复杂的信息环境当中，基于网络的技术使得现代信息资源当中最丰富，并且也是最重要的网络信息能够更加便利和高效地得到利用。

(2) 信息的智能化分析。信息抽取技术应用到情报分析当中，可以帮助对信息的智能化分析，同时从一定程度上来解决信息过载的问题。嵌入了信息抽取技术的情报分析系统，比传统的情报分析系统具有更强的分析和处理能力。

(3) 基于知识语义结构的推理。用语义网络的数学模型来进行情报的定量分析，可以处理复杂的数据。它的联想性和高效性，适用于情报分析当中的数据和事实的推理，而且提供了比较客观可靠的数据。

(4) 知识发现。数据挖掘技术可以从大量的实际应用数据中，找到其中的规律，提取出隐含在其中潜在有用的知识。

(5) 分析结果的有效解读。信息可视化技术使得情报分析的结果能够直观明了地表述，并且展现出数字和文本无法凸现的隐含信息、潜在的对象以及对象间关系。

4 尚需解决的问题

尽管新技术的应用有效解决了情报分析当中存在的一些问题，但是下列情况，还有待进一步研究。

(1) 文本信息是最容易获取和处理的信息资源，虽然非文本信息的获取处理起来难度会大一些，但也是不容忽视的，其中蕴含着许多有价值的情报。从作者的角度来说，由于图像图表甚至声音，都比文字表达的信息更完全和清晰。因此，将多媒体信息的处理技术应用到情报分析当中，将会使得信息的获取更加完整，同时分析处理功能也更加强大。

(2) 以上提到五个方向的计算机技术都已经应用到了情报分析当中，然而，有很多都只是试探性的研究，这五个方面的计算机技术，包括了情报研究当中，信息的收集、分析处理和结果表达的全过程，如果能够将这五个方面的技术当中已经成熟的算法和软件工具集成在一起，组成一个完整的情报分析平台，那么，将会让情报分析工作的效率得到很大的提高，并且得到更好的效果。

5 结 语

本文主要讨论了现代互联网环境当中信息资源发生了巨大变化的背景下，五个方向的计算机技术在情报分析工作当中的应用。在新的信息环境当中，信息量越来越大，信息的形式多样而复杂，与此同时，信息的获取方式更多更便利。本文正是基于这样一种变化趋势，分析了互联网环境下的五项计算机技术在国内外的研究与应用情况，总结了这些技术已经解决的情报分析问题，以及尚需解决的问题。每一项技术都有其特点，本文对各项技术的适用情况尚未深入研究。而且，各项技术只能解决各自对应的特定问题，怎样融合这些先进技术在情报分析系统当中，是后续的研究目标。

参考文献

1 鞠英杰. 网络竞争情报研究 ——竞争者网站的挖掘与监测[J]. 情报理论与实践,2005,(2):215 - 218.

2 Aleksander Kołcz, Abdur Chowdhury, Joshua Alspector. Improved robustness of signature - based near - replica detection via lexicon randomization[A]. Proceedings of the tenth ACM SIGKDD

© 1994-2008 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

international conference on Knowledge discovery and data mining [C]. New York ,USA :ACM Press. 2004 :605 - 610.

3 Jian - Chao Xu ,Da - You Liu ,Ming Hu. Feature selection and text classification for Chinese Web documents[A]. Proceedings of the Third International Conference on Machine Learning and Cybernetics[C]. Piscataway ,NJ ,USA : IEEE. 2004 ,(2) :1304 - 1309.

4 Longzhuang Li , Yi Shang , Wei Zhang. Improvement of HITS - based algorithms on web documents[A]. Proceedings of the 11th international conference World Wide Web[C]. New York ,USA : ACM Press. 2002 :527 - 535.

5 薛燕波. WEB 文本分类技术在企业竞争情报分析中的应用[J]. 情报科学 ,2004 ,(3) :378 - 381.

6 天下通企业情报门户系统(CIPS)产品介绍[EB/OL]. http ://chinawi. tixa. com/cpzx/cips. htm ,2007 - 09 - 11.

7 天下通专业网媒监测介绍[EB/OL]. http ://chinawi. tixa. com/cpzx/wmjc. htm ,2007 - 09 - 11.

8 余　丰 ,朱东华. 信息抽取技术在竞争情报研究中的作用[J]. 情报杂志 ,2006 ,(3) :25 - 27.

9 Maxime Morneau , Guy W. Mineau , PDan Corbett. SeseiOnto : Interfacing NLP and Ontology Extraction[A]. Proceeding of the 2006 IEEE/WIC/ACM International Conference on Web Intelligence (WI '06)[C]. Washington ,USA : IEEE Computer Society. 2006 :449 - 455.

10 马　静 ,倪辉峰. 基于模式匹配抽取技术的网上产品情报获取[J]. 情报理论与实践 ,2007 ,(2) :228 - 231.

11 Dawn G. Gregg , Steven Walczak. Adaptive web information extraction[J]. Communications of the ACM ,2006 ,(5) :78 - 84.

12 Thierry Poibeau , Dominique Dutoit. Generating extraction patterns from a large semantic network and an untagged corpus [A]. COLING - 02 on SEMANET : building and using semantic networks[C]. Morristown , NJ , USA : Association for Computational Linguistics. 2002 ,(11) :1 - 7.

13 熊静娴 ,李生红. 基于概念网络的文本信息监控技术[J]. 信息安全与通信保密 ,2005 ,(10) :57 - 59.

14 Thierry Poibeau , Dominique Dutoit. Inferring knowledge from a large semantic network[A]. Proceedings of the 19th international conference on Computational linguistics[C]. Morristown , NJ ,USA :Association for Computational Linguistics. 2002 ,(1) : 1 - 7.

15 顾永跟 ,朱玉　. 一种语义网络情报分析模型的研究和应用[J]. 计算机应用与软件 ,2000 ,(9) :51 - 55.

16 韩家炜 ,坎　伯. 数据挖掘 :概念与技术[M]. 北京 :机械工业出版社 ,2001 :3 - 6.

17 熊　雯. 竞争情报分析技术与数据挖掘[J]. 大众科技 , 2004 ,(12) :31 - 32.

18 Cherie R. Courseault. A text mining framework linking technical intelligence from publication databases to strategic technology decisions[D]. Georgia USA : Georgia Institute of Technology , 2004 - 04 - 12.

19 袁　冰 ,朱东华 ,任智军. 基于数据挖掘技术的专利情报分析方法及实证研究[J]. 情报杂志 ,2006 ,(12) :99 - 102.

20 Andrea Lau , Andrew Vande Moere. Towards a Model of Information Aesthetics in Information Visualization[C]. 11th International Conference Information Visualization[C]. Washington , USA : IEEE Computer Society. 2007 :87 - 92.

21 李　琦 ,陈少强. 走进信息可视化[J]. 中国计算机用户 , 2003 ,(2) :29 - 29.

22 Hayato Ohwada , Fumio Mizoguchi. Integrating information visualization and retrieval for WWW information discovery[J]. Theoretical Computer Science ,2003 ,(2) :547 - 571.

23 董献洲 ,胡晓峰 ,司光亚. 信息可视化技术在情报分析中的应用研究[J]. 计算机工程与应用 ,2006 ,(34) :175 - 177.

(责任编辑 :孙晓明)

(上接第 679 页)

产权人的合法权利。所以, IC 应采取措施, 避免侵权事件的发生。

参考文献

1 Donald Beagle. Conceptualizing an information commons[J]. Journal of Academic Librarianship ,1995 ,25(2) :82 - 89.

2 Mac Whinnie L. The information commons :the academic library of the future[J]. Portal ,2003 ,3(2) :241 - 257.

3 任树怀 ,孙桂春. 信息共享空间在美国大学图书馆的发展与启示[J]. 大学图书馆学报 ,2006 ,(3) :24 - 27.

4 Robert A. Seal. The information commons :New pathways to Digital Resources and Knowledge Management[　]. Preprint for the 3rd China/U. S conference on libraries ,Shanghai ,March 2005.

5 吴建中. 开放存取环境下的信息共享空间[J]. 国家图书馆学刊 ,2005 ,(3) :7 - 10.

6 张冬荣 ,戴利华 ,陈朝晖. 图书馆 Information Commons 建设实践研究[J]. 图书情报工作 ,2006 ,(10) :6 - 10.

7 任　静 ,周凤飞 ,杨丰全 ,马新蕾. 试论大学图书馆信息共享空间的建设[J]. 情报资料工作 ,2007 ,(2) :86 - 89.

8 刘晓霞. Information Commons 发展分析与启示[J]. 情报理论与实践 ,2007 ,(1) :128 - 130.

(责任编辑 :孙晓明)

© 1994-2008 China Academic Journal Electronic Publishing House. All rights reserved.　http://www.cnki.net