

·综述·评论·争鸣·

网络信息计量的数据挖掘技术研究

张艳青

(信阳师范学院, 河南 信阳 464000)

摘 要:网络信息计量的数据挖掘技术研究是运用数据挖掘技术方法发现网络信息计量中的数据特征模式,利用现代可视化软件工具,开发网络信息计量中的数据特征模式的可视化系统,满足网络信息分析人员使用更复杂、更为强大的数据挖掘技术分析、处理数据的要求,为网络信息计量学理论研究及其方法和技术的丰富与发展开拓新的方向,为网络信息计量学规律应用于数据挖掘领域提供平台。

关键词:网络信息计量;数据挖掘技术;数据特征模式

中图分类号: G35 **文献标识码:** A **文章编号:** 1003-0972 (2007) 04-0531-04

Research about Network Informatics Data Mining Technique

ZHANG Yan-qing

(Xinyang Normal University, Xinyang 464000, China)

Abstract: The research on network informatics data mining technique is to discover informatics data characteristic modes with data mining technique method, and to develop informatics data characteristic mode visualized system with modern visualized tool, to meet network information analyzer's demand that make use of more complex and more powerful data mining technique to analyze and deal with data, to exploit new way for webometrics theory research, method and technique development, and to provide support for webometrics law apply to data mining field

Key words: network informatics; data mining technique; data characteristic modes

网络信息计量学是采用数学、统计学等各种定量方法,对网上信息的组织、存贮、分布、传递、相互引证和开发利用等进行定量描述和统计分析,以便揭示网络信息数量特征和内在规律的一门新兴分支学科。它主要是由网络技术、网络管理、信息资源管理与信息计量学等相互结合、交叉渗透而形成的一门交叉性边缘学科,是信息计量学的一个新的发展方向 and 重要的研究领域。其根本目的主要是通过网上信息的计量研究,为网上信息的有序化组织和合理分布、为网络信息资源的优化配置和有效利用、为网络管理的规范化和科学化提供必要的定量依据,从而改善网络的组织管理和信息管理,提高其管理水平,促进其经济效益和社会效益的充分发挥^[1]。而数据挖掘是通过基于计算机技术的自动化工具,对大量数据进行探索和分析的过程,其目的是发现其中有意义的特征模式和规律,从而获得知识,支持决策^[2]。可以看到,网络信息计量学与数据挖掘有着天然的联系。然而,它们却是独立发展着的两个学科。他们都对大量数据进行探索和分析,似乎网络信息计量学强调数据特征模式和规律描述与解释,而数据挖掘更强调数

据特征模式和规律发现的技术与方法。这样,网络信息计量学有必要借鉴数据挖掘技术。

1 国内外研究现状与分析

1.1 国内相关研究现状

国内对基于数据挖掘的网络信息计量研究已有关注。如:邱均平^[3]论述了网络使用记录数据分析的作用和意义,介绍了网络使用记录数据的采集方法和网络使用记录计量分析的主要过程,讨论了有关网络使用记录计量分析的方式和方法;龚立群^[4]介绍了网络计量学的产生和发展,论述了网络计量学研究方法的分类及在网络环境中的具体应用;邱均平^[5]论述了网络使用模式的体系结构与挖掘过程;郭家义^[6]介绍了各种新技术在提高检索效率中的应用,并探讨了如何将传统信息检索策略应用到网络信息检索中的方法;赵黎明^[7]系统总结国内外有关基础研究与发明专利之间知识转移机制,将数据挖掘的有关方法和模式移植到了专利引文分析,建立应用数据挖掘对专利引文数据进行智能分析的创新方法体系,为定量测度基础研究与专利之

收稿日期: 2007-01-15; 修订日期: 2007-07-31; E-mail: zyq1720@163.com

作者简介: 张艳青 (1967-), 女, 河南鹿邑人, 馆员, 主要从事网络信息计量学研究。

间的知识转移机制提供了可能;李村合^[8]介绍了网络信息挖掘的含义,研究了网络信息挖掘的步骤、类型及其关键技术,并指出了网络信息挖掘的应用前提。但数据挖掘与网络信息计量两个学科的科学共同体之间缺少交流,尚无系统研究。

1.2 国外研究现状

国外基于数据挖掘的网络信息计量研究比国内要多。研究的问题主要集中在网络数据挖掘方法技术的具体运用。较为系统的研究见于 The Statistical Cybermetrics Research Group 所承担的欧盟基金研究计划 (EU funded projects) WISER、CREEN、NetReAct 的系列成果,主要对科技与创新的网络指标、网络跟踪工具、预测建模进行研究^[9]。如 TheWall^[10]对博客数据挖掘进行了研究;Prabowo^[11]对 RSS 集成反馈优化方法进行了比较研究;TheWall 对自动化网络问题分析方法^[12]、语言进化与思想传播网络模式^[13]、网络发行^[14]、google 信息搜索模式与挖掘^[15]进行了研究;Schamhorst^[16]对引证分析与超链接进行了比较;Thoms^[17]对研究性主页自我重建与信息挖掘进行了研究;Price^[18]对学术性网络中低频词聚类模式进行了研究。总的来看,国内一般较注重理论与方法研究。国外在理论方法研究的同时,侧重于应用与技术和对实践的指导。但数据挖掘方法技术如何系统的应用于网络信息计量,尚有许多工作要做。

2 网络信息计量及其数据挖掘面临的问题

网络信息计量学主要研究 3 个层面的问题:一是网上文献的直接计量问题;二是网上文献、文献信息及其相关特征信息的计量问题;三是网络结构单元的信息计量问题^[19]。网络信息计量中的数据特征模式,主要有网络信息分布计量特征模式、网络信息动态计量特征模式、网络信息使用计量特征模式等。如何发现这些模式,并实现其可视化,是需要解决的重要问题。网上数据库及搜索引擎为网络信息计量学提供了强有力的工具。也提出了如何从网络以及大型数据库获取计量数据的问题。而数据挖掘提供了大量获取数据的方法技术。同时,也应该注意到问题的另一面,在数据挖掘的过程中也需要信息计量,特别是在知识发现过程中信息计量方法理论有着广阔的应用前景,这也是网络信息计量学发展的必然方向。

目前,国内著名的数据库主要有 VIP 中文科技期刊全文数据库、万方数据库、CNKI 中国专利数据库及各种专业数据库等;国外的主要有 NSPEC、DIALOG、ISI Web of Knowledge 等。还有因特网上的各种搜索引擎是获取网络信息数据的有效工具,其在网络信息计量学中的应用十分重要,如 Alta Vista 和 Google 能够提供关于网站链接的搜索功能,可为网络信息计量研究提供有效的数据源,这些数据库与搜索引擎也为网络信息计量的数据挖掘技术研究提供了实证研究的平台。可以选择数据库或搜索引擎,运用数据挖掘技术,挖掘网络信息计量中的数据结构模式,并实现其可视化。

因此,有关网络信息计量的数据挖掘技术研究的内容可以归结为运用数据挖掘技术方法发现网络信息计量中的数据特征模式,并运用现代可视化软件工具支持,开发网络信息计量中的数据特征模式的可视化系统,具体有 3 个方面工作要做:

(1)对 3 个层面的网络信息中的数据特征模式的数据挖掘,研究数据挖掘技术在网络信息计量中的应用;

(2)对数据挖掘过程中的信息计量问题研究,探讨网络信息计量中的数据特征模式在数据挖掘中的作用;

(3)选择数据库或搜索引擎,对以上 2 方面进行实证性研究。

通过网络信息计量的数据挖掘技术研究,将对网络信息计量中的数据特征模式有更充分的认识,开发网络信息计量的数据挖掘方法和技术,并实现网络信息计量的数据挖掘的可视化系统。具体解决以下问题:

(1)如何从大型数据库及搜索引擎获取网络信息计量的相关数据。运用数据挖掘方法和技术发现网络信息计量中的数据特征模式。

(2)实现网络信息计量的数据挖掘的可视化系统。运用现代可视化软件工具支持,完成网络信息计量特征的数据预处理、可视化映射和呈现。

3 网络信息计量的数据挖掘技术研究的基本步骤与工具

有关网络信息计量的数据挖掘技术研究,应遵循数据挖掘的下列基本步骤进行:

(1)清晰描述研究的问题。设定网络信息计量中的数据特征模式的特征变量,这里可能会有几个可能的假设,需要筛选使问题与数据挖掘模型相结合,并运用比较分析方法选定网络数据库或搜索引擎作为实证研究对象。

(2)网络信息计量特征数据收集。这一步要研究如何从网络数据库及搜索引擎收集网络信息计量数据的问题,关注数据收集对取样理论分布的影响。

(3)网络信息计量特征数据的预处理。这一步要研究异常点的检测和去除,寻找不受异常点影响的健壮性建模方法。对数据进行比例缩放并使它们加权相同,以进行进一步分析,应用编码方法完成数据维度归约。

(4)网络信息计量特征数据的数据挖掘建模与评估。选择并实现适当的数据挖掘技术是这一步的主要任务。应用这些技术从数据中学习,并找到适当的模型。

(5)解释模型和得出结论。依照网络信息计量学原理对所得出的模型进行解释得出有益的结论。

在整个过程中都贯穿着网络信息计量学方法、计量分析方法、数据挖掘方法技术及工具的利用。特别是有强大的数据挖掘工具为研究提供了技术保障。如 SAS 不仅提供了完整有伸缩性的模块化工具,还在挖掘过程中提供了从数据取样 (Sample)、数据探索 (Explore)、数据调整 (Modify)、模型研发 (Model) 到模型综合解释和评价 (Assess) 的完整

科学的 SEMMA 方法论,它包括了智能型的数据挖掘集成工具: SAS/EM;和基于传统统计算法的数据挖掘工具: SAS/INSIGHT、SAS/STAT、SAS/ETS,以及专用于网络流量分析的最新软件产品 e-Discovery和 web Hound等.除了使用 SAS的开发工具以外,还可选择其他厂商的开发工具,将 SAS提供的数据挖掘功能通过界面集成起来,实现定制的数据挖掘应用,比如 CART是一种健壮性的数据挖掘工具,它可以自动搜索大型数据集中的重要模式和关系,迅速揭示隐藏结构.另外还可以从 <http://www.kdnugget.com> 和 <http://www.knowledgestorm.com> 上获得有关最新工具及性能的信息挑选更实用的工具^[20].

4 运用灰色趋势关联聚类进行网络数据挖掘的研究案例

在网络数据库中,存在一类“贫”信息的数据,其主要特点如下:

- (1)样本数据量不大或有残缺.这主要由于许多现实、历史或环境条件的影响导致数据缺少且不精确.数据输入失误,如空值数据等.
- (2)样本数据更新变换快.前期数据将对数据挖掘结果有负面影响,不应当如此计算.
- (3)整体数据规律相当复杂,但在某一时间或空间的数据却有很强的规律性.
- (4)数据概化以后出现的数据概念过于抽象,描述信息过少.

具有以上特点的数据,由于不具备大样本数据的特点,因此不能采用已有的方法来得出有意义的结论.但在这类数据的背后,同样可能隐藏着某种规律,如果能根据数据采集的背景和这类数据代表的主题,采用灰色系统的理论和方法,使用灰色趋势关联聚类进行数据挖掘,便可以得出非常有利于决策的结论.

具体步骤如下:

- (1)确定面向主题的数据集.根据决策者的需求,确定被研究对象的因子集 X ;
- (2)获取特征变量序列.通过数据处理得到所要研究的特征变量序列 $x_i(k)$, $k=1, 2, \cdots, n$; $i=1, 2, \cdots, m$;
- (3)计算灰色趋势关联度 r_{ij} 和趋势关联聚类矩阵 $\Gamma=\{r_{ij}\}_{mm}$, $i, j \in M=\{1, 2, \cdots, m\}$;
- (4)根据聚类阈值 $r_c \in [0, 1]$ 进行系统特征变量聚类;
- (5)对聚类结果 $V=\{V^\alpha\}$, $\alpha \in \Sigma=\{1, 2, \cdots, \sigma\}$ 进行分析和解释.

为准确掌握网络信息分布情况,需要对网络数据库信息分布趋势进行分析,根据分析结果,来寻找其中潜在的规律.这类问题属于数据挖掘的内容.当分析的每种数据库的历史数据较少时,则可采用灰色趋势关联聚类的方法进行挖掘.分析步骤如下:

- (1)确定面向主题的数据集.考察该数据库,其数据包括各种类别.因此确定对其中 12种类别的信息进行聚类分

- 析,即确定系统的因子集 $X=\{x_1, x_2, \cdots, x_{12}\}$.
- (2)确定数据序列.将列举的每种信息的历史数据作为一个序列值,通过对存入在该数据库中的所有相关记录进行数据清理,统计和概化,得到 2003 - 2006年的年收录信息概化数据,其结果如表 1所示.即因子序列 $x_i(k)$ 中, $i=1, 2, \cdots, 12$; $k=1, 2, \cdots, 4$ 由于每个序列所包含的数据过少,因此可采用灰色趋势关联聚类的方法进行分析.
- (3)计算趋势关联聚类矩阵 Γ 结果如表 2所示.

表 1 各类信息 2003 - 2006年的年收录量

Tab 1 2003 - 2006 years each kind of year of information registers quantity

数据库信息分类	2003	2004	2005	2006
X_1	59 543	132 052	199 498	204 676
X_2	2 478	3 340	2 176	2 086
X_3	3 531	4 911	7 443	8 913
X_4	3 408	7 029	14 218	21 709
X_5	375	762	1 115	1 885
X_6	26 248	48 673	73 739	93 872
X_7	174 622	365 540	547 621	677 834
X_8	1 821	2 432	2 411	2 626
X_9	14 301	25 865	30 287	25 023
X_{10}	91 159	162 502	215 892	223 299
X_{11}	830	3 685	5 674	9 748
X_{12}	16 742	25 864	38 326	38 028

表 2 趋势关联聚类结果

Tab 2 The trend connection gathers a result

r_{ij}	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	X_9	X_{10}	X_{11}	X_{12}
X_1	1	0.37	0.5	0.48	0.59	0.64	0.7	0.4	0.5	0.6	0.2	0.55
X_2		1	0.58	0.28	0.33	0.39	0.4	0.8	0.5	0.5	0.1	0.55
X_3			1	0.33	0.46	0.58	0.5	0.7	0.6	0.7	0.2	0.77
X_4				1	0.62	0.46	0.5	0.3	0.4	0.4	0.3	0.36
X_5					1	0.68	0.7	0.4	0.5	0.5	0.2	0.49
X_6						1	0.8	0.4	0.6	0.7	0.2	0.6
X_7							1	0.4	0.5	0.6	0.2	0.51
X_8								1	0.6	0.6	0.1	0.61
X_9									1	0.7	0.2	0.66
X_{10}										1	0.2	0.82
X_{11}											1	0.16
X_{12}												1

- (4)当取阈值 $r_c=0.7$ 为评价准则时,可在评价空间 V 中得到 $r_{ij} \geq 0.7$ 的趋势关联度有 $r_{1,7}=0.74$, $r_{2,8}=0.74$, $r_{3,10}=0.72$, $r_{3,12}=0.77$, $r_{5,7}=0.70$, $r_{6,7}=0.80$, $r_{9,10}=0.74$, $r_{10,12}=0.82$

至此,可得 12类信息增长趋势聚类为 $\{x_1, x_5, x_6, x_7\}$, $\{x_2, x_8\}$, $\{x_3, x_9, x_{10}, x_{12}\}$, $\{x_4, x_{11}\}$.

- (5)结果分析.每组对应的 x_1, x_5, x_6, x_7 的增长趋势相似; x_2, x_8 的地增长趋势相似; x_9, x_{10} 的增长趋势相似;而 x_4 和 x_{11} 各自为一种增长特点.以上分析结果与实际增长特点相吻合,为决策提供了良好的支持.

5 小结

现在是互联网、内联网、数据仓库和数据集市的时代,要求分析人员使用更复杂、更为强大的数据挖掘方法对网络信息进行计量分析和处理,并发现网络信息计量中的数据特征模式,这为网络信息计量的数据挖掘技术研究提供了社会需求.有关网络信息计量的数据挖掘技术研究,试图探讨运用数据挖掘方法和技术解决网络信息计量问题的一般方式,发现网络信息计量中的数据特征模式,并开发其可

视化系统.这是对网络信息计量学理论的补充和完善;是对网络信息计量研究的方法和技术的丰富和发展;也将为网络信息计量学规律运用于数据挖掘任务中提供平台.数据挖掘的技术路线与网络计量学的理论、规律的结合对两方面的发展都是非常有意义的.特别是网络信息计量中的数据特征模式的挖掘及其可视化,对网络信息计量学研究尤为重要.

参考文献:

- [1] 王宏鑫,邱均平.关于网络信息老化的若干问题[J].情报理论与实践,2004,27(4):433-435.
- [2] M CHAEL J A B, GORDON S L. *Mastering Data Mining: The Art and Science of Customer Relationship Management*[M]. New York: John Wiley & Sons Inc, 2000: 4-8.
- [3] 邱均平,黄晓斌.网络用户使用记录的计量分析[J].现代图书情报技术,2002(5):50-55.
- [4] 龚立群,朱庆华.网络计量学的研究方法及应用[J].新世纪图书馆,2003(6):6-9.
- [5] 邱均平,安璐.论网络使用模式的挖掘[J].情报理论与实践,2002,25(5):369-372.
- [6] 郭家义.网络信息检索效率研究[J].图书与情报,2003(2):60-62.
- [7] 赵黎明,李海霞.基于数据挖掘的专利引文研究与知识发现[J].预测,2002,21(6):6-9.
- [8] 李村合.网络信息挖掘技术及其应用研究[J].情报科学,2002,20(11):1212-1214.
- [9] THELWALL M. *The Statistical Cybermetrics Research Group*[DB/OL]. (2005-05-22) [2007-01-15]. <http://cybermetrics.wlv.ac.uk/>.
- [10] THELWALL M. *Bloggers during the London Attacks: Top information sources and topics WWW2006 blog workshop*[DB/OL]. (2006-12-23) [2007-01-15]. <http://www.blogpulse.com/www2006-workshop/papers/blogs-during-london-attacks.pdf>
- [11] PRABOWO R, THELWALL M. *A Comparison of Feature Selection Methods for an Evolving RSS Feed Corpus*[J]. Information Processing & Management (Informetrics special issue) (S0306-4573), 2006, 42(6): 1491-1512.
- [12] THELWALL M, THELWALL S, FAIRCLOUGH R. *Automated Web Issue Analysis: A Nurse Prescribing Case Study*[J]. Information Processing & Management (Informetrics special issue) (S0306-4573), 2006, 42(6): 1471-1483.
- [13] THELWALL M, PRICE E. *Language Evolution and the Spread of Ideas: A Procedure for Identifying Emergent Hybrid Word Family Members*[J]. Journal of the American Society for Information Science and Technology (S1532-2882), 2006, 57(10): 1326-1337.
- [14] THELWALL M, VANN K, FAIRCLOUGH R. *Web Issue Analysis: An Integrated Water Resource Management Case Study*[J]. Journal of the American Society for Information Science and Technology (S1532-2882), 2006, 57(10): 1303-1314.
- [15] THELWALL M. *Directing Students to New Information Types: A New Role for Google in Literature Searches*[J]. Internet Reference Services Quarterly (S1087-5301), 2005, 10(3/4): 159-166.
- [16] SCHARNHORST A, THELWALL M. *Citation and Hyperlink Networks*[J]. Current Science (S1523-3812), 2005, 89(9): 1518-1523.
- [17] THOMAS L, THELWALL M. *Academic Home Pages: Reconstruction of the self*[OL]. (2005-11-20) [2007-01-15]. http://www.firstmonday.org/issues/issue10_12/thoms/.
- [18] PRICE E, THELWALL M. *The Clustering Power of Low Frequency Words in Academic Webs*[J]. Journal of the American Society for Information Science and Technology (S1532-2882), 2005, 56(8): 883-888.
- [19] 邱均平,陈敬全.网络信息计量学及其应用研究[J].情报理论与实践,2001,24(3):161-163.
- [20] MEHMED K. *数据挖掘——概念、模型、方法和算法*[M]. 北京:清华大学出版社,2003:281-309.

责任编辑:任长江