网络论坛中中医药信息的聚类分析研究

田 野 贾李蓉 李园白 刘 静 刘丽红 李敬华 于 彤 杨 策 张竹绿
（中国中医科学院中医药信息研究所，北京市东城区东直门内南小街 16 号，100700）

摘要 随着 INTERNET 网络在国内的普及以及互联网用户的大量增加，以文本信息为载体的网络论坛已经成为人们获取信息，发表个人看法或评论，与其他人进行交流的重要平台。然而，在当前信息爆炸的时代背景下，信息固然重要，但更重要的是对信息内新知识、隐性知识，以及信息发布者潜在需求的深度解读，并基于信息和数据形成相应的思考并提供对策。

关键词 论坛；BBS；中医药信息；聚类分析

Cluster Analysis of the Chinese Medicine Online Forums

Tian Ye, Jia Lirong, Li Yuanbai, Liu Jing, Liu Lihong, Li Jinghua, Yu Tong, Yang Ce, Zhang Zhulu
(*Institute of Information on Traditional Chinese Medicine, China Academy of Chinese Medical Sciences, Add.: No.* 16, *South street, Dongzhimen, Dongcheng District, Beijing, Post code*: 100700)

Abstract With a substantial increase of internet networks as well as internet users, the online forums have become an important platform for people to obtain information, to express their personal views, and communicate with others. However, in the context of the current era of information explosion, it's more important than the information itself, to discover new knowledge, tacit knowledge, and potential demand of information releaser, and think of a solution based on information and data.

Key Words Forum; BBS; Information on Traditional Chinese Medicine; Cluster analysis

1 研究背景

近年来，随着 INTERNET 网络在国内的普及以及互联网用户的大量增加，以文本信息为载体的网络论坛已经成为人们获取信息[1]，发表个人看法或评论，与其他人进行交流的重要平台[2-3]。每天在线论坛中皆会涌现大量主题信息，这些信息的特点往往是数量巨大，难于持久，也就是说发表过的主题隔一段时间就会被后来出现的主题所替代，最终淹没在诸多主题信息中。这些信息虽然数量庞大，但往往内容杂乱，即使有一些信息是有效的，也很容易被淹没掺杂在无效垃圾信息之中。

从这些海量的、有噪声的数据中提取挖掘出隐含其内、但又有用的信息知识是我们要尝试进行探讨的问题。数据挖掘是一门新近的热门研究方法，它是从大型数据集中发现可行信息的过程，数据挖掘使用数学分析来派生存在于数据中的模式和趋势。通常，由于这些模式的关系过于复杂或涉及数据过多，因此使用传统数据浏览无法发现这些模式。它主要探讨如何在海量的、有噪声的、模糊的数据资料中，挖掘出潜在的有用信息，从而为相关决策人员提供数据参考。

2 研究方法

2.1 选择数据 数据的选择是根据需求设定数据抽取目标。对本文而言，数据的抽取目标就是当前热点中医药网站论坛中的词条。随着网络的极速发展，疾病、健康、养生，已不仅是医生才关心的问题，越来越多的普通大众开始予以关注。在绝大多数医药网站中都专门辟有沟通平台——网络论坛。各种角色的人们活跃在论坛上各抒己见。我们试对这些论坛上的词条进行抽取来作为数据的选择。需要注意的是，为了避免人为导向因素的影响，抽取时不对具体内容进行筛选，仅按顺序对词条进行抽取。抽取内容包括论题标题及所有回贴信息。

2.2 数据预处理 数据挖掘对数据的要求比较高，因此对未规范化的数据进行预处理就十分必要。数据的预处理是一个对数据进行格式转化的过程，它的一般过程包括数据清理、用户识别、会话识别、路径补充、事务识别等等[4]。这其中，数据清理是整个数据预处理工作的基础，在数据挖掘中起着至关重要的作用。在这一阶段，可根据挖掘任务的不同对抽取后的词条进行整理转化，如消除噪声、清除重复记录，并对不完整数据进行处理等等，使之成为一种可用形式。

人们在论坛中所使用的往往是自然语言，自然语

基金项目：全国中医医疗与临床科研信息共享关键技术及应用研究－中医药科技信息公共服务平台及文献数据库规范建设研究；科技部基础条件平台－医学科学数据共享网中医药学数据中心（编号：2005KDA32405）

通讯作者：张竹绿，中国中医科学院中医药信息研究所，北京市东城区东直门内南小街 16 号，100700

言因其用词不够规范，或者词汇的重复使用造成了查全率和查准率低下，这就对词条内信息全面抽取工作带来了一定的影响，因此对选择的数据进行预处理就成为了一项不可或缺的工作[5]。

2.3 数据转换 数据转换的主要目的是降维，也就是从初始特征中找出真正有用的特征。在此可以选择中文分词技术[6-7]。网络论坛的特点决定了使用者的用词遣句往往不是那么标准严谨，而是以自然语言为主。中文分词技术刚好属于自然语言处理技术范畴[8]。对于一句话，人可以通过自己的知识来明白哪些是词，哪些不是词，但如何让计算机也能理解[9]？这个处理过程就需要分词算法技术的支持。中文分词方法的基本原理是针对输入文字串进行分词、过滤处理，输出中文单词、英文单词以及数字串等一系列分割好的字符串[10]。

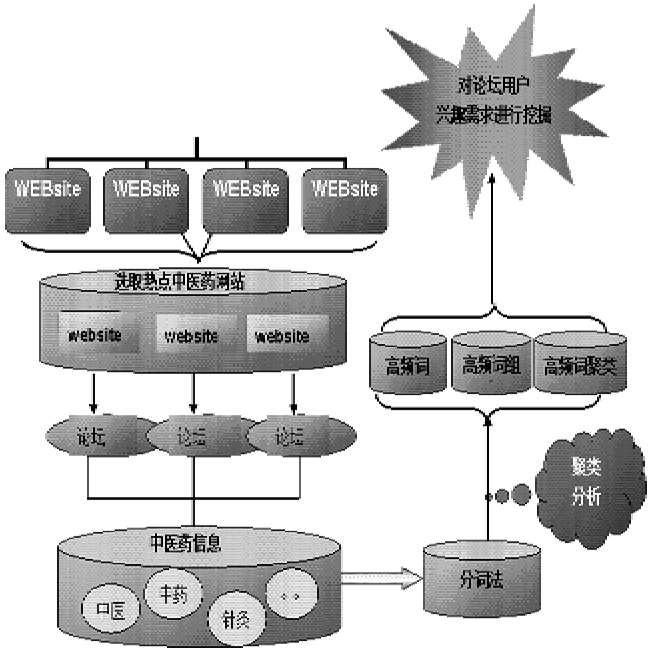

图 1

2.4 数据挖掘 聚类方法最早应用于 20 世纪 70 年代中后期，近年逐渐被广泛应用于各个领域，并为我们的研究提供了参考与借鉴[11]。针对本文研究的情况，对主题论坛中的论题标题及相关回贴信息中的信息进行分词处理后，对词语进行两两统计，得出其在同一论题中出现的次数，形成共词矩阵[12]。共词分析认为两个词语在同一篇文献中出现的频率越高，表示主题间的关系越紧密。以共同矩阵为基础进行聚类分析，通过分析确认与中医药信息相关的词语间的联系，进而分析学科结构的变化[13]。

要想找到词语之间真正的共现关系，需要对词语之间的共现强度按照特定公式进行计算。常用 Salton 指数表示词语之间的关联强度，其计算公式为：

$$S = n_{ij} / (n_1 \times n_j)^{u2}$$

通过 SPSS 软件对词语进行聚类分析，最终形成共词聚类树。把每一点作为一类，找出距离最小的 dij，从而得出距离最近的两类 i、j，把它们合并成为层次最高的一类。此动作重复进行，直到所有的点都并成一个大类。再根据实际需要确定以第几层的类为最终分类标准。

3 结语

我所作为中医药信息领域的研究机构，可以国内中医药领域的论坛为基础，通过对这些论坛数据的提取，汇聚，运用数据挖掘的技术及时获得中医药领域热点问题，掌握中医药科研机构的最新动向，以及普通民众对于中医药的关注话题[14]，为我所及数据库建设今后的工作提供一些新的思路[15]。

参考文献

[1] 熊淑君. 虚拟社区中信息交流的引导机制研究[J]. 图书馆学研究，2011，29(09)：45-47.

[2] 江雪生，续枫，储节旺. 专业虚拟社会区知识服务的概念及其机制研究[J]. 情报理论与实践，2011，34(5)：27-29.

[3] 谢珍，崔旭. 关于专业虚拟社区中个人知识管理的研究[J]. 情报杂志，2010，29(2)：105-109.

[4] 庄爱武，肖云，王军. Web 日志挖掘数据预处理的优化[J]. 计算机技术与发展，2011，21(01)：42-45.

[5] 刘红岩. 中文分词技术的研究[J]. 电脑开发与应用，2010，23(3)：1-3.

[6] 龙树全，赵正文，唐华. 中文分词算法概述[J]. 电脑知识与技术，2009，5(10)：2605-2607.

[7] 马婷婷. 中文自动分词系统概述[J]. 电脑知识与技术，2010，6(33)：9336-9338.

[8] 赵新辉，郭瑞. 基于数据挖掘技术的网络舆情智能监测与引导平台设计研究[J]. 电脑知识与技术，2012，8(1)：1-2，4.

[9] 韩月阳，邓世昆，贾时银，等. 基于字分类的中文分词的研究[J]. 计算机技术与发展，2011，21(7)：29-31，35.

[10] 席朝琼. 面向中文全文索引的中文分词策略[J]. 电脑知识与技术，2012，18(3)：722-726.

[11] 章成志，梁勇. 基于主题聚类的学科研究热点及其趋势监测方法[J]. 情报学报，2010，29(02)：342-349.

[12] 王珏，曾剑平，周荣华，等. 基于聚类分析的网络论坛意见领袖发现方法[J]. 计算机工程，2011，37(5)：44-46，49.

[13] 魏莎莎，熊海灵. 中文分词中的歧义识别处理策略[J]. 微计算机信息，2010，26(10)：190-192.

[14] 陈永刚，孙卉垚. 互联网舆情研究[J]. 情报杂志，2011，30(S1)：85-88.

[15] 于慧新，阮建海. 高校图书馆如何参与网络舆情监测工作[J]. 现代情报，2012，32(2)：71-72，106.

（2012-07-09 收稿）