文本挖掘技术研究进展*

袁军鹏[1]，朱东华[2]，李　毅[3]，李连宏[2]，黄　进[2]

(1. 清华大学 公共管理学院，北京 100084；2. 北京理工大学 管理与经济学院，北京 100081；3. 空军济南四站厂，山东 济南 250022)

摘　要: 文本挖掘是一个对具有丰富语义的文本进行分析从而理解其所包含的内容和意义的过程，已经成为数据挖掘中一个日益流行而重要的研究领域。首先给出了文本挖掘的定义和框架，对文本挖掘中预处理、文本摘要、文本分类、聚类、关联分析及可视化技术进行了详尽的分析，并归纳了最新的研究进展。最后指出了文本挖掘在知识发现中的重要意义，展望了文本挖掘在信息技术中的发展前景。

关键词: 文本挖掘；中文分词；特征选取；文本摘要；文本分类；文本聚类；关联分析；数据可视化

中图法分类号: TP311；TP18　　**文献标识码**: A　　**文章编号**: 1001-3695(2006)02-0001-04

Survey of Text Mining Technology

YUAN Jun-peng[1], ZHU Dong-hua[2], LI Yi[3], LI Lian-hong[2], HUANG Jin[2]

(1. *School of Public Policy & Management, Tsinghua University, Beijing* 100084, *China*; 2. *School of Management & Economics, Beijing Institute of Technology, Beijing* 100081, *China*; 3. *Manufactory of Sizhan, Jinan Air Force, Jinan Shandong* 250022, *China*)

Abstract: Text Mining, also known as intelligent text analysis, text data mining or Knowledge-Discovery in Text (KDT), is a rapidly emerging field concerned with the extraction of concepts, relations, and implicit knowledge from texts. As most information (over 80%) is stored as text, text mining is believed to have a high commercial potential value. Firstly, this review paper discusses the research status of text mining, then it lays out the framework of text mining and analyses techniques of text mining, such as feature selection, automatic abstracting, text categorization, text clustering, text association, data visualization. In the end, it shows the importance of text mining in knowledge discovery and highlights the upcoming challenges of text mining and the opportunities it offers.

Key words: Text Mining; Cutting Chinese Word; Feature Selection; Text Automatic Abstracting; Text Categorization; Text Clustering; Text Association; Data Visualization

1　引言

据数据挖掘著名网站 Kdnuggets 的调查，已有 60% 左右的人在利用软件工具进行文本挖掘，另有 12% 的人计划在六个月内进行文本挖掘，如图 1 所示。

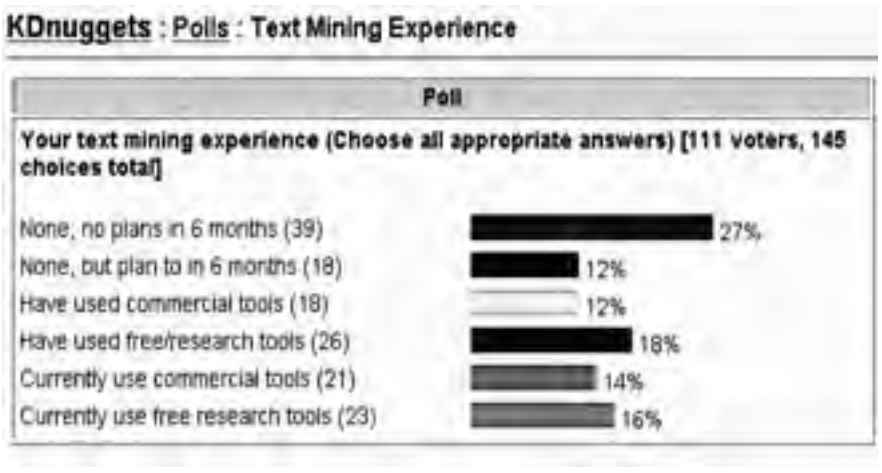

图 1　文本挖掘使用经验调查

由此可见，文本挖掘已经成为数据挖掘中一个日益流行而重要的研究领域。与一般数据挖掘以关系、事务和数据仓库中的结构数据为研究目标所不同的是，文本挖掘所研究的文本数据库，由来自各种数据源的大量文档组成，包括新闻文章、研究论文、书籍、期刊、报告、专利说明书、会议文献、技术档案、政府出版物、数字图书馆、技术标准、产品样本、电子邮件消息、Web 页面等。这些文档可能包含标题、作者、出版日期、长度等结构化数据，也可能包含摘要和内容等非结构化的文本成分[1]，而且这些文档的内容是人类所使用的自然语言，计算机很难处理其语义。因此传统的信息检索技术已不适应日益增加的大量文本数据处理的需要，人们提出文本挖掘的方法进行不同的文档比较，以及文档重要性和相关性排列，或找出多文档的模式或趋势等分析[2]。

收稿日期: 2005-06-22；修返日期: 2005-09-21

基金项目: 国家自然科学基金资助项目(70031010)；北京理工大学学校基金项目；北京理工大学育苗基金项目

2　文本挖掘概述

2.1　文本挖掘的定义

文本挖掘作为数据挖掘的一个新主题，引起了人们的极大兴趣，同时，它也是一个富于争议的研究方向，目前其定义尚无统一的结论，需要国内外学者开展更多的研究以便进行精确的定义。

借鉴 Choon Yang Quek 对 Web 挖掘的定义[3]，我们给出文本挖掘的定义:

定义 1　文本挖掘是指从大量文本的集合 C 中发现隐含

的模式 p。如果将 C 看作输入,将 p 看作输出,那么文本挖掘的过程就是从输入到输出的一个映射 ξ C→p。

2.2 文本挖掘的一般过程

文本挖掘的主要处理过程是对大量文档集合的内容进行预处理、特征提取、结构分析、文本摘要、文本分类、文本聚类、关联分析等。图 2 给出了文本挖掘的一般处理过程。

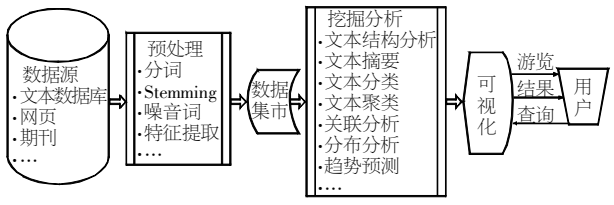

图 2 文本挖掘的一般过程

3 文本挖掘技术分析

文本挖掘不但要处理大量的结构化和非结构化的文档数据,而且还要处理其中复杂的语义关系,因此,现有的数据挖掘技术无法直接应用于其上。对于非结构化问题,一条途径是发展全新的数据挖掘算法直接对非结构化数据进行挖掘,由于数据非常复杂,导致这种算法的复杂性很高;另一条途径就是将非结构化问题结构化,利用现有的数据挖掘技术进行挖掘,目前的文本挖掘一般采用该途径进行。对于语义关系,则需要集成计算语言学和自然语言处理等成果进行分析。我们按照文本挖掘的过程介绍其涉及的主要技术及其主要进展。

3.1 数据预处理技术

预处理技术主要包括 Stemming(英文)/分词(中文)、特征表示和特征提取。与数据库中的结构化数据相比,文本具有有限的结构,或者根本就没有结构。此外,文档的内容是人类所使用的自然语言,计算机很难处理其语义。文本信息源的这些特殊性使得数据预处理技术在文本挖掘中更加重要。

3.1.1 分词技术

在对文档进行特征提取前,需要先进行文本信息的预处理,对英文而言需进行 Stemming 处理,中文的情况则不同,因为中文词与词之间没有固有的间隔符(空格),需要进行分词处理。目前主要有基于词库的分词算法和无词典的分词技术两种。

基于词库的分词算法包括正向最大匹配、正向最小匹配、逆向匹配及逐词遍历匹配法等[4]。这类算法的特点是易于实现,设计简单;但分词的正确性很大程度上取决于所建的词库[5]。因此基于词库的分词技术对于歧义和未登录词的切分具有很大的困难。文献[6]在分析了最大匹配法的特点后,提出了一种改进的算法。该算法在允许一定的分词错误率的情况下,能显著提高分词效率,其速度优于传统的最大匹配法。文献[7]中采用了基于词典的正向逐词遍历匹配法,取得了较好的效果。

基于无词典的分词技术的基本思想是:基于词频的统计,将原文中任意前后紧邻的两个字作为一个词进行出现频率的统计,出现的次数越高,成为一个词的可能性也就越大,在频率超过某个预先设定的阈值时,就将其作为一个词进行索引。这种方法能够有效地提取出未登录词[8,9]。文献[10]设计了一个基于无词典分词的算法,能比较准确地切分出文本中的新词。文献[11]基于层次隐马模型,设计开发了“汉语词法分析系统”,将分词、词语排歧、未登录词的识别三个过程融合到一个相对统一的理论模型中。

3.1.2 特征表示

文本特征指的是关于文本的元数据,分为描述性特征(如文本的名称、日期、大小、类型等)和语义性特征(如文本的作者、机构、标题、内容等)。特征表示是指以一定特征项(如词条或描述)来代表文档,在文本挖掘时只需对这些特征项进行处理,从而实现对非结构化的文本处理。这是一个非结构化向结构化转换的处理步骤[12,13]。特征表示的构造过程就是挖掘模型的构造过程。特征表示模型有多种,常用的有布尔逻辑型、向量空间模型(Vector Space Model, VSM)[14]、概率型以及混合型等。W3C 近来制定的 XML[15], RDF[16] 等规范提供了对 Web 文档资源进行描述的语言和框架。

3.1.3 特征提取

用向量空间模型得到的特征向量的维数往往会达到数十万维,如此高维的特征对即将进行的分类学习未必全是重要、有益的(一般只选择 2% ~5% 的最佳特征作为分类依据),而且高维的特征会大大增加机器的学习时间,这便是特征提取所要完成的工作。

特征提取算法一般是构造一个评价函数,对每个特征进行评估,然后把特征按分值高低排队,预定数目分数最高的特征被选取。在文本处理中,常用的评估函数有信息增益(Information Gain)、期望交叉熵(Expected Cross Entropy)、互信息(Mutual Information)、文本证据权(The Weight of Evidence for Text)和词频[17,18]。

3.2 挖掘分析技术

文本转换为向量形式并经特征提取以后,便可以进行挖掘分析了。常用的文本挖掘分析技术有:文本结构分析、文本摘要、文本分类、文本聚类、文本关联分析、分布分析和趋势预测等。

3.2.1 文本结构分析

其目的是为了更好地理解文本的主题思想,了解文本所表达的内容以及采用的方式。最终结果是建立文本的逻辑结构,即文本结构树,根节点是文本主题,依次为层次和段落[19]。

3.2.2 文本摘要

文本摘要是指从文档中抽取关键信息,用简洁的形式对文档内容进行解释和概括。这样,用户不需要浏览全文就可以了解文档或文档集合的总体内容。

任何一篇文章总有一些主题句,大部分位于整篇文章的开头或末尾部分,而且往往是在段首或段尾,因此文本摘要自动生成算法主要考察文本的开头、末尾,而且在构造句子的权值函数时,相应的给标题、子标题、段首和段尾的句子较大的权值,按权值大小选择句子组成相应的摘要。

假设文本 T, $S=\{S_1, S_2, \cdots, S_n\}$ 是其句子集合,将句子按权值从大到小排列 $P=\{P_1, P_2, \cdots, P_n\}$,其中 P_i 为整型数据,代表权值为第 i 位的句子的顺序号,LEN 为已生成的 ABSTRACT 的长度,LA 为即将生成的文本摘要的长度要求,ABSTRACT 为存放文本摘要的列表。下面给出其算法思想:

```
create ( ABSTRACT, [ Pi |P] , LEN) :
    gets ( S, Pi, TEMP) , length ( TEMP, L) ,
    IF LEN + L < LA,
      LEN1 = LEN + L,
      append ( ABSTRACT, [ TEMP] , ABSTRACT1) ,
      create ( ABSTRACT1, P, LEN1) .
    END IF
```

[注]: 初始值 LEN = 0

gets (S, Pi, TEMP): 从 S 集合中取出第 Pi 个句子 S_{pi}, 放入 TEMP 中

append (ABSTRACT, [TEMP] , ABSTRACT1): 将 TEMP 放至 ABSTRACT 表的末尾

文献[20 ~23] 进行了基于概念统计和语义层次分析的英文自动文摘和自动标引研究。文献[24] 提出了使用中心文档来代表文档集合, 使用中心词汇来表示文档的方法, 并给出了求取中心文档和中心词汇的算法。

3.2.3　文本分类

文本分类的目的是让机器学会一个分类函数或分类模型, 该模型能把文本映射到已存在的多个类别中的某一类, 使检索或查询的速度更快, 准确率更高。训练方法和分类算法是分类系统的核心部分。用于文本分类的分类方法较多, 主要有朴素贝叶斯分类(Native Bayes) [25]、向量空间模型[14]、决策树、支持向量机[26]、后向传播分类、遗传算法、基于案例的推理、K - 最临近、基于中心点的分类方法[27]、粗糙集、模糊集以及线性最小二乘(Linear Least Square Fit, LLSF) [28] 等。

文献[29] 指出传统特征提取的方法是基于词形的, 并不考察词语的意义, 忽略了同一意义下词形的多样性、不确定性以及词义间的关系, 尤其是上下位关系。该文的方法在向量空间模型(VSM) 的基础上, 以 "概念 "为基础, 同时考虑词义的上位关系, 使得训练过程中可以从词语中提炼出更加概括性的信息, 从而达到提高分类精度的目的。文献[30] 对智能文本分类进行了研究。

3.2.4　文本聚类

文本分类是将文档归入到已经存在的类中, 文本聚类的目标和文本分类是一样的, 只是实现的方法不同。文本聚类是无教师的机器学习, 聚类没有预先定义好的主题类别, 它的目标是将文档集合分成若干个簇, 要求同一簇内文档内容的相似度尽可能大, 而不同簇间的相似度尽可能小[31]。Hearst 等人的研究已经证明了 "聚类假设 ", 即与用户查询相关的文档通常会聚类得比较靠近, 而远离与用户查询不相关的文档[32]。目前, 有多种文本聚类算法, 大致可以分为两种类型: 以 G-HAC 等算法为代表的层次凝聚法[33] 和以 K-means 等算法为代表的平面划分法[34]。文献[35] 介绍了将 G-HAC 和 K-means 集合起来的 Buckshot 方法和 Fractionation 方法。

3.2.5　关联分析

关联分析是指从文档集合中找出不同词语之间的关系。Feldman 和 Hirsh 研究了文本数据库中关联规则的挖掘[36], 文献[37] 提出了一种从大量文档中发现一对词语出现模式的算法, 并用来在 Web 上寻找作者和书名的出现模式, 从而发现了数千本在 Amazon 网站上找不到的新书籍; 文献[38] 以 Web 上的电影介绍作为测试文档, 通过使用 OEM 模型从这些半结构化的页面中抽取词语项, 进而得到一些关于电影名称、导演、演员、编剧的出现模式。

3.2.6　分布分析与趋势预测

分布分析与趋势预测是指通过对文档的分析, 得到特定数据在某个历史时刻的情况或将来的取值趋势。文献[39] 使用多种分布模型对路透社的两万多篇新闻进行了挖掘, 得到主题、国家、组织、人、股票交易之间的相对分布, 揭示了一些有趣的趋势。文献[40] 通过分析 Web 上出版的权威性经济文章对每天的股票市场指数进行预测, 取得了良好的效果。

3.3　可视化技术

数据可视化(Data Visualization) 技术指的是运用计算机图形学和图像处理技术, 将数据转换为图形或图像在屏幕上显示出来, 并进行交互处理的理论、方法和技术。它涉及到计算机图形学、图像处理、计算机辅助设计、计算机视觉及人机交互技术等多个领域。文献[41 ~50] 对信息可视化技术进行了大量的研究, 运用最小张力计算、多维标度法、语义分析、内容图谱分析、引文网络分析及神经网络技术, 进行了信息和数据的可视化表达。

4　文本挖掘的现状分析

文本挖掘属于新兴的前沿领域, 国内对此研究相对较少, 我们于 2005 年 3 月 3 日 10 点 38 分在中国期刊网(CNKI) 以文本挖掘为检索词只检索到 96 篇文章。目前国内外学者主要在文本结构分析、文本摘要、文本分类、文本聚类、文本关联分析、分布分析和趋势预测等方面进行了研究, 中国学者在中文分词等领域取得了一些进展。由于目前计算机的运算能力还不能够达到文本挖掘研究的要求, 国内外学者在这方面的进展非常小。

5　结束语

文本挖掘最大的动机是来自于潜藏于电子形式中的大量的文本数据。将来所需要做的工作就是: ①如何将现存的数据挖掘技术应用与文本挖掘领域很好地融合, 那样文本挖掘就能够更有效地进行; ②发展全新的非结构化文本挖掘算法; ③将文本挖掘与自然语言处理、计算语言学等有效集成, 处理文档中的语义关系。

参考文献:

[1] iawei Han, Michaline Kamber. Data Mining Concepts and Techniques[M] . 北京: 高等教育出版社, 2001. 285-295.

[2] 王继成, 潘金贵, 张福炎. Web 文本挖掘技术研究[J] . 计算机研究与发展, 2000, (5): 513-520.

[3] Choon Yang Quek. Classification of World Wide Web Documents[D] . School of Computer Science, Camegie Mellon University, 1997.

[4] 刘源, 等. 信息处理用现代汉语分词规范及自动分词方法[M] . 北京: 清华大学出版社, 1994. 36-37.

[5] 蒋澄, 马范援, 蒋思杰. 中英文 WWW 搜索引擎的信息处理[J] . 计算机工程, 1999, 25(4): 37-38.

[6] 杨斌, 孟志青. 一种文本分类数据挖掘的技术[J] . 湘潭大学自然科学学报, 2001, 23(4): 34-37.

[7] 邹涛. WWW 上的信息挖掘技术及实现[J] . 计算机研究与发展, 2000, 36(8): 1020-1024.

[8] 严威, 赵政. 开发中文搜索引擎汉语处理的关键技术[J] . 计算机工程, 1999, 25(6): 5-6.

[9] 吴立德. 大规模中文文本处理[M]. 上海: 复旦大学出版社, 1997. 23-24.

[10] 胥桂仙, 苏筱蔚, 陈淑艳. 中文文本挖掘的无词典分词的算法及其应用[J]. 吉林工学院学报, 2002, 23(1): 16-18.

[11] 刘群, 张华平, 俞鸿魁, 等. 基于层次隐马模型的汉语词法分析[Z]. 2003.

[12] 陆玉昌, 鲁明羽, 李凡, 等. 向量空间法中单词权重函数的分析和构造[J]. 计算机研究与发展, 2002, 39(10): 1205-1210.

[13] 李凡, 鲁明羽, 陆玉昌. 关于文本特征抽取新方法的研究[J]. 清华大学学报(自然科学版), 2001, 41(7): 98-101.

[14] Salton G, Wong A, Yang C Sa. Vector Space Model for Automatic Indexing [J]. Communications of the ACM, 1975, 18(5): 613-620.

[15] Bray T, Paoli J, Sperberg-McQueen C M. Extensible Markup Language (XML) 1.0 Specification [EB/OL]. World Wide Web Consortium Recommendation, http://www.w3.org/TR/REC-xml, 1998.

[16] Lassila O, Swick R R. Resource Description Framework Model and Syntax Specification [EB/OL]. World Wide Web Consortium Recommendation, http://www.w3.org/TR/REC-rdf-syntax/, 1999.

[17] Koller D, Sahami M. Hierarchically Classifying Documents Using Very Few Words[J]. ICML'97, 1997, 170-178.

[18] 徐妙君, 顾沈明. 面向 Web 的文本挖掘技术研究[J]. 控制工程, 2003, 10(5): 44-46.

[19] 林鸿飞. 基于概念的文本结构分析方法[J]. 计算机研究与发展, 2000, 37(3): 324-328.

[20] 马颖华, 王永成, 等. 自动标引中基于概念层次树的主题词轮排选择的算法实现[J]. 高技术通讯, 2003, (6): 18-21.

[21] 史磊, 王永成. 英文文献自动摘要系统的研制与开发[J]. 高技术通讯, 1999, (11): 22-26.

[22] 万敏, 罗振声, 等. 基于概念统计的英文自动文摘研究[J]. 计算机工程与应用, 2002, (24): 7-9.

[23] 季姮, 罗振声, 等. 基于概念统计和语义层次分析的英文自动文摘研究[J]. 中文信息学报, 2003, (2): 14-20.

[24] Pirolli P, Schank P, *et al*. Scatter/Gather Browsing Communicates the Topic Structure of a Very Large Text Collection[C]. Proc. of the ACM Sig. Chi. Conf. on Human Factors in Computing Systems, 1996.

[25] Tom M Mitchell. 机器学习[M]. 北京: 机械工业出版社, 2003. 36-58.

[26] Joachims T. Text Categorization with Support Vector Machines: Learning with Many Relevant Features[R]. Ls Viii Technical Report, Na 23, University of Dortmand, 1997.

[27] Han E H, Karypis G. Centroid-based Document Classification: Analysis & Experimental Results[A]. Technical Report 00-017. Computer Science[R]. University of Minnesota, 2000.

[28] David Hand, Heikki Mannila, Padhraic Smyth. 数据挖掘原理[M]. 张银奎, 廖丽, 宋俊. 北京: 机械工业出版社, 2003. 186-207.

[29] 厉宇航, 罗振声, 程慕胜. 基于概念层次的英文文本自动分类研究[J]. 计算机工程与应用, 2004, (11): 75-77.

[30] 杨清, 杨岳湘, 瞿国平. 智能文本分类系统的研究与设计[J]. 计算机应用研究, 1999, 16(10): 15-17.

[31] 李聪, 张勇, 高智. 一种新的聚类算法[J]. 模式识别与人工智能, 1999, 12(2): 205-209.

[32] Hearst Ma, Pedersen J. Reexamining the Cluster Hypothesis: Scatter/Gather on Retrieval Results[C]. Proc. of the 19th Annual Int'l Acm/Sigir Conf. Zurich, 1996. 76-84.

[33] Willet P. Recent Trends in Hierarchical Document Clustering: A Critical Review[J]. Information Processing and Management, 1988, 24: 577-597.

[34] Cutting D, *et al*. Scatter/Gather: A Cluster-based Approach to Browsing Large Document Collections [C]. Proc. of the 15th Annual Int'l Acm/Sigir Conf. Copenhagen, 1992. 318-329.

[35] Rocchio J J. Document Retrieval Systems Optimization and Evaluation [D]. Harvard University, Cambridge, Ma, 1966.

[36] R Feldman, H Hirsh. Finding Associations in Collectionds of Text [M]. Machine Learning and Data Mining: Methods and Applications, John Wiley Sons, 1998. 223-240.

[37] Brin S. Extracting Patterns and Relations from the World Wide Web [C]. Proc. of Web Db Workshop at Edbt'98, Valencia, 1998.

[38] Wang Ke, Liu Huiqing. Schema Discovery from Semi-Structured Data [C]. Proc. of the 3rd Int'l Conf. on Knowledge Discovery and Data Mining, New Port Beach, 1997.

[39] Feldman R, Dagan I. Knowledge Discovery in Textual Databases (KDT) [C]. Montreal: Proc. of the 1st Int'l Conf. on Knowledge Discovery, 1995. 112-117.

[40] Wuthrich B, Permunetilleke D, Leung S, *et al*. Daily Prediction of Major Stock Indices from Textual WWW Data[C]. New York: Proc. of the 4th Int'l Conf. on Knowledge Discovery, 1998.

[41] D H Zhu, A LPorter. Automated Extraction and Visualization of Information [J]. Technological Intelligence and Forecasting, Technological Forecasting & Social Change, 2002, (69): 495-506.

[42] D H Zhu. Technology Mapping: Extracting Patterns from S&T Databases[R]. Tpac Internal Paper, 1998.

[43] Borner K, Chen C, Boyack K. Visualizing Knowledge Domains [J]. Annual Review of Information Science and Technology, 2003, (37).

[44] Chen C, Paul R J. Visualizing a Knowledge Domain's Intellectual Structure[J]. IEEE Computer, 2001, 34(3): 65-71.

[45] Chen C. Visualising Semantic Spaces and Author Co-Citation Networks in Digital Libraries [J]. Information Processing & Management, 1999, 35(3): 401-420.

[46] Chen C, Czerwinski M. Empirical Evaluation of Information Visualizations: An Introduction[J]. International Journal of Human-Computer Studies, 2000, 53(5), 631-635.

[47] Chen C. Information Visualization and Virtual Environments [M]. London: Springer-Verlag, 1999.

[48] S Morris, C DeYong, Z Wu, *et al*. DIVA: A Visualization System for Exploring Document Databases for Technology Forecasting [J]. Computers and Industrial Engineering, 2002, (43).

[49] Longitudinal Patent Analysis for Nanoscale Science and Engineering: Country[R]. Institution and Technology Field J. Nanoparticle Research, Kluwer Acad. Publ., 2003, (5): 1-47.

[50] JC Lamirel, S Al Shehabi, C Francois, *et al*. Intelligent Patent Analysis Through the Use of a Neural Network: Experiment of Multi-viewpoint Analysis with the MultiSOM Model[C]. Japan: The ACL 2003 Workshop on Patent Corpus Processing, Sapporo, 2003.

作者简介:

袁军鹏(1973-), 男, 山东人, 讲师, 博士生, 研究方向为管理信息系统、数据挖掘; 朱东华(1963-), 男, 福建人, 美国佐治亚理工学院技术政策与评估中心资深研究员, 研究员, 博士生导师, 研究方向为科技管理、技术预测、数据挖掘、人工智能、信息系统; 李毅(1972-), 男, 山东人, 工程师, 研究方向为管理信息系统; 李连宏(1967-), 男, 吉林人, 高级工程师, 博士生, 研究方向为管理信息系统、国民经济动员; 黄进(1978-), 男, 江西人, 博士生, 研究方向为数据挖掘、管理信息系统。