文本挖掘与中文文本挖掘模型研究

谌志群，张国煊

(杭州电子科技大学 计算机应用技术研究所，浙江 杭州 310018)

摘 要：文本挖掘，又称为文本数据挖掘或文本知识发现，是指在大规模的文本中发现隐含的、以前未知的、潜在有用的模式的过程。本文首先对文本挖掘进行了概述，给出了文本挖掘的定义、特点和研究现状。然后对国内中文文本挖掘的研究现状进行了分析，指出了当前中文文本挖掘研究中存在的主要问题和主要研究方向。最后提出了一个统一的中文文本挖掘模型 ——UCTMF。该模型具有层次性、开放性和可扩展性，为中文文本挖掘系统提供了基本体系框架。

关键词：文本挖掘；数据挖掘；中文文本挖掘模型；中文信息处理

中图分类号：G354　　**文献标识码**：A　　**文章编号**：1007 - 7634(2007)07 - 1046 - 06

Study on the Text Mining and Chinese Text Mining Framework

CHEN Zhi - qun， *ZHANG Guo - xuan*

(*Institute of Computer Application Technology*， *Hangzhou Dianzi University*， *Hangzhou* 310018， *China*)

Abstract: Text mining, also known as text data mining or knowledge discovery in texts, focuses on computerized exploration of large amounts of text and on discovery of implicit, previously unknown, and potentially useful patterns within them. Firstly, the text mining are introduced including its definition, its characteristics and its progress. Then, The problems and research direction of Chinese text mining are pointed out based on analysis for state - of - the - art of research on Chinese text mining. Finally, Unified Chinese Text Mining Framework (UCTMF) is presented. The framework are hierarchical, open, and scalable. It provide a unified and public frame for Chinese Text Mining System.

Key words: text mining; data mining; chinese text mining framework; chinese information processing

存储和交换信息的最自然形式是自然语言文本。研究表明，80 %左右的电子化信息是以无结构自由文本 (Unstructured Free - form Text) 的形式存在的，如 Web 页面、在线新闻、公司档案、研究论文、财经报道、医疗记录、E - mail 等。这些信息是海量的，非结构化的，具有模糊性和歧义性，人或计算机都难以使用或难以计算，但对公司、组织甚至个人来说，这些信息又都具有巨大的潜在价值，于是要求有大规模文本信息的自动处理与分析技术，需要文本挖掘技术。本文首先给出了文本挖掘的概念与特点、文本挖掘研究的主要内容与国外研究现状。然后对国内中文文本挖掘的研究现状进行了分析，指出了当前中文文本挖掘研究中存在的主要问题，以及中文文本挖掘进一步发展需要解决

收稿日期：2006 - 10 - 31

基金项目：浙江省自然科学基金项目 (M603025)

作者简介：谌志群 (1973 -)，男，江西南昌人，讲师，硕士，从事中文信息处理与数据挖掘研究。张国煊 (1945 -)，男，浙江温州人，教授，从事自然语言理解、人工智能、分布式系统研究.

的关键问题和主要研究方向。最后对汉语和中文文本的构成特点进行了分析，结合当前及可预见的将来的中文信息处理水平和中文文本挖掘的可能应用领域，提出了一个统一通用的中文文本挖掘模型——UCTMF（Unified Chinese Text Mining Framework），为中文文本挖掘提供了基本体系框架。

1　文本挖掘概述

文本挖掘[1]（Text Mining，TM），又称为文本数据挖掘（Text Data Mining，TDM）或文本知识发现（Knowledge Discovery in Texts，KDT），是指为了发现知识，从大规模文本库中抽取隐含的、以前未知的、潜在有用的模式的过程。

文本挖掘的特点：

(1) 文本挖掘处理的是大规模的文本集合，而不是一个或少量的文本文档。

(2) 文本挖掘发现的知识是隐藏在大量文本文档中的，是新的、以前未知的模式或关系。

(3) 文本挖掘抽取的知识是以真实世界为基础的，具有潜在价值的，是直接可用的，它或者是某个特定用户感兴趣的，或者对于解答某个特定问题是有用的。它将改变人或 Agent 的行为。

(4) 由于文本挖掘处理的是大规模的文本库，其挖掘算法复杂度必须在时间和空间上是多项式的，即若文本长度是 n，算法复杂度应该是 O（n）或 O（n log n）。

(5) 文本数据有大量的噪声和不规则的结构，因此文本挖掘算法应具有很强的算法鲁棒性。

(6) 文本挖掘是个多学科交叉的研究领域，涉及领域包括数据挖掘、机器学习、统计学、自然语言理解、信息检索、信息抽取、聚类、可视化、数据库技术等。

从 Feldman 在 1995 年正式提出文本挖掘的概念[2]到现在只有短短的 10 年左右时间，但文本挖掘研究在国外特别是拉丁语系国家发展迅速。国外的研究主要围绕文本挖掘模型[3-4]、文本特征抽取与文本中间表示[5-7]、文本挖掘算法（如关联规则抽取[8-9]、语义关系挖掘[10-11]、文本聚类与主题分析[12-13]、趋势分析[14]）等方面展开，已经形成了一套较成熟的理论体系与技术手段，并且在多个领域得到了应用。包括：网络聊天室文本流主题跟踪[12]、在线新闻实时监控[13]、股票价格预测[14]、专利数据分析[15]、分子生物学文献挖掘[16]、开放式问卷调查文本数据分析[17]等。

2　中文文本挖掘研究现状分析

我国学术界正式引入文本挖掘的概念并开展针对中文的文本挖掘研究是从最近几年才开始的。从公开发表的有代表性的研究成果[18-27]来看，目前我国文本挖掘研究还处在消化吸收国外相关的理论和技术与小规模实验阶段，还存在如下不足和问题。

(1) 没有形成完整的适合中文信息处理的文本挖掘理论与技术框架。目前的中文文本挖掘研究只是在某些方面（如特征抽取[18-19]）和某些狭窄的应用领域（如分类与聚类[20-26]）展开，是零散的、孤立的，远未形成理论体系和针对中文的技术框架。在技术手段方面主要是借用国外针对英文语料的挖掘技术，没有针对汉语本身的特点，没有充分利用当前的中文信息处理与分析技术来构建针对中文文本的文本挖掘模型，限制了中文文本挖掘的进一步发展。

(2) 中文文本的特征提取与表示大多数采用“词袋”法。目前中文文本挖掘中多采用“词袋”法[19-20]，即提取文本高频词构成特征向量来表达文本特征。“词袋”法没有考虑词在文本（句子）中担当的语法和语义角色，也没有考虑词与词之间的顺序，丢失了大量有用信息，加之汉语中同义词与多义词的普遍存在，更加减弱了高频词向量表达文本特征的可信度。而且用“词袋”法处理真实中文文本数据时，特征向量的维数往往会达到几千或几万，如此高的维数将使挖掘算法的效率大大降低。为了降低特征维数，也有些有益的工作，如通过奇异值分解技术[21]，分析特征词之间的语义相关性来压缩特征向量。但总的来说，这些方法还都局限于词一级，局限在语形层面，没有引入句子和篇章的语法分析和语义分析技术，没有涉及文本的语义特征。这大大限制了针对中文文本的深层次知识挖掘。

(3) 知识挖掘的种类和深度有限。由于中间表示没有充分反映文本的语法语义特征，造成不能进行深度的更有价值的知识挖掘，限制了中文文本挖掘的应用范围和应用效果。目前中文文本挖掘在挖掘种类和挖掘深度上都是有限的，一般只是进行文本的分类[25]、聚类[26]或者信息抽取[27]，而且针对开放语料的实验结果也不是很理想。

综上所述，中文文本挖掘研究还处于初级阶段，还有许多需要研究与解决的问题。本文认为，中文文本挖掘的研究一方面要重视基本理论与系统框架的研究，一方面要重视针对汉语本身的研究，要研究中文文本的构成特点与特征提取机制，只有在对中文文本的分析技术取得进展后才可能为后面的知识挖掘提供强有力的支持，才能为中文文本的深度挖掘提供可能。另外，在挖掘的种类方面，不能仅局限于分类、聚类等简单知识的发现，可以探索更丰富的知识挖掘方法，如语义规则发现、趋势分析、主题跟踪等。

3 统一的中文文本挖掘模型

根据上面对中文文本挖掘研究的分析，本文认为，目前的关键问题之一是根据汉语言本身和中文文本的构成特点，以及当前及可预见的将来的中文信息处理水平，结合通用的知识发现理论框架和中文文本挖掘的可能应用领域，建立一个统一通用的中文文本挖掘模型，为中文文本挖掘提供基本框架，为中文文本挖掘研究的进一步发展打下基础。基于以上认识，本文提出了一个分层、开放、可扩展的统一中文文本挖掘模型——UCTMF。

3.1 中文文本挖掘流程

中文文本挖掘是一个分析中文文本数据，抽取中文文本信息，进而发现中文文本知识的过程。一个完整的中文文本挖掘过程应包括中文文本集合的预处理（文本数据的选择、清洗、分类、特征提取等）、索引与存储、中间表示分析（知识挖掘）、后处理（知识的评价与取舍、知识的解释与知识的可视化表达）等步骤，参见图1。中文文本挖掘模型应涵盖这个过程的各个环节。

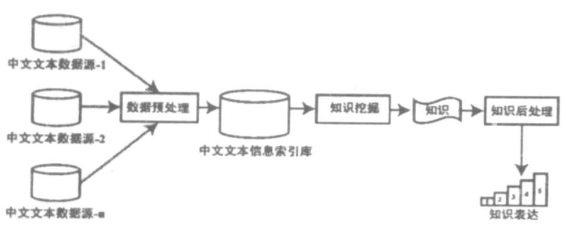

图1 中文文本挖掘流程

3.2 汉语特点与当前中文信息处理水平

中文文本挖掘处理的是汉语文本，因此在引入一般的文本挖掘方法时必须适应汉语的特点，在构建实用的中文文本挖掘系统时必须考察中文信息处理技术的研究现状和发展趋势。

汉语是一种语义型语言[28]，重“意合”，轻形式，而且语形、语法和语义等各层面的歧义现象非常严重。

(1) 汉语缺乏狭义的形态。拉丁语系语言的形态，对于计算机来说就是可以识别的标记；而汉语没有这种标记，于是需要人来总结规律，并且把这些规律形式化。在此基础上还要设计算法使得计算机能自动获得语言的形式化表达。

(2) 语法灵活。由于汉语缺乏狭义的形态，汉语句子中各个成分之间的关系主要靠词序、“意合”及虚词来体现。但是，词序虽同可能意义迥异；虚词常常可以省略，只能是解决词与词、句与句关系问题的辅助手段；“意合”则更为复杂，它与语境密切相关，而世界上还没有一个通用的语境模型，语境的形式化是个世界性的难题。

(3) 语义灵活。从词汇层面说，存在一词多义、同音词、同义词、近义词等；从句义层面说，同一结构可以表达不同的语义，同一语义可以用不同结构来表达。

汉语的这些特点使得对汉语文本的自动语法和语义分析变得非常困难。

目前面向大规模真实中文语料的词法分析技术已经成熟，词切分技术已经达到了实用的程度；而语法、语义分析技术尚未成熟，难以获得高质量的语言分析结果，因此要在中文文本挖掘的特征提取阶段对中文文本进行深入的分析与理解，获取文本的完整“语义”，无论是在技术可行性上还是在时间与空间效率上都是不现实的。本文认为，如果在中文文本挖掘中需要文本的语义信息，在目前的技术条件下，可行的方法是引入汉语的浅层分析技术[29]，对中文文本进行有限深度的分析，在兼顾实现代价的前提下，提取句子与文本的基本语法与语义特征作为整个文本的语义表达，从而在一定程度上实现中文文本的“语义”特征提取和“语义”索引。

3.3 中文文本挖掘模型的层次性、开放性与可扩展性

本文认为，一个统一的中文文本挖掘模型在架构上应是分层次的，并且应具有开放性和可扩展性，应能适应不同的文本特征、不同的知识类型、不同的应用领域和不同的中文信息处理水平。

统一中文文本挖掘模型的层次性应体现在以下几个方面。

(1) 分析粒度的层次性。根据不同的应用需求，中文文本集合的索引可以以文本为基本单位，也可以以每一文本内的词或概念为基本单位。以文本为单位可进行聚类、主题分析等，以词或概念为单位可挖掘关联规则、语义关系等。

(2) 文本中间表达的层次性。文本的中间表达应包含足够的文本特征信息来支持文本挖掘操作，文本中间表达的形式决定了能进行的挖掘深度。文本的中间表达应该是结构化或半结构化的，易于进行挖掘操作。文本的中间表达还应易于索引、易于存储、易于解析、易于在网络上进行传输。统一的中文文本挖掘模型可根据应用目标的不同和中文信息处理水平的高低采用不同的文本中间表达形式，可以由简单到复杂，由语形层面到语义层面，由词、词组、概念、语法结构特征、语义特征到完整的语义表达等。

(3) 系统资源库的层次性。进行知识挖掘操作需根据需要配置不同的系统资源。在中文文本挖掘模型中支持挖掘操作的系统资源由低级到高级可以是词库、义类词典、规则库、领域知识库、高级知识库等。

(4) 知识挖掘的层次性。中文文本知识挖掘获取的文本知识可以由低级到高级，包括聚类、关联规则、语义关系、趋势分析等。

中文文本挖掘模型的层次性进一步决定了它应具有开放性和可扩展性。模型可随着各类语料库、资源库的完备，中文文本分析技术的发展而不断充实、完善与提升功能。根据应用需求的不同，模型中的各层均是可扩展的，也是可拆卸的。各类中文文本挖掘系统将可以统一纳入这个体系框架中。

3.4 统一中文文本挖掘模型 ——UCTMF

基于以上分析，下面给出一个分层次的、具有开放性和可扩展性的统一中文文本挖掘模型 ——UCTMF (Unified Chinese Text Mining Framework)，其体系框架见图 2。

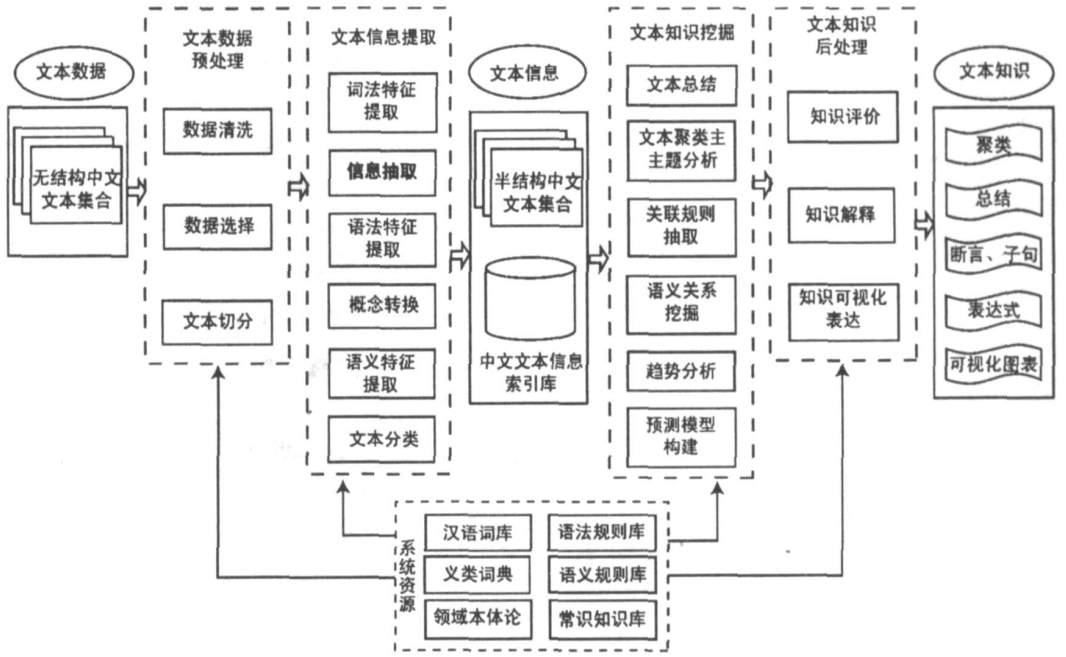

图 2　UCTMF 体系框架

中文文本挖掘处理的数据源是文本数据，是无结构中文文本的集合，可以是 Web 页面、文本文件、WORD 和 EXCEL 文件、PDF 文件、E - mail 等各种形式的电子文档。

在获取文本信息之前首先要对文本数据进行预处理，包括数据清洗，如去噪、去重；数据选择，即选择合适的、面向特定应用的、领域相关的文本数据；文本切分，如中文分词、段落切分等。数据预处理得到“干净”的文本数据后，必须提取中文文本的特征信息，包括关键词（高频词）提取、术语（词组、短语）提取、基于模板的信息抽取、基于义类词典的概念转换、基于浅层句法分析的语法

特征提取、基于浅层语义分析的语义特征提取、基于文本分类的文本类别信息获取等操作。

经过中文文本特征提取操作后，中文文本数据转换为中文文本信息。中文文本信息可以以结构化(如数据库) 或半结构化 (如 XML 文档) 形式进行存储或索引。存储或索引可以根据需要采用不同的颗粒度，可以以每类文本或单个文本为基本单位，也可以以每一文本内的某级语法或语义单位为基本单位。

知识挖掘在文本信息的基础上进行，还必须有系统资源的支持。支持挖掘操作的系统资源由低级到高级可以是汉语词库、义类词典、领域本体论(Ontology)、语法语义规则库、常识知识库等。知识挖掘包括文本总结，即对文本集合进行自动摘要；文本聚类与主题分析；关联规则抽取；语义关系挖掘；趋势分析；预测模型构建等。挖掘算法获取的知识可能不一致、不直观、难以理解，因此需要对文本知识进行必要的后处理，包括知识的评价与取舍，即消除知识的不一致性，评价知识的新颖性并舍弃不新颖的无价值知识；知识的解释与知识的可视化表达，即对抽象的、人难以理解的知识进行必要的转换，用易于人们理解的方式表达出来，如自然语言描述、表达式、走势图、表格等。

UCTMF 体系框架与文本知识挖掘的流程 (见图 1) 是一致的，涵盖了文本知识挖掘各个阶段的操作。图 2 中功能框和系统资源框 (虚线框) 中的模块是可拆卸、可装配和可扩展的，可以根据具体的文本挖掘系统、文本挖掘系统的应用领域进行选择与组合，体现了该模型的开放性与可扩展性。

4 结 语

从文本挖掘与传统数据挖掘的处理对象来看，文本挖掘可以看作是数据挖掘从结构化数据到无结构数据的一次飞跃，是知识发现领域的主流研究方向之一，具有重要的科学意义和广阔的应用前景。文本挖掘的可能应用领域包括：

(1) 客户模型分析，如自动分析客户反馈意见并总结；

(2) 网上有害信息的发现、过滤与跟踪；

(3) 主动个性化信息服务，如为客户提供商业新闻或报告的个性化信息服务；

(4) 公司资源计划，如挖掘公司的报告和商业信函；

(5) 科技文献分析，如为科研工作者提供面向特定研究领域的信息导航；

(6) 网上论坛的实时监控；

(7) 电子邮件分类与过滤；等等。

中文文本挖掘的研究尚处于起步阶段，无论是在挖掘模型、挖掘算法还是在应用系统开发等方面都落后于国际水平，这是与当前“知识经济”时代对知识的迫切需要不相符的。

本文在充分考察汉语作为一种意合型语言的特点及中文文本构成特点的基础上，构建了一个分层、开放、可扩展的中文文本挖掘模型，为中文文本挖掘提供了一个基本框架。对中文文本挖掘的进一步的研究可以在这个统一框架内展开。

参考文献

1 谌志群,张国煊. 文本挖掘研究进展[J]. 模式识别与人工智能,2005,18(1):65 - 74.

2 Feldman,R. and Dagan,I. Knowledge discovery in textual databases (KDT)[Z]. In:proceedings of the First International Conference on Knowledge Discovery and Data Mining (KDD - 95). Montreal,Canada. August 20 - 21,1995:112 - 117.

3 J. Mothe,C. Chrisment,T. Dkaki. Information mining - use of the document dimensions to analyse interactively a document set [Z]. European Colloquium on Information Retrieval Research, 2001:6 - 20.

4 Ghanem,M. Chortaras,A. Guo,Y. Rowe,A. Ratcliffe,J. A grid infrastructure for mixed bioinformatics data and text mining[J]. In:Computer Systems and Applications,2005,34(1):116 - 130.

5 H. Karanikas,C. Tjortjis,and B. Theodoulidis . An Approach to Text Mining using Information Extraction[Z]. In:Proceedings of the Fourth European Conference on Principles and Practice of Knowledge Discovery in Databases. Lyon,France. September, 2000:13 - 16.

6 M. Montes - y - Gómez,A. Gelbukh,A. López - López. Text Mining at Detail Level using Conceptual Graphs[Z]. In:Proceedings of the International Conference on Conceptual Structures. Borovetz,Bulgaria,2002:32 - 40.

7 Qinghua Hu, Daren Yu, Yanfeng Duan, Wen Bao. A novel weighting formula and feature selection for text classification based on rough set theory [Z]. In: Proceedings of Natural Language Processing and Knowledge Engineering, 2003: 638 - 645.

8 Catherine Blake and Wanda Pratt. Better Rules. Few Features: A Semantic Approach to Selecting Features from Text[Z]. In: proceedings of 2001 IEEE International Conference on Data Mining (ICDM '01). San Jose, California. November 29 - December, 2001 - 02.

9 Minoru Kawahara and Hiroyuki Kawano. An Application of Text Mining: Bibliographic Navigator Powered by Extended Association Rules[Z]. In: proceedings of 33rd Hawaii International Conference on System Sciences - Volume 2. Maui, Hawaii. January, 2000 - 07 - 04.

10 Roxana Girju and Dan Moldovan. Text Mining for Causal Relations[Z]. In: Proceedings of the International Florida Artificial Intelligence Research Society (FLAIRS 2002), Pensacola, Florida. May 2002.

11 Dekang Lin and Patrick Pantel. DIRT - Discovery of Inference Rules from Text[J]. Journal of Natural Language Engineering. Fall - Winter, 2001, (12): 22 - 31.

12 E. Bingham. Topic identification in dynamical text by extracting minimum complexity time components[Z]. In: Proceedings of ICA2001. 2001: 546 - 551.

13 Montes - y - Gómez, A. Gelbukh & A. López - López. Discovering ephemeral associations among news topics[Z]. In: Proceedings of IJCAI - 2001 Workshop on Adaptive Text Extraction and Mining. 2001: 216 - 230.

14 Pui Cheong Fung, G. Xu Yu, J. Wai Lam. Stock prediction: Integrating text mining approach using real - time news[Z]. In: Computational Intelligence for Financial Engineering, 2003 IEEE International Conference, 2003: 395 - 402.

15 Brian Lent, Rakesh Agrawal, and Ramakrishnan Srikant. Discovering trends in text databases[Z]. In: Proceedings of the Third International Conference on Knowledge Discovery and Data Mining. Aug, 1997. 227 - 230.

16 Harte, H. Lu, Y. Osborn, S. Dehoney, D. Chin, D. Refining the extraction of relevant documents from biomedical literature to create a corpus for pathway text mining[Z]. In: Proceedings of the 2003 IEEE: Bioinformatics Conference, 2003: 644 - 645.

17 Hang Li, Kenji Yamanishi. Mining from open answers in questionnaire data[Z]. In: Proceedings of the seventh ACM SIGKDD international conference on Knowledge discovery and data mining. San Francisco, California, 2001: 56 - 65.

18 林鸿飞, 战学刚, 姚天顺. 中文文本挖掘的特征导航机制[J]. 东北大学学报, 2000, 21(3): 240 - 243.

19 周　茜, 赵明生, 等. 中文文本分类中的特征选择研究[J]. 中文信息学报, 2004, 18(3): 17 - 23.

20 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, 37(5): 513 - 520.

21 刘昌钰, 等. 基于潜在语义分析的 BBS 文档 Bayes 鉴别器[J]. 计算机学报, 2004, 27(4): 566 - 572.

22 宋擒豹, 等. 基于关联规则的 Web 文档聚类算法[J]. 软件学报, 2002, 13(3): 417 - 423.

23 李颖基, 等. Web 日志中有趣关联规则的发现[J]. 计算机研究与发展, 2003, 40(3): 435 - 439.

24 阮备军, 等. 基于商品分类信息的关联规则聚类[J]. 计算机研究与发展, 2004, 41(2): 352 - 360.

25 卢娇丽, 等. 基于粗糙集的文本分类方法研究[J]. 中文信息学报, 2005, 19(2): 66 - 70.

26 刘云峰, 等. 基于潜在语义空间维度特性的多层文档聚类[J]. 清华大学学报(自然科学版), 2005, 45(1): 1783 - 1786.

27 袁毓林. 用动词的论元结构跟事件模板相匹配 —一种由动词驱动的信息抽取方法[J]. 中文信息学报, 2005, 19(5): 37 - 43.

28 鲁　川. 汉语语法的意合网络[M]. 北京: 商务印书馆, 2001: 1 - 38.

29 孙宏林, 等. 浅层句法分析概述[J]. 当代语言学, 2000, 2(2): 74 - 83.

(责任编辑: 徐　波)