Science Discovery
2017; 5(6): 438-443
http://www.sciencepublishinggroup.com/j/sd
doi: 10.11648/j.sd.20170506.18
ISSN: 2331-0642 (Print); ISSN: 2331-0650 (Online)

The Literature Review of Text Data Mining

Yanli Xu, Rong Zhao

School of Educational Technology Information, Central China Normal University, Wuhan, China

Email address:
13264723983@163.com (Yanli Xu), ruthzhao@163.com (Rong Zhao)

To cite this article:
Yanli Xu, Rong Zhao. The Literature Review of Text Data Mining. *Science Discovery.* Vol. 5, No. 6, 2017, pp. 438-443.
doi: 10.11648/j.sd.20170506.18

Received: September 30, 2017; **Accepted**: November 6, 2017; **Published**: November 21, 2017

Abstract: At present, the study of structured data analysis, researchers at home and abroad mainly focus on the learners in the network teaching environment, with diversified interactive learning mode, text based nonstructured data is generated continuously. In recent years, through the mining of text data to evaluate the learner's ability and knowledge of psychology and screening the behavior has become a new learning method. Firstly introduces the concept and technology of text data mining, then introduces the tools and methods of text mining in the mainstream, finally expounds the present situation of the application of text mining technology in natural and Social Sciences in the two fields and 6 application analysis, namely curriculum evaluation support learners, knowledge and ability, learning community groups, learning behavior of crisis early warning, forecasting learning effect and learning state visualization.

Keywords: Text Data Mining, Analysis Tools, Learning Analysis

文本数据挖掘综述

徐燕丽，赵蓉

教育信息技术学院，华中师范大学，武汉，中国

邮箱
13264723983@163.com（徐燕丽），ruthzhao@163.com（赵蓉）

摘要：当前，对于学习分析的研究，国内外研究者主要关注学习者在网络教学环境下产生的结构化数据，伴随学习交互模式的多元化，文本为主的非结构化数据正在不断生成。近年来，通过对文本数据的挖掘来测评学习者的知识能力以及甄别其心理与行为已成为一种新的学习分析方法。首先介绍了文本数据挖掘的概念和技术，然后介绍了文本挖掘主流的工具和方法，最后阐述了文本挖掘技术在自然科学和社会科学两大领域的应用现状以及在学习分析的6大应用，即课程评价支持、学习者知识与能力测评、学习共同体分组、学习行为危机预警、学习效果预测和学习状态可视化。

关键词：文本挖掘，分析工具，学习分析

1．引言

随着大数据时代的降临，文本挖掘成为了一个新兴的研究领域，它主要是从大量的、无结构的文本信息中发现潜在的、可能的数据模式、内在联系、规律、发展趋势等，抽取有效、新颖、有用、可理解的、散布在文本文件中的有价值知识，并且利用这些知识更好地组织信息的过程。近年来，文本挖掘在信息分析中的应用以及与特定领域的结合已经逐渐成为当前研究的重点[1]。

2．文本挖掘概述

2.1．文本挖掘的概念

文本挖掘的发展历史较短，是一个新兴的研究领域，1995年Feldman等将数据挖掘技术运用在非结构化数据上，第一次正式提出了文本挖掘的概念。在文本挖掘概念被正式提出后，有关文本挖掘的研究得到了快速的发展，研究主要围绕文本挖掘模型、文本特征抽取与文本表示模型、模式发现（如关联规则抽取、文本分类、文本聚类）等方面展开（陈建龙，论情报思维及其概念来源），文本挖掘的概念最早出现在20世纪80年代中期，它集成了自然语言处理和数据挖掘的部分技术与理念，至今已有30多年的历史。早期，文本挖掘经历了一个曲折而缓慢的起步过程，其科学性一度受到质疑和诟病。近20年来，随着计算机技术的突飞猛进和互联网技术的不断发展，文本挖掘这一领域取得了前所未有的进步和发展，逐渐成为一种主流方法论。

W. W. Cohen认为文本挖掘（Text Mining，TM）和文本数据库中的知识发现（Knowledge Discovery in Textual Database，KDD）具有相同的含义[2]。Pons-Porrata A等人认为文本挖掘，是指从非结构化的文本集合中提取兴趣信息和非检索信息或知识的过程[3]。文本数据挖掘隶属于数据挖掘这一交叉学科的一个具体研究领域，它的主要任务是在海量文本中发现潜在规律和趋势。与狭义的数据挖掘不同，文本挖掘有一个将自然语言文本转化为可为计算机处理的结构性数据的过程，这一过程在文本挖掘技术上称为文本的预处理，其主要目的是抽取代表文本特征的元数据[4]。

2.2．文本挖掘技术

文本挖掘技术本质上是一种人类语言技术，其以自然语言为分析对象，通过区分文本中可识别性要素和语法结构（包括术语、事实、断言以及其他语言形式），建立起包含于文本中的某种可为计算机处理的概念与关系类型。袁金鹏提出的文本挖掘的一般过程中认为，对于文本数据挖掘分为以下几个阶段：对文本数据预处理、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1为文本挖掘的一般过程的示意图[4]。

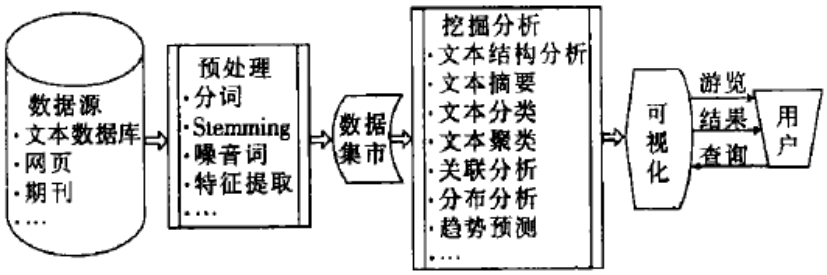

图1 文本挖掘的一般过程。

目前，常用的技术主要包括数据转换、去停用词、中文分词、特征提取和表示、文本的分类、文本的聚类、自动摘要和可视化等。

3．文本挖掘工具

3.1．商业文本挖掘工具

近年来，国内外文本挖掘技术发展较快，许多技术已经进入商业化阶段。各大数据挖掘工具的提供商也都推出了自己的文本挖掘工具。这些工具除具备常规的文本挖掘功能（如数据预处理、分类、聚类和关联规则等）外，针对庞大的、非结构化数据都能做出较好的应对，支持多种文档格式，文本解析能力强大，大部分支持通用数据访问[5]，但是价格都十分昂贵。由于个提供商的专注领域或企业背景不同，工具的定位和适用性也有所不同[6]。以目前市面上比较流行的10款商业文本挖掘工具为对象，针对其不同点进行简要的分析比较，如表1所示：

表1 10款商业文本挖掘工具。

工具名称	提供商	工具简介
Intelligent Miner for Text	IBM	挖掘结果展现能力较强，系统具有可扩展性，但是缺乏统计方法，限制了其本身的挖掘能力。在连接除DB2以外的数据库时，需要安装中间件。图形界面不友好且操作复杂，适合专业人员。
Text Miner	SAS	算法齐全，360°数据视图展示。提出SEMMA方法论。用户界面灵活友好，但是操作复杂，分析结果难以理解，适合专业人员。
Text Mining	IBM SPSS	提出Crisp-DM方法论。图形界面非常友好，易于操作，支持脚本功能，应用领域广泛且维护和升级成本较低。但是缺少最新的统计方法，且分析结果与其他软件的交互性较弱。
IDOL Server	Autono-my	基于贝叶斯概率论和香农信息论。工具性能较高，支持SOA，提供完全可配置的监控。但是系统的维护与管理缺乏相应的图形化应用界面，且工作过程中没有相关报告输出。
Darwin	Oracle	通过ODBC访问数据，提供wizard引导用户构建模型。可扩展性较高，模型能够作为C、C ++和Java代码导出并集成于其他应用，用户界面友好。但是工具的适用面窄，市场份额较小;数据展示需要额外的工具，交互性差。
SQL Server	Microsoft	基于OLAP，利用数据源系统对数据进行清洗、转换和加载。挖掘功能集成于SQL Server系列产品中，易于使用。但是由于算法不足，解决问题有限，只适合小型业务。

工具名称	提供商	工具简介
Clear Forest	Thomoson Reuters	基于文本挖掘的专利分析工具，有自己的专属领域。数据的前期处理能力较强，但在分类、聚类、关联等方面算法简陋。分析结果以列表、矩阵和聚类图呈现。缺少基本的统计方法和引证分析。
Themescpease	Cartia	同样是以专利文档为基础数据，通过标准的文本挖掘流程，生成强大的主题（词汇）地形图，拥有高级神经网络技术和统计分组技术。系统响应速度较快，分析结果交互性强。
方正智思	北大方正	技术研究院支持二次开发，具有良好的可扩展性。框架设计灵活，功能模块相对独立。
TRS文本挖掘软件	北京拓尔思（TRS）信息技术有限公司	基于统计原理的自动分类和基于语义规则的规则分类、自动过滤、政治常识校对以及标准的文本挖掘技术。系统性能较高，文本分析速度快。

3.2．开源文本挖掘工具

目前开源文本挖掘较多，但大部分工具由于其固定的算法只适用于特定的场景，应用范围较窄，与其相关的文献资料极少，故不纳入本文的比较范围。本文对10款较具普适性的主流开源工具进行了比较，如表2所示：

表2 10款主流开源工具。

工具名称	开发者	开发语言	操作系统
Weka	新西兰怀卡托大学	C/ C + +	跨平台
GATE	谢菲尔德大学自然语言处理研究小组	JAVA	跨平台
Orange	斯洛文尼亚卢布尔雅那大学计算机与信息科学学院人工智能实验室	C + +	跨平台
Bow	卡内基梅隆大学Mc Callum团队	C	Linux / Unix
Mallet	马萨诸塞大学Andrew Mc Callum团队	Java	跨平台
UIMA	Apache继承IBM UIMA	C + + / Java	跨平台
Ling Pipe	Alias	Java	跨平台
LIBSVM	台湾大学林智仁团队	Java、Matlab、C #、Ruby、Python、R、Perl、Common LISP、Labview	跨平台
Open NLP	Apache	Java	跨平台
ROST CM	武汉大学ROST团队	C + + / C#	Windows

大部分商业文本挖掘工具都对多语言、多格式的数据提供了良好的支持，且数据的前期处理功能都比较完善，支持结构化、半结构化和完全非结构化数据的分析处理。开源文本挖掘工具一般会有自己固有的格式要求，国外开源文本挖掘工具对中文的支持欠佳，而且大部分开源工具仍然停留在只支持结构化和半结构化数据的阶段。商业文本挖掘工具的分类、回归、聚类和关联规则算法普遍都较开源文本挖掘工具齐全，包含了目前主流的算法，只是每个工具在算法的具体实现上存在差异。同时，前者在处理庞大的数据量时依旧能够保持较高的速度和精度，后者则显得有些望尘莫及。

目前文本挖掘还处于探索发展的阶段，其中商业文本挖掘工具的发展要快于开源文本挖掘工具。不过，任何事物都有其两面性，大部分商业软件由于其高质量和稀缺性而非常昂贵，不适合小企业和科研机构。优秀的开源文本挖掘工具则能在最大程度上满足相关需求，并且还能够支持加载使用者自己扩充的算法，或者直接嵌入到使用者自己的程序当中去。

Weka以算法全面得到了许多数据挖掘工作人员的青睐，Ling Pipe是专门针对自然语言处理开发的工具包，LIBSVM是SVM模式识别与回归的工具包，ROST CM在各大高校应用面非常广，对中文的支持最好。ROST是由武汉大学沈阳博士ROST虚拟学习团队研发的一款内容挖掘软件，可以对数字化的材料进行组织、标引、检索和利用，具有海量性、智能性和客观性等特点，通过定量分析和定性分析的结合，ROST文本挖掘软件能从数字化的材料中归纳出具有说服力的普遍性结论。ROST文本挖掘软件可以对各类文本进行词频、聚类、分类、情感等分析[7]。

3.3．学习分析领域文本挖掘工具

通过对文献的梳理，在学习分析领域，国内学者研究文本挖掘技术的应用还比较少，但是在国外，文本挖掘技术的应用已比较成熟。通过对国外文献的梳理，国外目前在学习分析领域应用的文本挖掘工具主要有LIWC、Cohere、Sobek、Rapid Miner、Dissertation Browse、Edu Miner、GCS、Toreador等，这些工具各有特色，对这些工具的对比概述如表3所示：

表3 文本分析工具概述。

工具	特色点	目的
LIWC	语词分析、词频统计、心理评估	通过统计学习者文本表述中情感词的频率来识别其持有的态度和观点；可分析学习者不同形式写作内容中（论文、自我介绍短文、日记等）的语词构造和心理意义，揭示学习者自我管理学习和反思过程。
Cohere	语义连接、社交网络、可视化	使用该工具学习者可以将注释文本作为一种信息资源进行索引和检索，现语义连接；可发现持有相似观点的学习者，建学习者在线交互的协作网络概念模型，在集体智慧最大化。
Sobek	可视化	从学习者短文写作中实现相关概念提取，点、关系、事实和事件的发现，以交互式图形可视化呈现，而支持学习者和教学者任务的顺利开展。
Rapid Miner	观点挖掘	收集学习管理系统中学习者关于课程评论的文本数据来识别其对平台功能、教学设计、学习持有的态度和观点。

工具	特色点	目的
Dissertation Browse	相似性检测、可视化	可视化呈现不同学科间学术论文的相似性检测结果。
Edu Miner	实时反馈、可视化	实现对学习过程的自动形成性评估；在动态的追踪学习者交互过程中，用该工具不仅可减轻教学者的压力，且能够及时将交互绩效的可视化结果反馈给学习者，习者进行自我调整，助于促进在线学习者对话交流过程中完成更深层次的表达。
GCS	分类	对课程管理系统讨论区中产生的帖子进行分类，助教学者监控和调整其在线交互活动。
Toreador	难度评定、标注	根据学习者年龄、性别、先验知识水平的差异，估在线阅读资料的难度系数，测并标注对于学习者的难度词汇；主动推送与其阅读水平更加吻合的阅读内容，发其学习兴趣。

4．文本挖掘分析方法

文本挖掘是一个方法群，涉及众多领域，是典型的信息分析过程[8]，其在信息分析中的应用研究是一项跨学科的应用研究。近年来，文本挖掘在信息分析中的应用领域越来越广泛，特别在计算机、生物化学和社科情报等领域有着较多的应用，因而其研究方法也更多地借鉴了其他学科的方法，呈现多样化趋势，其中数理统计方法、机器学习方法等是该领域研究的主要方法[9]。

通过对文献的研究发现，学者们在对文本数据进行分析时，所用到的方法主要有以下几种：分类、聚类、关联规则分析、语义分析、可视化、话语分析、内容分析等7种方法。其中聚类的方法是应用最广泛的。

所谓文本聚类，实际上就是把一个文本信息构成的信息集合，执行内容上的分组处理.经过这个分组应该达到的情况是，组内的文本信息从内容上高度相似，而组间文本信息的关联性则应该尽可能的低[10]。对现有的文本聚类方法进行分类，大概可以分成5大类：第一种是基于划分的文本聚类方法，第二种是基于层次的文本聚类方法，第三种是基于密度的文本聚类方法，第四种是基于网格的文本聚类方法，第五种是基于模型的文本聚类方法。在这5种方法中，基于划分的方法原理简单并且易于实现，算法的复杂度也比较低，因此应用也最为广泛。

K均值聚类方法就是一种典型的基于划分的文本聚类方法，这种方法的思路是：在原始文本信息集合中，随机选取k个文本信息作为这个集合的中心，之后计算其它文本信息和这k个中心的距离并比较这些距离的远近，和哪个中心的距离最近，就将这个文本信息归入到以最近中心为标志的分类之中。

5．文本挖掘技术的应用

5.1．文本挖掘技术在科学领域的应用

当前，国际上文本挖掘的主要研究成果集中在自然科学领域，成绩令人瞩目。在人文社会科学领域的应用则相对薄弱，不仅数量上不及自然科学领域的25%，质量上也存在一定差距。在人文社科领域，国外学者及其研究领域是文本挖掘应用与实践的主流；反观国内，其应用研究在数量和质量上都与国外差距甚大。不过，CNKI反映的文献增长趋势与国外研究过往类似，说明该领域正越来越多地受到国内相关学者的重视。值得一提的是，就人文社会科学而言，国内外研究的侧重点区别较大。国外的相关研究主要集中于社会科学，目前应用比较成熟的学科包括经济学、管理科学、教育学、行为学、心理学、公共管理学、法学、社会学等；国内的研究侧重于文史哲等人文科学，社会科学领域的相关成果不多。

5.2．文本挖掘技术在学习分析中的应用

通过梳理文献可知，华中师范大学刘三女牙等人在文本挖掘应用于学习分析领域的研究成果较多，且提出了文本挖掘应用于学习分析的系统框架，如图2 [11]所示：

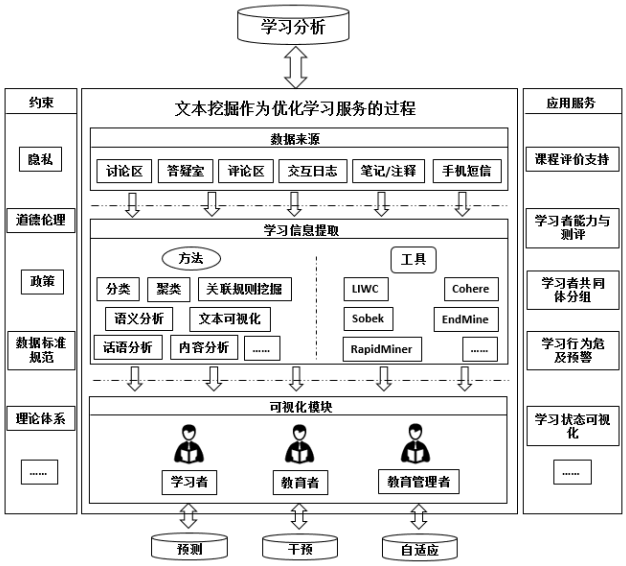

图2 文本挖掘应用于学习分析的系统框架。

通过对文献的梳理，文本挖据技术在学习分析中的应用主要体现在以下几个方面：课程评价支持、学习者知识与能力测评、学习共同体分组、学习行为危机预警、学习效果预测和学习状态可视化等6个方面。

5.2.1．课程评价支持

在国外有许多研究者应用文本挖掘技术来实施对课程的评价。如一项探讨挖掘学习者反馈的教学文本情感的研究[12]提到，在每门课程结束后，学习者通过移动设备发送短消息来评价课程教学。该研究旨在利用短文本内容和表情符号来呈现课程评价的积极和消极内容。与此类似的一项研究[13]中，Kontogiannis等人提出了一种新的课程教学评价框架来自动挖掘学习者情感观点，他们通过收集学习者在社区网络微博客中产生的有关学习课程活动的文本数据，并利用观点挖掘技术来判断学习者对每门课程所持有的积极或消极态度。

5.2.2．学习者知识与能力测评

相比传统研究中利用学习者的客观题解答结果来评估知识掌握程度，我们可以利用学习者在教学活动中产生的主观文本数据来测评其知识结构、高级思维技能等。如在智利大学，一项基于情境化模型和潜在语义特征的自动化文本理解分类器研究[14]中，以来自工程学和语言学学院的大学一年级学生为实验对象，通过分析学习者关于阅读资料问题的文本作答来检测发现其阅读理解能力的不足。此外，就学习者的阅读理解任务而言，也有相关研究描述了如何结合多维K-means聚类方法和布鲁姆教育分类理论来确定他们采取的积极和消极的认知策略。

5.2.3．学习共同体分组

在大规模在线学习的大环境之下，学习者共同体的组建对学习者的学习效果有着很大的促进作用。如加拿大英属哥伦比亚大学，一项辨识学习者在线交互行为模式的研究[15]中，研究者基于学习者与在线学习平台的交互日志进行分析，构建用户模型框架，对具有相同学习偏好、兴趣、主题等的学习者进行聚类和分组，供合适的学习支持和交互体验。

5.2.4．学习行为危机预警

根据言语行为理论[16]，人们在书写文字的同时也在实施某种行为，文字中传递出来的言语信息可反映个体的意志和心理行为现象，为教师及时掌握学习者的思想状况及行为危机预警提供可靠的决策支持。在澳大利亚昆士兰大学，研究者使用SNAPP工具来分析学习者与讨论区的交互活动，旨在识别高风险状态的学习者，引导教学者及时采取干预措施。

5.2.5．学习效果预测

研究者通过采集和分析学习者学习经历相关文本数据，探究学习者在不同阶段中学习效果的改变，发现学习者学习行为与学习效果的相关关系。如在西班牙科尔多瓦大学计算机科学与数据分析学院提出了从定量、定性和社交网络等三个测量角度来评判学习者在讨论区内的参与行为，并结合不同的文本挖掘方法来提高学习者最终学习效果预测的准确度。实验结果表明，于学习者在讨论区内产生的文本数据，用聚类及关联规则分析方法有助于预测学习者课程通过情况。

5.2.6．学习状态可视化

使用文本数据的交互式图形来评价学习者的知识结构、认知能力、情感态度等多维度的状态特征，助于教学者的教学决策和学习者的自我学习监控。如台湾一项分析学习者知识构建过程的研究[17]中，以56个来自信息管理专业的研究生为实验对象，通过分析处理学习者在在线学习社区交互过程中产生的文本内容，按照布鲁姆的教育目标分类法，知识层次、领会层次、应用层次、分析层次、综合层次和评价层次等6个方面为教学者和学习者呈现实时可视化的认知能力评估图形。研究结果指出，评估方法可通过有效激发学习者的学习动机来发布高阶认知层次的内容，促进学习者之间更深层次的对话交流。

6．结论

综上所述，文本数据挖掘最早被国外提出，因此对文本挖掘的研究，无论从数量上还是质量上来说，国外的研究成果已远远超出国内，但就对目前的文献研究发现，对文本数据挖掘的研究还尚未成熟，都是一个在不断摸索的过程之中，文本挖掘的概念已达成共识，但在文本挖掘工具的研究上还比较缺乏，普遍适用的工具少之又少，尤其是缺乏对中文数据进行处理的工具。在对文本数据进行分析的方法上还比较单一，对文本内容的分析比较浅显，不够深入，没有真正挖掘出文本内容的潜在价值。在学习分析应用方面，国外应用较为领先，国内在这方面的研究还比较缺乏，总之，从国内外的研究现状来看，文本数据挖掘还处于一个发展的初期。

参考文献

[1] 李尚昊，朝乐门. 文本挖掘在中文信息分析中的应用研究述评[J]. 情报科学，2016,(08):153-159。

[2] W. W. Cohen. What can we learn from the web? In proceedings of the Sixteenth International Conference on Machine Learning (ICML'99), 1999, 515-521.

[3] Pons-Porrata A, Berlanga-Llavori R, Ruiz-Shulcloper J. Topic discovery based on text mining techniques[J]. Informa‑tion Processing & Manmanagement,2007, 43(3): 752-768.

[4] 袁军鹏，朱东华，李毅，李连宏，黄进. 文本挖掘技术研究进展[J]. 计算机应用研究，2006,(02):1-4。

[5] 钱峰. 国内数据挖掘工具研究综述[J]. 情报杂志，2008,(10):11-13。

[6] 王敏，李海存，许培扬. 国外专利文本挖掘可视化工具研究[J]. 图书情报工作，2009,(24):86-90。

[7] 蔡溢，杨洋，殷红梅. 基于ROST文本挖掘软件的贵阳市城市旅游品牌受众感知研究[J]. 重庆师范大学学报(自然科学版)，2015,(01):126-134。

[8] 范并思. 社会科学信息分析中的文本挖掘[J]. 图书情报工作，2012,56(8):6-9。

[9] 李尚昊，朝乐门. 文本挖掘在中文信息分析中的应用研究述评[J]. 情报科学，2016,(08):153-159。

[10] 魏桂英，高学东，武森. 基于领域本体的个性化文本信息检索[J]. 辽宁工程技术大学学报(自然科学版)，2011,(02):316-320。

[11] 刘三女牙，彭晛，刘智，孙建文，刘林，郑年亨. 基于文本挖掘的学习分析应用研究[J]. 电化教育研究，2016,(02):23-30。

[12] Leong, C-K., ee, -H., ak, -K. Mining Sentiments in SMS Texts for Teaching Evaluation [J]. Expert Systems with Applications, 2012, 39(3): 2584~2589.

[13] Kontogiannis, Valsamidis Kazanidis et al. Course Opinion Mining Methodology for Knowledge Discovery, Based on WebSocial Media [A]. Proceedings of the 18th Panhellenic Conference on Informatics [C]. New York: ACM Press, 2014: 1~6.

[14] Bravo –Marquez, L'Huillier, Moya, et al. An Automatic Text Comprehension Classifier Based on Mental Models and Latent Semantic Features [A]. Proceedings of the 11th International Conference on Knowledge Management and Knowledge Technologies [C]. New York: ACM Press, 2011: 158~162.

[15] Kardan, Conati, A Framework for Capturing Distinguishing User Interaction Behaviors in Novel Interfaces [A]. EDM [C]. New York: ACM Press, 2011: 159~168.

[16] L. J. Austin. How to Do Things with Words [M]. Oxford: Oxford University Press, 1962.

[17] C Hsu. L., Chou, W., Chang,. H.. Edu Miner: Using Text Mining for Automatic Formative Assessment [J]. Expert Systems with Applications, 2011, 38(4): 3431~3439.