Wu, Y.-T., Chang, M., Li, B., Chan, T.-W., Kong, S. C., Lin, H.-C.-K., Chu, H.-C., Jan, M., Lee, M.-H., Dong, Y., Tse, K. H., Wong, T. L., & Li, P. (Eds.). (2016). *Conference Proceedings of the 20th Global Chinese Conference on Computers in Education 2016*. Hong Kong: The Hong Kong Institute of Education.

大数据背景下的学习分析及预测方法研究综述

The Review of Learning Analysis and Forecast Method under the Background of Big Data

高丹丹，马颖莹*，李曼曼
华东师范大学 教育信息技术学系
*yingyma@foxmail.com

【摘要】自21世纪以来，人、机、物三元世界的高度融合引发了数据规模的爆炸式增长，越来越多的学习研究开始向数据分析方面倾斜，但实际研究效果并不理想，针对大数据背景下的学习分析及预测的系统性介绍也相对较少，本文利用文献综述法、调查研究法找出大数据背景下学习分析及预测的方法，以期为未来的研究者对在线学习平台中学习者学习所产生的海量数据进行分析时提供借鉴。

【关键词】大数据；数据挖掘；学习行为分析；行为预测

Abstract: *Since the 21st century, human, machine, material highly integrated sparked the explosive growth of data scale. There are more and more study and research began to tilt to the data analysis, but the actual research results are unsatisfactory. For the systemic introduction of learning analysis and forecasting under the background of big data is relatively small, we use literature review, survey research study to find out the learning analysis and forecasting under the background of big data and specific implementation steps, with a view to future researchers for reference when they analysis huge amounts of learning data of the learning platform.*

Keywords: Big data; data mining; learning analytics; behavior predict

1. 前言

人、机、物三元世界的高度融合以及互联网的飞速发展，云计算、物联网、大规模在线课程、社交网络等新兴网络媒介服务的兴起，促使人们越来越多的行为在网络中发生，这直接导致互联网中人类相关数据呈爆炸式增长，人类在不知不觉中已经进入了一个“大数据”时代。2012年3月29日美国政府拨款2亿美元推出的“大数据的研究和发展计划”,2013年2月6日，美国国家科学基金会（National Science Foundation）宣布将额外投入1千万美元以激励社会及人文科学中的“大数据”研究的发展（徐鹏等，2013）。显然在一些发达地区大数据背景下的研究已提升到了国家战略性的地步。国内虽越来越多的学习研究开始向数据分析方面倾斜，但实际研究效果并不理想，截止2014年12月底，我国网民规模达6.49亿，其中学生群体是网民中规模最大的群体（中国互联网中心，2014），利用教育数据挖掘技术和学习分析技术对学生网上学习行为所产生的有效数据进行分析，能够更加深入地了解学生学习行为及趋势，从而建立学生学习自适应系统并实现真正意义上的个性化学习。

2. 基本概念的界定

2.1. 大数据的概念

大数据本身就是一个相对较抽象的名词，就字面上意义来看，很容易就将其理解为大规模的庞大数据集，但目前众说纷纭的定义中较有代表性的为3V定义（黄升民,&刘珊，2012），即认为大数据需满足3个特点：规模性（volume）多样性（variety）和高速性（velocity），也有学者认为符合4V的才能被称之为大数据，即在原来3V的基础上加上价值性（value），维基百科对大数据的定义：大数据是指利用常用软件工具捕获、管理和处理数据所耗时间超过可容忍时间的数据集（孟小峰和慈祥，2013）。但在搜集、查询及观看了众多文献之后笔者认

为：大数据不能简单地看作是大规模数据的集合，它最主要的功能在于基于庞大数据所展现的信息体征。

2.2. 学习分析的概念

在首届“学习分析技术与知识国际会议”上将学习分析被定义为“测量、收集、分析和报告有关学习者及其学习情景的数据集，以理解和优化学习及其产生的环境的技术”（徐 鹏等，2013），新媒体联盟将学习分析定义为“利用松散耦合的数据收集工具和分析技术，研究分析学习者学习参与度，学习表现张力以及学习过程中所产生的相关数据，从而对课程教学资源以及课程安排力度进行重新调整。”（顾小清,张进良,&蔡慧英等，2012）《2012NMC 地平线报告（高教版）》给出的定义为“学习分析技术是对学生在学习过程中产生的海量数据进行解释分析，以评估学生的学术进展，预测未来表现，并发现潜在的问题”（魏顺平，2013）。从这些定义可以看出学习分析实质上是对学习者在学习过程中所产生的海量数据，进行系统性的分析，从而总结出学习者在整个学习过程中的学习习惯、学习倾向，学习进度等来了解学习者的一切学习现状，以实现掌握学习者学习规律及预测学习者学习行为与表现的目的。

3. 大数据背景下学习分析方法

学习分析及预测的方法主要涉及内容分析、话语分析、社会网络分析等一系列的可将数据可视化的方法。

3.1. 内容分析法

内容分析法（Content analysis）是一种以系统客观的、量化方式对传播内容进行描述和解释的研究方法（李艳燕、马韶茜和黄荣怀，2012）。其通过对传播内容以及传播内容所起的作用加以归类，对学习者的学习过程数据进行编码量化，然后再对其所产生的海量数据进行定量分析，以数据结果来描述学习者的学习行为特征，从而寻求学习者的行为模式，同时还可以对其进行定性分析，运用已有累积的经验来预测当前的学习行为，为学习者提供一个个性化的学习资源服务，可视图如图1所示。

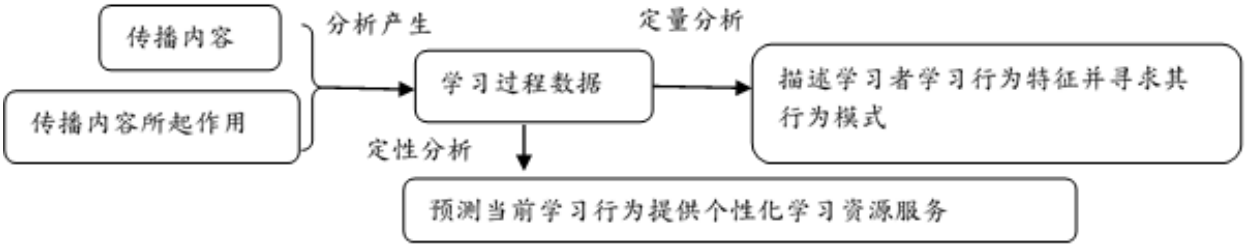

图1 内容分析法可视图

3.2. 话语分析法

话语分析法（Discourse Analysis）是对学习交流过程进行分析的方法（李艳燕、马韶茜和黄荣怀，2012）。其最早应用与语言学中，主要对话语结构形式，规则等进行研究，后经过发展也开始运用到教育研究中（李青和王涛，2012）。其分析对象为面对面的对话内容、网络课程与会议中产生的文本内容、通过各种网络媒介实时或异步传播的交流内容以及数字符号等。通过话语分析我们可以了解学习者在学习交流过程中话语的文本性含义，了解学习者是如何建立起自己的观点，从而来探究他们的知识建构过程，这将会使学习者学习行为发生的过程更加清晰，可视图如图2所示。

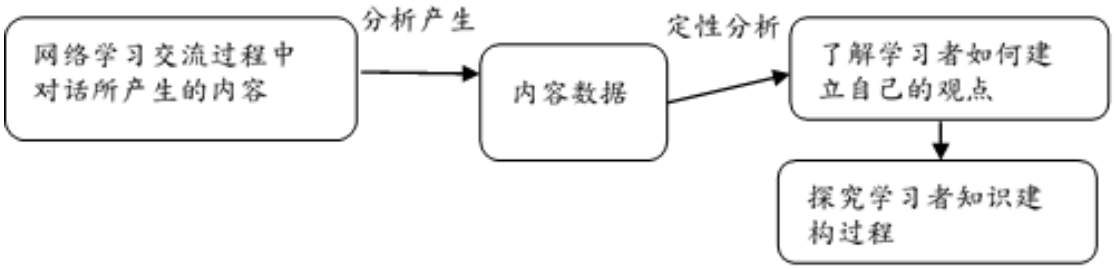

图2 话语分析法可视图

3.3. 社会网络分析法

社会网络分析法（Social Network Analysis）是由社会学家根据数学方法、图论等发展出的定量分析方法，其以关系作为基本分析单位（李青和王涛，2012），探究网络学习过程中的联系（ties）关系、角色以及网络形成的过程与特点，从而了解人们如何在网络学习中建立并维持关系从而为自己的学习提供支持。其以学习者与他人互动交流的频率及内容所产生的数据为基础，分析学习者与伙伴学习关系、学习方向取向、学习瓶颈等，了解并预测整体学习者学习信息分布以及学习进展情况等，可视图如图3所示。

图3社会网络分析法可视图

4.大数据背景下的学习预测方法

4.1. 时间序列预测法

时间序列预测法是一种根据动态数据揭示系统动态结构和规律的统计方法，时间序列的获取是通过对已有数据进行分类汇总后得到的，获取时间序列数据以后可以对其进行预测，从而准确地预测系统的演进（赵仁义&朱玉辉,2011）。在大数据背景下对学生的学习过程或学习结果进行预测时使用时间序列法可将随着时间t变化的量，即在t1<t2<…<tn<…时所产生的阶段性学习结果的观测值记为y（t1）,y（t2）,…,y（tn），…组成离散有序集合，称为一个时间序列，记作｛y（t）｝。由于不同的学生在学习过程中的时间序列是由某一随机过程产生的，因此可将基于时间序列的学习预测记作公式：y（t）=f（t）+p（t）+x（t）,其中f（t）为趋势项反映y（t）的变化趋势；p（t）为周期项，反映y（t）不同阶段的周期变化；x（t）为随机项，反映随机因素对y（t）的影响。

用时间序列法对学习过程或结果的预测是指以学习者在不同时间段的学习数据为基础，运用时间序列算法对数据进行分析预测，从而判断学习者下一步可能发生的学习行为，该方法具有一定的自适应性，能更好地预测学习者在学习过程中成功或者失败的可能。

4.2. 决策树预测法

决策树预测法是通过对学生在学习过程中所产生的活动过程序列节点数据、个人交互序列节点数据、学习内容序列节点数据、知识序列节点数据进行计算分析。在该预测算法上采用ID3算法，ID3主要针对的是属性选择的问题，是决策树算法中一种典型的算法。它主要包括四个步骤，第一创建节点，即对学习过程中所产生的数据结果进行归类，若数据样本都在同一类中则该节点为树节点，并标记该类；第二若数据样本不在同一类中，则选择一个能够很好的将样本集中分类的属性，将该属性作为测试属性；第三依据测试属性设置属性值划分样本数据；第四使用同样的过程自上向下进行递归。在选择树叉时ID3采用信息增益来进行计算，

其计算方式为I=-$\sum_{i=1}^{m} p_i log_2（p_i）$,其中i为目标变量，m为不同属性值，$p_i$是任意样本属于m个类别中的概率（武法提&牟智佳，2016）。

用决策树来对学习过程及结果进行预测，分析速度快，计算量小，同时也容易转化成分类规则，准确性也相对较高，从挖掘出来的节点分类及属性能够很明显的看出哪些字段比较重要，以及学生习过程中的关键点。

5. 结束语

我们已经进入了一个“以数据为基础，分析实现个性化学习”的大数据时代，大数据必将改变传统教育的面貌，建立在大数据上的学习分析能够更好更快更直接的分析学生学习特点，掌握学生学习规律，预测学生学习趋势，提供学生个性化的学习资源，以实现个性化学习，实现教育公平的目的。本研究主要通过对学习分析及预测的方法进行归类汇总，以期为以后研究着的研究提供借鉴。

参考文献

三川（2015）。Cnnic发布第35次《中国互联网络发展状况统计报告》。**中国远程教育**，**2**，31-31。

赵仁义和朱玉辉（2011）。关于时间序列预测法的探讨。**科技信息**，**15**。

李青和王涛（2012）。学习分析技术研究与应用现状述评。**中国电化教育**，**8**，129-133。

李国杰和程学旗（2012）。大数据研究：未来科技及经济社会发展的重大战略领域——大数据的研究现状与科学思考。**中国科学院院刊**，**27**（6）。

李艳燕、马韶茜和黄荣怀（2012）。学习分析技术：服务学习过程设计和优化。**开放教育研究**，**18**（5），18-24。

徐鹏、王以宁、刘艳华和张海（2013）。大数据视角分析学习变革*--美国《通过教育数据挖掘和学习分析促进教与学》报告解读及启示。**远程教育杂志**，**6**，11-17。

张杰夫（2013）。大数据大视野大教育。**中小学信息技术教育**，**10**，12-14.。

顾小清、张进良和蔡慧英（2012）。学习分析：正在浮现中的数据技术。**远程教育杂志**， **30**（01），18-25。

黄升民和刘珊（2012）。“大数据”背景下营销体系的解构与重构。**现代传播：中国传媒大学学报**，**34**（11）， 13-20。

孟小峰和慈祥（2013）。大数据管理：概念、技术与挑战。**计算机研究与发展**，**50**（01），146-169。

武法提和牟智佳（2016）。基于学习者个性行为分析的学习结果预测框架设计研究。**中国电化教育**，**1**。

魏顺平（2013）。学习分析技术：挖掘大数据时代下教育数据的价值。**现代教育技术**，**23**（02），5-11。

Feng, Y. X. (2013). Big data research. *Computer Technology & Development*.

Wever, B. D., Schellens, T., Valcke, M., & Keer, H. V. (2006). Content analysis schemes to analyze transcripts of online asynchronous discussion groups: a review. *Computers & Education*, *46*(1), 6-28.