基于文本挖掘的e-Learning学习评价研究

刘革平[1] 黄智兴[2] 李立新[2] 邱玉辉[2]

(西南师范大学网络教育学院 重庆400715)[1] (西南师范大学计算机与信息科学学院 重庆400715)[2]

摘 要 在e-Learning系统中,对学生学习的评价难以进行。在分析e-Learning环境下学习评价特征的基础上,本文引入电子学档评价方法,提出将文本挖掘技术运用于学习评价,依据学生学习评价量规,实现对学生学习过程的评价。

关键词 文本挖掘,e-Learning,学习评价,评价量规,电子学档

Study of Text Mining Based Learning Evaluation of e-Learning System

LIU Ge-Ping[1] Huang Zhi-Xin[2] Li Li-Xin[2] Qiu Yu-Hui[2]

(College of Net Education, Southwest China Normal Vniversity, Chongqing 400715)[1]

(College of Computer Information Science, Southwest China Normal University, Chongqing 400715)[2]

Abstract It is very difficult to evaluate the learning of the students in e-Learning system. Based on analyzing the characters of learning evaluation in e-Learning system, this research fetched in the assessment method of e-Portfolio. And then, an idea has been put forward applying Text Mining to learning evaluation. According to the rubric, the evaluation of learning will be accomplished.

Keywords Text mining, E-Learning, Evaluation of learning, Rubric, e-Portfolio

1 引言

"知识爆炸"的信息社会要求人们不断更新知识。e-Learning为人们在信息技术条件下获取知识提供了一条行之有效的途径:人们可以不离职、不离岗,借助计算机、网络等信息工具进行学习。e-Learning现已成为学校教育、企业培训、机关研修的有效形式。

学生自主学习、分散学习是e-Learning的显著特征,这一特征体现了现代教育理论以学生为中心的理念,符合建构主义学习理论意义建构的观点,可以更好地适应学生的个性心理特征。然而,由于学生与教育者在地域上的分离,在e-Learning系统中难以进行对学生学习的评价。

本研究试图将文本挖掘技术应用到e-Learning系统的学习评价中来,利用文本特征抽取、关联等方法,对记录学生学习过程的电子学档进行挖掘与分析,提取学习过程的关键信息,参照学生学习评价量规,从而实现对学生学习的评价。

2 e-Learning环境下的学习评价

研究表明,电子学档(e-Portfolio)正逐渐成为e-Learning系统的有效学习模式和评价方式。从评价的角度来看,电子学档是运用数据库和超文本技术清晰地展现学生学习目标、学习要素、学业作品与自我反思的一种评价手段[1,2]。

本研究中,为体现e-Learning环境下电子学档的特点,我们为电子学档设计了学习计划、学习记录、自评与互评和学习统计四个模块,主要包含以下项目:本学期学习计划、阶段学习记录、学习方法、学习媒体、学生作品、自我评价、他人评价等。

基于电子学档的评价是一种质性评价,利用它可以记录学生在e-Learning系统中的学习过程,进而通过教师对电子学档的评定实现对学生学习过程的评价。然而,在学生数量巨大的e-Learning系统中,这种通过教师对学生的电子学档进行手工评价的方法显然是没有实用价值的。如果引入文本挖掘技术,对学生记录信息进行分析与评判,则可以大大提高学习过程评价的效率。

3 面向中文的文本挖掘研究

电子学档中的记录信息以文本为主,而文本是非结构化的数据。与关系数据库中的结构化数据相比,这种文本数据没有结构、缺乏组织规律性。可以通过技术手段,将这些文档转化为一种类似关系数据库中记录的、较规则且能反映文档内容特征的中间形式,一般采用文本特征向量表示法。然后,利用数据挖掘的相关算法来提取知识模式。针对英语的这方面工作已经取得一定的进展。面向中文的文本挖掘,由于中文本身的形式化建模程度低,目前尚存在很多困难。本研究中,先对中文文本进行分词等预处理,然后再采用相关算法进行文本挖掘[4]。

3.1 中文分词

中文分词[3]是将中文句子中的词语用切分标志分隔开来。本研究中采用目前常用的正向最大匹配算法(Forward Maximum Matching, FMM)。

设句子:$S=c_1c_2\cdots c_n$,n为句子中的汉字数。

假设词:$w_i=c_1c_2\cdots c_m$,m为词典中最长词的字数。令$i=0$,当前指针p_i指向输入字串的初始位置,执行下面的操作:

1)计算当前指针p_i到字串末端的字数(即未被切分字串的长度)n,if $n=1$,转3)。否则,令$m=$词典中最长词语的字数,if $n<m$,$m=n$;

2)从当前指针p_i起取m个汉字作为词w_i,作如下判断:

i)如果w_i确定是词典中的词,则在w_i后添加一个切分标志,转iii);

ii)如果w_i不是词典中的词且w_i的长度大于1,将w_i从

右端去掉一个字，转2)中的i)步；否则(即 w_i 的长度等于1)，则在 w_i 后添加一个切分标志，将 w_i 作为单字词添加到词典中，执行iii)；

iii)根据 w_i 的长度修改指针 p_i 的位置，如果 p_i 指向字串末端，转3)，否则，$i=i+1$，返回1)；

3)输出切分结果，结束分词程序。

本研究的系统实现时，将正向最大匹配算法与基于统计模型的分词方法结合在一起来使用，可以达到较高的切分精度。

3.2 文本的特征表示

文本特征表示[5,6]是指以一定的规则和描述表示文本或文本集，是文本挖掘的基础。本研究采用的文本特征表示法是TFIDF向量表示法。所有的词从文本中抽取出来，而不考虑词间的顺序和文本的结构，从而构成一个二维数据表，其中列集为特征集，每一列是一个特征；行集为所有的文档集合，每一行为一个文档的特征集合。

设 D 为一个包含 m 篇文档的集合，d_i 为第 i 个文档的特征向量，则有

$$D=\{d_1,d_2,\cdots,d_i,\cdots,d_m\}\ i=1,2,\cdots,m$$

该文档集 D 中的任一文档 d_i 共包含 n 个词汇，这些词汇构成了一个 n 维向量：

$$d_i=\{d_{i1},d_{i2},\cdots,d_{ij},\cdots,d_{in}\}\ i=1,2,\cdots,m,j=1,2,\cdots,n$$

其中，d_{ij} 为文档 d_i 中第 j 个词条 t_j 的权值，它一般被定义为 t_j 在 d_i 中出现的频率 t_{ij} 的函数，常用的函数有布尔函数、平方根函数、对数函数、TFIDF函数等。在实际应用中一般采用TFIDF函数：

$$d_{ij}=t_{ij}\times\log(m/n_j)$$

其中，m 为文档数据库中的文档总数，n_j 是文档数据库含有词条 t_j 的文档数目。TFIDF函数保证了出现频率较低的关键字拥有较高的分量值。

3.3 文本特征抽取

特征抽取从文本提取能够代表文本价值的特征项的过程。文本特征抽取有许多方法，基于因子分析的特征抽取机制采取改进特征项的表示，如采用词串或词序列等方式，利用贝叶斯分类器，通过训练集来寻找比较重要的文本特征。

4 基于文本挖掘技术的学习评价

4.1 评价量规

电子学档评价总体上是一种质性评价方法，但质性评价并不绝对排斥量化范式。本研究中，我们采用评价量规实现对电子学档的评价。

量规(rubric)是一种结构化的定量评价标准，是将评价目标的各个方面详细制订出评级指标。量规具有操作性好、准确性高的特点，可以有效降低评价的主观随意性。

量规指标体系和指标权重的确定分别采用“头脑风暴法”和“专家会议法”。最终，评价量规被确定为学习态度、学习方法、学习过程三个部分，共18项指标。

4.2 评价系统设计

基于文本挖掘技术的学习评价系统结构如图1所示。

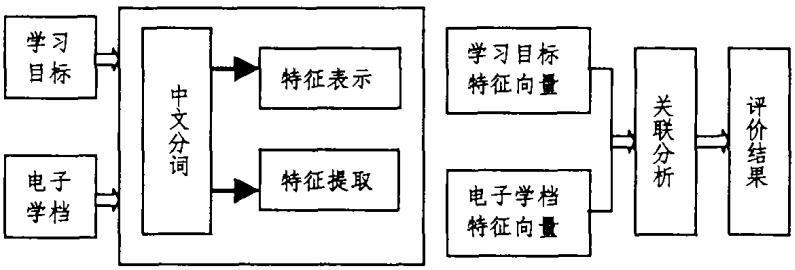

图1 基于文本挖掘的学习过程评价系统结构示意图

运用中文分词和文本挖掘方法，对课程学习目标进行特征表示和特征抽取，形成学习目标特征向量；对电子学档中的记录信息进行文本挖掘形成电子学档特征向量。根据评价量规各项目的要求，利用关联算法对两组向量进行关联分析，得出两向量的相关度，即学生学习记录与学习目标的符合程度，以此作为对学生学习过程评价的重要依据。

假设，学习目标的文档向量为 Q：

$$Q=\{q_1,q_2,\cdots,q_i,\cdots,q_n\}\ i=1,2,\cdots,n$$

如前所述，电子学档中某记录文档为 d_i：

$$d_i=\{d_{i1},d_{i2},\cdots,d_{ij},\cdots,d_{in}\}\ j=1,2,\cdots,n$$

根据向量空间模型(Vector Space Model，VSM)，文档 d_i 与文档 Q 相关度可以用两向量之间的夹角来度量，夹角越小说明相关度越大。相关度计算公式如下：

$$sim(d_i,Q)=\cos(d_i,Q)=\frac{\sum_{j=1}^{n}d_{ij}\cdot q_j}{\sqrt{\sum_{j=1}^{n}d_{ij}^2\cdot\sum_{j=1}^{n}q_j^2}}$$

讨论 基于以上分析，我们在远程教育系统中实现了学习评价系统并已开始试验运行。学习评价系统还可以运用用户使用记录挖掘(Web Usage Mining)技术，对学生访问e-Learning系统的行为进行分析，了解学生在线学习时间、作业完成情况、交互学习参与程度(讨论时发起/回复问题的次数)等信息，进一步丰富学习评价的内容[7]。

参考文献

1 Barrett H C. Create Your Own Electronic Portfolio (using off-the-shelf software), Learning & Leading with Technology. April, 2000. 14～21

2 王佑镁. 基于ePortfolio的信息化教学评价策略研究[J]. 电化教育研究，2003，(12)：61～66

3 宗成庆. 自然语言理解(讲义6)[EB/OL]. http://www.nlpr.ia.ac.cn/English/cip/ZongReport_and_Lecture/Report%20and%20Lecture-Index.htm，2005-01-06

4 黄晓斌. 网络信息挖掘[M]. 北京：电子工业出版社，2005

5 邹涛，黄源，张福炎. 基于WWW的文本信息挖掘[J]. 情报学报，1999(4)：289～293

6 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(7)：98～101

7 黄智兴，刘革平，程静，邱玉辉. WEB数据挖掘与远程教育的紧密集成[A]. 见：第七届全球华人计算机教育应用大会论文集[C]. 南京：南京师范大学出版社，2003. 690～694