浙江科技学院学报，第19卷第2期，2007年6月
Journal of Zhejiang University of Science and Technology
Vol. 19 No. 2, June 2007

潜在语义分析方法在主观题评判中的应用

陈明晶
（浙江财经学院 信息学院，杭州 310018）

摘　要：主观题的自动评判是实现在线考试功能的一个关键技术，由于具有相当的难度，故目前国内外在这方面的研究还不多，真正实用的系统则更少。传统的主观题评判系统是基于关键字匹配的，仅比较关键字就得出结论，不仅结果不准确，而且存在隐患。为此，在引入传统的潜在语义分析方法的同时对其进行改进，并为不同的关键字使用不同的权重以体现语义的重要性，由此设计出一种加权的语义分析方法。经实验证明，该方法的准确率有所提高，达到了预期的效果。

关键词：潜在语义分析(LSA)；语义相似度；主观题；自动评判
中图分类号：TP311.13　　**文献标识码**：A　　**文章编号**：1671-8798(2007)02-0093-04

Applying of Latent Semantic Analysis in Automated Assessment of Subjective Tests

CHEN Ming-jing
(Information School, Zhejiang University of Finance and Economics, Hangzhou 310018, China)

Abstract: Automated assessment of subjective tests is the key technology in the reality of the online exam. Because of its hardness, little research has been done in this area, and less real practical system. Traditional automated assessment of subjective tests is based on key words matching. It only matches all of the key words to draw a conclusion. The results are inaccurate, and troubles hidden. Based on the introduction of the traditional method of latent semantic analysis, the author improves it, and gives different weights to different key words, and designs a weighted semantic analysis. The experiment proved that the accuracy rate has been increased, so as to achieve the expectant results.

Key words: latent semantic analysis; semantic similarity; subjective test; automated assessment

目前，在线考试系统已在各个领域或地区广泛应用，在这些系统中，试题的管理、组卷、客观题评判等基本得到解决，但对主观题的评判还存在很多问题。主观题的自动评判涉及到人工智能、模式识别以及自然语言理解等方面的理论知识，需要解决很多技术上的问题，因而成为在线考试系统中的一个技术难点。主观题在答题时一般采用语言叙述的方式，而每个人对知识的理解程度不同，表达方式也不

收稿日期：2007-03-21
作者简介：陈明晶(1978—　)，男，江苏句容人，讲师，主要从事电子商务系统和商业智能的研究。

一致，即便学生的答案准确，也很难与标准答案完全一样，若要像对客观题批改那样准确地对主观题进行自动批改是非常困难的[1]。

当前的主观题自动批改算法中，比较简单的方法是根据检索到的关键字给予评分[1]，但由于汉语语义的复杂性，导致这种方法的精确度受到极大的影响。另一种方法是用基于向量空间模型的方法和基于语义的方法，定义出句子相似度算法[2]，并计算学生答案与标准答案的相似度，但整个算法比较复杂，且准确率不能完全达到适用的标准。

在参考上述文献及其他相关文献的基础上，笔者尝试将中文信息处理中的中文文本聚类方法应用到主观题自动评判中，比较学生答案与标准答案的语义相似性，以期能在主观题自动批改准确率上有所突破。

1 潜在语义分析(LSA)理论

潜在语义分析(Latent Semantic Analysis, LSA)是 Landauer, Dumais 等人提出的[2]，其基本思想是文档中的词与词之间存在某种联系，即存在某种潜在的语义结构。同义词之间具有基本相同的语义结构；多义词的使用具有多种不同的语义结构。词汇之间的这种语义结构与其在文档中的出现频率有关，因此，可以通过统计学方法提取并量化这些潜在的语义结构，进而消除同义词、多义词的影响，提高文档表示的准确性。

潜在语义分析是利用数学方法中矩阵奇异值分解(SVD)理论来实现的。

1.1 词条矩阵(Term Matrix)

为了实现 LSA 思想，首先要构造一个 $m\times n$ 的词条矩阵 $\boldsymbol{D}$[3,4]，其中 n 表示文本集中的文本数，m 表示文本集中包含在所有不同的词的个数。也就是说，每一个不同的词对应于矩阵 $\boldsymbol{D}$ 的一行，而每一个文档则对应于矩阵 $\boldsymbol{D}$ 的一列。矩阵 $\boldsymbol{D}$ 表示为：

$$\boldsymbol{D}=[d_{ij}]_{m\times n} \tag{1}$$

式(1)中：d_{ij} 为非负值，表示第 i 个词在第 j 个文档中出现的次数。由于每个词只会出现在少量文档中，故 d_{ij} 通常为高阶稀疏矩阵。

1.2 奇异值分解 SVD(Singular Value Decomposition)

在词条矩阵 $\boldsymbol{D}$ 中($m\geqslant n$)，存在下式 SVD 分解：

$$\boldsymbol{D}=U\sum V^{\mathrm{T}} \tag{2}$$

式(2)中：$\boldsymbol{U}$ 是 $m\times m$ 的正交矩阵(即 $\boldsymbol{U}\boldsymbol{U}^{\mathrm{T}}=\boldsymbol{U}^{\mathrm{T}}\boldsymbol{U}=\boldsymbol{I}_m$，$\boldsymbol{I}_m$ 是 $m\times m$ 单位矩阵)，称为 $\boldsymbol{D}$ 的左奇异值向量(即 $\boldsymbol{U}$ 是 $A^{\mathrm{T}}A$ 的正交特征向量)；$\boldsymbol{V}$ 是 $n\times n$ 的正交矩阵，称作 $\boldsymbol{D}$ 的右奇异值向量(即 $\boldsymbol{V}$ 是 AA^{T} 的正交特征向量)；$\sum$是 $m\times n$ 的对角矩阵，对角元素为 a_1，a_2，…，$a_{\min(m,n)}$，且 $a_1\geqslant a_2\geqslant\cdots\geqslant a_{\min(m,n)}>0$，$a_1$，$a_2$，…，$a_{\min(m,n)}$ 是 $\boldsymbol{D}$ 的奇异值(即是 $A^{\mathrm{T}}A$ 的非负平方根)，$\boldsymbol{D}$ 的 k 秩近似矩阵 $\boldsymbol{D}_k$ 为：

$$\boldsymbol{D}_k=\boldsymbol{U}_k\sum_k\boldsymbol{V}_k^{\mathrm{T}} \tag{3}$$

式(3)中：$\boldsymbol{U}_k$ 是 $m\times k$ 的矩阵，由 $\boldsymbol{U}$ 的前 k 列组成，是压缩到 K 维空间的词向量，m 是词的数量；$\sum_k$ 是 $k\times k$ 矩阵，是由 $\sum$的前 k 行、前 k 列组成，是奇异值，k 为因子数，r 为 $\boldsymbol{D}$ 的秩；$\boldsymbol{V}_k$ 是 $k\times n$ 矩阵，由 $\boldsymbol{V}$ 的前 k 行构成，是压缩到 K 维空间的文档向量，n 为文档数。

这样定义的 $\boldsymbol{D}_k$ 是矩阵 $\boldsymbol{D}$ 的 k 秩近似矩阵，这种近似保持了 $\boldsymbol{D}$ 中所反映的词和文档之间联系的内在结构(潜在语义)，同时去除了大量因词汇同义或多义而产生的“噪声”。例如，经常出现在类似上下文中的词相似度比较大，在 K 维词空间中也会比较接近。将此 K 维词空间理解为概念空间，那么这些词在概念上是相近的或同义的。依据于矩阵 $\boldsymbol{A}_k$，就可将文档的词空间转化为语义概念空间。

2 文本语义相似度计算

中文文本包括 3 个层次：词语、句子和段落。在主观题的评判中，应该侧重于判断词语和句子的相似度，因此，可以使用潜在语义分析方法进行相似性的判断。潜在语义分析出发点就是文本中的词与词之间存在某种联系，即存在某种潜在的语义结构。这种潜在的语义结构隐含在文本中词语的上下文使用模式中。因此采用统计计算的方法，对大量的文本进行分析来寻找这种潜在的语义结构，它不需要确定的语义编码，仅依赖于上下文中事物的联系，并用语义结构来表示词和文本，达到消除词之间的相关性[5]。

在本文方法中，笔者使用潜在语义分析的方法，计算出学生的答案与标准答案的相似度，并以此相似度为依据，为学生的答案进行评分，以达到准确评判的目的。以下对主要步骤和要点进行阐述。

2.1 构造文本集

将标准答案与学生的答案并列，构成文本集 $S=\{S_1,S_2,\cdots,S_n\}$，其中标准答案位于首位 S_1，其他按自然顺序排列，此时，对学生答案的评判，实际

上是求解各个文本与首行文本的相似度，相似度大的说明答案越准确；反之亦然。

2.2 选择特征词

在进行潜在语义分析之前，先要构造词条矩阵。首先选取若干特征词 $\boldsymbol{W}=\{W_1,W_2,\cdots,W_m\}$ 作为统计的对象，特征词的选取方法可以从标准答案中人工抽取具有代表性的关键字，也可以事先为该门课程建立关键词库，再从中词库中抽取在标准答案中出现频率较高的词。为了能区别相似的概念，还需要从一些相似的其他概念中引入一些“不正确”的词，作为“误差”，即标准答案中没有这些词，而如果学生的答案中包含了这些词，说明与另外的一些概念混淆，应当抵消一部分正确性。

2.3 关键词匹配

在进行关键词匹配时，通常采用符号匹配法，即按字符进行比较。考虑到语言的多样性和同义词等因素，可以事先定义一个同义词库，列举出在该学科中的常见同义词，构成同义词表，当某关键字的同义词出现时，也计入词条矩阵。这种方法虽然增加了系统的计算时间，但能显著提高结果的准确性。

2.4 加　权

传统的潜在语义分析方法在构建词条矩阵时，只考虑到关键词出现的次数，并不关心这些关键词在上下文中的重要性问题，构成的词条矩阵为 $\boldsymbol{D}_1=[\boldsymbol{d}_{ij}]_{m\times n}$，其中 $i=1,2,\cdots,m,j=1,2,\cdots,n$。而实际上，这些关键词的作用是不同的，因此可为一些比较重要的关键词设置较大的权重，以扩大其在结果中的影响。设置权重的方法，可根据该关键词在上下文中的出现次数，也可以人工指定，但要求权重之和等于 1，即 $\sum w_i=1$。设权重为 $w=\{w_1,w_2,\cdots,w_m\}$，则加权后的词条矩阵为：$\boldsymbol{D}_2=[d_{ij}\times w_i]_{m\times n}$。

2.5 语义相似度计算

对矩阵 $\boldsymbol{D}_2$ 进行 SVD 后，可得到 $\boldsymbol{U}_k$ 和 $\boldsymbol{V}_k$ 向量，其中 $\boldsymbol{V}_k$ 为文本向量，保持着与词条矩阵对应列向量相似的特性，因此，可以使用 $\boldsymbol{V}_k$ 的行向量来计算文本之间的相似度。文本的相似度计算一般采用余弦距离公式[3]：

$$\mathrm{Sim}(j)=\frac{\sum_{m=1}^{k}W_{1m}\times W_{jm}}{\sqrt{\sum_{m=1}^{k}(W_{1m})^2\times\sum_{m=1}^{k}(W_{jm})^2}} \tag{4}$$

$\mathrm{Sim}(j)$ 为文本 j 与文本 1 的相似度（$-1\leqslant \mathrm{Sim}\leqslant 1$），$W_{jm}$ 表示矩阵 V_k 的第 m 列的第 j 行的值。通常情况下，语义相似的文本，其 $\mathrm{Sim}(j)$ 值较大，接近于 1；而语义不相似的文本，其 $\mathrm{Sim}(j)$ 值较小；若 $\mathrm{Sim}(j)<0$，则说明答案误差较大。

2.6 简单实例分析

假设考试系统中的一道简答题，题目为：什么是电子政务。设置标准答案为：S_1：高效、开放的政府，凭借计算机技术、现代通信技术等信息技术，在安全可靠的网络平台上行使管理职能、开展政务活动。

选择 7 个不同的词为特征词：W_1：政府；W_2：信息；W_3：安全；W_4：网络；W_5：管理；W_6：政务；W_7：商务。

显然，前 6 个词是标准答案中的关键词，而最后 1 个词则是为了防止学生回答了其他概念而引入的“误差”。

假设学生的答案有以下几个：S_2：政府部门利用安全计算机网络信息平台进行政务管理活动；S_3：在开放的网络环境中，政府部门以门户网站为平台，开展相关活动；S_4：使用信息技术和设备，提高政府部门的办事效率；S_5：政府信息化；S_6：在开放的网络环境中，买卖双方进行的商务活动。

可以得到一个 7×6 的词条矩阵，如下所示：

$$\boldsymbol{D}_1=\begin{bmatrix}1&1&1&1&1&0\\1&1&0&1&1&0\\1&1&0&0&0&0\\1&1&1&0&0&1\\1&1&0&0&0&0\\1&1&0&0&0&0\\0&0&0&0&0&1\end{bmatrix}$$

此外，d_{ij} 为词条 i 在文本 j 中出现的频度。

由于在特征词中，各个词的重要性是不同的，根据实际的含义，定义加权向量 $\boldsymbol{W}$ 为 $\boldsymbol{W}=\{0.3,0.1,0.1,0.1,0.2,0.1,0.1\}$。则词条矩阵为：

$$\boldsymbol{D}_2=\begin{bmatrix}0.3&0.3&0.3&0.3&0.3&0\\0.1&0.1&0&0.1&0.1&0\\0.1&0.1&0&0&0&0\\0.1&0.1&0.1&0&0&0.1\\0.2&0.2&0&0&0&0\\0.1&0.1&0&0&0&0\\0&0&0&0&0&0.1\end{bmatrix}$$

$\boldsymbol{D}_2$ 的秩为 6，经过奇异值分解计算后，得到 $\boldsymbol{U}$、Σ、$\boldsymbol{V}$ 如下：

$$U=\begin{bmatrix}-0.880\,3 & 0.455\,6 & 0.100\,9 & 0.052\,3 & 0.067\,6 & -0.000\,0 & 0\\ -0.290\,1 & -0.670\,5 & 0.643\,4 & -0.224\,3 & -0.045\,5 & -0.000\,0 & 0\\ -0.142\,8 & -0.256\,3 & -0.254\,7 & 0.201\,7 & 0.091\,6 & 0.894\,4 & 0\\ -0.197\,2 & -0.119\,2 & -0.471\,3 & -0.622\,7 & -0.580\,5 & 0.000\,0 & 0\\ -0.285\,6 & -0.512\,5 & -0.509\,3 & 0.403\,4 & 0.183\,2 & -0.447\,2 & 0\\ 0 & 0 & 0 & 0 & 0 & 0 & -1.000\,0\\ -0.003\,6 & -0.016\,8 & -0.171\,6 & -0.596\,5 & 0.783\,9 & -0.000\,0 & 0\end{bmatrix}$$

$$\Sigma=\begin{bmatrix}0.747\,3 & 0.000\,0 & 0.000\,0 & 0.000\,0 & 0.000\,0 & 0.000\,0\\ 0.000\,0 & 0.284\,8 & 0.000\,0 & 0.000\,0 & 0.000\,0 & 0.000\,0\\ 0.000\,0 & 0.000\,0 & 0.193\,5 & 0.000\,0 & 0.000\,0 & 0.000\,0\\ 0.000\,0 & 0.000\,0 & 0.000\,0 & 0.143\,0 & 0.000\,0 & 0.000\,0\\ 0.000\,0 & 0.000\,0 & 0.000\,0 & 0.000\,0 & 0.050\,9 & 0.000\,0\\ 0.000\,0 & 0.000\,0 & 0.000\,0 & 0.000\,0 & 0.000\,0 & 0.000\,0\\ 0.000\,0 & 0.000\,0 & 0.000\,0 & 0.000\,0 & 0.000\,0 & 0.000\,0\end{bmatrix}$$

$$V=\begin{bmatrix}-0.591\,8 & -0.218\,1 & -0.252\,1 & 0.091\,2 & -0.110\,6 & -0.223\,6\\ -0.475\,4 & 0.011\,8 & -0.745\,1 & -0.379\,5 & 0.157\,2 & -0.223\,6\\ -0.379\,8 & -0.438\,0 & -0.087\,1 & -0.325\,9 & -0.741\,7 & 0.000\,0\\ -0.392\,2 & -0.244\,5 & -0.488\,8 & -0.047\,2 & 0.308\,6 & -0.670\,8\\ -0.353\,4 & 0.479\,9 & 0.156\,4 & 0.109\,7 & 0.097\,9 & -0.670\,8\\ -0.026\,9 & -0.047\,7 & -0.332\,2 & -0.852\,8 & 0.399\,3 & 0.000\,0\end{bmatrix}$$

使用公式(4)对 V_k 进行计算，得到 Sim 为：

Sim={1,0.821 4,0.541 0,0.713 3,0.350 5,−0.274 5}

由 Sim 的结果可知，语义相似的语句能得到比较大的相似度，若本题分值为 5 分，则对应的答案得分为(除去标准答案的项)：

Score={4,2.5,3.5,2,0}

与实际情况基本相符，说明本方法具有一定的效果。

3 实验结果分析

利用本文的方法，对浙江财经学院《电子政务》课程的 120 份期中考试试卷进行了测试，其中 4 道名词解释和 3 道简答，每道题选择 7～10 个关键字，结果分析如下。

3.1 准确率测试

不同题型在准确率上的表现不尽相同，准确率结果统计如表 1 所示。

其中准确率计算方法如下：

$$\text{准确率}=\left(1-\frac{|\text{人工评分}-\text{系统评分}|}{\text{人工评分}}\right)\times 100\%$$

可见，名词解释和简答题准确率在 90%以上的个体数分别为 77 和 69，占总数的 64.2%和 57.5%，基本达到了要求。但准确率在 70%以下的个体数也分别有 26 和 34，占总数的 21.7%和 28.3%。另外，名词解释的准确率明显高于简答题。

表 1 不同题型的准确率分布

准确率	100%	≥95%	≥90%	≥80%	≥70%	≥60%	<60%
名词解释	14	41	22	10	7	12	14
简答	11	43	15	14	3	15	19

3.2 原因分析及改进方法

从以上数据可以看出，高准确率和低准确率的个体都比较多，而且名词解释的准确率高于简答题。经分析实验数据后发现，由于在选择关键词时，一般选择比较典型的词语，而且类似的词也只选择一个，但学生在答题时经常会使用一些近义词，因此造成系统评判的偏差。另外，由于名词解释的字数较少，并且涉及到的名词比较集中，而简答题则涉及的内容比较广泛一点，因此系统评判名词解释比简答题准确。

经实验证明，选择准确的关键词或者适当分配各关键词的权重，可以提高系统评判的准确率。另一方面，适当增加关键词的个数可以使检索的内容更全面，还可以提高准确率。但由于增加一个关键词，会使词条矩阵增加 n 项，它所带来的计算量代价非常可观，因此并非关键词越多越好。一般 7～10 个关键词比较适中。

(下转第 132 页)

(上接第96页)

4 结 语

在主观题自动批改方法中,笔者尝试了改进传统潜在语义分析方法,为选择的关键词分配适当的权重,以体现部分关键词的重要性,再构建词条矩阵,并进行奇异值分解,来计算中文文本之间的相似度,进而对在线考试系统中的主观题进行评判。试验结果显示,该方法是比较成功的,为考试系统中主观题自动批改提出了一种新的解决途径。在使用不同的关键词进行对比分析时发现,该方法在使用7～10个关键词时效果比较明显。

但同时也注意到,该语义相似度的方法只考虑了关键词出现的次数和上下文,却没有考虑中文文本信息中的其他因素对语义的影响,如否定词、语气等,因此准确率还不能完全达到适用的标准,在实际的应用中,还需要对结果进行复核和调整。而如何使评判结果更加准确,笔者将在后续的工作中进行进一步研究。

参考文献:

[1] 孟爱国,卜胜贤,李鹰,等.一种网络考试系统中主观题自动评分的算法设计与实现[J].计算机与数字工程,2005,33(7):147-150.

[2] 高思丹,袁春风.主观试题的计算机自动批改技术研究[J].计算机应用研究,2004,21(2):181-185.

[3] 王国勇,徐建锁.TCBLSA:一种中文文本聚类新方法[J].计算机工程,2004,30(5):21-22,37.

[4] 余正涛.基于潜在语义分析的汉语问答系统答案提取[J].计算机学报,2006,29(10):1889-1893.

[5] 盖杰,王怡,潜在语义分析理论及其应用[J].计算机应用研究,2004,21(3):9-12,20.