文章编号：1674－9057（2010）03－0426－05

基于自然语言处理的智能评分系统

麦范金[1]，岳晓光[2]，赵子强[3]，路 英[4]，王 挺[5]，岳砥柱[6]

（1. 桂林理工大学 现代教育技术中心，广西 桂林 541004；2. 广西民族师范学院 数学与计算机科学系，广西 崇左 532200；3. 山西中医学院 网络信息中心，太原 030024；4. 莫纳什大学 管理学院，澳大利亚 墨尔本 3800；5. 西交利物浦大学 计算机科学与软件工程系，江苏 苏州 215123；6. 北京大学 哲学系，北京 100871）

摘 要：针对目前考试系统和主观题评分算法的研究现状，引入了自然语言处理技术。在对已有的相似度计算的算法进行改进的基础上，提出了一种新的排序和划分区间的评分方法，并在.NET平台上设计一个智能评分系统。实验表明该智能评分算法及智能评分系统有效且可行。

关键词：自然语言处理；智能评分系统；.NET

中图分类号：TP182 **文献标志码：**A

自然语言处理中的相似度计算和人工阅卷中的主观题评分有一些相似之处：相似度计算是通过一定的计算方法来比较词、语句、段落等，从而计算出一个确定的值来判定它们之间相似度的高低；人工阅卷中的主观题评分通常是在参考答案的基础之上，给学生答案一个相对客观的分数。因此通过对学生答案和参考答案进行相似度计算来判定学生分数是一个可行的思路。

1 自然语言处理与智能考试系统

自然语言处理主要分为语言学、数据处理、人工智能和认知科学、语言工程等4个方向[1]。研究自然语言处理的关键是研究其计算模型，其建模可以用图1表示[2]。

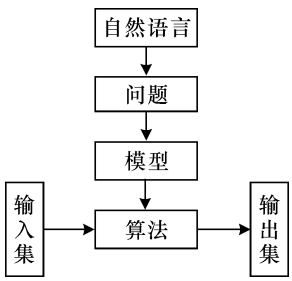


图1 自然语言处理计算模型

Fig. 1 Computational model of natural language processing

众所周知，传统考试的主观题人工评分存在诸多弊病，迫使人们研究计算机上机考试及计算机智能评分，使考试趋于更客观、公正、高效[3]。早在20世纪60年代，国外开始研究计算机自动评分技术，如：Braun、Bennett、Frye及Soloway开发的CAS系统[4]。目前美国思而文公司的计算机化认证考试系统涉及IT领域证书认证、学业考试（GRE、GMAT、TOEFL）和职业证书考试[5]。国内对这方面也有研究，但不足之处显而易见，主要表现为不够灵活、以评测客观题为主、没有通用性、缺乏考试结果分析等[6]。目前国内的考试系统只是针对客观题，而对于主观题的自动阅卷仍然没有解决[7]。因此，考试系统如何对学生的成绩进行评分一直困扰着教育领域的各个方面。从某种角度来看，它甚至比组织和使用一个合理的测试都要困难和复杂[8]。

智能考试系统的实质[9]是利用先进的现代计算机技术，用计算机组卷来代替人工活动，解决在传统的人工考试环境下不能解决的问题，达到

收稿日期：2009－09－24

基金项目：广西自然科学基金项目（桂科自0991254）

作者简介：麦范金（1963—），男，副教授，研究方向：中文信息处理。

引文格式：麦范金，岳晓光，赵子强，等. 基于自然语言处理的智能评分系统［J］. 桂林理工大学学报，2010，30（3）：426－430.

提高工作质量和工作效率的目的。智能考试系统的核心智能评分算法面对两个方面的问题：一个是客观题的评分，另一个是主观题的评分。客观题较为容易解决，通过比对参考答案就可以实现。文字类主观题的自动评分是在线考试系统的一个关键技术，由于其自动评判具有相当难度[10]，所以这里引入自然语言处理（natural language processing，NLP）及其相关技术，在 .NET 平台下设计一个智能评分算法，以期能够解决文本主观类型题目的评分问题，从而实现一个智能考试系统。

2　智能评分算法的核心技术

本文主要讨论中文分词技术和相似度计算相结合的主观题评分算法。

2.1　中文分词技术

2002 年之前对自动分词方法的研究可分为基于规则和基于统计的两大类[11]。2002 年以后多种中文分词技术出现。在智能评分系统中考虑到分词算法实现的难度及简答题通常回答需要的字数不多，使用最大匹配分词算法[12]。

匹配原则[13]：正向最大匹配（MM）对于文本中的字串 $ABC, A \in W, AB \in W, ABC \notin W, W$ 为字典，那么就取切分 AB/C；逆向最大匹配（RMM）对于文本中的字串 $ABC, C \in W, BC \in W, ABC \notin W, W$ 为字典，那么就取切分 A/AC。这种匹配方法可能会出现歧义和未登录词的问题，但是考虑到分词方法已经达到了较高的精度，再结合相似度计算，可以很大程度的降低这方面的影响。词表设计采用文献[14]中提出的分词规范进行设计。

2.2　相似度计算

相似度计算的公式是综合和参考文献[3]、[10]、[15]里面的计算方法，同时，通过引入对于关键词相似度和关键句相似度的计算，将相似度计算加以改进。

首先引入一个概念——词形相似度计算。其计算公式为

$$\mathrm{simWords}(A,B) = \frac{\mathrm{Same}(A,B)}{\mathrm{Num}(A) + \mathrm{Num}(B)}。\tag{1}$$

对于句子 A 和句子 B，计算其词形相似度 $\mathrm{simWords}(A,B)$，就是用两个句子相同的词的个数 $\mathrm{Same}(A,B)$ 去除两个句子词的总数（$\mathrm{Num}(A) + \mathrm{Num}(B)$）。

那么，如果一份试卷有 $k(k = 1,2,\cdots,z)$ 道主观题，则对于第 k 道主观题，计算其相似度 $\mathrm{sim}(k)$，可以用下式进行计算

$$\mathrm{sim}(k) = \alpha \cdot \mathrm{sim}SR(k) + \beta \cdot \mathrm{sim}RS(k) + \theta \cdot KW(k) + \delta \cdot KS(k)。\tag{2}$$

其中：①α、β、θ、δ 都是可以调整的系数，通过系数来调整 4 个单项数据的权重。

②$\mathrm{sim}SR(k)$ 是用来计算学生答案 $S(k,m)(m = 0,1,2,\cdots,a)$ 与参考答案 $R(k,n)(n = 0,1,2,\cdots,b)$ 的相似度，计算学生答案中的每个词与参考答案所有词的相似度的最大值，并求其最大值之和的平均值，计算如下式

$$\mathrm{sim}SR(k) = \sum_{m=0}^{a} \max(\mathrm{simWords}(S(k,m),R(k,0)), \mathrm{simWords}(S(k,m),R(k,1)),\cdots,\mathrm{simWords}(S(k,m),R(k,b)))/a。\tag{3}$$

③$\mathrm{sim}RS(k)$ 是用来计算参考答案 $R(k,n)(n = 0,1,2,\cdots,b)$ 与学生答案 $S(k,m)(m = 0,1,2,\cdots,a)$ 的相似度，计算参考答案中的每个词与学生答案所有词的相似度的最大值，并求其最大值之和的平均值，计算如式(4) 所示：

$$\mathrm{sim}RS(k) = \sum_{n=0}^{b} \max(\mathrm{simWords}(R(k,n),S(k,0)), \mathrm{simWords}(R(k,n),S(k,1)),\cdots,\mathrm{simWords}(R(k,n),S(k,a)))/b。\tag{4}$$

④$KW(k)$ 是关键词相似度，可以表示核心关键词的匹配程度，对于每道主观题，不同的核心关键词分配不同的权值，学生考试能够写出几个核心关键词，则将这几个核心关键词的权值累加求和，其计算公式用式(5) 表示，其中 $\mathrm{KeyWords}(k,w)$ 表示第 k 道题有 c 个核心关键词

$$KW(k) = \sum_{w=0}^{c} \mathrm{KeyWords}(k,w)。\tag{5}$$

⑤$KS(k)$ 是关键句相似度，可以表示核心关键句的匹配程度，对于每道主观题，不同的核心关键句则分配不同的权值，学生能够写出几个核心关键句，则将这几个核心关键句的权值累加求和，其计算公式用式(6) 表示，其中 $\mathrm{KeySentences}(k,s)$ 表示第 k 道题有 d 个核心关键句

$$KS(k) = \sum_{s=0}^{d} \mathrm{KeySentences}(k,s)。\tag{6}$$

2.3　主观题评分算法

设置最大切分长度为 l，分词的词典为 M，并且

用“/”符号标识作为词与词之间的间隔符。对于第 k 道题，总分为 Score(k)，得分划分为 ScoLevel(k, e) 个等级，相似度划分为 simLevel(k,e) 个等级，其中 $e = 0,1,\cdots,$Score(k)，有：(1) 切分学生答案，并将学生答案 $S(k,m)$($k = 1,2,\cdots,z$; $m = 0,1,2,\cdots,a$) 以词为单位存入数据库。

① 获取学生答案字符串 StudentString，其字符串长度为 ssl，去除标点符号。

② 对 StudentString 的末端选择长度为 l 的待切分字符串 CurrentSS，其字符串长度为 css。

③ 将 CurrentSS 与 M 进行比较。

④ 如果 M 中存在这个字符串，则 CurrentSS 是一个词，将这一段字符串从学生答案字符串中去除，即 StudentString = Studentstring − CurrentSS，得到一个长度为 $ssl = ssl - css$ 新的 StudentString，返回到 ②。

⑤ 如果 M 中不存在这个字符串，并且 $css \geqslant 1$，则字符串做自减运算，即 $css = css - 1$，得到一个新的 CurrentSS，返回到 ③；如果 $css = 1$，则得到单字，这里作为词切分，然后返回到 ②。

⑥ 切分完毕 StudentString 之后，以 StudentString 的间隔符为界，将 $S(k,m)$ 存入数据库中学生答案词表 StudentAnswerList。

(2) 切分参考答案，并将参考答案 $R(k,n)$($k = 1,2,\cdots,z$; $n = 0,1,2,\cdots,b$) 以词为单位存入数据库。步骤(2) 和(1) 的过程类似，不再赘述，下面采用与(1) 类似的变量标明。

(3) 计算 StudentString 和 ReferenceString 的相似度。

① 读取 $S(k,m)$ 和 $R(k,n)$，按照式(2) 进行计算，得到 $\alpha \cdot \text{sim}SR(k) + \beta \cdot \text{sim}RS(k)$。

② 通过 $S(k,m)$ 和 Keywords(k,w) 进行比较，如果存在一致的话，则计一次权值，不重复计算，按照式(2) 进行计算，得到 $\alpha \cdot \text{sim}SR(k) + \beta \cdot \text{sim}RS(k) + \theta \cdot KW(k)$。

③ 读取 Keywords(k,s) 和 StudentString 比较，如果存在一致，则计一次权值，不重复计算，按照式(2) 进行计算，得到 $\alpha \cdot \text{sim}SR(k) + \beta \cdot \text{sim}RS(k) + \theta \cdot KW(k) + \delta \cdot KS(k)$，即 sim($k$)。

(4) 根据试题的总分，得分划分为 ScoLevel(k, e) 个等级：ScoLevel(k,e) = Score(k) − e；相似度划分为 simLevel(k,e) 个等级：simLevel(k,e) = (ScoLevel(k,e) − e/ScoLevel(k,e))/sim(k)。

3 系统设计及实现

3.1 系统总体功能模块

限于篇幅，系统界面和程序不再附上。系统总体功能模块主体如图 2 所示。

3.2 评分模块

对于抽取分词和相似度计算，设置 $\alpha = 1.5$、$\beta = 1$、$\theta = 2$、$\delta = 5$，核心关键词和核心关键句的赋值数据如表 1 所示。对于不同的考生答案，得到不同的结果，相似度计算结果如图 3、图 4 所示。这里的相似度最大值是超过 1 的，因为这样有利于划分评分区间。

4 实验结果及分析

4.1 实验条件

学生通过机房的主页链接访问该智能评分系统，同时，将计算机基础测试试卷导入到系统之中，通过比对计算机评分和教师评分的数据，设置合理的评分系统，可以比较出该智能评分算法和手工评分的优劣。

α、β、θ、δ 是测试的经验值（表 2），每道题的核心关键词和核心关键句权值的取值范围如表 3 所示。

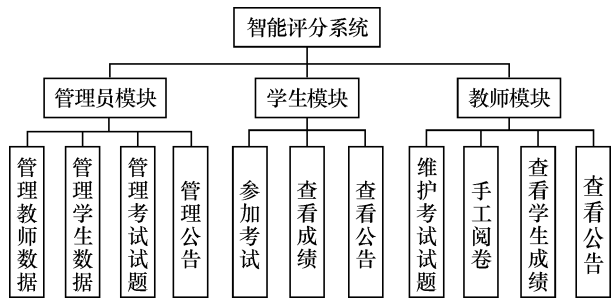


图 2 系统总体功能模块图

Fig. 2 Diagram of overall system functional block

表 1 赋值数据

Table 1 Data of assignment

核心关键词(权值)	核心关键句(权值)
软件系统(0.5)、硬件系统(0.5)	计算机系统是由软件系统和硬件系统组成(5)
计算器(0.5)、运算器(0.5)、控制器(0.5)、输入设备(0.5)、输出设备(0.5)	硬件系统由计算器、运算器、控制器、输入设备、输出设备组成(3)
程序(0.5)、数据(0.5)	软件系统由程序、数据及其相关文档组成(3)

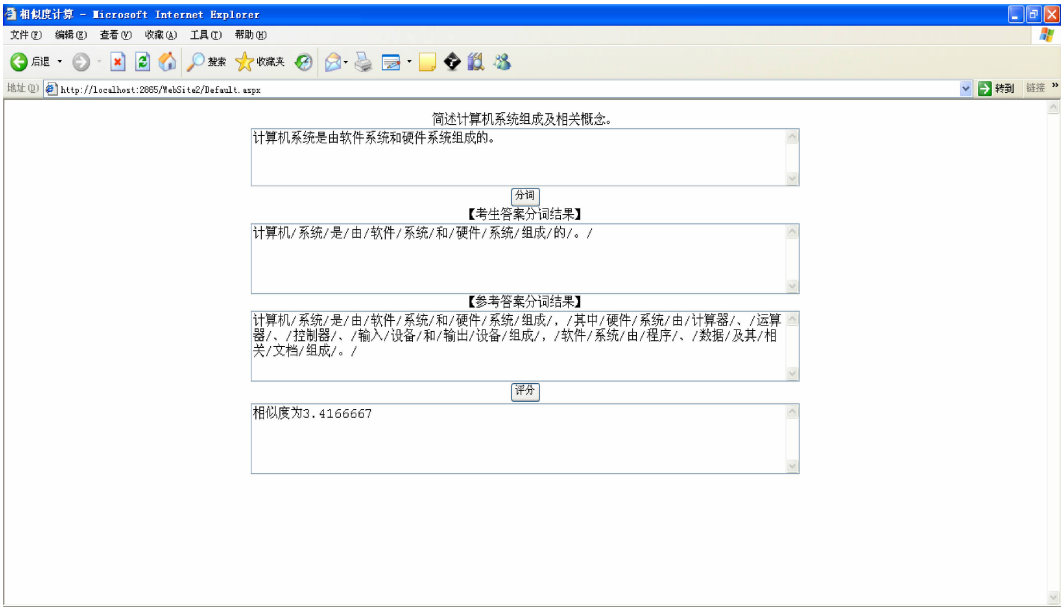


图 3　计算结果 1

Fig. 3　First result of calculation

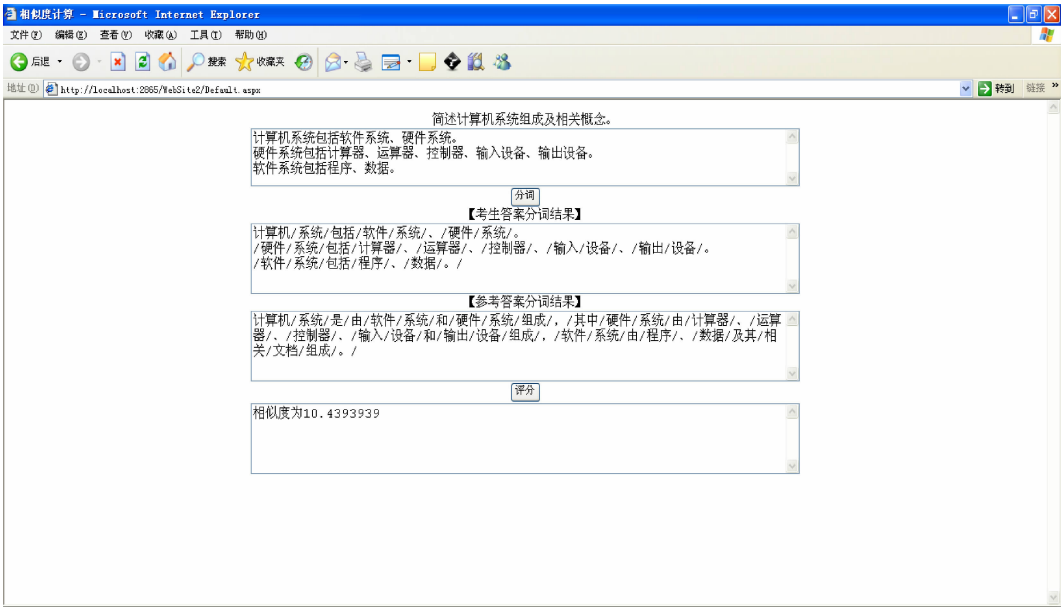


图 4　计算结果 2

Fig. 4　Second result of calculation

4.2　实验数据

实验数据有 2 组，2 组数据的试卷结构是一致的，难度是相似的：第 1 组是抽取 100 份教师已评分的学生试卷，并导入评分系统进行评分；第 2 组，请 100 位学生在上机的时候进行测试，教师对主观题再评分；用该系统和人工对主观题评分的总分进行分差比较（表 4）。

由表 4 可知，智能评分系统和人工阅卷的分差主要集中在 5 分左右，并且智能评分系统和人工阅卷通过考试人数的差别较小，不会影响考试的公平性，由此表明：该智能评分算法是有效的。

5　结　论

研究智能评分算法和智能评分系统的目的就是为了能够减轻教师的工作负担，同时避免主观因素影响考试的公平性。本文综合了自然语言处

表 2 参数取值
Table 2 Value of the parameter

参数	经验值
α	1.5
β	1
θ	2
δ	5

表 3 权值的取值范围
Table 3 Weights range

类型	权值
核心关键词	0.5 ~ 2
核心关键句	2 ~ 5

表 4 评分差距比较
Table 4 Comparison of score gap

条件	试卷份数						
组别差距比较	分差 =1	分差 =2	分差 =3	分差 =4	分差 =5	分差 >5	通过考试人数差别
第 1 组	0	7	15	40	32	6	2
第 2 组	1	1	10	34	37	17	1

理的中文分词技术和相似度技术的理论，提出通过相似度高低的分级，来实现考生成绩高低的区分，只要考生答案大体上和参考答案相似就能够通过考试。后续还有很多工作要做，在进一步提高中文分词算法精度和优化相似度计算算法性能、设置一个合理的参数及权值范围等方面，还有待于进一步拓展。

参考文献：

[1] 冯志伟. 自然语言处理的学科定位 [J]. 解放军外国语学院学报，2005，28 (3)：1 – 8.
[2] 张钹. 自然语言处理的计算模型 [J]. 中文信息学报，2007，21 (3)：3 – 7.
[3] 南铉国. 基于语句相似度计算的主观题自动评分技术研究 [D]. 延吉：延边大学，2007.
[4] 葛诗利，陈潇潇. 国外自动作文评分技术研究 [J]. 外语电化教育，2007，117 (10)：25 – 29.
[5] 刘毅. 网络考试系统的设计与实现 [D]. 重庆：重庆大学，2008.
[6] 欧阳炜昊. 基于 .NET 的网络考试系统的设计与实现 [D]. 长沙：中南大学，2007.
[7] 寇丽华. 基于 Web 的考试系统的研究与设计 [D]. 哈尔滨：哈尔滨工程大学，2009.
[8] Ebel R L. Marks and Marking Systems [J]. IEEE Transactions on Education，1974，17 (2)：76 – 92.
[9] 袁鑫攀. 基于自适应遗传算法的智能考试系统研究 [D]. 长沙：中南大学，2008.
[10] 贾电如,李阳明. 基于语句结构及语义相似度计算主观题评分算法的研究[J]. 信息化纵横,2009(5):5 – 7.
[11] 黄昌宁，赵海. 中文分词十年回顾 [J]. 中文信息学报，2007，21 (3)：8 – 19.
[12] 宗成庆. 统计自然语言处理 [M]. 北京：清华大学出版社，2008.
[13] 翟风文，赫枫龄，左万利. 字典与统计相结合的中文分词方法 [J]. 小型微型计算机系统，2006，27 (9)：1766 – 1771.
[14] 许嘉璐，傅永和. 中文信息处理现代汉语词汇研究 [M]. 广州：广东教育出版社，2006.
[15] 南铉国，崔荣一. 基于多层次融合的语句相似度计算模型 [J]. 延边大学学报：自然科学版，2007，33 (3)：191 – 194.

Intelligent Scoring System Based on Natural Language Processing

MAI Fan-jin[1], YUE Xiao-guang[2], ZHAO Zi-qiang[3], LU Ying[4], WANG Ting[5], YUE Di-zhu[6]

(1. *Modern Education Technology Center*, *Guilin University of Technology*, *Guilin* 541004, *China*; 2. *Department of Mathematics and Computer Science*, *Guanxi Normal University for Nationalities*, *Chongzuo* 532200, *China*; 3. *Network Information Center*, *Shanxi University of Traditional Chinese Medicine*, *Taiyuan* 030024, *China*; 4. *Department of Management*, *Monash University*, *Melbourne 3800*, *Australia*; 5. *Department of Computer Science and Software Engineering*, *Xi' an Jiaotong-Liverpool University*, *Suzhou* 215123, *China*; 6. *Department of Philosophy*, *Peking Univeristy*, *Beijing* 100871, *China*)

Abstract: The introduction of natural language processing technology is put forward by analyzing the current research of the examination system and the subjective questions assessment algorithm. A method of improved similarity calculation algorithm and a new scoring is proposed by sorting and divided interval. An intelligent scoring system is designed based on the. NET platform. Experiments prove the intelligent scoring algorithm and perform system.

Key words: natural language processing; intelligent scoring system; . NET