第 38 卷 第 11 期
Vol.38 No.11

计 算 机 工 程
Computer Engineering

2012 年 6 月
June 2012

·开发研究与设计技术· 文章编号：1000—3428(2012)11—0274—03 文献标识码：A 中图分类号：TP391

两级相似度计算在主观题机器阅卷中的应用

秦学勇，张润梅

(安徽建筑工业学院电子与信息工程学院，合肥 230601)

摘 要：针对问答题类文字描述性主观题机器阅卷的复杂性和困难性，提出一种用于机器阅卷的两级相似度计算算法。综合考虑答案的关键词、句子语法和语义信息，并结合分数微调规则设计算法。实验结果表明，该算法在词语级系数 α 取值约 0.7 时，阅卷系统具有最低的无效阅卷比例和较快的速度，符合人工阅卷的要求。

关键词：相似度计算；向量空间模型；机器阅卷；自然语言处理；依存文法；规则

Application of Two Level Similarity Computation in Subjective Machine Marking

QIN Xue-yong, ZHANG Run-mei

(School of Electronic and Information Engineering, Anhui University of Architecture, Hefei 230601, China)

【Abstract】According to answers' keywords, sentence grammar semantic informations and adjusting rules, this paper proposes a two level similarity computation algorithm which integrates the advantage of word level and sentence level similarity computation. It focuses on questions of text descriptive subjective examinations complexities and machine marking difficulties. Experimental results show that the system can obtain the minimum invalid marking ratio and fast speed, while word level coefficient α value is about 0.7. This algorithm complies with the effect of manual marking better.

【Key words】similarity computation; vector space model; machine marking; natural language processing; dependency grammar; rule

DOI: 10.3969/j.issn.1000-3428.2012.11.083

1 概述

计算机技术的飞速发展推动了传统考试方式正逐步向现代的考试方式转变，即试卷由纸质试卷向电子试卷转变，阅卷方式由人工阅卷向机器阅卷转变。国外很多学者自 20 世纪 60 年代就开始研究基于任意文本答案的主观题计算机自动阅卷技术，取得了一定的成果，并研制出一些实用的系统。国内关于该课题的研究起步较晚，主要是通过词语的相似度计算来获得语句的相似度。词语相似度计算主要采用关键词匹配、词语矢量相似度计算、词语语义相似度计算等方法处理。文献[1]基于模糊数学中隶属度进行词语相似度计算；文献[2]基于《知网》[3]进行词语相似度计算；文献[4]根据词语之间的语义信息提出了基于语义的主观题阅卷算法；文献[5]运用匈牙利算法建立句子和句群的相似度计算模型；文献[6]采用矢量来表示答案文本进行词语级相似度计算。由于中文自然语言处理的复杂性，在进行文本相似度计算时应充分考虑词语的相似度，同时也应该密切关注语句的语法结构，并结合考试过程的特点制定一些规则进行分数调整和约束，本文以该思想为基础，提出一种两级相似度计算方法，即相似度由词语级和语句级两级相似度加权求和来确定，使得机器阅卷结果更加接近人工阅卷效果。

2 两级相似度计算

相似度是描述 2 个事物或者个体在形式或者内容上的相似接近程度，是一个复杂的概念，在哲学、语义学和信息理论中得到广泛研究和讨论。在主观题自动批改系统中，相似度表示标准答案和学生答案在表达内容意思上的符合程度，计算结果用 0~1 之间的数值来表示，数值越大表示学生答案越接近标准答案，反之越远。

2.1 词语级相似度计算

词语级相似度计算首先对答案进行中文分词、停用词去除和同义词替换，然后对处理后的答案文本采用向量空间模型(Vector Space Model, VSM)来表示[7]。答案是相互独立的词条组 $\{T_1,T_2,\cdots,T_n\}$ 的集合，并且根据词条组 T_i 在文本中的重要程度赋予其一个权值 W_i(T_i 在文本中的重要程度越高，W_i 的值越大)，其中，W_i 由 TF/IDF 方法确定，于是文本匹配问题就转化为向量空间中矢量的匹配问题。本文采用夹角余弦方法来进行相似度计算，计算公式如下：

$$similarity(\boldsymbol{stuans},\boldsymbol{stdans})=\frac{\boldsymbol{stuans}\cdot\boldsymbol{stdans}}{\|\boldsymbol{stuans}\|\times\|\boldsymbol{stdans}\|}=\frac{\sum_i(wstuans_i\times wstdans_i)}{\sqrt{\sum_i wstuans_i^2\times\sum_i wstdans_i^2}}$$

其中，***stuans***、***stdans*** 表示学生答案和标准答案向量，$wstuans_i$ 和 $wstdans_i$ 为向量 ***stuans*** 和 ***stdans*** 中的第 i 个分量的权值。

2.2 语句级相似度计算

VSM 主要考虑统计答案中词的信息，通过使用 TF/IDF 方法进行相似度计算，并没有考虑到词序、句子的语法和语义信息。也就是说，在词语级相似度计算中即使句子的语法有问题，或者说语义有问题，只要相关的关键词在答案中出现并且和位置无关，那么就会认为 2 个答案是相似的。所以，词语级相似度计算虽然简单直观并且效率较高，但是有些时

基金项目：安徽省高校优秀青年人才基金资助项目(2009SQRZ101)

作者简介：秦学勇(1974－)，男，讲师、硕士，主研方向：人工智能，自然语言处理，数据挖掘；张润梅，教授

收稿日期：2011-10-14 **E-mail：**qinxueyong@aiai.edu.cn

候会对评阅产生误导从而导致误判。在网络考试系统中，答案是由句子组成，而句子才由一个个词组成的，在进行评判的时候不仅要考虑关键词的匹配，也要考虑句子的匹配。

2.2.1　语句缩句

汉语的句子具有完整的语法和语义信息，每个句子的骨架部分描述句子的基本结构和知识点，能够基本表达句子的意思。句子骨架在句法结构上首先是一个句子各种成分，可以是并列关系、从属关系或者其他属性描述关系，其在意义表示上具有相对独立性，在骨架内部具有相对完整的句法结构。也就是说，句子的骨架抽取就相当于对句子进行缩句处理，主要是对句子进行主谓宾成分的抽取，缩短后的句子并不改变原句的意思并且句子长度得到减少。汉语句子的基本语法结构可以概括如下：基本成分主谓宾，连带成分定状补，定语必居主宾前，谓前为状谓后补。根据语法特点，缩句方法就是对句子进行中文分词和词性标注后，去除句子的辅助成分，保留句子的骨干成分。

2.2.2　语句骨架成分抽取

针对汉语语句的特点，为了正确地抽取句子的主干成分，一般来说，在句子中名词或者代词作主语和宾语、动词和形容词做谓语。依据句法理论，通常谓语中心词通常有一主语与之对应，谓语中心词在句子中起绝对支配作用[8]。谓语中心词获取不是本文研究重点，在此不再赘述。本文采用文献[9]提出一种利用句子的主语和谓语之间的句法关系来识别谓语中心词的方法，当语句的谓语部分确定下来以后，通过查询语义词典，明确在句子中主语和谓语以及谓语和宾语的常用搭配关系，参考汉语缩句方法进行句子的骨架成分，即主谓宾成分的抽取。通过以上的工作，就可以获得句子的骨架成分，即一个复杂句子的主谓宾成分，既保留了句子的意思又减少了句子的长度，方便后面的相似度计算。

2.2.3　语句级相似度计算

标准答案和学生答案都是由一个或者多个句子组成，句子经过骨架抽取以后，形成只包含主谓宾成分的简单句子。语句级相似度计算也就是对学生答案中的每个简单的句子到标准答案中找匹配的句子，一般来说学生答案只要回答出标准答案知识点就可以评判为正确了，不一定要和标准答案语句顺序一致。每个答案由多个句子组成，语句级的计算首先要找出二维数组(矩阵)中每行的最大值元素。在二维数组中，每一行最多且只有一个最大值元素，因为标准答案不会出现重复的知识点，但是同一列可能会出现多个最大值，因为学生答题的知识点可能会重复或者反复回答某一个知识点，导致该列有多个最大值即学生答案的每句话都和标准答案的一个知识点匹配，所以对每一列来说只记其中一个最大值。然后再把所有最大值不在同一列的元素相加并除以标准答案的知识点语句数。因为学生答案的知识点语句数可能会超过标准答案的知识点语句数，即可得到语句级上的相似度。以下用 ***stuans*** 来表示学生答案；***stdans*** 来表示标准答案。答案句子相似度二维数组如下：

$$\begin{pmatrix} sim(stuans_1,stdans_1), sim(stuans_1,stdans_2),\cdots, sim(stuans_1,stdans_n) \\ sim(stuans_2,stdans_1), sim(stuans_2,stdans_2),\cdots, sim(stuans_2,stdans_n) \\ \vdots \\ sim(stuans_m,stdans_1), sim(stuans_m,stdans_2),\cdots, sim(stuans_m,stdans_n) \end{pmatrix}$$

二维数组中每个元素 $sim(stuans_i,stdans_j)$表示学生答案中的第 i 个子句和标准答案中第 j 个子句的相似度。在二维数组中的子句都是经过处理后的简单句，为了减少计算量，相似度值用 0 和 1 来表示，也就是说，$stuans_i$ 子句和 $stdans_j$ 子句完全匹配则相似度为 1，否则为 0。有了二维相似度数组后，下面就可以方便地进行语句级的相似度计算了。

2.2.4　分数微调规则制定

经过两级相似度计算后，就可以得到学生考试的得分。通过观察和分析大量的学生考试试卷，得出以下规则对学生的分数进行修正和微调，弥补好的学生在机器阅卷中的无辜丢分损失，惩罚差的学生侥幸获得的分数，具体规则如下：

规则 1 如果学生答案长度>标准答案长度，该题分数增加 γ_1；否则分数减少$|\gamma_1|$；长答案包含更多的信息量。

规则 2 如果学生客观题得分较高，分数增加 γ_2；否则分数减少$|\gamma_2|$；客观题得分较高说明学生基础知识比较扎实。

规则 3 如果学生平时成绩较好，分数增加 γ_3；否则分数减少$|\gamma_3|$；平时学习好的同学一般来说可以考出好的成绩。

规则 4 如果学生答案完全抄写题目内容，分数减少$|\gamma_4|$；完全抄写原题目说明该学生不会做题。

规则 5 如果学生提前很多时间交卷并且得分一般，采用另外一种阅卷算法进行 2 次阅卷，进行折中计分；短时间交卷的学生一般都是非常好或者非常差的学生，其他情况可能是阅卷系统误判。

通过采用规则进行分数微调，可以取得一定效果但也会增加阅卷前的工作量，微调分值 γ_1、γ_2 和 γ_3 可能取正数或负数，而 γ_4 只能是负数。针对规则 1~规则 4 可以在数据库中增加相应字段来记录修正值。对规则 4 来说，为了兼顾阅卷效率和准确率，还要在数据库中设计一个字段用来记录答案的长度，在执行规则 4 时，只需要读取学生答案长度和题目长度一致或者是倍数关系的学生试卷就可以了。这样做，根据经验来看效果好漏判少，因为完全抄题的学生根本不会答题，这样做的目的是为了不留空白或者是消耗考试时间。对于规则 5 来说，出现的可能性不是很大，主要是为了减少误判，毕竟对学生来说考试分数是非常重要的。总的来说，规则 1 和规则 4 是针对每道题目来进行分数修正的，而其他规则是从试卷主观题总分上进行调整。根据上述规则特点，在使用时候必须满足以下约束条件：$\gamma_1 \in [-per\times score_i, per\times score_i]$，$\gamma_4 \in [-per\times score_i, 0]$，$(\gamma_1+\gamma_4) \in [-per\times score_i, per\times score_i]$，$2\times\gamma_2$，$2\times\gamma_3$，$(\gamma_2+\gamma_3)$和$(\gamma_2+\gamma_3+\gamma_5)$取值约束条件一样，都为$[-per\times score, per\times score]$，其中，$\gamma_5$ 为满足规则 1 或规则 4 的学生所有主观题修正分数之和，修正后的分数必须在该题目或者试卷主观题总分数的取值范围内；$score_i$ 表示第 i 道主观题的分值；$score$ 表示主观题的总分值；per 是微调参数表示修正分数所占的百分比。鉴于不少学生可能会同时满足规则 2 和规则 3，在此设定 γ_2 和 γ_3 的微调参数取值为 per 的一半。如果修正后的分数超过值域，那么取值域上下限的值作为修正值，上述约束条件必须满足，否则很可能会出现误判或错判。

2.3　两级相似度计算

通过采用词语级相似度计算可以判断学生答案和标准答案中词汇的相似度，主要是关键词的相似度，并没有考虑到词序、语法和语义方面的问题；通过采用语句级的相似度计算，主要考虑答案的整体语法和语义的信息，可能会疏忽对关键词的判断。综合 2 种相似度计算的特点，提出采用词语级和语句级两级相似度算法，可以结合两者的优势给出较为符合人工阅卷效果的评判。两级相似度计算算法如下：

1 挑选出空白答案，分数置零，对规则 2 和规则 3 的修正分值 γ_2、γ_3，以及满足规则 1 或规则 4 的学生所有主观题修正分数之和 γ_5

初始值置零;

2 for(m=1;m≤number;m++)//m 为题号，number 题数，依次处//理试卷的每道主观题

{2.1 规则 2 和规则 3 的修正分值 γ_1 和 γ_4 置零;

2.2 对标准答案和学生答案进行中文分词，通过查停用词表和同义词表，去停用词和同义词消解，把学生答案和标准答案中所有词汇加入关键词词典，用 TF/IDF 方法产生答案向量;

2.3 for(i=1;i≤n+1;i++)

按照依存文法理论，对答案中每个句子抽取句子中心动词谓语，并结合汉语缩句方法抽取句子骨架内容，形成具有主谓宾结构的简单句;

2.4 for(j=1;j≤n;j++)

{采用夹角余弦方法进行学生答案和标准答案间词语级相似度计算，产生词语级相似度 sw_{jm};

(1)对每个学生答案，产生句子相似度二维数组，用完全匹配算法计算二维数组中每个元素值，扫描二维数组每一列，如果某列中非零元素个数多余一个，则只保留一个其余置 0。找出所有非零元素并求和，结果记为 sum_{jm}，然后除以标准答案的语句数 l_m(知识点数)，产生学生答案和标准答案语句级相似度 ss_{jm}，$ss_{jm}=sum_{jm}/l_m$;

(2)计算两级相似度 stl_{jm}，$stl_{jm}=sw_{jm}\times\alpha+ss_{jm}\times\beta$;

//α 与 β 是系数，且 α+β=1

(3)计算第 j 个学生第 m 题得分 $score_{jm}=stl_{jm}\times val_m$;

//val_m 为第 m 题分值

(4)if 学生答案满足规则 1 或规则 4;

分数微调，给出第 j 个学生第 m 题微调分值 $\gamma_1+\gamma_4$，累加第 j 个学生微调分数，即 $\gamma_5=\gamma_1+\gamma_4+\gamma_5$，对 $\gamma_1+\gamma_4$ 和 $score_{jm}+\gamma_1+\gamma_4$ 的值约束合法性进行检查修正；//超过值域的，取值域上下限的值作为修正值

}}

3 计算每位同学主观题总成绩 $score_i$;

4 对满足规则 2 或规则 3 的学生试卷进行分数微调，修正 γ_2 和 γ_3 分值;

5 计算每位学生微调后的最后成绩 $score_i$，$score_i=score_i+\gamma_2+\gamma_3+\gamma_5$，检查 $score_i$ 和 $\gamma_2+\gamma_3+\gamma_5$ 取值约束合法性并修正不合法取值，得到每位同学的最后成绩 $score_i$;

//超过值域的，则取值域上下限的值作为修正值;

6 对满足规则 5 的学生进行二次阅卷，取两次阅卷的平均成绩作为该学生的最后成绩。

3 实验结果与分析

鉴于主观题机器阅卷的复杂性，没有标准的测试集，以计算机专业 100 位同学《操作系统》课程试卷为实验数据，将每个学生答案和标准答案进行两级相似度计算，并且和其他 5 种相似度计算方法在阅卷准确率和速度 2 个方面进行比较和分析。算法效果好坏评价标准自然是和人工阅卷效果相比，借鉴高考等重要考试的阅卷标准，即误差在一定的范围之内就可以认为该评分是有效的。定义误差率为评价标准，且误差率在 20%之内的机器阅卷是有效阅卷，否则为无效阅卷。其中，误差率=|人工阅卷得分−机器阅卷得分|/满分，无效阅卷比例=无效阅卷数目/试卷总数。6 种相似度计算算法无效阅卷比例对比实验结果如图 1 所示，系数 α 表示词语级相似度计算在两级相似度计算中权重大小。6 种相似度计算算法执行速度排序如表 1 所示。由图 1 可知，语句级相似度无效阅卷比例最高，达到 65%，简单关键词匹配仅进行关键词的比较，算法简单速度快但无效阅卷比例很高，达到 45%，词语级相似度计算和语义相似度计算无效阅卷比例为 30%左右，匈牙利算法无效阅卷比例较低，而两级相似度计算在系数 α 取适当范围值的时候无效阅卷比例最低。通过分析学生答案特点和进行人工阅卷后，可以发现语句级相似度计算主要考虑句子的主谓宾整体结构，忽略了一些作为句子辅助结构的重要关键词，这些关键词是答案的知识点；词语级相似度和语义相似度整体来说准确率较高，抓住答案中的关键词和语义信息，但是并没有很好考虑到答案中词序、句序、语法和语义结构，效果并不是非常理想；匈牙利算法基于知网进行义原相似度计算，通过词语的相似度来表达由词语组成的句子相似度，效率较高但较少考虑词序以及句子的语法语义结构；两级相似度计算结合词语级和语句级相似度计算的优点和特点，既考虑到关键词对评分的主导作用，又考虑到句子语法和语义方面的对知识理解的重要作用，只要取合适的系数 α，两级相似度计算方法无效阅卷比例达到最低，并具有较快的执行速度，更加符合人工阅卷的效果。

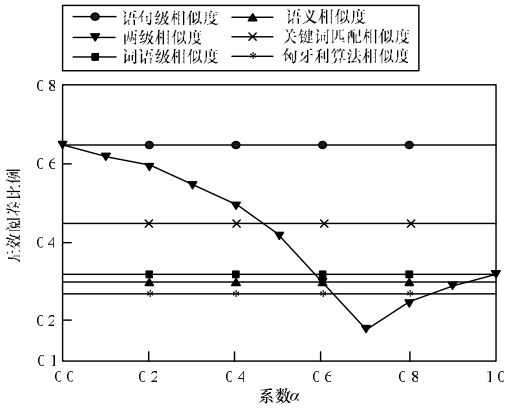

图 1 6 种相似度计算算法无效阅卷比例对比

表 1 6 种相似度计算算法执行速度排序

算法名称	执行速度排序
关键词匹配相似度算法	1
语句级相似度算法	2
词语级相似度算法	3
两级相似度算法	4
语义相似度算法	4
匈牙利算法相似度算法	4

答案主要由一些句子组成，在人工评判试卷的过程中主要是抓住关键词的信息，因为关键词是答案的要点和知识点，是决定答案正确与否的主要因素，而由关键词组成的句子才能真正表达答案的语义信息，句子的骨架成分表达了句子的基本含义，是评判过程中不能忽略的重要因素，只有把词语和句子表达的信息结合起来才能做出更加客观的评判。通过认真观察和深入分析考试过程、标准答案和学生答案的特点和实验数据，可以发现词语级系数 α 在两级相似度计算中权重比例为 70%左右时，阅卷效率最高，具有最高的准确率。最后，通过使用上述规则对学生综合得分进行微调，目的在于修正机器阅卷中的随机误差。由于主观题的复杂性和多样性，即使人工进行阅卷也有误差，在实验中，对规则 1 来说，当学生的答案长度超过(少于)标准答案长度三分之一时 γ_1 取值为该题分数的 5%(−5%)，一半以上 γ_1 取值为该题分数的 10%(−10%)。对满足规则 2 和规则 3 的学生试卷来说，根据学生客观题答题情况和平时表现按比例给出适当的修正分值。对满足规则 4 的答案来说，完全抄写题目通过词语相似度计算是有可能获得分数的，在微调时惩罚该学生侥幸获得的分数，γ_4 取值为该题分数的−10%。规则 5 出现的概率较小影响不大，在本文实验中没有出现这样的情况。

由表 1 可知，语句级相似度、关键词匹配相似度和词语级相似度 3 种简单的相似度计算算法速度较快，而另外 3 种

(下转第 280 页)