文本聚类在大学英语作文自动评分中应用

葛诗利[1],陈潇潇[2]
GE Shi-li[1],CHEN Xiao-xiao[2]

1.华南理工大学 外国语学院,广州 510640
2.广东金融学院 外语系,广州 510520
1.School of Foreign Languages,South China University of Technology,Guangzhou 510640,China
2.Department of Foreign Languages,Guangdong University of Finance,Guangzhou 510520,China
E-mail:geshili@gmail.com

GE Shi-li,CHEN Xiao-xiao.Cluster analysis of college English writing in automated essay scoring.Computer Engineering and Applications,2009,45(6):145-148.

Abstract: The automated essay scoring for the teaching of college English writing requires that the scoring method should have the feature of generality,namely,without pertinency of specific subjects.In the aspect of content evaluation,document clustering can put eassys together according to the similarity of their contents to form a clustering tree which has a higher similarity in the core than in the peripheral area of the tree.A few essays that locate in the peripheral area are quite different from most others in content.These essays are possibly off the topic and will be submitted to teachers for further examination. By this way,eassy contents can be evaluated accurately with only minor labor expense.Experiment shows that this method can identify essays off the topic effectively with a reasonable threshold value of content similarity.

Key words: document clustering;automated essay scoring;college English;writing teaching

摘 要: 面向大学英语写作教学的自动作文评分要求评分方法具有针对非特定作文题目的通用性。在作文内容评价方面,文本聚类能够把作文按内容的相似程度聚集到一起,从而形成一棵内密外疏的聚类树。位于聚类树外围的少数与其它作文内容差异较大,即可能跑题的作文可以反馈给教师进行人工判断,从而花费较少的人力即可做出较准确的作文内容评价。实验表明,通过设置合理的相似度阈值,该方法能够有效识别跑题作文。

关键词: 文本聚类;自动作文评分;大学英语;写作教学
DOI: 10.3778/j.issn.1002-8331.2009.06.041 **文章编号:** 1002-8331(2009)06-0145-04 **文献标识码:** A **中图分类号:** TP391

1 引言

自动作文评分(Automated Essay Scoring,AES)近年来已渐成为自然语言处理研究中的热点问题。自动作文评分研究中通常都包含内容方面的评价,只是针对不同人群作文的研究中,内容所占比重各有不同,有的研究甚至只依靠内容对作文进行评价。文秋芳[1]的最新研究表明"作文内容能够解释作文总体质量56%的差异"。虽然她的研究对象是中国英语专业学生的英语作文,但也明确说明了作文评价中内容的重要性。

国外AES领域有代表性的几个系统包括PEG、IEA、E-rater、IntelliMetric和BETSY等[2],国内相关研究如文献[3-4]等。这些研究中除了PEG未涉及内容[5]之外,其他研究均包含内容方面的评价。但对于面向大学英语写作教学的自动评分来讲,上述这些内容方面的研究都难以达到应用的要求,因为所有内容评价相关的研究都具有题目针对性,而大学英语写作教学中的自动评分研究面向不定题目的作文,必须具有通用性。

为了研发面向大学英语写作教学的通用自动评分系统,采用了语言质量和内容分别评价的方法,其中的内容评价采用文本自动层级聚类这种试探性数据分析(exploratory data analysis)来识别跑题作文,并辅以人工鉴别。这种内容评价方法的特点是不需要事先基于大规模标注训练集构建评价模型,并且有着层级聚合聚类法的突出优点,即能够生成比较规整的类集合,聚类结果不依赖文档的初始排列或输入次序,与聚类过程的先后次序无关,聚类结果比较稳定,不易导致类的重构。并且对于作文评价来讲,得到的结果比较容易解释。实验结果表明,该方法能比较清晰地识别与大多数作文内容不同的作文,再辅以人工鉴别,可准确识别跑题作文,从而在通用自动作文评价中实现作文内容的测量。

首先回顾了当前AES领域对作文内容评价的长处和弱点,然后提出了针对大学英语写作教学的通用自动作文评分中的内容评价方法,最后通过实验检验了该方法的效果。

2 当前自动作文评分中的内容研究

根据对作文中语言和内容评价的侧重,当前的AES研究

基金项目:国家自然科学基金(the National Natural Science Foundation of China under Grant No.60272055,No.60572159)。
作者简介:葛诗利(1969-),男,博士,讲师,主要研究领域为自然语言处理;陈潇潇(1975-),女,博士生,主要研究领域为应用语言学。
收稿日期:2008-08-27 修回日期:2008-09-24

可以划分为3类:侧重语言形式的自动评分,如PEG[6];侧重语义,或者说内容的自动评分,如IEA[7];二者兼顾的自动评分,如E-rater[8]。其他研究基本上也可以划归其中的一类,如Intelli Metric属于二者兼顾[9];基于文本分类技术的自动评分系统BETSY虽然说可以兼顾包括语言使用的各方面特征,但目前研究的分类方法都是基于词汇,而且主要是实义词汇,所以应该划归第二类,即偏重内容的自动评分[10];文献[3]的研究属于第三类;文献[4]的研究非常类似于IEA,是基于潜伏语义分析(Latent Semantic Analysis)的汉语作文评分,明显属于第二类。

PEG虽然很早就达到了较高的评分准确率,但由于其对语义,也就是内容方面的忽视和更多地注重表面结构而遭受指责[11]。也就是说,自动作文评分必须考虑内容,否则,即使评分准确率较高也难以推广使用。

IEA以及类似的只基于潜伏语义分析的作文评分实际上测量的是"文本的内容和学生作文中所传达的知识,而不是作文的风格或语言"[12]。这种方法用于语言使用熟练,基本没有较大语言失误的本族语作文效果较好,但用于需要衡量语言质量的外语作文就会出现很大的问题。因为潜伏语义分析只是分析词汇以及在词汇使用的情况下,分析词汇间共现关系,而作文写作,尤其是外语作文写作,不仅仅是词汇的堆砌,更要考察由词汇构成的搭配和短语、并进而构成更高级的句子是否准确和适当。所以,单独的潜伏语义分析用于外语作文评分显然不够。

类似BETSY的基于文本分类技术的自动评分研究也存在同样的问题。因为文本分类就是把文章看成词汇的集合,进而构建词汇向量,用以代表文章本身做相似性比较,而词汇的顺序根本不予考虑,更不用说句法结构了,所以这种方法也不适合单独用于外语作文评分。

真正能够而且目前正用于外语作文自动评分的只有语言质量和内容兼顾的AES系统,如E-rater和IntelliMetric[8-9]。E-rater的内容分析采用了向量空间模型。事先经人工评分的训练集中每篇作文首先转换成为词频的向量,待评分作文也转换成为词频的向量,并与训练集作文向量一一比较,与待评分作文最相似的训练集作文的分数就是待评分作文的内容得分。然后合并语言质量得分,可计算得到作文的最终得分。这种方法既考虑到语言质量,同时也兼顾了作文内容,再加上二者权重的调节,对于外语作文评分可达到较好的评分效果。

3 作文聚类研究

尽管目前语言和内容兼顾的AES系统对外语作文评分已可达到较高的准确率,但对于我国大学英语写作教学上的应用仍存在诸多的问题。首先,较成熟的商用系统研究细节并不公开,使用起来价格昂贵;然后,这类系统多由国外研究机构开发,对我国大学生英语写作评价的针对性不强;最后也是最关键的问题,目前AES研究通行做法都是收集已评分的某一题目作文若干篇作为训练集,建立并训练评分模型,最后用于同一题目其它作文的评分。这样就可以把待评分作文与作为标准的训练集作文相比较,包括语言和内容两方面的比较,从而得出最终的评分。这种方法用于大规模考试效率较高,人工评阅数百份试卷,训练模型后可用于数万份甚至数十万份试卷的自动评分。但是大学英语写作教学中,一位教师所教授的学生通常不超过200人,布置一个题目的作文,收集起来尚不足以训练评分模型。因此,提出了面向大学英语写作教学的、非特定题目的、通用自动作文评分方法。这种评分方法中,作文语言质量和内容分别评价。语言质量方面,由于不需要考虑特定题目,可以采用一次训练多次使用的语言评价模型;但内容评价方面,由于不针对某一特定题目,也就难以事先训练评分模型。不经训练而能够对内容进行分析的只有信息检索中常用的文本自动聚类。聚类研究能通过对文档词汇的分析和计算,把语义或者内容相似或者相近的文档自动聚集到一个类别中。对于同一题目的多篇作文,自动聚类同样可以把它们按照内容的相似程度聚合到一起。少数与其它作文内容差异较大,即可能跑题的作文可以反馈给教师进行人工的判断,从而花费较少的人力即可做出较准确的作文评价。

4 聚类综述

4.1 聚类的定义与过程

聚类技术已成为文本挖掘和信息检索等多个领域的研究热点,其"目标是将一组对象划分成若干组或类别,简单地说就是相似元素同组、相异元素不同组的划分过程"[13],它是一种无指导学习的基本方法。

给定一数据样本集$X=\{X_1,X_2,\cdots,X_n\}$,根据对象间的相似程度将样本集合分成k簇:$\{C_1,C_2,\cdots,C_k\}$的过程称为聚类。

聚类的结果使划分为一个簇(cluster)的成员具有较大的相似性,而不同簇之间的成员具有较大的相异性。簇就是数据对象的集合。一个好的聚类方法生成的簇具有高的类内相似度和低的类间相似度。

文档聚类(Document Clustering)就是对文本信息的聚类,最初用来研究提高信息检索的查准率和查全率,或作为查找最相似文档的有效方法。要进行文档聚类,首先要把文档表示成计算机能够理解的形式,也就是信息检索中常用的向量空间模型。要把文档表示成向量空间模型形式,首先要对文档进行预处理,提取能够代表文档内容的特征,一般来说是词汇。需要的话,同时进行一些特殊词汇的识别,如地名、人物名、地址、电话等。然后,过滤掉停用词并对单词进行词根还原(stemming)。最后,删除出现频次过低而对聚类作用不大的词汇,构成文档向量。

有了代表文档的向量,就可以计算文档间的相似度了。理想情况下,是比较文档间的语义或者内容相似度。但语义和内容很难计算,所以就采用词汇统计的方法来近似求取文档间语义或内容相似度。这已是信息检索领域比较成熟的方法了,E-rater的内容比较也是采用此法。相似度的计算以向量间的欧式距离或者向量间的夹角余弦来衡量,距离越近或者夹角越小,两个向量代表的两个文档就越相似。

4.2 硬聚类和软聚类

根据聚类后的元素属于所划分的类别的概率,可以将聚类分为硬聚类和软聚类两种。

硬聚类(Hard Clustering)指一个元素只能属于一个类别。

软聚类(Soft Clustering)中,一个元素可以同时属于几个类别,用概率表示属于每个类别的程度。

4.3 层级聚类与非层级聚类

按照聚类的具体实现方法,聚类的算法分为层级聚类(hierarchical clustering)和非层级聚类(non-hierarchical clustering)两种。层级聚类的结果可以表示为一个树的图形,每个节点都是父节点的一个类。非层级聚类结构比较简单,类别之间没有层级关系,其算法是一个迭代的过程,首先经过初始聚类,

通过不断的迭代重新分配数据的类别。

(1)层级聚类

层级聚类随着类别层次的变化，类别中的元素也同样发生变化。层级聚类形成一棵类别树。每个类结点还包含若干子结点，按照类别树的生成方式，可将层级聚类法分为自底向上法和自顶向下法。自底向上(Bottom-up)，也称为合并聚类；自顶向下(Top-down)，也称为分割聚类。

自底向上的算法中，每个对象都被初始化为一个类别集合，然后反复合并两个或多个合适的类别，减少类别的数目，当只存在一个包含所有对象的类时或满足某个终结条件时，算法结束。

自顶向下的算法正好相反，先将整个集合看成一类，然后逐渐反复从结点分裂出新的子结点，增加类别的数目，直至每个对象都成为一个类别，或者满足某个终结条件，算法停止。

(2)非层级聚类

非层级聚类从一个初始划分开始，任何一个样本都可以作为种子(seed)，作为初始聚类的中心点，也叫质点，在此基础上进行迭代，将样本数据进行再分配，重新划分出新的类别。当迭代的结果不能起到提高分类效果的时候，迭代过程结束。

4.4 聚类流程

本研究的任务是将同一题目的多篇作文按内容或者说所用词汇的相似程度聚类。虽然以文章内容为标准进行聚类会产生一定的分歧，即一篇文章从一个角度看属于 A 类，但从另外一个角度看属于 B 类，即存在软聚类情况。但是本研究只以文章所用词汇为标准进行层级聚类，最终根据文章所用词汇的相似程度形成一棵类别树，该树上与大多数节点相似程度较低的个别节点会被反馈给教师做最后的判断。因此，本研究中聚类标准是统一的，属于硬聚类中的层级聚类。本研究聚类的具体流程如图 1 所示。

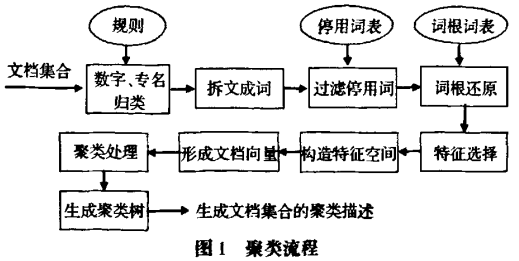

图 1　聚类流程

5 聚类方法

5.1 聚类步骤

基于向量空间模型的常见文档聚类算法有分割算法中的 k-means 算法和层级算法中的凝聚层级算法(Hierarchical Agglomerative Clustering，HAC)。

k-means 算法以 k 为参数，把 n 个对象分为 k 个簇，以使簇内具有较高的相似度，而簇间的相似度较低。其过程如下：

(1)在 n 个对象中随机的选取 k 个对象作为初始的聚类中心；

(2)把其余 $n-k$ 个对象归到距离最近的聚类中；

(3)重新计算每一个聚类的中心；

(4)重复(2)和(3)，直到每一聚类的中心不再改变。

这种算法实质是一种多次迭代的方法，把每篇文档看成一个对象，利用文档与聚类中心之间的相似度来进行聚类。

HAC 算法将数据对象组成一棵聚类的树。根据层级分解是自底向上还是自顶向下形成，层级的聚类方法可以进一步分为凝聚的(Agglomerative)和分裂的(Divisive)层级聚类，但凝聚的层级算法更多用于文档的聚类。其过程如下：

(1)把 n 个对象作为 n 个聚类，计算所有聚类两两之间的相似度；

(2)合并最为相似的聚类；

(3)重新计算更新后所有聚类两两之间的相似度；

(4)重复(2)和(3)，直到只余下一个聚类。

这种算法尽管方法简单，聚类效果一般比 k-means 算法要好，因此，本研究采用此算法。

HAC 聚类融合过程就产生一个树状层级结构。考虑 n 个样本聚成 c 类的情况。首先，将所有样本分成 n 类，每类正好含有一个样本。其次，将样本分为 $n-1$ 类，接着是 $n-2$ 类，这样下去直到所有样本都被分为一类。称聚类数目 $c=n-k+1$ 对应层次结构的第 k 层，因此第 1 层对应 n 个类别，而第 n 层对应一个类别。图 2 是 4 个样本的聚类情况示例。底层一行是作文在集合中的标号，其中左面的数字是作文的评分。

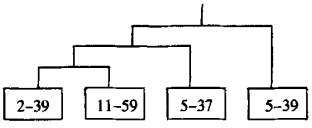

图 2　聚类情况示例

5.2 特征选择和相似度计算

(1)特征选择

依照图 1 的流程，首先识别数字和专名，然后把所有作文拆分成词汇，根据停用词表去除 320 个冠词、介词、代词和表示属格的's 等停用词，再根据词根词表进行词根还原(stemming)。如：accept，accepts，accepted，accepting，acceptable，acceptability 等都识别为 accept。最后去除所有作文中词根还原后只出现 3 次及以下，对聚类效果贡献不大的低频词，剩下的词汇作为聚类研究的特征项。统计这些单词分别在每篇作文中出现的次数，构成数量等于作文数、维数等于特征数的向量组。

(2)聚类计算

聚类中的计算包括三个方面：第一个是单独一篇作文特征向量的建立，第二个是两个特征向量的相似度计算，第三个是聚成的类的特征向量计算。

建立特征向量的计算采用信息检索中常用的 TF-IDF (Term Frequency-Inverse Document Frequency)方法。首先计算每篇作文中词语权重，然后通过计算两篇作文向量夹角余弦得到它们之间的相似度，最后通过比较相似度来进行聚类。

即文档向量：$D=\{TF_1, TF_2, \cdots, TF_n\}$；特征项(Term)权值：

$$W_{ik}=TF_{ik}\times IDF_{ik}=TF_{ik}\times[\mathrm{lb}(\frac{N}{n_k})+1] \quad (1)$$

TF-IDF 中 TF 表示某个特征项 T 在文档 D 中出现的次数，如 university 在某篇作文中出现的次数。DF 表示整个作文集合中，包含特征项 T 的作文篇数在整个作文集合中所占的比例，如出现单词 university 的作文篇数在某一集合的作文中占多大比例。n 为作文集合中的作文总数量。n_k 为包含特征项 T_i 的作文总数，如出现 university 这个单词的作文篇数。

两个特征向量的相似度(Similarity)通常用余弦值来测量。如公式(2)所示。

$$\mathrm{sim}(D_i,D_j)=\cos(D_i,D_j)=\frac{D_i\cdot D_j}{|D_i||D_j|}=\frac{\sum_{k=1}^{n}W_{ik}W_{jk}}{\sqrt{\left(\sum_{k=1}^{n}W_{ik}^2\right)\left(\sum_{k=1}^{n}W_{jk}^2\right)}} \quad (2)$$

两篇及以上作文会产生一个聚类，这里采用聚类中心点来表示。其特征向量的特征值计算如公式(3)所示。

$$WC_i=\frac{1}{n}\sum_{j=1}^{n}W_{ij} \quad (3)$$

公式中 WC_i 是第 i 个特征项在聚类中心的权值；W_{ij} 是第 i 个特征项在第 j 个文档向量中的权值；n 是聚类中作文总数。

具体聚类算法如下(以全部数量为 n 篇的某题目作文集合为例)：

Begin

令 $k=n$；$c_i=D_i$，$1\leqslant i\leqslant n$。初始化阶段认为每篇作文是一个聚类；

While $k>1$ do

计算所有的聚类之间的相似度 $s_{ij}=\mathrm{sim}(c_i,c_j)$，其中 $1\leqslant i<j\leqslant k$，找出两个聚类，它们之间的相似度最大，不失一般性，记这两个聚类为 c_i,c_j；

把 c_i,c_j 融合成一个新聚类。不失一般性，计算其聚类中心点，并用 c_i 表示这个新聚类；

删除 c_j，置 $k=k-1$；

End while

结束

End

作文聚类研究的目标是自动发现同一题目作文中跑题或者说与其它大多数作文内容不同的作文。但由于本研究所采用的作文样本来自中国学习者英语语料库(CLEC)的大学英语四级考试子库(st3)。该子库中 1319 篇作文中，未发现明显偏离题目内容的作文，因此，单独某一题目作文的自动聚类很难说明问题。本实验从全部作文集合中选取数量为 55 篇的一个题目作文，题目为 health gains in developing countries，另外选取两篇题目较为相近的作文作为参照，题目为 health and wealth。共 57 篇作文，构成聚类作文集合。这样就可以有一个较为客观的衡量标准。

6 聚类结果及反馈

作文的内容聚类不存在标准测试集。本研究以两篇题目相近但不相同的作文模拟完全跑题的作文，采用层级聚类方法，聚合成为一棵聚类树。

观察该聚类树可以清楚地看出，标号为 5-315 和 8-221 的两篇不同题目作文之间相似度较低(0.497 150 909 883 04)，在 56 次聚类循环中，到第 44 次才聚到一起成为一类。而这一类直到最后一个聚类循环才以极低的相似度(0.097 590 820 141 136 1)合并到整个聚类之中。此实验结果表明，尽管题目相近，但内容相差确实较大的作文，还是非常容易以该方法自动发现的。但是，在学生的作文写作中，尤其是日常主动的写作练习中，彻底跑题的作文非常少见。而作文从完全扣题到彻底跑题之间存在着很长一段过度区间，这之间的作文是否符合题意或者说是否跑题往往见仁见智，教师之间也不一定能达成一致。也就是说，作文是否符合题意存在一个程度的问题。程序对此向教师提供反馈的时候，可以适当设定一个相似度阈值，低于此阈值的聚类需要反馈给教师人工作出进一步的判断。如在本实验中，阈值可以设为 0.4 或者 0.3。这样，两个不同题目的作文就会被自动识别并提供反馈。

当然，内容聚类的判定不仅有程度的问题，另外一个不容忽视的问题是同一题目作文，有时也会产生内容相差很大的作文，如 2007 年 6 月的大学英语四级考试作文"Join the club"。在该题目作文中，大部分学生讨论了加入英语俱乐部或者英语角能够获得的益处，但也有学生探讨加入足球队或者篮球队的得失，甚至还有学生议论了加入音乐团队或者集邮小组的是是非非。虽然由于缺少语料，未能对该题目作文进行聚类实验，但仅凭经验也可推测这种题目作文的整体自动聚类反馈效果不会具有参考意义。但不同内容的作文会分别被聚到不同的大类中。对每一个大类，可以反馈该类中个别与大类相似度较低的作文。所以，自动聚类方法在计算机作文评分中作为为内容把关的最后一道手续，还是能够起到应有的作用。

参考文献：

[1] 文秋芳."作文内容"的构念效度研究[J].外语研究，2007(3)：66-71.

[2] 葛诗利，陈满潇.国外自动作文评分技术研究[J].外语电化教学，2007(5)：25-29.

[3] 梁茂成.中国学生英语作文自动评分模型的构建[D].南京：南京大学，2005.

[4] 曹亦薇，杨晟.使用潜语义分析的汉语作文自动评分研究[J].考试研究，2007，3(1)：63-71.

[5] Valenti S，Neri F，Cucchiarelli A.An overview of current research on automated essay grading[J].Journal of Information Technology Education，2003(2)：319-330.

[6] Page E B.Project essay grade：PEG[M]//Shermis M D，Burstein J.Automated essay scoring：A cross-disciplinary perspective.Mahwah，NJ：Lawrence Erlbaum Associates，2003：43-54.

[7] Landauer T K，Laham D，Foltz P W.Automated essay scoring and annotation of essays with the intelligent essay assessor[M]//Shermis M D，Burstein J.Automated Essay Scoring：A Cross Disciplinary Perspective.Mahwah，NJ：Lawrence Erlbaum Associates，2003：87-112.

[8] Burstein J.The e-rater scoring engine：Automated essay scoring with natural language processing[M]//Shermis M D，Burstein J.Automated essay scoring：A cross-disciplinary perspective.Mahwah，NJ：Lawrence Erlbaum Associates，2003：113-122.

[9] Elliot S.IntelliMetric：from here to validity[M]//Shermis M D，Burstein J.Automated essay scoring：a cross disciplinary perspective.Mahwah，NJ：Lawrence Erlbaum Associates，2003：71-86.

[10] Rudner L M，Liang T.Automated essay scoring using Bayes' theorem[J].The Journal of Technology，Learning and Assessment，2002(2)：3-21.

[11] Kukich K.Beyond automated essay scoring[C]//Hearst M A.The debate on automated essay grading，2000(5)：27-31.

[12] Foltz P W，Kintsch W，Landauer T K.The measurement of textual coherence with Latent semantic analysis[J].Discourse Analysis，1998，25：285-308.

[13] Manning C D，Schutze H.统计自然语言处理基础[M].苑春法，译.北京：电子工业出版社，2005：310.