LSA 在中文短文自动判分系统中的应用研究

李 莉,张太红
LI Li,ZHANG Tai-hong

新疆农业大学 计算机与信息工程学院,乌鲁木齐 830052
College of Computer and Information Engineering,Xingjian Agricultural University,Urumqi 830052,China
E-mail:25580920@sina.com.cn

LI Li,ZHANG Tai-hong.Application researches of latent semantic analysis in Chinese essay auto scoring system construction.Computer Engineering and Applications,2007,43(20):177-180.

Abstract: The basic idea of Latent Semantic Analysis(LSA) is described in the introduction of the paper.The application of LSA to the Chinese writing essays' auto scoring is studied in this paper.The influence of machine auto scoring results of different data standard processing and different semantic space construction methods are compared basing on the analysis of 136 college students' writing test paper.The effect of Singular Value Decomposition (SVD) and the rules for definition of the K numbers of singular values are discussed.In the last part of the paper,two teachers validate the results.It shows that it is feasible to score Chinese subjective test by machine auto scoring system.The method proposed in this paper offers a new option for variety test questions of auto test system.

Key words: Latent Semantic Analysis(LSA);Singular Value Decomposition(SVD);machine auto scoring in subjective test

摘 要:对潜在语义分析(Latent Semantic Analysis,LSA)的理论基础进行了介绍,研究了潜在语义分析在中文短文写作自动评分领域的应用方法。从136名大学生的短文写作试卷着手,对比了不同的语义空间构造方法和不同数据标准化方法对机器自动评分结果的影响,探讨了SVD的作用和奇异值个数K的取值规律,比较了LSA对不同类型学生的短文写作自动评分结果的差异。通过与两名教师对学生短文写作评分的比较表明,使用机器对主观题进行自动评分是可行的,该方法为自动化考试系统试题多样性提供了有效的解决方案。

关键词:潜在语义分析;奇异值分解;主观题自动判分
文章编号:1002-8331(2007)20-0177-04 **文献标识码**:A **中图分类号**:TP311

考试作为考查学生学习和掌握知识的程度及评估学校教学水平的手段由来已久,并且还会在今后相当长的一段时间内存在下去。目前我校自主开发的基于B/S模式的考试系统"网上在线考试系统"(主要面向计算机专业课程的测评)主要包含四种类型的测试:单选题、多选题、判断题及填空题。该系统虽然在一定程度上缓解了教师手工方式评测的弊端,如评改工作量大、工作效率不高、教学反馈周期长等缺陷,但是它所使用的考评方式均以客观题为主,测评方式单一,无法满足考试类型多样性的要求。为此笔者计划使用潜在语义分析(Latent Semantic Analysis,LSA)技术为考试系统加入主观题(即对中文短文自动评分)的测评,并论证LSA在中文短文自动评分系统构建中的可行性。

1 理论基础

潜在语义分析(Latent Semantic Analysis,LSA)是一种用于知识获取和表示的理论和方法。LSA通过分析大量的文本集,自动生成关键字/概念(语义)之间的映射规则,并通过统计方法提取和量化这些潜在的语义结构,进而消除同义词、多义词的影响,提高了文本分类的准确性[1]。

LSA方法在提取信息处理上可分为两个阶段:第一个阶段为预处理阶段。在这个阶段,应用奇异值分解(Singular Value Decomposition,SVD)建立一个术语对文档的语义空间。即构造一个训练集$m\times n$词条/文本矩阵$\boldsymbol{A}=[a_{ij}]$,其中m为提取单词数,n为文本数。对矩阵$\boldsymbol{A}$进行截取SVD分解(设$m>n$,$rank(A)=r$,存在K,$K<r$且$K\ll\min(m,n)$=。此时矩阵$\boldsymbol{A}$可以近似表示为:$\boldsymbol{A}\approx\boldsymbol{A}_K=\boldsymbol{U}_K\sum_K\boldsymbol{V}_K^{\mathrm{T}}$。其中,$\boldsymbol{U}_K^{\mathrm{T}}\boldsymbol{U}_K=\boldsymbol{V}_K^{\mathrm{T}}\boldsymbol{V}_K=\boldsymbol{I}_K$,$\boldsymbol{U}_K$和$\boldsymbol{V}_k$分别被称为矩阵$\boldsymbol{A}_k$的左、右奇异向量,$\sum_k$是奇异值按递减排列的对角矩阵,对角元素被称为矩阵$\boldsymbol{A}_k$的奇异值[2]。第二个阶段为求取相关系数阶段。即应用相关系数测量来分析词与词、词与文档、文档与文档之间的相似程度。两个阶段是同等重要的,预处理阶段完全独立于第二个阶段。

为将潜在语义分析(Latent Semantic Analysis,LSA)方法应用到中文写作的短文中,本研究将首先将文本转化为空间向量矩阵,其次对形成的空间矩阵进行数据标准化分析,接着使用LSA/SVD方法对生成的矩阵进行处理,最后是对实验结果进行对比,即机器自动评分与两位教师评分及两位教师评分的平均分之间进行相关系数对比并论证LSA在中文短文自动评分系统构建中的可行性,为今后在线考试系统中加入主观题

作者简介:李莉(1980-),女,硕士,主要研究方向:数据库技术及应用。

(短文自动评分或问答题自动评分等方式) 的测评奠定良好的理论和现实的基础。

2 实验设计

在实验中,首先要求新疆农业大学计算机信息与工程学院的信息管理与信息系统专业和计算机科学与技术专业的 134 人和新疆农业大学科学技术学院的计算机科学与技术专业的 18 人共计 152 名在读三年级大学生参与此次测试。但是实际测试中剔除未参加实验测试的人员 13 人, 在测试中未作答即成绩以零分计算的 3 人,因此,实际参加测试的人数为 136 人。在做测试前, 他们每人都写过一篇约 3 000 字左右的文献综述,综述主要阐述了有关数据库发展简史的内容。因此,假设他们均已具备进行实验测试所需的相关背景知识。

接下来,这些学生在阅读完一篇参考资料后,收回阅读资料并要求学生在规定的时间内独立完成关于数据库发展简史内容相关的 250 字左右的短文书写。然后,将这些短文交给新疆农业大学计算机信息与管理工程学院的两位教师 T1 和 T2,由他们对学生书写的短文分别进行评分。评改过程中教师 T1 与 T 之间的评分是不可见,即这两位教师之间的评分并不相互影响。他们在参阅标准参考答案的基础上,对每一篇短文给出评分。分数在一定程度上反映两位教师对学生关于该主题了解并正确传达了多少知识的评估。

为应用潜在语义分析技术对学生书写的短文进行评分,需要通过数学方法构建潜在语义空间模型。这是潜在语义分析实现的一个关键性的问题,直接影响潜在语义分析的性能。

2.1 建立训练集

首先是训练集的产生。LSA 利用训练集分析获得的语义知识,对自然语言文本进行分析并确定文本的主题,从而自动提取文本的信息。这些文本中的主题信息,在语义空间中可用一个个向量来表示。当文本提供了关于主题的新的信息时,可自己潜在地修改和扩充语义空间。由于中文语言自身的特殊性,势必决定其训练集的设定将是一个非常复杂的处理过程。

选取《数据库综合大辞典》[3]、《数据库系统概论》[4]和《中国大百科全书》[5]相关部分的内容及 IT 专家网 _What is、www.ynfn.gov.cn、www.zhirui.com 等网络上的相关内容作为训练集。该训练集的六篇文章总计 23 740 个字符, 平均每篇文章字符数为 3 957,平均段落数为 42 段。

大家知道词是最小的能够独立活动的有意义的语言成分,而汉语是以字为基本的书写单位,词语之间没有明显的区分标记。因此,中文词法分析是中文信息处理的基础与关键。为此,采用中国科学院计算技术研究所研制出的基于多层隐马模型的汉语词法分析系统 ICTCLAS(Institute of Computing Technology,Chinese Lexical Analysis System)。该软件分词正确率高达 97.58%(最近的 973 专家组评测结果), 假设该软件的分词精度已符合设计实验所要求的精度,即对由分词产生的误差不做任何操作性的处理, 使用 ICTCLAS1.0 版本对短文进行分词处理并添加入后台数据库中。

对这六篇文章进行分词处理后,由分成的词和这六篇文档组成了由 6 个列向量和 1 736 个唯一的词/词语构成的行向量组成的训练矩阵。矩阵中行向量和列向量的交叉处为该词(行向量)在该文档(列向量)中出现的次数。这六篇文档分别用 doc1、doc2、doc3、doc4、doc5、doc6 表示,这里选取第六篇文档即 doc6 作为实验测试用的阅读参考资料。

2.2 文本预处理

在文档预处理中,使用停词表(Stop List)切除非信息词。这是提高结果准确度,降低计算冗余的常用方法。非信息词一般由停词表定义, 其中主要有构成语法的词和一些高频词,中文中构成语法、高频词的如“的,地,得,很,了”等等。大多数文本信息处理系统使用的停用词表含有的停用词基本相同,使用相同的停用词表可以安全地切除文档中的一般冗余词,很少会产生系统的处理精度下降,但也很难显著地提高系统效率。

实验数据测试中,直接对原始矩阵进行数据处理的方法称为方法一;选取至少在两篇或两篇以上的文章中出现过的行向量构成的矩阵进行数据处理的方法称为方法二。选取方法二的目的是,假设对于只出现一次的词,其对推出语义空间贡献很小,并且为了减少计算量。其中这两种方法构成的 LSA 语义空间矩阵为:由 143 列和 2 522 行组成的方法一的语义空间矩阵和由 143 列和 1 313 行组成的方法二的语义空间矩阵。其中前六列为训练集文档列,第七列为标准答案列,其余各列为相应人数的学生短文。对这两种方法产生的矩阵使用两种不同的方式进行分别处理,即使用和不使用 LSA 方法,以验证 LSA 方法是否对实验结果产生有效的影响。

3 实验处理过程

LSA 方法在提取信息处理上可分为两个阶段:即预处理阶段和分析阶段。预处理阶段过程如本文一中所述,这里主要介绍分析阶段过程。

本课题实验在分别完成了将学生短文及其分词录入到后台数据库中以后,即实验准备工作完成后,首先使用几种数据标准化处理方法分别对 LSA 空间矩阵进行使用和不使用 LSA 方法进行分析,并相应给出机器自动评判的学生成绩。在学生成绩处理中仅简单的将相关系数乘以 100 作为百分制的成绩结果,即:学生成绩=相关系数 *100。

3.1 数据标准化处理方法

权重也称权值,一个指标权重的大小反映该指标在整个评价指标体系中的重要程度,权重越大说明其越重要。给予各种词不同的权重,其目的就是区别各种词在文档中的重要性。权重的统计结果可以表明它们所反映的词所在文档的分布状态。矩阵转换通常使用“权重”,其目的是为了让矩阵中(i,j)的内容更好地接近术语和文档之间的相互关系:列相关于文档,行相关于术语(词或重要的短语)[6]。其中,主要介绍权重测量中的词频与倒文档频度方法、最大正规化方法、对数词频法和余弦正规化法四种方法的具体内容及其对空间矩阵的影响。

3.1.1 词频与倒文档频度方法

在索引词的权重计算中最为成功的和广泛应用的方法称为“词频与倒文档频度”(Term Frequency*Inverse Document Frequency,TF·IDF)方法。该方法将一个词在单个文档中的重要性和整个数据全集中的重要性结合起来, 成为一个统一的量度。

一个词 ki 在文档 d_j 中的权重 $\omega_{i,j}$ 由下式计算:$\omega_{i,j}=TF_{i,j}\cdot IDF_i=freq_{i,j}*\log(n/n_i)$,其中,$freq_{i,j}$ 表示词 k_i 在文档 d_j 中的频度,n 为数据全集中文档的总数,n_i 为包含词 i 的文档总数。函数 log 的底可以取 10,自然对数 e 或者 2。在实验中统一选取自然

对数 e 为底的 log 函数。由该公式说明一个在单文档中频度很高，而在整个数据全集中频度很低的词是更加重要的词。

3.1.2　最大正规化法

针对词频的改进主要是将词频进行正规化处理，将它反映为一个在区间[0,1]中的量。改进方法之一是将词频除以某个与包含该词文档的索引词总数相关的因子，如文档中词的总数或者文档中具有最大频度的词的频度等，即 $TF_{i,j}=freq_{i,j}/\max f_k\{freq_{k,j}\}$。这类改进方法称为“最大正规化”(Maxfimum Normalization)法。

3.1.3　对数词频法

另一个经常使用的词频正规化称为“对数词频”(Logarithmic Term Frequency)法，该方法使用如下公式计算词频值：

$TF_{i,j}=log(freq_{i,j}+1)$

3.1.4　余弦正规化法

另一类正规化方法是通过整个文档向量的长度来实现。当一个文档向量构造完成后，该向量的每一维都设定了对应词的 TF·IDF 值，将这个向量的所有维上的这些值都除以该文档向量的欧氏长度，即得到经过正规化的文档向量。一个向量的欧氏长度是该向量所有分量平方和的平方根。由于经过正规化后的向量具有单位长度，而且在每一维上的值恰好是该向量及其在这一维上相应坐标轴上投影的夹角的余弦值，因此，这种正规化方法又称为余弦正规化法(Cosine Normalization)。余弦正规化法解决了文档中少数高频词对其他词权值扰动过大这一问题。

3.2　LSA 在 MATLAB 中的实现过程

在 MATLAB 中，使用 xlsread 函数导入数据；使用 corrcoef 函数生成数据数组的相关系数矩阵；使用 svds 函数进行 LSA/SVD 的实现运算。其中，使用相关系数的方法对 SVD 分解后的数据进行运算时，过程如下所示，其中矩阵 $\boldsymbol{X}$ 是 $m\times n$ 的矩阵：

```
>>for k=1:1:n
[U,S,V]=svds(X,k);A=U*S*V';c=corrcoef(A);YX(:,k)=c((1:n),1);
end
```

计算后的数据使用百分制的公式对其进行分数计算，即 $\boldsymbol{cj}=YX*100$。对 $\boldsymbol{cj}$ 矩阵中的负值全部使用零值来替换，因为学生成绩评判没有负值。接下来，进行人机数据对比。

4　实验结果及分析

通过使用相关系数方法对上述标准化方法处理后的数据进行分数评判后，可以求取语义空间矩阵及其数据标准化处理后产生的分数与两名教师评分之间的相关系数。在使用 LSA 方法中选取 K 从第一维一直到最大一维(136 维)中的最大的相关系数列在表中，表中对数词频方法即 log 列的所有数据结果为负值，选取其绝对值最大的相关系数列出在表中。其实验结果如表 1 所示，其中，F1 表示方法一，F2 表示方法二，all 表示所有学生；cipin 表示直接对空间矩阵进行运算的相关系数，tfidf 表示使用 TF·IDF 方法进行运算的相关系数，max 表示使用最大正规化方法进行运算的相关系数，log 表示使用对数词频方法进行运算的相关系数，cos 表示使用余弦正规化方法进行运算的相关系数；t1 表示教师 T1，t2 表示教师 T2，avg 表示教师 T1 和教师 T2 评分的平均分产生的相关系数，t1&t2 表示教师T1 和教师T2 评分之间的相关系数。

表 1　对所有学生数据的人机数据对比

			cipin	tfidf	max	log	cos	t1&t2
不使用 LSA 方法	F1-all	t1	0.43	0.32	0.22	0.37	0.32	
		t2	0.54	0.53	0.28	0.26	0.53	
		avg	0.55	0.47	0.28	0.26	0.53	
	F2-all	t1	0.43	0.24	0.25	0.33	0.24	
		t2	0.53	0.45	0.24	0.27	0.45	
		avg	0.54	0.39	0.28	0.34	0.39	
使用 LSA 方法	F1-all	t1	0.43	0.32	0.25	0.42	0.45	
		t2	0.54	0.53	0.32	0.30	0.60	
		avg	0.55	0.47	0.31	0.41	0.59	
	F2-all	t1	0.43	0.33	0.25	0.38	0.42	
		t2	0.53	0.51	0.32	0.31	0.60	
		avg	0.54	0.48	0.31	0.39	0.57	0.56

对比表 1 中方法一的相关系数，可以明显看出使用和不使用 LSA 方法评判的成绩与两位教师成绩之间的相关系数对于 cipin 和 tfidf 这两种方法没有变化，都是 0.43、0.54、0.55 和 0.32、0.53、0.47；而 max 方法中不使用 LSA 方法比使用 LSA 方法的相关系数低；使用 log 方法中不使用 LSA 方法比使用 LSA 方法的绝对值的相关系数要低；使用 cos 方法中不使用 LSA 方法比使用 LSA 方法的相关系数要低。

对比表 1 中方法二的相关系数，可以明显看出使用和不使用 LSA 方法评判的成绩与两位教师成绩之间的相关系数对于 cipin 方法没有变化，都是 0.43、0.53、0.54；而除 max 方法中 t1 的相关系数相等外，tfidf 方法、max 方法、log 方法中的绝对值和 cos 方法中不使用 LSA 方法的相关系数均低于使用 LSA 方法的相关系数。

对比表 1 的方法一和方法二中的相关系数在不使用 LSA 方法时，方法一的相关系数大都较方法二的相关系数要高；而在使用 LSA 方法时，除 max 方法的相关系数没有变化外，其余方法中方法一的相关系数大都较方法二的相关系数要高一些。这说明对 LSA 语义空间不做任何处理比提取出现在两篇及两篇以上的行向量构成的语义空间更能提高相关程度，且使用 LSA 方法对提高相关程度有明显的改善，其中使用 cos 方法产生的结果最理想——接近或大于人类评分的相关系数。

因此，认为使用 LSA 方法对方法一的数据进行标准化处理中的余弦正规化法可以满足机器自动评判中文短文的要求——与人类评分的相关系数相接近。实验数据可以清楚地表明：使用机器自动评判方式对中文写作的短文进行评测是可行的，也即是说机器自动评判方式今后完全可以应用到主观题如短文书写、简答题、问答题等方式的中文测试评判中。

5　实验结论及讨论

5.1　实验结论

综上所述，就以上所述的这几组实验数据与人类评判之间的相关系数的对比，可以得出如下实验结论：

(1)使用机器自动评判方式与人类评判之间的相关系数很接近，甚至更高；

(2)对 LSA 语义空间中的行向量不做任何处理的矩阵，使用 LSA 方法给出学生成绩与两位教师成绩及其平均分之间的相关系数，大都比取出现在两篇或两篇以上文档中的行向量分别构成的矩阵进行运算的实验结果对提高相关程度有较明显的变化，即使用方法一产生的实验结果对提高相关程度要更

好些;

(3)测试中,对数据标准化处理的几种方法中的余弦正规化方法、cipin 和 TF·IDF 方法使用相关系数方法进行测试的结果大都与人类评判之间的相关系数相接近或更高,但相对而言使用余弦正规化方法得到的实验结对提高相关程度要更好些;

(4)本研究基本实现了研究目的——论证了使用机器自动评判方式对中文短文的写作进行评测的可行性,也即是说机器自动评判方式将来可以应用到主观题如短文书写、简答题、问答题等方式的中文测试评判中。

5.2 实验讨论

这里对几种数据标准化处理方法进行讨论,为什么余弦正规化法能够得出较好的实验结果?通过数据标准化方法部分对几种方法的介绍,可以知道,这几种方法其实都是对向量空间中向量的长度进行缩放,这种缩放将直接影响空间矩阵中行列的交叉点处的值, 而该值经过 SVD 分解后又对 LSA 方法的语义空间产生影响。具体影响如下所述:

(1)TF·IDF 方法的不足主要表现在它没有考虑文档中词的总数。该方法说明一个在单文档中频度很高,而在整个数据全集中频度很低的词是更加重要的词,在对空间矩阵直接求得的相关系数的值恰好是经过该数据标准化处理后的矩阵的相关系数。换而言之,本实验矩阵中在单文档中频度很高,而在整个数据全集中数频度很低的词不是更加重要的词;

(2)最大正规化方法的不足是当文档中的某个词的词频很大,而其他词词频相对较小的时候,会出现多数词汇词频值较小,并且彼此差别不大的结果,因此,该方法得出的实验结果对相关程度的提高大都不如其他几种方法产生的实验结果对相关程度的提高;

(3)对数词频法通过对数函数降低了词频取值的影响,从而减少了文档中少数高频词对计算的影响,降低了低频词的取值,而且减轻了文档长度的变化对这一取值的变化影响,而本实验的空间矩阵为稀疏矩阵.使用该方法虽在一定程度上缓解了高频词对矩阵的影响,但不能有效地对稀疏的空间矩阵产生影响,反而使得起积极作用的高频词降低了对稀疏的空间矩阵的影响,而作用不大的低频词反而提高了对稀疏的空间矩阵的影响,就产生其实验结果大多为负相关的现象;

(4)余弦正规化法中经过正规化后的向量具有单位长度,而且在每一维上的值恰好是该向量及其在这一维上相应坐标轴上投影的夹角的余弦值,因此它解决了文档中少数高频词对其他词的权值扰动过大这一问题。使用该方法可以对少数不起积极作用的高频词降低其对稀疏的空间矩阵的影响,同时向量分布在一个单位长度的圆上,数据的比较具有更为有效的可比性,因此该方法对相关程度的提高有较明显的效果。

(收稿日期:2006 年 11 月)

参考文献:

[1] Landauer T K,Dumais S T.A solution to Plato's problem:the latent semantic analysis theory of the acquisition,induction,and representation of knowledge[J].Psychological Review,1997,104:211-240.

[2] Deewester S,Dumais S T,Harshman R,et al.Indexing by latent semantic analysis [J].Journal of the Society for Information Science,1990,41(6):391-407.

[3] 萨师煊,何守才.数据库综合大辞典(A Comprehensive Dictionary of Databases)[M].上海:上海科学技术文献出版社,1995.

[4] 萨师煊,王珊.数据库系统概论[M].3 版.北京:高等教育出版社,2004.

[5] 刘伯根,岑红.中国大百科全书(中国百科全书 24CD-19)[M].北京:中国大百科全书出版社,2004.

[6] 王晓关,关毅.计算机自然语言处理[M].北京:清华大学出版社,2005.

(上接 165 页)

4 结论

由结果看出,此方法能够较准确地定位字幕的区域,取得了一定的成果,但也存在着不足的地方。因为维文笔画比较尖锐,有些笔画会有缺失,但是并不妨碍阅读即识别出来文字的辨认,效果令人满意。

对文本图象进行了切分、分类、提取外围特征,综合全局信息和局部信息的分析和应用,使维文字符的识别获得了比较满意的结果。

对于噪声比较大的文字采取了抖动方法,使之其辨认出来的维吾尔文字更加清晰,提高文字的识别率,达到了预期的效果。(收稿日期:2007 年 3 月)

参考文献:

[1] 胡小锋,赵辉.Visual C++/MATLAB 图像处理与识别实用案例精选[M].北京:人民邮电出版社,2004.

[2] 余成波.数字图像处理及 MATLAB 实现[M].重庆:重庆大学出版社,2003.

[3] 胡小峰,周勇,叶庆泰.复杂背景彩色图像中的文字分割[J].光学技术,2006,32(1):141-147.

[4] 田春娜,高新波,哈力旦·A.一种基于相对模糊连通度的交互式序列图像快速分割算法[J].电子与信息学报,2005,27(10):1549-1554.

[5] 高新波,雷云,姬红兵.一种复杂背景下模板检测与定位的新方法[J].系统工程与电子技术,2004,26(1):87-90.

[6] Haritaoglu I.Scene text extraction and translation for handheld de2vices[C]//Proceedings of the IEEE Computer Society Conferenceon Computer Vision and Pattern Recognition,USA,Kauai,2001,2:408-413.

[5] Li H,Doberman D,Kia O.Automatic text detection and tracking in digital video[J].IEEE Trans on Image Processing,2000,9(1):147-156.