潜在语义分析在中文信息处理中的应用

刘云峰　齐　欢　代建民
（华中科技大学系统工程研究所，武汉 430074）
E-mail：liu_yun_feng@hotmail.com

摘　要　潜在语义分析是一种关于自然语言信息提取和再现的理论方法，它通过代数的方法提取语义空间中潜在结构。论文叙述了潜在语义分析的基本理论方法，概述了这种方法所建立的潜在语义空间的数学意义；然后通过一个简单示例说明 LSA 在中文信息处理中的分析方法，并通过分析结果中文本间、词汇间关联度的变化来说明 LSA 在中文信息处理中的重要意义。

关键词　潜在语义分析　潜在语义空间　中文信息处理　奇异值分解

文章编号 1002-8331-(2005)03-0091-03　**文献标识码** A　**中图分类号** TP393

Applying Latent Semantic Analysis in Chinese Information Processing

Liu Yunfeng　Qi Huan　Dai Jianmin
(Institute of Systems Engineering, Huazhong University of Science and Technology, Wuhan 430074)

Abstract: Latent Semantic Analysis is a theory and method about extracting and representing information of nature language.LSA retrievals the latent semantic structure from semantic space by mathematical method.Firstly, this paper presents the underlying idea of LSA and introduces the mathematical means of the Latent Semantic Space which is built by LSA.In the following subsection, the paper introduces the application of LSA in the field of Chinese information processing though a sample example.The variation of the similarities between documents, or between terms, in the example analysis result, shows the important meaning of LSA.

Keywords: Latent Semantic Analysis, Latent Semantic Space, Chinese information processing, Singular Value Decomposition

1　引言

1990 年，来自 University of Chicago、Bell Communications Research 和 University of Western Ontario 的 Scott Deerwester、Thomas K.Landauer 等五位学者共同提出了潜在语义分析（Latent Semantic Analysis，缩写为 LSA）这一自然语言处理的方法，在自然语言理解、文本分析、信息过滤、情报检索等领域得到了广泛的应用。包括日本、英国、德国等国家在内的计算机科学和认知科学领域的学者正在积极研究潜在语义分析，丰富 LSA 的理论方法，扩展 LSA 的应用领域。美国 Berkeley 实验室的 Ding 等人用双重概率模型来解释 LSA 的理论方法[2]；孟菲斯大学认知科学实验室的 Arthur C.Graesser 等人用 LSA 方法建立了一种类似于人类辅导员的自助教学辅导系统 AutoTutor[3]。LSA 的最早提出者之一的 Thomas K.Landauer 和他的 K-A-T 团队开发的 Intelligent Essay Assessor（IEA）是 LSA 的应用之一，IEA 对人们撰写的文章中上下文概念上的合理性给予评估和建议，被 Discover 杂志评价为一个“创新性的进步”[4]。

潜在语义分析将每个文档视为以词汇为坐标系的空间中的一个点，认为一个包含语义的文档出现在这种空间中，它的分布绝对不是随机的，而是服从某种语义结构。同样地，也将每个词汇视为以文档为坐标系的空间中的一个点。文档是由词汇组成的，而词汇又要放到文档中去理解，体现了一种“词汇-文档”双重概率关系。LSA 利用奇异值分解降秩的方法达到信息过滤和去除噪声的目的，LSA 不同于向量空间模型（VSM）中文档的高维表示，而是将文档的高维表示投影在低维的潜在语义空间中，缩小了问题的规模，并且使得原本稀疏的数据变得不再稀疏，从而呈现出一些潜在的语义结构。

2　LSA 的理论方法

假设有一个文本集，包含 n 个文档，用到了 m 个词汇，构造“词汇-文档矩阵”，$X_{m\times n}=[x_{ij}]=(doc_1,doc_2,\cdots,doc_n)=(term_1,term_2,\cdots,term_m)^T$，$x_{ij}$ 表示词汇 i 在文档 j 中出现的频数，有时 x_{ij} 还加入了词汇的权重，词汇的权重说明了不是所有词汇在语义空间中的重要性都是相同的，定义词汇权重的原则是，识别文档语义能力强的词汇的权重高于识别能力弱的词汇，目前定义 $term$ 权重的方法包括传统的 tf-idf 方法、term 熵（entropy）方法、微软提出的 Okapi 方法等等（很多方法是从向量空间模型中继承过来的）。$term_i$ 和 doc_j 分别是代表词汇和文档的列向量。由于任意一个文档总是由有限个词汇，而不是由所有 m 个词汇构成，所以 X 必是一个稀疏矩阵。

LSA 的关键思想是将文档和词汇映射到一个低维的向量空间，即潜在语义空间。LSA 利用奇异值分解的（Singular Value Decomposition，缩写为 SVD）方法实现这种降维。下面介绍奇异值分解定理。

定理 1：任何一个矩阵 $X_{m\times n}$，X 的秩记为 r，均可分解为两个正交矩阵和一个对角矩阵的乘积：

$$X=TSD^T \tag{1}$$

$T_{m\times r}=(t_1,t_2,\cdots,t_r)$ 为正交矩阵，其中 $t_1,t_2,\cdots,t_r$ 为 X 的左奇异向量，并且是 XX^T 的特征向量；$S_{r\times r}=diag(\sigma_1,\sigma_2,\cdots,\sigma_r)$ 为对角

作者简介：刘云峰（1977-），男，在读博士，研究方向为系统分析与集成。齐欢，博士，博士生导师。代建民，在读博士。

表 1　原始文档

编号	文档
Phy1	力是物质间的一种相互作用,运动状态的变化是由这种相互作用引起的
Phy2	死力可用物体的质量和该物体由静止状态转入运动状态时所获得的速度的乘积来量度
Phy3	牛顿首先引入了质量的概念,而把它和物质的重力区分开来,物质的重力只是地球对它的引力作用
Phy4	任何两个物体都是相互吸引的,引力的大小跟两个物体质量的乘积成正比
PCA1	主成分分析(PCA)法寻找较少的综合指标来代表原来较多的指标
PCA2	主元分析法是一种数据压缩并从中提取有用的信息的方法
PCA3	主成分分析法是解决数据相关并降低数据维数的一种非常有效的统计方法
PCA4	PCA 是一种将分散在一组变量上的信息集中到某几个综合指标(主成分)上的探索性统计分析方法
PCA5	当变量间的相关关系量度不明显时,作主成分分析意义不大

阵,$\sigma_1,\sigma_2,\cdots,\sigma_r$ 为 X 的所有奇异值,同时也是 XX^T 或 X^TX 所有特征值的平方根,并且满足关系 $\sigma_1\geq\sigma_2\geq\cdots\geq\sigma_r>0$;$D_{n\times r}=(d_1,d_2,\cdots,d_r)$ 为正交矩阵,其中 $d_1,d_2,\cdots,d_r$ 为 X 的右奇异向量,并且是 X^TX 的特征向量。因此,矩阵 X 可以用下式表达:

$$X=\sigma_1 t_1 d_1^T+\sigma_2 t_2 d_2^T+\cdots+\sigma_r t_r d_r^T \tag{2}$$

SVD 的优势是通过一种简单的方法,就可以使原矩阵塌陷、找到一个规模大大减小的近似矩阵。LSA 在 SVD 的基础上保留最大的 k 个奇异值,而忽略其他较小的奇异值,k 就是低维空间的维数。然后进行奇异值分解反运算,得到原始矩阵的近似阵。k 应当足够小,去除掉不该保留的噪声,又要足够大以保留语义空间中的主要框架,通常根据经验来说,当 X 的秩为几千的时候,k 的取值为几百。

由于矩阵的奇异值正好对应由 $EL=\{Xy:\|y\|_2=1\}$所定义的超椭球 EL 的各半轴之长,所以 LSA 中降秩的过程可以视为去除了语义空间中代表低信息量(即噪声)的自由度,而保留了代表语义空间中主要信息的自由度。

令 $S_k=diag(\sigma_1,\sigma_2,\cdots,\sigma_k)$,$T_k=(t_1,t_2,\cdots,t_k)$,$D_k=(d_1,d_2,\cdots,d_k)$,则:

$$\hat{X}_k=T_k S_k D_k^T \tag{3}$$

为 X 的秩为 k 的近似阵,那么 $\hat{X}_k$ 与 X 到底有多近似呢?可以用下面的定理来说明[5]:

定理 2:对于任意一个秩不大于 k 的 $C_{m\times n}$ 矩阵,下式始终成立:

$$\|X-C\|_F\geq\|X-\hat{X}_k\|_F \tag{4}$$

其中 $\|\cdot\|_F$ 是矩阵的 Frobenius 范数,$\hat{X}_k$ 为用上述方法计算得来的 X 的近似阵。从数学分析的角度来看,与其他降维方法相比,塌陷的 SVD 式是原矩阵 X 在 k 维子空间上最佳的近似[6]。

经过降秩得到近似矩阵 $\hat{X}_k=(doc_1^*,doc_2^*,\cdots,doc_n^*)=(term_1^*,term_2^*,\cdots,term_m^*)^T$,可以比较任意两个文档、任意两个词汇间的相似度。常用的方法是求两文档向量间(或两词汇向量间)的点乘、夹角余弦值或者是相关系数。根据(2)式,也可以把 $\hat{X}_k$ 写成:

$$\hat{X}_k=\sigma_1 t_1 d_1^T+\sigma_2 t_2 d_2^T+\cdots+\sigma_k t_k d_k^T \tag{5}$$

对于其中的每一个文档向量,可以用下式表示:

$$doc_i^*=\sigma_1 d_{1,i}\cdot t_1+\sigma_2 d_{2,i}\cdot t_2+\cdots+\sigma_k d_{k,i}\cdot t_k \tag{6}$$

其中 $d_{h,i}$ 表示向量 d_h 的第 i 个分量。可以将 $doc'_i=(\sigma_1 d_{1,i},\sigma_2 d_{2,i},\cdots,\sigma_k d_{k,i})^T$ 视为文档 i 在 k 维向量空间中的表示,因为 $t_1,t_2,\cdots,t_k$ 是正交向量,这样就得到了文档在低维空间中的向量表示。可以证明向量 doc_i^* 和 doc_j^* 的点乘或夹角余弦值与向量 doc'_i 和 doc'_j 的点乘或夹角余弦值是相同的。doc'_i 实际上是 D_kS_k 中第 i 个行向量的转置。因此可以得到如下结论:LSA 处理后,可以把矩阵 D_kS_k 中的行视为代表文档的向量,换句话说原 X 中的列向量被投影在 T_k 中的列向量所张成的低维空间中。同样,可以把矩阵 T_kS_k 中的行视为在低维空间中代表词汇的向量,X 中的行向量被投影在 D_k 中的列向量所张成的低维空间中。这里把这两个低维空间合称为潜在语义空间。如果有一查询文档 $qdoc$,这个文档不包含在 X 当中,如果想比较这个查询文档与 X 中任一文档的相似度,那么也要将查询文档投影到这个空间中去,而 D_k 中没有代表查询文档的行,可以用下述方法增加这一代表 $qdoc$ 的行:

$$D_q=qdoc^T T_k S_k^{-1} \tag{7}$$

3　LSA 应用示例

下面用一个简单的示例来说明 LSA 在中文信息处理中应用的效果。由于该例采用的样本数较少,无法体现出词汇权重的统计意义,因此在此不考虑词汇在文档中的权重问题,且在比较文档之间、词汇之间相似度时,采用它们在高维下表示的相关系数来度量。现有 9 个文档,其中 4 个是关于物理学的,另 5 个是关于主成分分析的,文档内容如表 1。

提取 9 个文档中的 14 个关键词,建立词汇-文档矩阵 X,如表 2 所示。

表 2　词汇-文档原始矩阵

词汇	Phy1	Phy2	Phy3	Phy4	PCA1	PCA2	PCA3	PCA4	PCA5
物质	1	0	2	0	0	0	0	0	0
运动	1	1	0	0	0	0	0	0	0
作用	2	0	1	0	0	0	0	0	0
物体	**0**	**1**	**0**	**2**	**0**	**0**	**0**	**0**	**0**
质量	0	1	1	1	0	0	0	0	0
量度	0	1	0	0	0	0	0	0	1
引力	0	0	1	1	0	0	0	0	0
主成分分析	**0**	**0**	**0**	**0**	**1**	**0**	**1**	**0**	**1**
指标	0	0	0	0	2	0	0	1	0
主元分析	**0**	**0**	**0**	**0**	**0**	**1**	**0**	**0**	**0**
数据	0	0	0	0	0	1	2	0	0
信息	0	0	0	0	0	1	0	1	0
统计	0	0	0	0	0	0	1	1	0
相关	0	0	0	0	0	0	1	0	1

原始矩阵中,“主成分分析”和“主元分析”两个词汇的相关系数为-0.25。“物体”和“主元分析”的相关系数为-0.177。计算奇异值分解,并保留 2 个奇异值,再进行奇异值分解反运算得

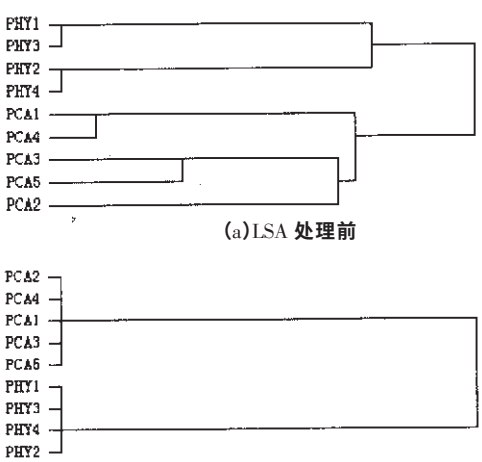
到 X 的近似阵 $\hat{X}_2$，如表 3 所示。

表 3　原始矩阵在二维空间中的重构矩阵

词汇	Phy1	Phy2	Phy3	Phy4	PCA1	PCA2	PCA3	PCA4	PCA5
物质	0.998	0.588	1.284	0.753	−0.033	−0.030	−0.067	−0.030	0.036
运动	0.444	0.266	0.571	0.337	0.010	0.0049	0.020	0.005	0.036
作用	0.918	0.541	1.181	0.692	−0.033	−0.030	−0.067	−0.030	0.031
物体	**0.585**	**0.356**	**0.754**	**0.448**	**0.037**	**0.026**	**0.077**	**0.025**	**0.068**
质量	0.734	0.441	0.946	0.559	0.018	0.010	0.038	0.010	0.061
量度	0.173	0.129	0.226	0.147	0.136	0.103	0.281	0.102	0.124
引力	0.570	0.339	0.734	0.432	−0.002	−0.005	−0.005	−0.005	0.034
主成分分析	**−0.025**	**0.09**	**−0.019**	**0.046**	**0.546**	**0.417**	**1.125**	**0.412**	**0.451**
指标	−0.032	0.056	−0.032	0.022	0.387	0.296	0.798	0.293	0.319
主元分析	**−0.010**	**0.015**	**−0.010**	**0.005**	**0.107**	**0.082**	**0.221**	**0.081**	**0.088**
数据	−0.055	0.099	−0.054	0.040	0.687	0.524	1.415	0.518	0.565
信息	−0.020	0.029	−0.020	0.011	0.214	0.163	0.440	0.161	0.175
统计	−0.032	0.057	−0.032	0.023	0.396	0.302	0.816	0.299	0.326
相关	−0.014	0.070	−0.008	0.038	0.406	0.310	0.835	0.306	0.335

在初始矩阵 X 中，向量“主成分分析”和“主元分析”的相关系数只有−0.250，表现不出两者的相似性。在降秩矩阵 $\hat{X}$ 中，向量“主成分分析”和“主元分析”的相关系数已非常接近 0.997，可见降秩后含义相近的词相关性得到加强。许多因特网搜索引擎采用准确匹配关键字的方式查询数据库中相关网页，例如用“雅虎中国”以“主成分分析”作关键词搜索可以得到 2374 项相关中文网站，而搜索“主元分析”只能得到 123 项相关中文网站。然而实际上“主成分分析”和“主元分析”同是英文“Principal Component Analysis”一词的中文叫法，只不过是翻译方法不同、习惯叫“主成分分析”的人更多一些而已。这样用“主元分析”作为关键字搜索的用户就看不到用“主成分分析”作为关键字可以搜索到的大多数网站了（只有少数是重叠的）。既然两者的含义相同，完全可以将以“主成分分析”为关键词搜索到的网页提供给以“主元分析”为关键词搜索的用户。而另一方面，在初始矩阵中向量“物体”和“主元分析”相关系数为−0.353，降秩后，相关系数降到−0.765，可见含义不相关的词汇之间的相关性被减弱。

同样的现象也出现在文档与文档之间的相关性变化上，也可以用上述分析词汇间相似度变化的方法分析文档间相似度的变化。这里用降秩前和降秩后的文档聚类分析结果的比较来说明。聚类分析的目的简单来说是分组，使组内距离尽量小，而组间距离尽量大。这里聚类分析采用组间均联法，组间距离采用皮尔逊相关系数。聚类结果如图 1 所示。



(a)LSA 处理前

(b)LSA 处理后

图 1　LSA 处理前后文档聚类分析结果比较

由图 1 可见，降秩前文档聚类分析的结果中，组间距离不明显，组内距离很大，虽然可以到达分类的目的，但效果不好；而降秩后文档聚类分析结果中，组间距离很大，组内距离很小，聚类效果远优于降秩前。说明 LSA 使得文档之间的关联值根据语义的相关性发生了变化。

文中的示例比较简单，还不能完全达到区别词义和文档语义的目的。例如仍然有多对词汇之间相关度非常高，这主要是因为选取的文档数量较少，“词汇−文档”双重概率关系下，不能明显地区别词义之间的差别。要更准确地鉴别文档，需要在语义空间中加入更多的关键词，同样的，要更准确地区别词汇就要加入更多的包含这些词汇的文档。当选取的文档数增多，相应选取的词汇量也增多时，区别文档语义和词义的精确度会有所提高。尽管如此，这个例子仍说明了 LSA 能够加强相关词汇之间的关联值，而削弱非相关词汇之间的关联值。文档之间的关联值的变化也可以得到同样的解释。

4　结论

随着计算机和因特网逐渐成为人们日常生活中的一部分，目前，大多数的信息都是通过文档的方式记录和保存的，有效的信息提取和过滤技术成为人们应对爆炸式增长的信息量的必不可少的措施。

LSA 的过程，实际上是将高维空间中的文档向量（词汇向量）投影到低维的潜在语义空间中，原来在高维中比较稀疏的向量表示在潜在语义空间中变得不再稀疏。即使原来没有任何共同项的两个文档（词汇），LSA 处理后仍然可能找到它们之间比较有意义的关联值。潜在语义分析的思想在某些程度上类似于多元统计分析中的主成分分析和因子分析，因此主成分分析和因子分析中的很多概念和方法可以应用到 LSA 中去。

LSA 的理论基础还不是很完全，例如潜在语义空间维数 k 的取值，目前主要依靠人的经验和实际检验来确定，文档长度如何选取直接影响潜在语义分析提取、再现语义的效果，目前也只能依赖经验而定。所有这些都是将来要研究的内容。

（收稿日期：2004 年 6 月）

参考文献

1.Deerweter S，Dumais S T，Furnas G W et al.Indexing by Latent Semantic Analysis[J].Journal of the American Society for Information Science，1990；41(6)：391~407

2.Ding C.A Similarity-Based Probability Model For Latent Semantic Indexing[C].In：Proc of ACM SIGIR Conf，1999

3.Graesser A C et al.A Simulation of A Human Tutor[J].Journal of Cognitive Systems Research，1999；1：35~51

4.Peter W Foltz，Darrell Laham，Thomas K Landauer.The Intelligent Essay Assessor：Applications to Educational Technology[J].Interactive Multimedia Electronic Journal of Computer-Enhanced Learning，1999；1(2)

5.Papadimitriou C，Raghavan P，Tamaki H et al.Latent Semantic Indexing：A Probabilistic Analysis[J].Journal of Computer and System Sciences，2000；61(2)：217~235

6.Golub G，Loan C V.Matrix Computations(2-nd ed)[M].John-Hopkins，Baltimore，1986

潜在语义分析在中文信息处理中的应用

作者： 刘云峰， 齐欢， 代建民， Liu Yunfeng， QI Huan， Dai Jianmin
作者单位： 华中科技大学系统工程研究所,武汉,430074
刊名： 计算机工程与应用 ISTIC PKU
英文刊名： COMPUTER ENGINEERING AND APPLICATIONS
年，卷(期)： 2005,41(3)
被引用次数： 17次

参考文献(6条)

1. Deerweter S;Dumais S T;Furnas G W Indexing by Latent Semantic Analysis[外文期刊] 1990(06)

2. Ding C A Similarity-Based Probability Model For Latent Semantic Indexing[外文会议] 1999

3. Graesser A C A Simulation of A Human Tutor[外文期刊] 1999

4. Peter W Foltz;Darrell Laham;Thomas K Landauer The Intelligent Essay Assessor:Applications to Educational Technology 1999(02)

5. Papadimitriou C;Raghavan P;Tamaki H Latent Semantic Indexing:A Probabilistic Analysis[外文期刊] 2000(02)

6. Golub G;Loan C V Matrix Computations(2-nd ed)John-Hopkins 1986

本文读者也读过(8条)

1. 杨翠 潜在语义分析理论及其在文本检索与聚类中的应用研究[学位论文]2008

2. 任纪生.王作英.Ren Jisheng.Wang Zuoying 一种新的潜在语义分析语言模型[期刊论文]-高技术通讯2005,15(8)

3. 王剑锋.乔冬.麻丽娜.李新叶 基于潜在语义分析的网页文本分类研究[期刊论文]-应用能源技术2009(11)

4. 盖杰.王怡.武港山 基于潜在语义分析的信息检索[期刊论文]-计算机工程2004,30(2)

5. 卢健 潜在语义分析在文本信息检索中的应用研究[学位论文]2005

6. 刘云峰.齐欢.Xiangen Hu.Zhiqiang Cai.LIU Yun-feng.QI Huan.Xiangen Hu.Zhiqiang Cai 潜在语义分析权重计算的改进[期刊论文]-中文信息学报2005,19(6)

7. 廖文彬.孙逊 矩阵奇异值的随机抽样方法[期刊论文]-科技资讯2008(13)

8. 周巧云.ZHOU Qiao-yun 面向计算机的深度语义分析[期刊论文]-喀什师范学院学报2009,30(2)

引证文献(17条)

1. 李湘东.巴志超.黄莉 基于加权隐含狄利克雷分配模型的新闻话题挖掘方法[期刊论文]-计算机应用 2014(5)

2. 吴昊.耿焕同.吴祥 一种基于聚类分析的BBS主题发现算法研究[期刊论文]-安徽师范大学学报（自然科学版） 2009(1)

3. 吴昊.耿焕同 基于潜在语义分析的BBS主题发现算法研究[期刊论文]-电脑知识与技术 2008(29)

4. 金小峰 一种大容量文本集的智能检索方法[期刊论文]-计算机工程与应用 2011(7)

5. 杨鹤标.刘志然 基于语义事务信息聚类的用户概貌构建[期刊论文]-计算机工程与设计 2010(20)

6. 王永智.滕至阳.王鹏.聂江涛 基于LSA和SVM的文本分类模型的研究[期刊论文]-计算机工程与设计 2009(3)

7. 张元虹.郭剑毅.龚华明.薛征山 基于DF与LSA相结合的降维法的文本分类系统的研究[期刊论文]-山西电子技术 2008(4)

8. 毕晓冬 基于语义和学习机制的信息过滤模型研究[期刊论文]-潍坊学院学报 2006(6)

9. 龚主杰 潜在语义索引在图像检索中的应用[期刊论文]-图书馆学刊 2009(5)

10. 赵连朋 基于关联规则的医疗处方智能监督方法的研究[期刊论文]-计算机工程与应用 2006(32)

11. 舒鑫柱 基于资源的汉语词汇语义网的实现[期刊论文]-山西大学学报（自然科学版） 2008(1)

12. 于一.廖睿.叶大田 电子病历结构化方法概述[期刊论文]-北京生物医学工程 2007(1)

13. 罗进军 当前计算语言学研究的发展态势[期刊论文]-湖南工业职业技术学院学报 2006(4)

14. 杨滨 网络考试系统的设计与开发[学位论文]硕士 2006

15. 胡旭昶 文本聚类分析研究及在中文新闻系统中的应用[学位论文]硕士 2006

16. 沈贺丹 核心能力评价系统的分类模块研究[学位论文]硕士 2005

17. 孙海霞. 成颖 潜在语义标引(LSI)研究综述[期刊论文]-现代图书情报技术 2007(9)

引用本文格式：刘云峰. 齐欢. 代建民. Liu Yunfeng. QI Huan. Dai Jianmin 潜在语义分析在中文信息处理中的应用[期刊论文]-计算机工程与应用 2005(3)