维普资讯 http://www.cqvip.com

中 文 信 息 学 报
第19卷 第6期 JOURNAL OF CHINESE INFORMATION PROCESSING Vol.19 No.6

文章编号:1003 - 0077(2005)06 - 0064 - 06

潜在语义分析权重计算的改进

刘云峰[1], 齐 欢[1], Xiang'en Hu[2], Zhiqiang Cai[2]

(1.华中科技大学 系统工程研究所, 湖北 武汉 430074;
2. University of Memphis, Institute of Intelligent Systems, USA, Tennessee, Memphis, TN 38152)

摘要:自从潜在语义分析方法诞生以来,被广泛应用于信息检索、文本分类、自动问答系统等领域中。潜在语义分析的一个重要过程是对词语文档矩阵作加权转换,加权函数直接影响潜在语义分析结果的优劣。本文首先总结了传统的、已成熟的权重计算方法,包括局部权重部分和词语全局权重部分,随后指出已有方法的不足之处,并对权重计算方法进行扩展,提出文档全局权重的概念。在最后的实验中,提出了一种新的检验潜在语义分析结果优劣的方法——文档自检索矩阵,实验结果证明改进后的权重计算方法提高了检索效率。

关键词:计算机应用;中文信息处理;潜在语义分析;权重;文档全局权重;文档自检索矩阵

中图分类号:TP391 **文献标识码**:A

A Modified Weight Function in Latent Semantic Analysis

LIU Yun-feng[1], QI Huan[1], Xiang'en Hu[2], Zhiqiang Cai[2]

(1. Huazhong University of Science and Technology, Institute of Systems Engineering, Hubei, Wuhan 430074, China; 2. University of Memphis, Institute of Intelligent Systems, USA, Tennessee, Memphis, TN 38152)

Abstract: Since the first paper about Latent Semantic Analysis (LSA) was published, LSA has been applied to many fields, such as information retrieval, text classification, automatic question answering, etc. . One important factor that affects the quality of LSA is the weighting scheme to the term - document matrix. In this paper, we first summarize the traditional and well - studied methods of weighting, including local weighting and global weighting. We then point out some inadequacy of original methods, modify these methods, and present the concept of global weighting of document. In the last part of this paper, we construct an experiment to compare the results of LSA with different types of weighting, in which we present a new measure to evaluate the result of LSA. We call this new measure self - indexing matrix. The result of the experiment confirms that the modified method of weighting can improve the efficiency of retrieval.

Key words: computer application; Chinese information processing; latent semantic analysis; weight; global document weight; self - indexing matrix

1 引言

基于概念的信息检索(IR)技术近年来发展迅速,潜在语义分析(Latent Semantic Analysis)就是这样一种基于潜概念索引的检索技术。潜在语义分析与其他基于概念的检索技术相比较的优势在于,需要人的参与性少,可计算性和可操作性强。在国外,潜在语义分析已经进入商业化应用阶段,在国内,近年来这方面的研究也取得了许多进展,但是针对汉语的一些特点,尚有

收稿日期:2004 - 10 - 10 **定稿日期**:2005 - 05 - 10

作者简介:刘云峰(1977—),男,博士研究生,主要研究智能信息检索、文本分类.

维普资讯 http://www.cqvip.com

多个难点有待进一步解决。

潜在语义分析基于这样一个假设:词语出现在某一个文档中以及两个词语出现在同一段上下文中不是完全随机的,而是存在某种潜在语义结构在起作用。如果能把这种潜在语义结构提取出来,建立词与词之间的语义关系,就可以消除词语用法的多样性和词语使用的随意性对检索产生的偏差。LSA 利用奇异值分解(Singular Value Decomposition)生成的潜概念(Latent Concept)索引来进行信息检索,而不再是传统的基于检索词匹配的检索方式。我们曾经用物理学科普文本(来自"大科普"网站)建立了一个简易的 IR 系统,文档长度的选取借鉴了 University of Memphis 的 AutoTutor[1] 开发团队应用 LSA 时所采用的文档长度,并考虑到中文文本的特点,选择一个完整的句子或者词语个数在 30 个左右的段落作为一个文档。我们用"电磁感应"作为检索词来查找相关文档,结果中排在前 10 位的文档中包含如下一个文档:"这个概念的核心思想是:变化着的电场能产生磁场;变化着的磁场也能产生电场。"这个文档中尽管不包含"电磁感应"一词,但是内容仍然是阐述电磁感应的现象,因此是"电磁感应"的相关文档,可见潜在语义分析检索是基于语义的检索。

潜在语义分析的基本方法是,对于一个的 $m \times n$ 词语文档矩阵(term - document matrix),它第 i 行第 j 列的元素 x_{ij} 表示第 i 个词语在第 j 个文档中出现频数,根据奇异值分解定理,X 可分解为 $X = TSD^T$,其中 T 和 D 为正交矩阵,S 为对角矩阵,对角线上元素称为 X 的奇异值,按数值由大到小在对角线上排列,T 和 D 中列向量依次对应各个奇异值,分别称为左奇异向量和右奇异向量。然后保留 T 和 D 中前 K 个列向量和 S 中的前 K 个奇异值,分别得到 T_k、D_k 和 S_k,$X_k = T_k S_k D_k^T$ 是原矩阵 X 在秩为 k 条件下的最小二乘意义上的最优近似,上述过程被称为截断的奇异值分解。$T_k S_k$ 和 $D_k S_k$ 当中的行向量就是分别代表词语和文档的向量,词语之间和文档之间的相关度可以用这些向量之间的余弦值来表示。$T_k S_k$ 中 k 个列向量所张成的空间称为潜在语义空间,潜在语义空间被视为对原向量空间的压缩。$T_k S_k$ 中的 k 个列向量是潜在语义空间的基,含义类似与因子分析中的因子,这 k 个列向量被认为分别对应 k 个潜概念,*LSA* 的本质是,用词语和文档在潜概念上的投影作为它们的向量表示[2,3]。在应用阶段,设 d 为检索文档的词频向量,通过如下运算获得文档向量:$\hat{d} = d^T T_k S_k$。

多数情况下,潜在语义分析并非直接对词频矩阵 X 进行奇异值分解,为了突出各个词语和文档对语义空间不同的贡献程度,需要定义一种权重函数 $W(i,j)$,对词频矩阵 X 进行加权转换,得到一个加权后的矩阵 X^*,然后对 X^* 进行奇异值分解及以后的运算。向量空间模型(VSM)方法的研究多集中在选择权重计算方法上[4,5],在潜在语义分析中,权重计算也同样非常重要。潜在语义分析中权重计算方法很多继承于向量空间模型,但因此往往忽视了潜在语义分析与向量空间模型基本思想上的不同之处。VSM 本质上将词语看作空间的维度,将文档根据其所包含的词语看作该空间中的一个点;LSA 中不再将词语看作单独的维度,潜在语义空间中的维度被认为是对应着各个"潜概念"(Latent Concept),词语向量被看作是它们在各个"潜概念"上的投影,文档向量是其所包含的词语向量之和。LSA 通过 SVD 将 VSM 的空间作了基变换生成了新的空间基,而词语在这些基向量上的投影体现出词语间的语义关系。在潜在语义分析中定义权重,体现出一种信息归约的作用,会使潜在语义空间的基更能呈现出主要的语义结构。因此针对潜在语义分析的特殊性,专门研究适用于 LSA 的权重计算方法是非常有意义的。本文在现有工作的基础上,对权重计算方法进行了改进和扩展,并对几种典型权重计算方法的效果进行比较分析。

维普资讯 http://www.cqvip.com

2 潜在语义分析权重计算方法的现状及扩展

传统的权重计算方法一般将权重分解为两部分分别定义，一部分称为局部权重(记作 $LW(i,j)$)，用来记录词语 i 在文档 j 中词频 tf_{ij} 信息；一部分称为词语全局权重(记作 $GWT(i)$)，词语全局权重说明了不同的词语在文档集中区别分辨文档语义的能力是不同的。LSA 的重要任务就是提取语义结构，即词语之间潜在的语义关系。两个权重较大的词语之间隐含的语义关系，更容易被 LSA 认为是重要的语义关系而被保留。因此权重函数一般可以用下式表示：

$$W(i,j) = LW(i,j) * GWT(i) \tag{1}$$

上述定义中，仅考虑局部权重和词语全局权重，而忽略文档在区别和分辨词语时的贡献也是各不相同。传统权重定义方法受向量空间模型方法的影响而没有考虑文档权重的重要性。LSA 定义了词语全局权重后生成的空间基突出了权重较大的词语间的语义关系，文档权重对潜在语义空间的作用是基于如下一种想法：能提供给词语更多信息量的文档，其对潜在语义空间基向量的影响应当被放大，仅能提供给词语较少信息量的文档，其对潜在语义空间基向量的影响应当被削弱，体现出一种信息归约的作用。

词语全局权重 $GWT(i)$ 完成了对矩阵 X 横向信息的归约，我们再定义一种文档全局权重 $GWD(j)$，用以完成对矩阵 X 纵向信息的归约。因此权重公式可以扩展为：

$$W(i,j) = LW(i,j) * GWT(i) * GWD(j) \tag{2}$$

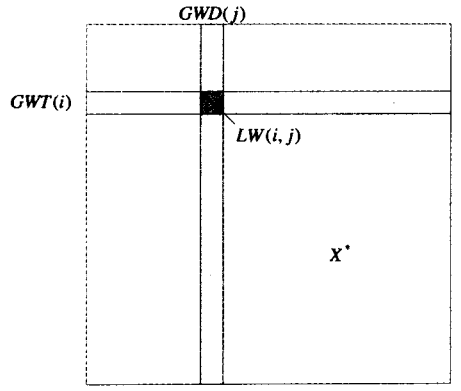

图 1 改进的权重计算结构

由公式(2)定义的权重计算结构由图 1 所示。

根据上面对权重含义的分类，下文分别讨论权重定义中的三部分，并讨论各种定义下的统计意义和对潜在语义空间的影响。这里先定义几个统计量的记号：

tf_{ij}——词语 i 在文档 j 中出现的频数；

df_i——文档集中出现词语 i 的文档个数；

gf_i——在整个文档集中词语 i 总共出现频数；

sgf——在文档集中所有词语出现频数之和，有 $sgf = \sum_{i=1}^{m} gf_i$。

2.1 局部权重定义方法

局部权重(Local Weight)强调某一词语在某一文档中的重要性[6]。最简单的就是将词频作为局部权重的定义。

$$LW_1(i,j) = tf_{ij} \tag{3}$$

上面这种定义下，有时某一文档中的某一高频词语的权重会过分突出，为了有效地消减高频词对潜在语义空间的影响，可以取词频的对数作为局部权重的定义。

$$LW_2(i,j) = \log_2(tf_{ij} + 1) \tag{4}$$

公式(4)中对数底取 2。根据矩阵的稀疏程度，非零元素的分布状况，或者原始文档的平均长度，可以选取不同的对数底。

2.2 词语全局权重定义方法

词语全局权重(Global Weight of Term)强调某一词语在整个文档集中的重要性。它在一定程度上代表了某一词语在区别和分辨文档时所扮演角色的重要程度。词语全局权重一般通过

维普资讯 http://www.cqvip.com

统计方法计算获得。

如果不考虑词语全局权重在整个权重定义中的作用，那么直接取 $GWT_1 = 1$。

向量空间模型中常用的 IDF(Inverse Document Frequency，倒排文档频)方法在潜在语义分析中同样经常使用，它的定义如下：

$$GWT_2(i) = \log_2 \frac{n}{df_i} \tag{5}$$

在信息论中，熵(entropy，$H(X)$)是信源 X 不确定性的度量(也是平均自信息量)。熵及其相关概念的特殊的统计性质非常适合作为权重定义方法[4,7]。条件熵 $H(X \mid Y)$表示即使 Y 确定，对 X 仍存在的平均不确定度，因此可以用 $H(doc \mid term_i)$表示 $term_i$ 确定出现后，文档集中的文档变量 doc 仍然存在的不确定性。平均互信息量 $I(X;Y) = H(Y) - H(X \mid Y)$表示 Y 确定后，关于 X 消除的不确定性，由此我们可以把 $H(doc) - H(doc \mid term_i)$看成 $term_i$ 是确定出现后，文档集中的文档消除了的不确定性，即 $term_i$ 提供 doc 给变量的信息量。一般认为，某一词语提供给文档集的信息量越大，说明它区别分辨文档的作用能力越强，因此它的全局权重应当越高。由此我们可以用某一词语和训练文档集的平均互信息量来定义这个词语的全局权重，这里稍作变化，定义词语全局权重如下：

$$GWT_3(i) = \frac{H(doc) - H(doc \mid term_i)}{H(doc)} = 1 - \frac{H(doc \mid term_i)}{H(doc)} \tag{6}$$

根据信息论中条件熵的定义，用下式计算 $H(doc \mid term_i)$：

$$H(doc \mid term_i) = -\sum_j p(j \mid i) \log_2 p(j \mid i)$$

其中 $P(j \mid i)$是条件“词语 i 出现”成立的情况下，“文档 j 出现”的概率，可用下式计算：

$$p(j \mid i) = \frac{tf_{ij}}{gf_i}$$

公式(6)中 $H(doc)$是一个常数，且存在不等式关系：$H(doc \mid term_i) \leqslant H(doc) \leqslant \log_2 n$。由于 $H(doc)$的计算较复杂，且当训练文档集中文档的长度比较均匀时，$H(doc)$的值非常接近 $\log_2 n$，因此权重公式可以改写为：

$$GWT_3(i) = 1 + \frac{\sum_j p(j \mid i) \log_2 p(j \mid i)}{\log_2 n} \tag{7}$$

词语全局权重除了对潜在语义空间基向量的影响以外，对文档向量的计算也有重要作用。上文曾指出为参加 SVD 的文档通过 $\hat{d} = dT_k S_k$ 计算获得向量形式。若在此式中引入词语全局权重，能够突出重要的词语对该文档的语义贡献，因此将上式改写为 $\hat{d} = dW_t T_k S_k$，其中 W_t 是对角矩阵，对角线上分别是对应各个词语的全局权重。

2.3 文档全局权重定义方法

文档的语义是由它包含的所有词语的语义来区别和分辨的，而词语的语义与它所出现文档的主题密切相关。如果一个文档包含多个重要词语，那么每个词语和这个文档的互信息量就比较少；相反，若一个文档仅包含少量重要词语，那么每个词语和它的互信息量就比较多。在迄今为止的权重定义方法中，一般都只考虑了词语在建立文档之间语义关系时信息量的贡献，而忽视了文档在建立词语之间语义关系时的贡献，而潜在语义空间中最重要的就是词与词之间的语义关系。为了放大能提供给词语更多信息量的文档，对潜在语义空间基向量的影响作用，这里我们提出文档全局权重的概念，用来记录文档重要性的不同。

维普资讯 http://www.cqvip.com

如果不考虑文档全局权重在整个权重定义中的作用，那么直接取 $GWD_1(j)=1$。

正如上文关于文档与词语互信息量的讨论，我们可以借鉴词语全局权重中的熵定义方法，采用平均互信息量的基本概念作为文档全局权重的定义方法。$H(term)-H(term\mid doc_j)$可以看成 doc_j 是确定出现后，文档集中的词语消除了的不确定性，即 doc_j 提供给 $term$ 变量的信息量。其中 $H(term\mid doc_j)$是文档确定条件下的词语条件分布熵，描述某个文档对消除词语不确定程度的贡献。因此我们可以采用下式定义文档全局权重

$$GWD_2(j)=\frac{H(term)-H(term\mid doc_j)}{H(term)}=1-\frac{H(term\mid doc_j)}{H(term)} \tag{8}$$

其中

$$H(term)=-\sum\nolimits_i P(term_i)*\log_2 P(term_i)=-\sum_{i=1}^{m}\frac{gf_i}{sgf}*\log_2\frac{gf_i}{sgf}$$

$$H(term\mid doc_j)=-\sum\nolimits_i p(i\mid j)*\log_2 P(i\mid j)$$

$p(i\mid j)$是条件“文档 j 出现”成立的情况下，“词语 i 出现”的概率，计算方法如下：

$$p(i\mid j)=\frac{tf_{ij}}{dl_j}$$

3 不同权重计算方法的实验分析

为了比较各种权重定义方式下计算得到的潜在语义空间的优劣，我们分别采用 $LW_1*GWT_1*GWD_1$（不考虑权重）、$LW_2*GWT_2*GWD_1$、$LW_2*GWT_2*GWD_2$、$LW_2*GWT_3*GWD_1$、$LW_2*GWT_3*GWD_2$ 五种权重定义方法，生成潜在语义空间。

实验中我们采用引言中提到的物理学文档（2899 篇文档，保留 3467 个检索词）作为训练文档集，分别用前述五种权重定义生成五个潜在语义空间。

为了检验潜在语义空间的优劣，这里我们提出一种文档自检索矩阵方法。文档自检索矩阵有别于其他文献采用的文档自相关矩阵[6]，主要是考虑到自相关矩阵是对称矩阵（因为 $Sim(doc_i,doc_j)=Sim(doc_j,doc_i)$），但是在 IR 系统中，如果将 doc_i 作为输入可以检索到 doc_j，并不意味着将 doc_j 作为输入一定可以检索到 doc_i，这取决于 doc_j 与其他文档的相关度。作者这样定义文档自检索矩阵：设有 N 个文档组成文档集，其中包含 O 个主题，如果将文档集中的某一主题下任一文档 doc_i 作为输入，比较该文档与文档集中所有 N 个文档的相关度$\{S_{i1},S_{i2},\Lambda,S_{iN}\}$，$N\times N$ 的文档自检索矩阵中第 i 行等于，将 doc_i 作为输入后，得到的 N 个相关度的排序$\{R_{i1},R_{i2},\Lambda,R_{in}\}$（$R_{ij}$表示 S_{ij}在$\{S_{i1},S_{i2},\Lambda,S_{iN}\}$中由小到大排序中的序号），可以取某一阈值，矩阵中阈值以下的元素置为 0。如果认为理想情况下 IR 系统应当输出本主题下的所有文档，那么最终生成矩阵中的非零元素的分布大致上是可以预测的，由此实验生成的矩阵与我们预测的矩阵进行比较，就可以得出潜在语义空间优劣的判断。这里在上一步实验的基础上，在物理学科普文档集中抽取天体力学、声学、光学、材料力学、电磁学、热力学、流体力学七个主题各 50 篇文档，分别基于上一步实验中五种权重定义下的语义空间生成自检索矩阵。如果认为每一次检索的结果中，同一主题下的文档为相关文档，非同一主题

表 1 五种权重定义下最优维数及平均查准率

权重定义方法	保留维数	平均查准率
$LW_1*GWT_1*GWD_1$	155	42.37%
$LW_2*GWT_2*GWD_1$	47	70.04%
$LW_2*GWT_2*GWD_2$	37	72.32%
$LW_2*GWT_3*GWD_1$	42	75.26%
$LW_2*GWT_3*GWD_2$	33	78.07%

维普资讯 http://www.cqvip.com

下的文档为误差，将阈值取为该主题下的文档数，那么可以计算查准率。表 1 中列出五种权重定义下的平均查准率，以及能达到最高查准率的维数。

由实验结果表明：①使用权重（后四种）后比不使用权重（第一种）效果提高很多；②词语全局权重的两种定义方法中，熵权重定义方法优于 IDF 权重方法；③采用作者提出的权重扩展模型，加入文档全局权重后，检索效果又有所提高。

为了便于更宏观的比较上述几种权重计算方法的检索效果，图 2 列出了分别基于由 $LW_2 * GWT_2 * GWD_1$、$LW_2 * GWT_2 * GWD_2$、$LW_2 * GWT_3 * GWD_1$、$LW_2 * GWT_3 * GWD_2$ 这 4 种权重定义生成潜在语义空间建立的 IR 系统的查准率和查全率关系图。

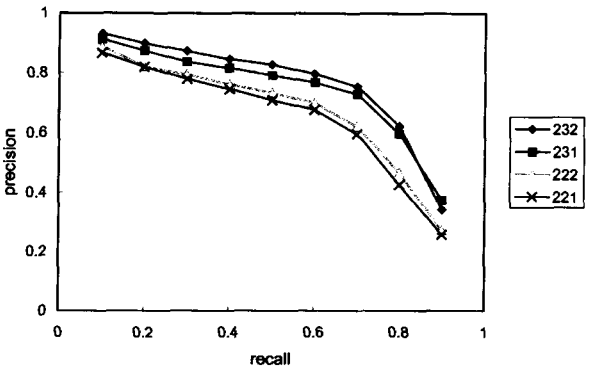

图 2　4 种权重定义的潜在语义空间查全率和查准率对比

4　总结

本文首先对潜在语义分析权重计算的传统方法作了总结，随后对权重的统计含义进行了扩展，提出文档全局权重概念。本文借鉴定义词语全局权重时熵方法的应用，提出了一种文档全局权重的定义方法，并对扩展后的权重计算方法与现有成熟方法进行比较，通过文档自检索矩阵对比，证明加入文档全局权重后，潜在语义空间的检索效率明显提高。当然，文档全局权重的定义根据不同的统计特性的需要，应有更多种定义方法，这正是我们进一步研究的方向。

参　考　文　献：

[1] Graesser A. C., Wiemer-Hastings K., Wiemer - Hastings P., et al.. AutoTutor: A Simulation of A Human Tutor [J]. Journal of Cognitive Systems Research , 1999, 1: 35 - 51.

[2] Hu, X., Cai, Z., Louwerse, M., et al.. A Revised Algorithm for Latent Semantic Analysis [A]. In: Proceedings of the 2003 International Joint Conference on Artificial Intelligence [C], 2003, 1489 - 1491.

[3] Cai, Z., McNamara, D. S., Louwerse, M. M., Hu, X., Rowe, M., & Graesser, A. C. (2004). NLS: A Non - Latent Similarity Algorithm [A]. K. Forbus, D. Gentner, & T. Reiger. Proceedings of the 26th Annual Meeting of the Cognitive Science Society [C]. NJ: Erlbaum, 2004, 180 - 185.

[4] 鲁松，李晓黎，白硕.文档中词语权重计算方法的改进[J].中文信息学报，2000，14[6]:8 - 13.

[5] 吴科，石冰，卢军，et al.. 基于文本集密度的特征选择与权重计算方案[J].中文信息学报，2004，18[1]: 42 - 47.

[6] Preslav Nakov, Antonia Popova, Plamen Mateev. Weight Functions Impact on LSA Performance[A]. In: Proceeding of EuroConference RANLP'2001 [C].Tzigov Chark, Bulgaria, 187 - 193.

[7] 刁倩，张惠惠.文本自动分类中的词权重与分类算法[J].中文信息学报.2000，14[3]:25 - 29.

[8] Hu, X., Cai, Z., Franceschetti, D., et al. . LSA: The First Dimension and Dimensional Weighting [A]. R. Alterman and D. Hirsh. Proceedings of the 25th Annual Conference of the Cognitive Science Society [C]. Boston, MA: Cognitive Science Society, 2003, 1 - 6.