第 33 卷 第 5 期 计 算 机 工 程 2007 年 3 月
Vol.33 No.5 Computer Engineering March 2007
·人工智能及识别技术· 文章编号：1000—3428(2007)05—0193—04 文献标识码：A 中图分类号：TP311

潜在语义标引在中文信息检索中的研究与实现

居 斌
（浙江省科技信息研究院网管中心，杭州 310006）

摘 要：随着网络信息的迅猛发展，信息检索已经成为人们获取信息不可缺少的工具。基于向量空间模型的检索方法是语义检索的重要研究方向，潜在语义标引模型是向量检索方法的一个有力扩展。对 LSI 中所涉及的关键技术，包括传统的向量空间模型的原理，以及潜在语义索引模型的原理、设计、实现，进行了研究和探讨，同时开发了一个适合中文信息检索的系统原型。对系统进行了测试，取得了较好的实验效果。

关键词：潜在语义标引；向量空间模型；信息检索；中文

Research and Application for Chinese Information Retrieval Based on Latent Semantic Indexing

JU Bin
(Network Management Center, Institute of Scientific and Technological Information of Zhejiang Province, Hangzhou 310006)

【Abstract】 As Internet information has been exploding, information retrieval has facilitated getting knowledge to the people. The approach based on the vector space is the important research direction on information semantic retrieval. Latent semantic indexing(LSI) is an important development for vector space model. The author has done research on how to use LSI technique to develop an application which is proved to perform well in experiment for Chinese information retrieval. The author illustrates some results of experiment.

【Key words】 Latent semantic indexing(LSI); Vector space model; Information retrieval; Chinese

Internet 的迅猛发展导致网上信息正以爆炸的速度增长，人们迫切需要搜索引擎能够从巨大的 Web 信息中找到自己想要的内容。目前 Web 上的搜索引擎基本上还是基于布尔逻辑的关键字检索技术，这种检索技术因为程序执行速度比较快，所以已经在互联网上广为使用并取得了巨大的成功。但是，基于布尔逻辑的检索技术也有不足之处，即它的使用太过刚性，有效地使用它的前提是需要用户能够准确地提炼出文档中的关键字，并且能运用布尔关系加以组合。从以人为本的角度看，这种检索方式不能带给用户良好的操作体验。

在这种背景下，研究者对信息检索提出了基于概念的检索问题。检索过程是从概念层次上进行文档原文信息和用户查询信息的相似度匹配，匹配结果不仅能包含查询中的关键词，还能检索出包含那些与该词同属一类概念的词汇。比如，实现同义词扩展检索(如查询“自行车”，也能查询到包含“脚踏车”、“单车”之类的文档)、语义蕴涵扩展(查询“操作系统”时，也能查询到“计算机软件”、“应用软件”)以及语义相关扩展(查询“微软”时，也能查询到“微软视窗”、“Windows NT”)[1]。本文主要探讨了基于潜在语义标引模型的概念检索技术和应用在中文信息环境下的系统模型，并根据本人单位信息库的实际情况，实现了一个中文检索系统的原型。

1 潜在语义标引的检索模型

1.1 概述

基于概念的信息检索主要包括 3 种模型：布尔模型(Boolean Model)，向量空间模型(Vector Space Model，VSM)及概率模型(Probabilistic Model)[2]。康奈尔大学的Salton等人提出的向量空间模型将文献内容和查询内容表示成标引项(也称标引词、关键词)及其权重的向量。每个文献 D_i 被表示为一个n维向量：$D_i = \{d_{i1}, d_{i2}, d_{i3}, \ldots, d_{in}\}$。其中，$d_{ij}$ 代表第j个标引词在文献 D_i 中的权重。如果把每个标引词看做是向量的一维，那么由这些标引词集合就定义了一个空间，文献[2]中有详细说明。文献集合中的任一文献都可以表示为这个多维空间中的一个向量，这个空间就称为“文献空间”，如图 1 所示，这个文献空间在数学上可以表示为项-文档矩阵。

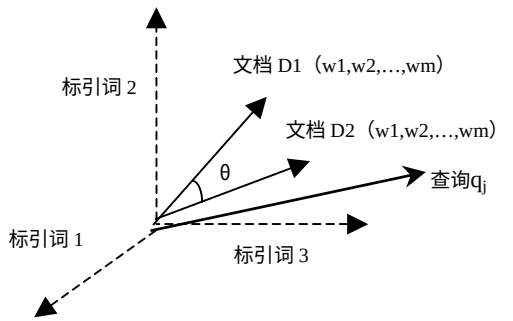

图 1 三维文献空间

用户的查询也可以像文献那样来处理。实际上，一个查询可以虚拟成一篇由各种标引词构成的文献。在经过“禁用”

作者简介：居 斌(1975－)，男，工程师、硕士，主研方向：软件工程，数据挖掘
收稿日期：2006-03-10 **E-mail：**jubin@zjinfo.gov.cn

词剔除，抽取词根等自然语言处理之后，转化成一个带有权重的项集。因此，这个查询就类似于文献空间中的一个文献向量。最后，通过向量之间的内积大小或余弦距离[3]来比较查询信息和每个文献之间的相似程度。相似度值越高，该文献贴近用户查询的可能性就越大。

向量空间模型的优点在于将非结构化的文本表示为向量形式，使得各种数学处理成为可能。在传统的向量空间模型中，因为简化模型的缘故，假定标引词之间不存在任何的联系，反映在文献空间中为：构成的基向量是两两正交的。而在实际中，这种正交的假设是很难满足的。由于在自然语言文献中，常常使用同一个词来表达多种不同的概念，或使用不同的词来表达同一个意思，这就是常说的多义和同义现象。如果在一篇文献中，使用了大量的多义词和同一词，则从该文献中抽取的标引词之间就不可避免地存在着相互的联系，从而，标引词向量之间存在着“斜交”的情形。若全然忽略这样的斜交可能，即忽略项之间存在的相互联系，必然使得检索效果产生极大的偏差[4]。

潜在语义标引模型(LSI)是向量检索模型的一个扩展，在LSI 方法中，假定在文献中使用的词中存在着一些隐含的或“潜在”的语义结构。比如，同义词之间具有基本相同的语义结构，多义词的使用必定具有多种不同的语义结构，而词语之间的这种语义结构体现为它们在文本中的出现频率上也具有一定的联系。通过统计学方法，提取并量化这些潜在的语义结构，进而消除同义词、多义词的影响，提高文本表示的准确性。LSI 将每个文献从一个基于标引词词频的向量表述，映射到一个较低维度的向量空间中，该空间是由一组正交的基本向量构成的；同样地，项也可以映射成这个低维空间中的向量，在这个低维空间中，两个相联系的文献即使没有使用任何相同的关键词，也能得到相似的文献表示。

如何构建这个低维的空间(LSI 空间)，从而得到新的项和文献向量表示，潜在语义标引使用矩阵奇异值分解来模拟项和文献之间的联系，文献[5]中有比较详细的说明。像在标准向量方法中一样，潜在语义标引也是开始于一个较大的项-文档矩阵，矩阵中每个元素都代表某一标引词在某一文献中的权值，在实际使用中，可以根据各词在描述文献上的重要性进行加权。因为一个词不会出现在每个文献中，所以该矩阵一般是一高阶稀疏矩阵。如果两个项之间的出现没有任何联系，则无法使用项-文档矩阵中的数据来改善检索。另外，如果在这一矩阵中存在着大量的结构，一些词的出现能够强烈地联系到其他词可能的出现情况。利用奇异值分解可以捕捉这样的“潜在”结构。

1.2 关键技术

奇异值分解(singular valued decomposition,SVD)是普遍使用的LSI空间的构造算法[6]。任何矩阵，例如一个 t× d 阶项-文档矩阵，可以被分解为 3 个矩阵的乘积：

$$X = T_0 S_0 D_0^T$$

其中，T_0 和 D_0 的各列均是相互正交的，并且都是单位长度的，即满足 $T_0^T T_0 = I$ 和 $D_0^T D_0 = I$。S_0是对角矩阵，其对角线元素为分解得到的各奇异值，并且按 $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_n$ 来顺序排列。t和d是矩阵X的行和列；m为矩阵X的秩，满足 $m \leq \min(t, d)$。这样的分解称为“X的奇异值分解”。图 2 给出了一个 t× d 阶矩阵X的奇异值分解示意。

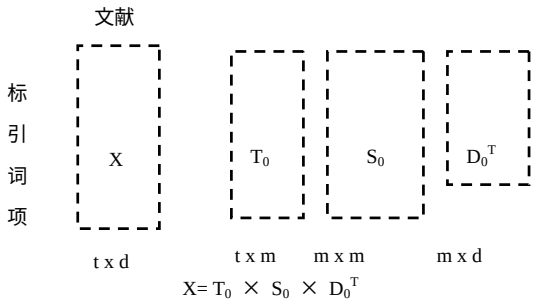

图 2 项-文档矩阵的分解

将矩阵 S_0 对角线上的元素按重要性来排列，若选择保留最大的k个，剔除剩下较小的奇异值，同时，删除矩阵 T_0 和 D_0 最右边的一些列，保留 T_0 和 D_0 中相应的k列，从而产生了 3 个新的矩阵 T_k、S_k 和 D_k，将这 3 个矩阵相乘，得到矩阵 X'，即 $X' = T_k S_k D_k^T$。将文献集合投影到这个k维 X' 矩阵时，与原来文献的距离平方差最小。在这个意义上，潜在语义标引是最优的降低维度的方法。检索时，如一个查询q，通过预处理可将它投影成项-文档矩阵X中的某一列X_q。为了比较q与其他的文献，需要把X_q投影到LSI空间，于是得到一个数学表达式 $D_q = X_q^T T_k S_k^T$。这样就可在k维空间中将q和其他的文献向量进行相似度计算。

从某种意义上说，SVD 是一种提取特征向量的技术，它把原来矩阵中有关词语具体的、非主要的信息都忽略掉，只关注该文本集的主要语义信息。

2 设计方案

2.1 系统框架

虽然国内外对潜在语义标引的理论研究已经取得了相当多的成果，而且也产生了一些实验性的检索系统，但是真正能用于中文环境下的检索系统是非常少的。本文试图在前人已有的研究基础上，提出一个潜在语义标引技术的可行性方案，并且用 Visual C++实现了一个中文信息检索系统，该系统的原型结构框架如图 3 所示。

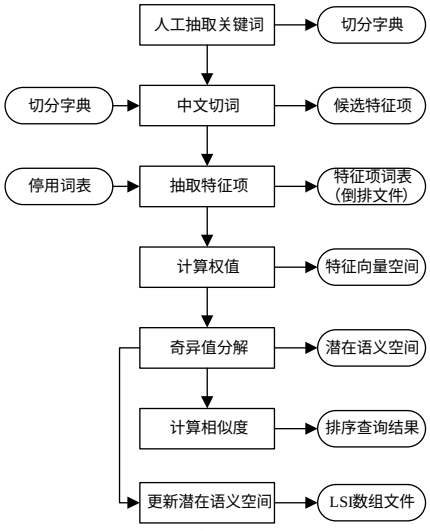

图 3 系统原型设计

在系统原型中，矩形方框代表系统要实现的功能模块(除了预处理模块)，弧形方框代表输入输出的数据。从图 3 中可以看到该系统原型分成几个步骤完成。

(1)人工抽取出文档集的关键词，形成一个切分字典，用于指导文档分词。

(2)中文切分词，程序选取文档中的名词形成候选特征项。选择名词做候选特征项的原因是认为名词所包含的文档意义比别的词性大。

(3)特征项除噪。用停用词表做指导，从候选标引项中剔除噪声词，最终形成倒排索引文件。

(4)计算特征项权重。根据改进型的 TFIDF 权值公式计算项-文档矩阵，产生文献空间。

(5)奇异值分解。利用 MATLAB 中的 SVD 工具包对项-文档矩阵进行奇异值分解，根据实验效果选取一个 k 值，产生潜在语义空间。

(6)计算相似度。把用户查询生成查询向量，用奇异值分解投影到语义空间上，然后和文档进行内积相似度计算，最后根据匹配程度降序输出结果。

(7)更新潜在语义空间。当有新的文本或词汇加入时，可采用 fold-in 方法和重新建立和项-文档矩阵这两种方式对潜在语义空间进行更新。

2.2 关键技术的实现

2.2.1 预处理

首先一个最基本的问题就是要对文本进行词的切分，在英文系统中，词之间由空格隔开，词的识别处理非常方便。而中文系统中，句子中的词之间没有任何分隔标志，加上中文词数量极多，仅靠人工挑选标引词显然不现实。本系统中采用了人工和自动切分相结合的方法，即人工抽取出文档集的关键词，形成一个切分字典，然后调用中科院计算所的词法分析系统接口模块 ICTCLAS.dll 进行中文切词。由于 ICTCLAS 能够进行词性分析，因此可以过滤出其中的名词(因为名词所含的信息最为丰富，最能反映文献内容)作为候选标引词。再根据停用词表做指导，从候选标引项中剔除噪声词。通过对文档进行一系列的 parse 处理，按照倒排索引的数据结构形成文件存储，整个倒排索引文件就是项-文档矩阵的存储结构。

2.2.2 权值计算

权值的计算基于 4 个直觉常识：词频(TF)，文献频率(IDF)，词的位置(出现在标题中的词的重要性大约是正文中词的 N 倍，文献资料称 N 在 5～10 左右)，规范化处理(长文档对词频的贡献要比短文档大，需要对权值进行余弦规范化处理)。在这 4 个因素的作用下，得到改进型 TF*IDF 的权值公式：

$$\frac{TF * IDF * N}{\sqrt{w_1^2 + w_2^2 \ldots + w_n^2}}$$

2.2.3 潜在语义空间的构造

首先，要从倒排文件中恢复项-文档矩阵。假定这个项-文档矩阵是 m × n 的矩阵，则倒排文件中的词表 Term_1,…, Term_m 是矩阵的行，文档表 Doc_1…Doc_n 是它的列，倒排文件中 positing list 就是矩阵每列的数值(代表着某个文档在 m 维向量中的权值)，positing list 中没有权值的则填入 0。由此构造出来一个二维数组，然后调用 MATLAB 包中的 SVD 算法即可得到“文献”矩阵 D_0、“项”矩阵 T_0、奇异值的对角矩阵 S_0。最后，将 S 的元素沿着对角线从大到小排列，S 的 m 个对角线的前 k 个保留，后 $m-k$ 个置 0，同时，删除矩阵 T_0 和 D_0 最右边的一些列，保留 T_0 和 D_0 中相应的 k 列，对 D_0 再次调用 MATLAB 中的矩阵转秩算法，最后将 3 个分解矩阵 $T_k S_k D_k^T$ 相乘，于是得到 k 维的潜在语义空间 X'。

2.2.4 潜在语义空间的更新

在现存的 LSI 空间中增加新的文档或标引项的最直接的方法是重新计算新的项-文档矩阵，然后再进行 SVD。然而这种方法需要大量的计算时间和内存，而导致实际操作有困难。

因此，必须考虑其他方法，其中一个比较简单的方法就是所谓的“折叠”方法(folding-in)[5]。该方法对于加入的新文本，首先利用分词程序生成反映该文本特征信息的文本向量 d，利用已经存在的语义空间 T_k 和 S_k^{-1} 矩阵，将其表示成k维空间向量 $d_k = d^T T_k S_k^{-1}$，每个新生成的文本向量直接附加到 D_k 的列上；类似地，对于新加入的特征项，将其对应的词向量t转化为k维语义空间中的向量 $t_k = t^T D_k S_k^{-1}$，并将其附加到 T_k 的行上。不过，当加入大量的新文本和词汇后，语义空间结构可能发生较大的变化，如果继续使用该方法进行SVD更新，将会影响检索的正确性，需要重新构造语义空间。

2.2.5 用户查询

用户通过交互界面输入查询语句后，系统以处理文本的同样方式构造查询向量，投影到 LSI 空间(其数学表示，前文已具体说明)，然后，利用查询向量和各个文本向量之间内积(w_{jk} 表示语义矩阵中的第 j 列)：

$$\mathrm{Sim}(d_q, d_j) = \sum_{k=1}^{m} w_{ik} \times w_{jk}$$

进行相似程度计算，并根据相似度的大小对所有的文本 ID 进行降序排序，然后将所有的相似度大于用户预先设定的阈值的文本列表通过用户界面提交给用户，如图 4 所示。

图 4 测试运行结果截图

3 实验分析

检索系统在一个具有 2 355 篇中文文本的小规模语料库上进行测试，语料库中的文本都是来自中国浙江网上技术市场技术难题的数据库。系统运行环境为 PIII 1.0 GB，内存 384MB。这个总容量达到 9.6MB 的语料库，经过中文切分词和禁用词过滤等的预处理，产生了不到 1M 的倒排索引文件，压缩比约为 10:1，索引耗时 5min 左右。由索引文件生成 8648×2355 的项-文档矩阵，使用 MATLAB 进行 SVD 分解生成语义空间，取 K 值为 100。在检索试验中，选取了基于传统向量空间模型(VSM)的 SMART 系统(权值采用 TF*IDF)进行对比。

首先定义对比实验中非限定输入和限定输入的概念。所谓限定输入是指允许用户把认为需要过滤的关键词加入到禁用词表中，之后再做相似度计算获得输出结果。非限定输入是指对输入信息不做任何限制操作。两种检索方法均按照自然语言的形式提交查询，且要求结果列表按相关性倒序排列。

这里，还需要给出查全率和查准率的定义。设 n 为文献总量，m 为检索输出的文献量，a 为 n 中与检索主题有关的文献量，b 为 m 中与检索主题有关的文献量。令 R 表示查全

率、P 表示查准率，则 R、P 定义为

$$R = b / a * 100\% \quad ; \quad P = b / m * 100\%$$

根据上述定义，做了 10 次对比实验，在结果列表上进行查准率和查全率的统计评测(表 1)。

表 1 对比试验

算法	非限定输入查全率	非限定输入查准率	平均查询耗时
	限定输入查全率	限定输入查准率	
VSM	100%	9.09%	1 220ms
	90.92%	20.44%	
LSI	100%	13.20%	2 268ms
	83.29%	51.42%	

由表 1 可见，系统在不限定输出的情况下，两种算法的查全率都是 100%，但是查准率却不尽如人意。原因是查询语句中的自然语言包含了一些不具备内容含义的词，或者是切分词不当，这些“噪音词”干扰了算法，导致基于概念的检索转变为基于关键词的检索。当采用限定输出的方式查询后，两种算法的查全率均有下降，这是因为舍弃了一些本应该命中的文本。LSI 的查全率在和 VSM 比较中表现一般，查准率(尤其是相关度排序)却比它要好得多。但是也要看到，LSI 算法的耗时是最高的。另外，比较限定输入和非限定输入可见，正确选择标引词对两种算法都十分重要。

4 结束语

LSI 是向量空间模型的扩展，它从词语之间的相关性出发，通过分析大量的文本中词语的使用关联，提取出潜在的语义空间结构，比较有效地解决了中文环境下自然语言的检索问题，发挥了语义检索的优势。

本人实现的中文信息检索系统，证明了 LSI 模型比 VSM 模型具有更高的查准率，同时，它们受中文切分词效果的影响非常大。最后，如果将 LSI 检索系统原型运用到大型文献库，甚至是作为 Web 搜索引擎，它还必须在潜在语义空间更新、文本向量高维索引组织、快速相似度匹配计算、相关反馈机制等方面做进一步研究。

参考文献

1 李国辉, 汤大权, 武德峰. 信息组织与检索[M]. 北京: 科学出版社, 2003.

2 Salton G, McGill M J. Introduction to Modern Information Retrieval[M]. New York: McGraw Hill Book Co., 1983.

3 庞剑锋, 卜东波, 白 硕. 基于向量空间模型的文本自动分类系统的研究与实现[J]. 计算机应用研究, 2001, 18(9): 23-26.

4 Dumais S T. Using LSI for Information Retrieval, Information Filtering, and Other Things[C]//Proc. of Talk at Cognitive Technology Workshop. 1997: 4-5.

5 盖 杰, 王 怡, 武港山. 基于潜在语义分析的信息检索[J]. 计算机工程, 2004, 30(2): 58-60.

6 Landauer T K, Foltz P W, Laham D. Introduction to Latent Semantic Analysis[J]. Discourse Processes, 1998, (25): 259-284.

（上接第 186 页）

表 1 自适应边缘检测阈值

图像	微分算子	自适应阈值方法	阈值
3(b)	Roberts	矩量保持法	21
3(c)		最大熵法	38
3(d)		大津法	51
3(e)	Sobel	矩量保持法	88
3(f)		最大熵法	131
3(g)		大津法	51

由图 3 及表 1 可知，矩量保持法得到的阈值对本文选取的原图进行边缘检测是比较理想的，较好地保留了原图的基本特征。针对本原图，矩量保持法对其它两种微分算子也是适用的。同时，对其它不同特征的图像进行了实验，实验结果也是比较理想的。该方法克服了常用边缘检测方法在阈值选取存在的主观性，解决了一种算子只对一种阈值选取方法的局限性，说明了本文的的自适应边缘检测方法是有效的、实用的。

4 结束语

研究了二值化阈值选取的几种方法，针对经典边缘检测方法在阈值选取中存在的问题，结合实际研究需要，提出了自适应边缘检测方法，在 C++Builder 开发平台中实现了它们的应用。采用该方法大大地缩短了实验的时间，提高了边缘检测的实时性。下一步研究的方向，将该方法运用到数学形态学边缘检测中去。

参考文献

1 Yang X G, Miao D. Fast Scene Matching Algorithm Based on Image Edge Detection[C]//Proceedings of International Conference on Navigation Guidance and Control. 2001: 391-394.

2 章毓晋. 图像处理和分析[M]. 北京: 清华大学出版社, 1999.

3 曹 菲, 杨小冈. 一种通用的边缘检测方法及其实现[J]. 计算机工程, 2005, 31(11): 30-31.

4 Otsu N. A threshold Selection Method from Gray-level Histogram[J] . IEEE Trans. on Systems Man and Cybemet, 1989, (8): 62-66.

5 陆宗骐. C/C++图像处理编程[M]. 北京: 清华大学出版社, 2005.

6 席卫文. C++ Builder6 程序设计与实例[M]. 北京: 冶金工业出版社, 2003.

（上接第 192 页）

2 Tsutsui S, Yamamura M, Higuchi T. Multi-parent Recombination with Simplex Crossover in Real Coded Genetic Algorithms[C]// Proceedings of the Genetic and Evolutionary Computation Conference. 1999: 657-664.

3 张丽萍, 柴跃廷. 遗传算法的现状及发展动向[J]. 信息与控制, 2001, 30(6): 531-536.

4 郑金华, 蔡自兴. 基于 agent 的并行 GA[J]. 湘潭大学自然科学学报, 2004, 26(1): 23-27.

5 王成栋, 张优云. 基于实数编码的自适应伪并行遗传算法[J]. 西安交通大学学报, 2003, 37(7): 707-710.

6 张 蓉, 彭 宏. 一种快速模拟退火算法及其在数据聚类中的应用[J]. 计算机工程与应用, 2001, 37(15): 85-87.