一种基于主题的概率文档相关模型*)

贾西平　彭　宏　郑启伦　石时需

(华南理工大学计算机科学与工程学院　广州 510640)　(广东技术师范学院计算机科学学院　广州 510665)

摘　要　现有文档关系分析模型难以从主题层次上判别文档相关性。为此,提出了一个基于主题的概率文档相关模型(TPDC)。TPDC借助 Latent Dirichlet Allocation 模型学习文档的主题结构;在计算出主题后验概率和主题相似度的基础上推导出文档后验概率;基于文档后验概率构建文档相关性分析模型。实验结果证明,TPDC 模型在文档检索精度和文档压缩程度两方面优于向量空间模型,因而更能胜任实际应用中的文档检索任务。

关键词　主题,主题相似性,文档相关性,文本挖掘

Topic-based Probabilistic Document Correlation Model

JIA Xi-ping　PENG Hong　ZHENG Qi-lun　SHI Shi-xu

(School of Computer Science and Engineering, South China University of Technology, Guangzhou 510640, China)

(School of Computer Science, Guangdong Polytechnic Normal University, Guangzhou 510665, China)

Abstract　Existing models on document relationship analysis have a difficulty in learning document correlation from topic level. To overcome this difficulty, a topic-based probabilistic document correlation model (TPDC) was proposed. The model learns the topic structure of a document through the latent dirichlet allocation model, infers the posterior probability of a document by computing the posterior probability of its topics and topic similarity, and then constructs the document correlation model based on the document posterior probability. Experimental results show that the TPDC model outperforms the vector space model in retrieval precision and document compression. So the TPDC model is more competent for document retrieval tasks in application.

Keywords　Topic, Topic similarity, Document correlation, Text mining

1　引言

计算机技术及 Internet 的快速发展带来了电子文档的快速增长,使文档的检索、分类和管理变得越来越困难,传统依靠人工的处理方法已经无法满足现实的需要。因而,实现文档关系的自动判别已经成为文档管理和语料库知识发现中亟待解决的重要问题。在这方面,近年来的研究工作主要集中在文档相似性分析。

向量空间模型[1-3](VSM)是信息检索领域最为经典的分析模型之一。给定一个文献集合,VSM 用一个 n 维向量表示每个文档,向量的每一个分量表示一个词项在文档中的权重。VSM 用一个相似性函数计算文档向量之间的相似度,并将它作为文档之间的相似度。向量夹角余弦是常用的相似函数之一。

基于主题的向量空间模型[4](TVSM)是另一个较为流行的文档相似性分析模型。它假定了一个 d 维空间,其每一维代表了一个基本主题,所有基本主题彼此独立。TVSM 用一个 d 维向量表示每个词项,向量的每个分量代表了该词项和对应的基本主题之间的相关性。在 TVSM 中,每个文档用一个基于所有词项基础之上的向量表示,而文档相似性表示为对应文档向量之间的夹角余弦值。

另外,文献[5]借助于 TextTiling 方法[6]和优化匹配理论[7]构建了一个文档相似性模型。该模型用 TextTiling 方法将每个文档切分为前后衔接的几个分段,将每对文档看作一个二分图(Bipartite Graph)的两个部分,而将每个分段看作二分图中的一个节点,不同分图节点(分段)之间的相似性表示相应边的权重。该模型用二分图的最佳匹配表示文档之间的相似性。

从文档的产生过程来看,作者往往会先有一个贯穿全文的中心思想,并围绕这个中心思想扩展出多个主题,最后选择不同的词项将这些主题描述清楚。另一方面,从人类理解文档的实践来看,过程则正好相反,通过阅读具体词项和文字提炼出一系列主题,最后,再归纳出文章的中心思想。而在人类判断文档之间关系的实践中,往往并不会拘泥于某个具体的词项或者文字,而是从主题的层面上判断。例如,两个文档如果都描述了某个相同的主题,人们会认为这两篇文档相关,相同主题越多,文档之间的相关性越强。

相似性和相关性是文档关系中两个密切相关而又有所区别的概念。文档相似性关注文档关系的外在表现,例如两个文档使用了相同的词项或结构等。文本相关性更侧重于强调文档关系的内在特征,例如两个文档都包含了某个相同的主题。显然,文档相关性更接近人类对文档关系的认知习惯,更应该作为文档关系分析研究的重点。

然而,在自然语言处理领域,现有关于文档关系分析方面

*)广东省自然科学基金项目(07006474),广东省科技攻关项目(2007B010200044)。**贾西平**　博士生,主要研究方向为自然语言处理、数据挖掘;**彭　宏**　教授,博士生导师,主要研究方向为智能网络技术、智能商务与数据挖掘;**郑启伦**　教授,博士生导师,主要研究方向为数据挖掘;**石时需**　博士生,主要研究方向为数据挖掘。

的主要研究大多集中在文档相似性分析，而文档之间的相关性却远远没有得到足够的重视。这正是本文研究的出发点。

2 相关工作

2.1 LDA 模型

Latent Dirichlet Allocation[8]（LDA）是近年来广泛应用于文档主题建模的一种三层贝叶斯产生式模型（如图 1）。在 LDA 模型中，文档 d_x 被看作是一系列潜在主题 $t_{x1}, t_{x2}, \cdots, t_{xK}$ 的随机混合，混合比例服从多项式分布，而多项式分布的参数由 Dirichlet 分布产生。每个主题被看作是一个词汇表 V_x 中所有词项的随机混合，混合的比例服从多项式分布，分布的参数在 LDA 中通过 EM 算法进行估计。在 LDA 模型中，词汇表 V_x 和文档中包含的主题数 K 是假定已知和固定的。

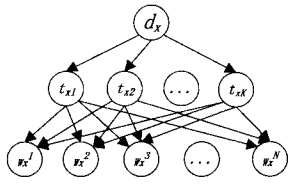

图 1 LDA 模型

本文按照文献[9]中的描述，用 LDA 模型和 Gibbs 抽样从文本中提取主题。

2.2 VSM 模型

向量空间模型[1-3]（VSM）假定了一个 n 维特征空间 R^n，R^n 中的每一维表示一个特征词项。每个文档 d_j 和检索请求 q 分别表示为 R^n 上的 n 维向量 $\vec{d}_j=(w_{1j}, w_{2j}, \cdots, w_{nj})$ 和 $\vec{q}=(w_{1q}, w_{2q}, \cdots, w_{nq})$。$w_{xj}$ 和 w_{xq}（$x=1,2,\cdots,n$）分别表示第 x 个特征词 t_x 在 d_j 和 q 中的权重。通常取 $w_{xj}=\frac{f_{xj}}{\sum_x f_{xj}} * \log \frac{M}{m_x}$，$w_{xq}=(0.5+\frac{0.5 * f_{xq}}{\sum_x f_{xq}}) * \log \frac{M}{m_x}$，$f_{xj}$，$f_{xq}$ 分别表示 t_x 在 d_j 和 q 中出现的频率，M 表示语料库中所有文档的总数，m_x 表示语料库中包含 t_x 的文档数。VSM 一般用向量 $\vec{d}_j$ 和 $\vec{q}$ 的夹角余弦值表示文档 d_j 和检索请求 q 的相似度。

3 TPDC 模型

3.1 定义

定义 1（词汇表） 词汇表是一个有序的词项集合，记为

$$V=\{w^1, w^2, \cdots, w^N\} \tag{1}$$

w^n 表示 V 的第 n 个词项。$\forall m \neq n$，有 $w^m \neq w^n$（$m,n \in [1,N]$）。本文用 V_x 和 w_x^n 分别表示与文档 x 对应的词汇表和词项。

定义 2（主题） 主题是一个由一系列特征词项按照一定比例混合而成的语义单元，记为 t。对于每个主题 t，有唯一的主题向量 T 与其对应。且

$$T=(v^1, v^2, \cdots, v^N) \tag{2}$$

$v^n=p(w^n|t)$，表示给定主题 t 时词项 w^n 出现的后验概率。$n \in [1,N]$，$v^n \in [0,1]$，$\sum_{n=1}^N v^n=1$。本文用 t_{xk} 和 T_{xk} 分别表示文档 x 的第 k 个主题及其对应向量，v_{xk}^n 表示 T_{xk} 的第 n 个分量。

定义 3（主题向量的投影） 给定主题向量 $T_{xi}=(v_{xi}^1, v_{xi}^2, \cdots, v_{xi}^M)$，$V_x=\{w_x^1, w_x^2, \cdots, w_x^M\}$ 为其对应的词汇表，假定有 $V_p=\{w_p^1, w_p^2, \cdots, w_p^R\}$ 且 $V_p \supseteq V_x$，显然 $M \leqslant R$。如果向量 $T_{pj}=\{v_{pj}^1, v_{pj}^2, \cdots, v_{pj}^R\}$ 满足

$$v_{pj}^r=\begin{cases} v_{xi}^m, w_p^r=w_x^m \\ 0, \text{other} \end{cases} \tag{3}$$

$r \in [1,R]$，$m \in [1,M]$，则称向量 T_{pj} 为向量 T_{xi} 在 V_p 上的投影。

定义 4（主题相似度） 给定主题 t_{xi} 和 t_{yj}，T_{xi} 和 T_{yj} 为其对应的主题向量，V_x 和 V_y 分别为 T_{xi} 和 T_{yj} 对应的词汇表。设词汇表 $V_p=V_x \cup V_y$，T_{pk} 和 T_{pl} 分别为 T_{xi} 和 T_{yj} 在 V_p 上的投影向量，定义主题 t_{xi} 和 t_{yj} 的相似度为

$$s(t_{xi}, t_{yj})=\frac{T_{pk} \cdot T_{pl}}{(|T_{pk}| * |T_{pl}|)} \tag{4}$$

$|T_{pz}|=\sqrt{\sum_r (v_{pz}^r)^2}$（$z=k,l$）为向量 T_{pz} 的模。$s(t_{xi}, t_{yj}) \in [0,1]$。当主题 t_{xi} 和 t_{yj} 不相似时，$s(t_{xi}, t_{yj})=0$，当 t_{xi} 和 t_{yj} 完全相同时，$s(t_{xi}, t_{yj})=1$。

定义 5（文档） 每个文档是由一定数量的主题按照一定比例混合而成的语义单位，记为 d。本文用 d_i 表示数据集的第 i 个文档。每个文档 d_i 有唯一的向量 D_i 与其对应：

$$D_i=(\theta_{i1}, \theta_{i2}, \cdots, \theta_{iK}) \tag{5}$$

$\theta_{ik}=p(t_{ik}|d_i)$，表示在文档 d_i 给定时主题 t_{ik} 的后验概率，$k \in [1,K]$，$\theta_{ik} \in [0,1]$，$\sum_{k=1}^K \theta_{ik}=1$。

定义 6（文档相关度） 给定两个文档 d_x 和 d_y，其相关度定义为

$$c(d_x, d_y)=\frac{(p(d_x|d_y)+p(d_y|d_x))}{2} \tag{6}$$

其中，$p(d_x|d_y)$ 表示给定 d_y 条件下 d_x 的后验概率，$p(d_y|d_x)$ 含义与之类似。$c(d_x, d_y) \in [0,1]$，当两个文档不相关时，$c(d_x, d_y)=0$；当两个文档相同时，$c(d_x, d_y)=1$。由（6）式可知，$c(d_x, d_y)=c(d_y, d_x)$。

3.2 TPDC 模型

本节将根据方程（6）中的文本相关度定义，推导出一个 TPDC 模型，用于识别文档的相关性。这里假定任意两个文档 d_x，d_y 的主题已经通过 LDA 模型得到。根据 LDA 的假定，同一文档的各个主题之间是彼此独立的。

根据方程（6），为了计算 $c(d_x, d_y)$，需要先求出 $p(d_x|d_y)$ 和 $p(d_y|d_x)$。考虑到每个文档是一系列相互独立主题的混合，对文档 d_x 中的所有主题求和，可得

$$p(d_y|d_x)=\sum_i p(d_y|t_{xi}) p(t_{xi}|d_x) \tag{7}$$

式中，$p(d_y|t_{xi})$ 和 $p(t_{xi}|d_x)$ 分别表示 d_y 和 t_{xi} 的后验概率，$p(t_{xi}|d_x)$ 可以通过 LDA 和 Gibbs 抽样求得，$p(d_y|t_{xi})$ 未知。

根据贝叶斯公式：

$$p(d_y|t_{xi})=\frac{p(t_{xi}|d_y) p(d_y)}{p(t_{xi})} \tag{8}$$

$p(t_{xi}|d_y)$ 表示文档 d_y 已知条件下主题 t_{xi} 的后验概率，$p(d_y)$ 表示 d_y 的先验概率，可以理解为在 d_y 中的任何词项或主题未被观察之前 d_y 出现的概率。$p(t_{xi})$ 表示主题 t_{xi} 的先验概率，可以解释为 t_{xi} 中的任何词项未被观察之前 t_{xi} 出现的概率。

在 LDA 模型中，超参数 α_{xi} 常被看作主题 t_{xi} 的先验概率。为了简化计算，本文假定所有主题具有相同的先验概率，即假定 dirichlet 分布是对称的，具有唯一参数 α（$0<\alpha<1$）。于是对所有 i,j，有

$$p(t_{xi})=p(t_{yj})=\alpha \tag{9}$$

根据人们判断相关文本的经验，文档之间的相关度只和每个文档的内容有关，而与每个文档出现的先验概率关系不大。据此，本文在 TPDC 模型中假定所有文档具有相同的先

验概率，设为 $\lambda(0<\lambda<1)$，于是对任意文档 d_m 有

$$p(d_m)=\lambda \tag{10}$$

根据方程(8)，(9)，(10)可得

$$p(d_y|t_{xi})=(\lambda/\alpha)p(t_{xi}|d_y) \tag{11}$$

对方程(11)中文档 d_y 的主题求和，可得

$$p(t_{xi}|d_y)=\sum_j p(t_{xi}|t_{yj})p(t_{yj}|d_y) \tag{12}$$

式中 $p(t_{xi}|t_{yj})$ 是已知主题 t_{yj} 条件下主题 t_{xi} 的后验概率，待求。$p(t_{yj}|d_y)$ 是文档 d_y 已知的条件下主题 t_{yj} 的后验概率，可以通过 LDA 模型得到。

根据方程 (7)，(11)，(12)可得

$$p(d_y|d_x)=(\lambda/\alpha)\sum_i\sum_j p(t_{xi}|t_{yj})p(t_{yj}|d_y)p(t_{xi}|d_x) \tag{13}$$

根据对称性，有

$$p(d_x|d_y)=(\lambda/\alpha)\sum_j\sum_i p(t_{yj}|t_{xi})p(t_{xi}|d_x)p(t_{yj}|d_y) \tag{14}$$

根据方程(6)，(13)，(14)，有

$$c(d_x,d_y)=\frac{1}{2}(\lambda/\alpha)\sum_i\sum_j[p(t_{xi}|t_{yj})+p(t_{yj}|t_{xi})]p(t_{yj}|d_y)p(t_{xi}|d_x) \tag{15}$$

根据贝叶斯公式

$$p(t_{xi}|t_{yj})=\frac{p(t_{xi},t_{yj})}{p(t_{yj})} \tag{16}$$

$$p(t_{yj}|t_{xi})=\frac{p(t_{xi},t_{yj})}{p(t_{xi})} \tag{17}$$

由方程(9)，(15)，(16)，(17)，可得 TPDC 模型：

$$c(d_x,d_y)=(\lambda/\alpha^2)\sum_i\sum_j p(t_{xi},t_{yj})p(t_{yj}|d_y)p(t_{xi}|d_x) \tag{18}$$

其中 $p(t_{xi},t_{yj})$ 是 t_{xi} 和 t_{yj} 的联合概率，待求。

考虑到如下事实：主题 t_{xi} 和 t_{yj} 越相似，它们共同出现的概率越高。于是这里用主题的相似度 $s(t_{xi},t_{yj})$ 作为 $p(t_{xi},t_{yj})$ 的近似，结合方程(18)有

$$c(d_x,d_y)\approx(\lambda/\alpha^2)\sum_i\sum_j s(t_{xi},t_{yj})p(t_{yj}|d_y)p(t_{xi}|d_x) \tag{19}$$

由于常数因子 λ/α^2 不会影响文档之间相关度大小的比较，为了便于计算，这里对所有 $c(d_x,d_y)$ 同时除以常数因子 λ/α^2，得到 TPDC 模型的简化表达式：

$$c^*(d_x,d_y)=\sum_i\sum_j s(t_{xi},t_{yj})p(t_{yj}|d_y)p(t_{xi}|d_x) \tag{20}$$

4 实验

本文的实验设计如下：分别利用 TPDC 和 VSM 模型构建两个相关文本搜索引擎，用“检索精度”和“文档压缩率”作为模型的评估标准，并通过相关文档检索的结果比较来验证 TPDC 模型。

表 1 模型参数

模型	参数
LDA	$\alpha=50/n_{\text{topic}}$，$\beta=0.01$，$SEED=3$
	$n_{\text{topic}}=3$，$l_{\text{topic}}=5$
Gibbs sampler	$n_{\text{iteration}}=100$

注：n_{topic}是每个文档中的主题个数；l_{topic}是主题长度(每个主题包含的关键词个数)；$n_{\text{iteration}}$是抽样遍历次数。

本文采用的实验数据为 20 Newsgroups 新闻组数据集，涵盖了计算机、医学、电子、宗教、政治等 20 个领域，共约 20,000条新闻记录。该数据集作为公认的标准数据集，已被广泛用于自然语言理解、搜索引擎、文本挖掘等方面的实验研究。

本文的所有实验是在一台 PC 机上完成的，其主要配置为：AMD Athlon™ 64 X2 Dual Core 4200+ 2.2 GHz 处理器；2.0GB 内存。

实验中的主要模型参数参见表 1。

4.1 检索精度

查准率(Precision)和查全率(Recall)是信息检索中常用的两个检验检索模型的标准。前者表示检索模型返回的文档中相关文档所占的比率；后者表示所有相关文档中实际被模型检索到的部分所占的比率。由于目前还没有非常理想的专用于相关文档检索的语料库，而对于像 20 Newsgroups 这样庞大的语料库，每个检索文档究竟有多少真实相关的文档不得而知。在这种情况下，查全率是无法精确计算的。查准率的准确与否，有赖于检索模型判断文档相关与否的相关度阈值的正确选取。阈值选取过大，会导致相关文档的漏检；阈值选取过小，则无关文档过多。相关度阈值的选取本身就是一项复杂的工作，已经超出本文的研究范围。

为了尽量避免因相关度阈值设置不当导致对模型检索精度的错误判断，本文参考文献[5]用 N 个返回文档的查准率作为相关文本检索模型检索精度的度量标准，记为 $P@N$：

$$P@N=\frac{|C\cap R|}{|R|} \tag{21}$$

式中 C 代表与检索文档相关的文档集合；R 表示由检索模型返回的文档集合；$C\cap R$ 表示由检索模型返回且与检索文档相关的文献集合。$|C\cap R|$和$|R|$表示相应集合中包含的文档数量。

为了验证 TPDC 模型的检索精度，本文从 20 Newsgroup 的 20 个新闻组中随机选取一组，并从该组的 1,000 份文档中随机选取 100 份作为搜索引擎的输入，分别采用 TPDC 和 VSM 模型进行相关文档检索。检索结果由 3 名计算机专业研究生按照如下方法进行评分：输出文档与检索文档相关得 1 分，不相关得 0 分；按照方程(21)计算每次检索的精度，3 人评分的平均值作为每项检索的最终评分。考虑到后期处理的工作量，本文从“sci. electronics”组随机选取了 100 份文档进行实验分析，结果参见图 2。

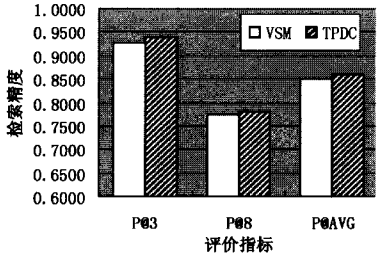

图 2 检索精度比较

图 2 显示，在返回文档数 $N=3$ 和 $N=8$ 的检索实验中，TPDC 模型的检索精度略高于 VSM；平均这两种检索结果，TPDC 模型的检索精度比 VSM 模型约高出 0.67%。另外，在 $N=8$ 的文档检索实验中，两个模型的检索精度都有所“下降”，这是因为有些文档的实际相关文档数小于 8，从而导致 P@8 的平均值下降。然而，这种"下降"并不会影响两个模型检索精度的相对高低。

4.2 文档压缩率

TPDC 模型最大的特点是，将文本相关性分析建立在文

(下转第 218 页)

[5] Liu D, Hosechek J. GC^1 contunuity conditions between adjacent rectangular Bézier surface patches. Computer Aided Geometric Design, 1989, 21(4): 194-200
[6] 白鸿武. 双三次 Bézier 曲面片光滑拼接条件的一个推导[J]. 咸阳师范学院学报, 2004, 12(6): 6-7
[7] DeRose T D. Necessary and sufficient conditions for tangent plane continuity of Bézier surfaces[J]. CAGD, 1990, 7(1/4): 165-180
[8] Liu D. GC^1 continuity conditions between two adjacent rational Bézier surface patches[J]. Computer Aided Geometric Design, 1990, (7): 151-163
[9] Degen W L F. Explicit continuity conditions for adjacent Bézier surface patches[J]. Computer Aided Geometric Design, 1990, 7(20): 181-189
[10] 曲学军, 宁涛, 席平. B 样条曲面的光滑拼接[J]. 计算机辅助几何设计与图形学学报, 2004, 16(1): 138-141
[11] 施锡泉, 赵岩. 双三次 B 样条曲面的连续条件[J]. 计算机辅助几何设计与图形学学报, 2002, 14(7): 676-682
[12] Du W H, Francis J M. On the G^1 continuity of piecewise Bézier surface: A review with new results[J]. Computer-Aided Geometric Design, 1990, 22(9): 556-573
[13] 周西军, 杨海成. NURBS 曲面 G^1 光滑拼接算法[J]. 计算机辅助几何设计与图形学学报, 1996, 8(3): 227-233
[14] Konno K, Tokuyama Y, Chiyokura H. A G^1 connection around complicated curve neshes using C^1 NURBS Boundary Gregory Patches[J]. Computer Aided Geometric Design, 2001: 293-306
[15] 赵庶丰. NURBS 曲面 G^1/G^2 光滑拼接方法[J]. 工程图学学报, 2003(2): 105-115
[16] Lai M J. Geometric interpretation of smoothness conditions of triangular polynomial patches[J]. Computer Aided Geometric Design, 1997, 14(2): 191-199
[17] 丁金扣. 三角域上 B-B 插值曲面片的拼接条件[J]. 北京邮电大学学报, 1994, 17(1): 71-78
[18] Farin G. Triangular Bernstein-Bézier Patches, CAGD, 1986, 3(2): 83-127
[19] 王相海. 三角 Bézier 曲面的一种 GC^1、GC^2 混合及一类隐式代数曲面参数话研究. 博士论文. 长春: 吉林大学, 1999
[20] 赵东福. Bézier 三角组合曲面的局域设计[J]. 工程设计学报, 2002, 12(5): 261-264
[21] Liu D Y, Hoschek J. GC^1 continuity condition between adjacent rectangular and triangular Bézier surface patches[J]. Computer-Aided Geometric Design, 1989, 21(4): 194-200
[22] 吴晓勤, 唐运海. 曲率连续的三角 B 样条曲线与曲面[J]. 计算机应用与软件, 2005, 22(1): 118-120
[23] Greiner G, Seidel H P. Modeling with triangular B-splines. IEEE Computer Graphics and Applications, 1994, 14(2): 56-60

(上接第 180 页)

本主题的基础之上，这从一定程度上降低了描述文档所需的词汇量，起到了文档内容压缩和降维的作用。本文采用文档压缩率(R)表示文档内容被压缩的程度，其计算方法如下：

$$R=n_f/l_d \tag{22}$$

其中，n_f 表示文档特征词的数量；l_d 表示文档中不同词项的数目。在 VSM 中，$n_f=l_d$；在 TPDC 中，$n_f=n_{\text{topic}} * l_{\text{topic}}$，显而易见，$R$ 越小，用于表示一篇文档的特征词越少，文档被压缩的程度越大。

在表 1 所列的参数条件下，对 20 Newsgroups 中所有 20,000条新闻记录按照方程(22)计算压缩率，按组平均结果如表 2 所示。

表 2 压缩率比较

20 Newsgroups	R_{VSM}	R_{PDC}
alt. atheism	1.00	0.50
comp. graphics	1.00	0.59
comp. os. ms-windows. misc	1.00	0.70
comp. sys. ibm. pc. hardware	1.00	0.68
comp. sys. mac. hardware	1.00	0.73
comp. windows. x	1.00	0.58
misc. forsale	1.00	0.92
rec. autos	1.00	0.64
rec. motorcycles	1.00	0.73
rec. sport. baseball	1.00	0.67
rec. sport. hockey	1.00	0.66
sci. crypt	1.00	0.42
sci. electronics	1.00	0.64
sci. med	1.00	0.54
sci. space	1.00	0.50
soc. religion. christian	1.00	0.44
talk. politics. guns	1.00	0.44
talk. politics. mideast	1.00	0.37
talk. politics. misc	1.00	0.40
talk. religion. misc	1.00	0.48
Average	1.00	0.58

实验表明，与 VSM 模型相比，TPDC 模型有着更小的压缩率。如果忽略特征词长度的区别，假定存储每个特征词所需的空间是相同的，则由表 2 可以看出，在文档的特征词存储方面，TPDC 模型的空间开销远小于 VSM，平均约为 VSM 的 58%。因此，TPDC 模型对文档内容的压缩能力明显优于 VSM。

结束语 本文提出一个 TPDC 模型，将文档的相关性建立在文档后验概率的基础之上，并通过概率推理和合理近似，把求解文档之间的相关性转化为计算主题向量之间的相似性，使问题得以简化和解决。实验结果显示，与 VSM 模型相比，TPDC 模型主要有两方面的优点：1)TPDC 模型有较高的检索精度；2)TPDC 模型存储文档特征词的空间开销较少。因此，TPDC 模型在文档检索中有更好的应用前景。下一步工作将重点研究如何根据具体文档自动选取最佳的主题个数和主题长度，并通过更多的语料库实验检测 TPDC 模型的性能。

参 考 文 献

[1] Salton G, McGill M J. Introduction to modern information retrieval. New York: McGraw-Hill, 1983
[2] Baeza-Yates R, Ribeiro-Neto B. Modern information retrieval. New York: ACM Press and Addison Wesley, 1999
[3] van Rijsbergen C J. Information retrieval. London: Butterworths, 1979
[4] Becker J, Kuropka D. Topic-based vector space model// Proceedings of Sixth International Conference on Business Information System. Colorado Springs, 2003: 7-12
[5] Wan Xiao-jun, Peng Yu-xin. A new retrieval model based on TextTiling for document similarity search. Journal of Computer Science and Technology, 2005, 20(4): 552-558
[6] Hearst M A. Multi-paragraph segmentation of expository text// Proceedings of 32nd Meeting of the Association for Computational Linguistics. Los Cruces, 1994: 9-16
[7] Lovasz L, Plummer M D. Matching Theory. Amsterdam: Elsevier Science Publishers B V, 1986
[8] Blei D M, Ng A Y, Jordan M I. Latent dirichlet allocation. Journal of Machine Learning Research, 2003, 3: 993-1022
[9] Griffiths T L, Steyvers M. Finding Scientific Topics// Proceedings of the National Academy of Sciences. 2004: 5228-5235