基于概念空间的文本分类研究*)

黄海英　林士敏　严小卫
(广西师范大学计算机科学系　桂林 541004)

A Study of Text Classification Based on Concept Space

HUANG Hai-Ying　LIN Shi-Min　YAN Xiao-Wei
(Department of Computer Science, Guangxi Normal University, Guilin 541004)

Abstract　Following the expanding of VSM and LSI, a text classification based on Concept Space is proposed in this paper. Information gaining is applied to acquire concepts based on large training set. Concept Space is built by acquiring latent semantic indexing data, building a latent semantic space by LSI, and then adding the class-basis vector. The calculating method of the word-similarity, the text-similarity, the similarity of the text vector and the class-basis vector in Concept Space are presented. Experiment results show the Concept Space method is superior to Vector Space Model. This paper also discusses the future work——the problem of concept space learning.

Keywords　Text classification, Vector space model(VSM), Latent semantic indexing(LSI), Concept space

1. 引言

随着文本信息的快速增长，特别是 Internet 上在线信息的增加，文本(网页)自动分类已成为一项具有较大实用价值的关键技术，是组织和管理数据的有力手段[7]。

文本分类的方法分为两类：一是基于知识的分类方法；二是基于统计的分类方法。基于知识的文本分类系统应用于某一具体领域，需要该领域的知识库作为支撑。由于知识提取、更新、维护以及自我学习等方面存在的种种问题，使得它适用面较窄。基于统计的分类方法由于采用纯粹的数学运算，不苛求复杂的语言学知识和领域知识，在实际应用中收到的良好效果，成为目前流行的文本分类方法。现在广泛应用的基于统计的模型有向量空间模型、Naive Bayes 模型、实例映射模型和支撑向量机模型。

G. Salton 等人在 20 世纪 60 年代提出的向量空间模型(Vector Space Model-VSM)把文本表示为以项的权重为分量的向量，作为向量空间的一个点，然后通过计算向量间的距离决定向量类别的归属。由于把分类简化为空间向量的运算，使得问题的复杂性大大减低。向量空间模型对项的权重评价、相似度的计算都没有作出统一的规定，只是提供一个理论框架，可以使用不同的权重评价函数和相似度计算方法，使得此模型有较广泛的适应性[5]。目前已提出了许多基于 VSM 的统计和机器学习的分类方法。Apte 用决策树技术来建立分类器[11]，Yang 构造了近邻算法[12]，Lewis 设计了线性分类器[13]，Cohen 提出了建立在权值更新基础上的休眠专家算法[14]。但 VSM 模型的不足之处主要有三：一是标引词权值较难确定；二是相似度计算量巨大；三是没有解决词的同义词和多义性对分类的干扰问题[4]。

LSI(Latent Semantic Indexing，潜在语义索引)方法是 1988 年 S. T. Dumains 等人提出的一种新的信息检索代数模型，用于文本的检索有着较好的效果。LSI 利用统计计算导出的概念索引进行信息检索，而不再是传统的索引字、词。LSI 基于这样一种断言：文本中的词与词之间存在某种联系，即存在某种潜在的关于词使用的语义结构。这种语义结构由于部分地被文本中词的多义性和形式上的多样性所掩盖而不明显[7~9]，因此采用统计的方法来寻找该语义结构，并且用语义结构来表示词和文本，以达到化简文本向量，提高信息检索质量的目的。LSI 通过对原文本库的词语——文本矩阵进行奇异值分解(Singular Value Decomposition)计算，并取前 k 个最大的奇异值及其对应的奇异矢量，得到一个 k 秩新矩阵来近似表示原文本库的词语——文本矩阵。新矩阵一方面使词语、文本向量空间大大缩减，另一方面消减了原词语-文本矩阵中包含的因用词习惯或语言的多义性等带来的“噪声”，因而可以提高文本检索的效率和质量[2,3]。直观地说，文本潜在的语义结构保留下来，而词义上的细微区别被忽略掉了。例如，和“电脑”这个词同在文本中出现的词(即上下文或语境)，其中有许多也会出现在“计算机”出现的文本中，如“硬件”、“软件”、“网络”、“操作系统”、“显示器”、“键盘”、“CPU”、“程序员”等，因而它们在 k 维空间中会有类似的表示；而“程序”的上下文会和“计算机”、“电脑”在某种程度上相似，但“大象”的上下文就会完全不一样。因而在 k 维空间表示中，“程序”和“电脑”、“计算机”更接近，和“大象”会相差较远。也就是说，利用潜在语义索引，通过潜在语义结构而不是词形去匹配文本，不但能够发现包含相同词汇的文本，而且还能够发现那些包含同义或近义词汇的文本，能够较好地解决同义词和近义词匹配问题。

2. 基于概念空间的文本分类步骤

2.1　概念空间的定义

首先，挑选一训练文本集，构造词语-文本矩阵 $A=[a_{ij}]_{m\times n}$(其中 a_{ij}为关键词 i 在文本 j 中出现的频率，通常还要根据关键词的重要性加上权值)。因为一个词不会出现在每个

*)中国科学院计算技术研究所智能信息处理开放实验室开放课题资助(IIP2001-4)。黄海英　硕士生，主要研究机器学习。林士敏　教授，硕士导师，主要研究知识系统，机器学习，数据采掘。严小卫　教授，硕士导师，主要研究数据库系统，数据采掘。

文本中，所以 A 一般是高阶稀疏矩阵。

然后，对矩阵 A 应用奇异值分解(Singular Value Decomposition，SVD)，得到词语向量矩阵 U_k 和文本向量矩阵 V_k^T，计算 A 的 k 秩近似矩阵——潜在语义矩阵 A_k($k \ll \min(m, n)$)。

再将每个类别的类基准向量投影到潜在语义矩阵空间中，等效为附加到文本向量矩阵 V_k^T 的若干列。定义加入类基准向量集的 V_k^T 为 $L^T(=(V_k^T, Z^T))$，其中 Z^T 为类基准向量的集合。定义矩阵 $G=U_k\Sigma_kL^T$ 为概念空间。即：

$$
\begin{array}{ccccc}
 & = & & & \\
 & & & V_k^T & Z^T \\
G & & U_k & \Sigma_k & L^T \\
m\times(n+p) & & m\times k & k\times k & k\times(n+p)
\end{array}
$$

图 1　在潜在语义空间中加入 p 个类基准向量

在概念空间中，每个词语均可表示成相应的词语向量 $t^*, t^*=tV_k\Sigma_k^{-1}$，t 为向量空间中的 $1\times n$ 词语向量，每个文本亦可表示成相应的文本向量 $d^*, d^*=d^TU_k\Sigma_k^{-1}$，d 为向量空间中的 $m\times 1$ 文本向量。

2.2　分类步骤

基于概念空间的文本分类方法其分类标准通过概念空间给出，分类过程分为四个步骤：

①给定作为训练集的文本集，在训练集中根据信息增益提取关键词；

②将训练集中的每一篇文本均用等长的向量表示：$D_j=(T_1, T_2, \cdots, T_k, \cdots, T_m)$ 为训练集中的第 j 篇文本；T_k 表示向量中的第 k 个分量，即文本中所含的第 k 个关键词。这样，就可以构成一个向量空间。本文使用词语-文本矩阵 $A=[a_{ij}]_{m\times n}$ 来表示此向量空间。其中 $a_{ij}=L(i,j)\times G(i)$，$L(i,j)$ 为词 i 在文本 j 中出现的频率，$G(i)$ 为关键词 I 的权值。

③通过 LSI 方法获取潜在语义索引数据。利用奇异值分解，计算 A 的 k 秩近似矩阵 A_k，用 A_k 表示潜在语义空间，称 A_k 为潜在语义矩阵；

④通过构造和加入类的表示——类基准向量，建立用于分类的概念空间。将待分类文本与类的匹配处理过程转化为概念空间中文本向量与类基准向量的相似度计算问题。

概念空间解决的是文本分类中的文本集合与类集合的表示问题，下面讨论文本分类中的文本集合与类集合的匹配计算问题。

3. 相似度的计算

相似度是表示词与词、文本与文本、文本向量与类基准向量的相似程度的度量[1]。

3.1　词的相似度计算

设词 t_i 和词 t_j 分别对应词语-文本矩阵 A 的第 i 行和第 j 行，在概念空间中，分别对应矩阵 G 的第 i 行和第 j 行。记 $G=[g_{pq}]$，则词 t_i 和词 t_j 在概念空间中的向量分别为

$t_i=(g_{i1}, g_{i2}, \cdots, g_{in})$

$t_j=(g_{j1}, g_{j2}, \cdots, g_{jn})$

其相似度 $\mathrm{sim}(t_i, t_j)$ 定义为 t_i 和 t_j 的点积，即

$$\mathrm{sim}(t_i, t_j)=t_i\cdot t_j=\sum_{s=1}^{n} g_{is}\cdot g_{js}$$

而对全部 m 个词，其两两之间的相似度为

$$G\cdot G^T=(U_k\Sigma_kL^T)\cdot(U_k\Sigma_kL^T)^T=(U_k\Sigma_k)\cdot(U_k\Sigma_k)^T$$

即 $\mathrm{sim}(t_i, t_j)$ 的计算可由矩阵 $U_k\Sigma_k$ 的第 i 行和第 j 行的点积得到。因为 Σ_k 是一对角阵，所以对概念空间中的坐标进行适当的缩放即可用 U_k 代替 $U_k\Sigma_k$，构造词语在 k 维概念空间中的向量而不影响各词语向量间的相似度。因而在概念空间中，定义 U_k 为词语向量矩阵。

3.2　文本的相似度计算

设文本 d_i 和文本 d_j 分别对应词语-文本矩阵 A 的第 i 列和第 j 列，在概念空间中，分别对应矩阵 G 的第 i 列和第 j 列，$i, j\leqslant n$。记 $G=[g_{pq}]$，则文本 d_i 和文本 d_j 在概念空间中的向量分别为

$d_i=(g_{1i}, g_{2i}, \cdots, g_{mi})^T$

$d_j=(g_{1j}, g_{2j}, \cdots, g_{mj})^T$

其相似度 $\mathrm{sim}(d_i^T, d_j^T)$ 定义为 d_i^T 和 d_j^T 的点积，即

$$\mathrm{sim}(d_i^T, d_j^T)=d_i^T\cdot d_j^T=\sum_{s=1}^{m} g_{si}\cdot g_{sj}$$

由于 $G=U_k\Sigma_kL^T=U_k\Sigma_k(V_k^T, Z^T)=(U_k\Sigma_kV_k^T, U_k\Sigma_kZ^T)$，

定义 $G_V=U_k\Sigma_kV_k^T, G_z=U_k\Sigma_kZ^T$

则对全部 n 个文本，其两两之间的相似度为

$$G_V^T\cdot G_V=(U_k\Sigma_kV_k^T)^T\cdot(U_k\Sigma_kV_k^T)=(\Sigma_kV_k^T)^T\cdot(\Sigma_kV_k^T)$$

即文本 i 与文本 j 的相似度可由矩阵 $\Sigma_kV_k^T$ 的第 i 列和第 j 列的点积得到。同样因为 Σ_k 是一对角阵，因而可用 V_k^T 代替 $\Sigma_kV_k^T$ 来构造文本在概念空间中的向量。在概念空间中，定义 V_k^T 为文本向量矩阵。

3.3　文本向量与类基准向量的相似度计算

设文本 d_i 对应词语－文本矩阵 A 的第 i 列，在概念空间中，对应矩阵 G 的第 i 列，$i\leqslant n$。记 $G=[g_{pq}]$，则文本 d_i 在概念空间中的向量为

$d_i=(g_{1i}, g_{2i}, \cdots, g_{mi})^T$

设共有 p 个类基准向量，z_j 在概念空间中对应矩阵 G 的第 j 列，$n<j\leqslant n+p$，则类基准向量 z_j 在概念空间中的向量为

$z_j=(g_{1j}, g_{2j}, \cdots, g_{mj})^T$

文本 d_i 与类基准向量 z_j 的相似度 $\mathrm{sim}(d_i^T, d_j^T)$ 定义为 d_i^T 和 d_j^T 的点积，即

$$\mathrm{sim}(d_i^T, z_j^T)=d_i^T\cdot z_j^T=\sum_{s=1}^{m} g_{si}\cdot g_{sj}$$

由于 $G=U_k\Sigma_kL^T=U_k\Sigma_k(V_k, Z)^T=(U_k\Sigma_kV_k^T, U_k\Sigma_kZ^T)$，

定义 $G_V=U_k\Sigma_kV_k^T, G_z=U_k\Sigma_kZ^T$

则对全部 n 个文本和 p 个类基准向量之间的相似度为

$$G_V^T\cdot G_z=(U_k\Sigma_kV_k^T)^T\cdot(U_k\Sigma_kZ^T)=(\Sigma_kV_k^T)^T\cdot(\Sigma_kZ^T)$$

即文本 i 与类 j 的相似度可由矩阵 $\Sigma_kV_k^T$ 的第 i 列和 Σ_kZ^T 第 j 列的点积得到。同样因为 Σ_k 是一对角阵，因而可用 Z^T 代替 Σ_kZ^T 来构造类在概念空间中的向量。在概念空间中，定义 Z^T 为类基准向量矩阵。

4. 概念空间的建立

为了减少计算量，在给定类别集 $S=\{\Psi_1, \Psi_2, \cdots, \Psi_w\}$ 后，为每一类 $\Psi_T\in S$ 建立一个概念子空间，合并概念子空间并作一定调整后，便建立了用于文本分类的概念空间。

概念子空间的建立可由核心词的获取、相关词的扩充和类基准向量的构造三个步骤实现。我们将核心词与相关词合称为关键词。

4.1　获取核心词

信息增益(information gain)是机器学习中经常使用的衡量某预示词重要程度的准则，通过预示词是否在文本中出现

对类别的影响来计算所获得的信息量。最初的全部预示词均要通过评估函数获得一个分值，然后按其大小排序，按某阈值大小决定最终保留的较好反映类别特点的预示词，评价函数如下所示：

$$\mathrm{InfGain}(F)=P(F)\sum_{T}P(\Psi_T|F)\log\frac{P(\Psi_T|F)}{P(\Psi_T)}+P(F)\sum_{T}P(\Psi_T|F')\log\frac{P(\Psi_T|F')}{P(\Psi_T)}$$

其中，F 表示一个词，F′表示词 F 不出现，P(F)表示词 F 出现的频率，$P(\Psi_T)$是第 T 类 Ψ_T 出现的概率，$P(\Psi_T|F)$为词 F 出现时 Ψ_T 出现的概率，$P(\Psi_T|F')$为词 F 不出现时 Ψ_T 出现的概率。

定义 InfGain(F)＞ω 的词为核心词，InfGain(F)定义为类别隶属度。通过取较大的 ω 较严格地限制了核心词的数量。此时，核心词较少但用它分类准确率很高。

设类别 Ψ_T 获取了 p 个核心词，则记核心词集合为 $F_T=\{F_{1T},F_{2T},\cdots,F_{pT}\}$，核心词权重记为

$G_F(i)=\mathrm{InfGain}(F_{iT})$，(i=1，2，…，p)。

4.2 扩充相关词

为了解决同义词与近义词影响文本分类的问题，在核心词的基础上，搜索与核心词相似度较大的词，即相关词。

给定类别 Ψ_T 的 n 个文本，构造一个类别 Ψ_T 的词语-文本矩阵

$A_T=[a_{ij}]_{m\times n}$，$a_{ij}=L(i,j)\times G(i)$，

其中 L(i，j)为关键词 i 在文本 j 中出现的频率，G(i)为关键词 i 的权值，若词语 $i\in F_T$，则 $G(i)=G_F(i)$，否则，G(i)＝1。设核心词对应 A_T 的前 p 行向量。

经 SVD 分解后，计算 m 个词的两两之间的相似度为

$$A_{Tk}\cdot A_{Tk}^{T}=(U_{Tk}\Sigma_{Tk}V_{Tk}^{T})\cdot(U_{Tk}\Sigma_{Tk}V_{Tk}^{T})^{T}=(U_{Tk}\Sigma_{Tk})\cdot(U_{Tk}\Sigma_{Tk})^{T}$$

这是一个 m×m 方阵，定义为词语相似度矩阵 $S=[s_{ij}]_{m\times m}$，它反映了类别中各词之间的词语相似度。

第一步：计算核心词与非核心词的相似度。在词语相似度矩阵 S 中，取前 p 行的后(m－p)列，形成核心相似度矩阵，记为 $HS=[h_{ij}]_{p\times(m-p)}$，它反映了核心词与非核心词之间的词语相似度。

定义词语 i 的语义关联度为 $CON_i=\max_j h_{ij}$，此时若 CON_i 值越大，在类别 Ψ_T 中，词语 i 与核心词的语义关联程度就越大。

第二步：搜索与核心词语义关联度较大的词。通过给定一个相关度阈值 ω_1，在核心相似度矩阵 HS 中，取 $CON_i>\omega_1$ 的列(设有 q_1 列)，形成词语相关度矩阵，记为 $XS=[x_{ij}]_{p\times q1}$，它反映了与核心词语义关联度较大的词语与核心词之间的词语相似度。

定义词语 i 对类别 Ψ_T 的支持度为 $SUP_i=\min_j x_{ij}$，通过一个支持度阈值 Φ_1 进行限制。如果 $SUP_i<\Phi_1$，则即使 CON_i 再大，也不认为词语 i 与核心词存在着语义关联，因为词语仅在很少的上下文共现也许并不相关。

第三步：搜索与核心词语义关联度较大且对类别 Ψ_T 支持度也较大的词。通过在词语相关度矩阵 XS 中，取 $SUP_i>\Phi_1$ 的列(设有 q_2 列)，形成词语相关支持度矩阵，记为 $ES=[e_{ij}]_{p\times q2}$，它反映了与核心词语义关联度较大且对类别 Ψ_T 的支持度也较大的词语与核心词之间的词语相似度。

此时，已经对与类别 Ψ_T 的核心词相似度较大的相关词进行了第一次扩充，获得 q_2 个相关词。记类别 Ψ_T 的相关词集合为 $C_T=\{C_{1T},C_{2T},\cdots,C_{Q2T}\}$，权重 $G_C(i)=CON_i$，(i=1，…，q_2)。

类似地，在核心词与相关词的基础上进行第二次扩充。根据一次相关词的权重 $G_C(i)$调整词语－文本矩阵 A_T。此时，若词语 $i\in F_T$，则 $G(i)=G_F(i)$；若词语 $i\in F_C$，则 $G(i)=G_C(i)$；否则，G(i)＝1。与获取一次相关词类似，调整相关度阈值和支持度阈值为 ω_2 和 Φ_2，即可进行第二次相关词扩充。

经过两次相关词的扩充，类别 Ψ_T 的相关词已够用。设类别 Ψ_T 相关词共有 q 个，则记相关词集合为 $C_T=\{C_{1T},C_{2T},\cdots,C_{qT}\}$，对应的权重为 $G_C(i)=\{CON_{1T},CON_{2T},\cdots,CON_{qT}\}$。

4.3 构造类基准向量

记类别 Ψ_T 的类向量为 z_T，则 $z_T=[t_j]_{m\times1}$，其中 t_j 为关键词 j 在类向量中的权值。若词语 $j\in F_T$ 则 $t_j=G_F(i)$，若词语 $j\in C_T$ 则 $t_j=G_C(i)$，否则 $t_j=0$。现在，将 z_T 在概念空间中表示出来，即

$z_T^*=z_T^TU_{Tk}\Sigma_{Tk}^{-1}$

这就是类别 Ψ_T 的类基准向量。至此，类别 Ψ_T 的概念子空间已经形成，$G_T=U_{Tk}\Sigma_{Tk}T_kL_T^T$，其中 $L_T^T=(V_{TK}^T,Z_T^{*T})$。

5. 分类机制

在关键词的基础上，重新训练词语-文本矩阵 $A=[a_{ij}]_{M\times N}$，其中 M 个词为不同类别中的关键词，分别按不同类别的词权重排列，N 个文本为训练文本，有：

$a_{ij}=L(i,j)\times G(i)$，

其中 L(i，j)为关键词 i 在文本 j 中出现的频率，G(i)为关键词 i 的权重。对 A 进行 SVD 分解，再将每个类别基准向量投影到潜在语义矩阵空间中，等效为附加到文本向量矩阵 V_K^T 的若干列。定义加入类基准向量集的 V_K^T 为 $L^T(=(V_K^T,Z^T)^T)$，其中 Z^T 为类基准向量的集合。定义矩阵 $G=U_K\Sigma_KL^T$ 为概念空间。

最后，待分类的文本 D 可视为关键词的序列，从而构成一个向量

$D=(T_1,T_2,\cdots,T_m)^T$

将向量 D 置入概念空间中，在概念空间中表示为

$D^*=D^TU_K\Sigma_K^{-1}$

在概念空间中，将待分类文本向量 D^* 与类基准向量进行相似度计算。设定一个分类阈值 ϕ，若存在 $\mathrm{sim}(D^{*T},z_T^{*T})>\phi$ 则取 $\max(\mathrm{sim}(D^{*T},z_T^{*T}))$的类别 Ψ_T 作为文本 D 的类别。

6. 概念空间的学习

无论预先设计多么大的概念空间，都会遇到新的词语，因而需要系统学习新词语、动态维护概念空间，使整个系统具有自适应性和动态性。构造一个能学习的概念空间将是今后的主要工作。这个工作可以这样进行：

在类别 Ψ_T 的 n 个训练文本进行关键词获取时，设定一个类别 Ψ_T 的平均关键词权重 ξ_T，

$$\xi_T=\frac{\sum_{t}^{p+q}(\sum_{j}^{n}L(t,j))/n}{p+q}$$

其中词语 $t\in F_T\cup C_T$，L(t，j)为词 t 在文本 j 中出现的频率。

将文本 D 视为词序列，$D=(T_1,T_2,\cdots,T_t,\cdots,T_m)^T$。若文本 D 在概念空间中与类基准向量进行匹配后归入类别 Ψ_T，且存在 T_t 不属于任一 $F_T\cup C_T$ 且 $T_t>\xi_T$。则此时设其相应的

$1\times n$ 词语向量为 t'，在概念空间中的向量表示成 $t'^{*}=t'V_k\Sigma_k^{-1}$。计算

$$minm(t)=\min_{词语j\in FTUCT}(sim(t'^{*},j))$$

$$maxm(t)=\max_{词语j\in FTUCT}(sim(t'^{*},j))$$

若 $maxm(t)>\omega_2$ 且 $minm(t)>\Phi_2$，则将词语 t 加入类别 Ψ_T 的相关词集合 C_T。新的相关词集合设为 C'_T。新加入的词语向量应附加到 U_k 的行上，则新的词语向量为 U'_k，则词语 t 的权重：

$$G_C(t)=\max_{词语j\in FTUCT}(sim(t'^{*},j))。$$

相应地，类基准向量也进行更新，即新的类基准向量为

$$z_T^{*}=z_T^{T}U'_K\Sigma_k^{-1}$$

但大量新词语的加入会使概念空间上的分类性能下降，因而要求当新加入的词太多时要重新进行 SVD 计算。

7. 实验结果

由于标题是文本内容的浓缩，与文本主题密切相关，利用标题进行文本分类，可以起到事半功倍的效果。本文对基于概念空间的分类算法进行的是标题分类的测试。建立 8 个类别，并以《电脑报》、《微型计算机》等几个关于计算机硬件的光盘出版物的部分文本标题作为文本源，从中整理出约 1400 条文本标题，其中的 1000 条作为训练集，400 条作为测试集。

文本分类实验，分为训练过程和测试过程。在训练过程中，通过训练集，建立相应的词语-文本矩阵，进行核心词获取和相关词扩充，生成相应的概念空间，并在此基础上形成相应的索引数据库，供计算时调用；在测试过程中，利用上述文本分类算法和索引数据库，计算测试集的类别归属。

表 1 为分别利用 400 篇测试集文本对基于概念空间和基于词索引的算法分别进行分类测试的结果（表中 N 表示属于此类的测试文本数，TN 为分类系统正确识别出的文本数，FN 表示误判为此类的文本数，C 和 W 分别表示基于概念空间测试和基于词索引测试）。

表 1 测试结果

类别名称	N		TN		FN		Recall %		Precision %	
	C	W	C	W	C	W	C	W	C	W
主板	60	60	53	53	0	1	88.33	88.33	100	98.15
CPU	60	60	50	31	0	8	83.33	51.67	100	79.49
显示器	60	60	55	55	2	3	91.67	91.67	96.49	94.83
MODEM	40	40	34	34	0	3	85	85	100	91.89
硬盘	60	60	50	48	0	0	83.33	80	100	100
打印机	40	40	36	36	2	3	90	90.3	94.74	92.31
扫描仪	40	40	32	32	2	2	80	80	94.12	94.12
数码相机	40	40	38	38	0	0	95	95	100	100
合计	400	400	348	327	4	20	Avg		Avg	
							87	81.75	98.86	94.24

实验结果表明，基于概念空间的文本分类与基于向量空间的文本分类相比，在平均查全率和平均查准率上都较占优势，分别达到 87%和 98.86%，比基于向量空间的文本分类方法分别高 5.25 和 4.62 个百分点。可见，上述讨论的基于概念空间的文本分类算法能较好地进行文本分类。

与基于向量空间的文本分类方法相比，基于概念空间的文本分类方法的优势表现在：

①压缩了一定的计算量，待分类文本与类的相似度计算由向量空间中的 m 维减少为概念空间中 k 维（$k\ll m$）；

②消减了向量空间中包含的因用词习惯或语言的多义性等带来的“噪声”。能够很好地解决同义词问题，不但能够发现包含相同词汇的文本，而且还能够利用概念发现那些包含同义或近义词汇的文本。

③基于概念层而不是基于词汇层分析词与词、文本与文本或文本与类之间的相似关系，比直接利用向量空间中的向量进行计算，具有良好的效果。

结束语 随着 Internet 的不断普及和发展，各类电子可读文本日益增多，文本自动分类是处理海量文本的良好手段。本文提出的基于概念空间的文本分类机制，表现出明显的性能优势，是文本自动分类值得研究的方法。今后进一步的工作有二：在理论上研究面向动态变化的文本库的概念空间学习问题和实时更新算法；以及在实践上对更多、更大规模的文本库进行基于概念空间的文本分类实验研究。

参 考 文 献

1 陈家辉．一种基于潜在语义索引的“垃圾”邮件过滤方法．计算机应用研究，2000，(10)

2 周水庚，关佶红，胡运发．隐含语义索引及其在中文文本处理中的应用研究．小型微型计算机系统，2001，(2)

3 林鸿飞．基于示例的文本标题分类机制．计算机研究与发展，2001，(9)

4 李晓黎，刘继敏，史忠植．概念推理网及其在文本分类中的应用．计算机研究与发展，2000，(9)

5 赖茂生，王延飞，赵丹群编著．计算机情报检索，北京大学出版社，1993

6 倪国熙．常用的矩阵理论和方法，上海科学技术出版社，1984

7 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．计算机应用研究，2001，(9)

8 Papadimitriou C H, et al. Latent Semantic Indexing : A Probabilistic Analysis [C]. In: Proc. of PODS'98, Seattle, WA. 1998

9 Deerwester S, Dumains S T, et al. Indexing by Latent Semantic Analysis[J]. Journal of the American Society for Information Science, 1990, 41(6): 391～407

10 Dumains S T. Latent Semantic Indexing(LSI)and TREC-2[C]. In D. Harman, ed. The Second Text Retrieval Conference(TREC2). National Institute of Standards and Technology Special Publication, 1994. 105～116

11 Apte C, Damerau F, Weisa S. Automated learning of decision rules for text categorization. ACM Transaction on Information System, 1994, 12(3): 233～251

12 Yang Y. Expert network: Effective and efficient learning from human decisions in text categorization and retrieval. In: Proc. of the Seventeenth Int'l ACM SIGIR Conference on Research and Development in Information Retrieval. Dublin, 1994. 13～22

13 Lewis D D, Schapore R E, Callan J P, et al. Training algorithms for linear text classifiers. In: Proc. of the 19th Int'l ACM SIGIR Corf. on Research and Development in Information Retrieval. Zurich, 1996. 298～306

14 Cohen W W, Singer Y. Context-sensitive learning methods for text categorization. In: Proc. of the 19th Int'l ACM SIGIR Corf. on Research and Development in Information Retrieval. Zurich, 1996. 307～315