

基于潜在语义分析的中文文本层次分类技术^{*}

王 怡, 盖 杰, 武港山, 王继成

(南京大学 软件新技术国家重点实验室; 南京大学 计算机系 江苏 南京 210093)

摘 要: 从网络文本自动分类的需求出发, 针对基于 VSM 模型的分处理中词条无关假设和词条维度过高等问题, 对基于类中心向量的分类方法进行了改进。利用 LSA 分析中的 SVD 分解获得 Web 文档的语义特征向量, 并在此基础上进行分类处理, 在不损害分类精度的同时提高了分类及其后处理速度, 并设计实现了一个原型系统。

关键词: 潜在语义分析; 类重心分类; 向量空间模型; 文本分类; 特征向量

中图法分类号: TP311 文献标识码: A 文章编号: 1001-3695(2004)08-0151-04

Technology of Chinese Documents Multi-hierarchy Categorization Based on Latent Semantic Analysis

WANG Yi, GAI Jie, WU Gang-shan, WANG Ji-cheng

(State Key Laboratory for Novel Software Technology, Dept. of Computer Science & Technology, Nanjing University, Nanjing Jiangsu 210093, China)

Abstract: To satisfy the need of the categorization of Chinese Web documents, expands text categorization based on category centroid to solve the problem of term independence hypothesis and dimension nimity of documents. Uses the SVD technology of LSA to get semantic eigenvectors of documents, and categorize documents based on them. From the theoretical viewpoint, this method improves the categorization and post-categorization speed meal while accuracy is guaranteed.
Key words: Latent Semantic Analysis(LSA); Text Categorization Based on Category Centroid; Vector Space Model; Text Categorization; Eigenvector

1 引言

20 世纪 90 年代以来, Internet 以惊人的速度发展起来, 它容纳了海量的各种类型的原始信息, 其中文本信息占了主要地位。如何在浩若烟海的文本中掌握最有效的信息始终是信息处理的热点问题。基于人工智能技术的文本分类依据文本的语义和预先定义的主体类别, 给待分文本自动确定主题类别, 与其他各种信息处理技术结合, 有效地提高了信息服务的质量。

当前文本分类的常用方法有基于统计的方法、基于神经网络的方法和知识工程的方法等。从总体上讲, 文本分类的方法可以分为两个类型^[2, 3]:

(1) 基于外延的分类方法。根据文本的外在特征进行分类。最常见的方法是基于向量空间模型的方法, 其基本思想是将文本表示成特征向量, 通过相似度比较来确定文档分类。

(2) 基于语义的分类方法。根据全部或部分理解文本的寓意进行分类。其中基于概念的归类技术是一种重要方法, 其基本思想是抽取短语周围的文本和潜在的语义概念进行文本类别的确定。

基于外延的分类方法如向量空间模型方法所计算出的文档的表征其实是文档中独立词之间关系的浅层次概念集合, 而非深层次语义特征, 这种特征是不准确的。这是因为: 一方面,

一个概念可以有不同的表达方式, 这就是词的同义性问题, 它会造成用户进行分类的文档中的特征词可能与相关类的文档中的词不匹配; 另一方面, 语言中的许多词包含有不同的含义, 即词的多义性问题, 由于词义的不确定性, 这使得用户进行的分类文档可能不会被匹配到最相关类上。

潜在语义分析(Latent Semantic Analysis, LSA) 可以看作一种扩展的向量空间模型, 在向量空间模型中加入了语义分析^[1, 4]。LSA 基于这样一种断言, 文本描述中存在概念之间隐含的语义结构, 文本和词之间的关系可以通过这种结构表示出来。LSA 通过将原来的文本和词的向量矩阵进行奇异值分解, 将文本的关键词空间用更小的语义空间进行表示。LSA 生成的新语义空间中相关文档更为接近, 而且在对解决降低分类精度的同义词和多义词问题非常有用。

2 潜在语义分析的基本思想和特点

潜在语义分析是通过对大批文本进行统计分析, 基于概念语义, 提取和表示词的含义的理论和方法^[6 ~ 8]。其隐含的思想是, 通过语义处理给定词的所有上下文, 同时提供了决定词含义的相似性的相互限制。在 LSA 处理中, 文档首先被抽词, 表示成词频的集合, 一个文档库可以表示为一个 $m \times n$ 词 - 文档矩阵 A , 这里每个不同的词对应于矩阵 A 的一行; 而每一个文档则对应于矩阵 A 的一列。 A 表示为: $A = [a_{ij}]$, 其中 a_{ij} 为非负值, 表示第 i 个词在第 j 个文档中的权重。在实验中, 对于单个词的权重主要考虑其对文本的表征程度和所带的文本的信

息量,所以对权重的处理我们主要考虑了两方面的贡献,即局部权值和全局权值。局部权值和全局权值有不同的取值方法,取值方法的不同会对最后分类的结果产生一定的影响。这里给出我们所选用的方法:

$$W_i = tf_i \cdot idf_i = tf_i \cdot \log_2(1 + N/n_i)$$

其中, W_i 表示该词条在矩阵中的权重; tf_i 表示该词条在文本出现的频率; idf_i 表示该词条的反比文本频率, N 是整个文档集的文档个数, n 是包含该词条的文档个数。

大多数文本只含有一部分词,所以经过处理的矩阵还是典型的稀疏矩阵;同时由于矩阵中的每个词都在每个文章项中有所表示,造成矩阵中含有很多不能表征文本信息的项。通过对此矩阵的奇异值变换可以降低矩阵的纬度,将文档在更少、更能表示其特征的语义空间表示出来。通过奇异值分解,矩阵 A 可以表示为三个矩阵的乘积:

$$A \approx A_k = U_k \sum_k V_k^T$$

其中, $U_k^T U_k = V_k^T V_k = I_k$, U_k 和 V_k 的列分别被称为矩阵 A_k 的左、右奇异向量, $\sum_k$ 是对角矩阵, 对角元素被称为矩阵 A_k 的奇异值。

U_k 矩阵中的行向量对应原矩阵 A 的词向量, V_k 矩阵中的行向量则对应原矩阵 A 的文档向量。这里 U_k 矩阵和 V_k 矩阵中的单个项不一定是非负数,词与词以及文档与文档之间的关系是通过整行之间的相关关系来获得。笔者认为,其中产生的负数不能表示文档的负相关性,因为对矩阵 A 的奇异值分解处理过程中,由于在矩阵 A 中同义词和反义词的表示并没有区分,所以对文档中的同义词和反义词的处理是相同的,这样生成的新矩阵中产生的负值应该与负相关性无关。

$\sum_k$ 是奇异值按递减排列的对角矩阵。因此,我们可以将 $\sum_k$ 中最大的 K 个奇异值提取出来,同时留下 U_k 和 V_k 中相应的奇异向量,构建 A 的 K -维近似矩阵。这里参数 K 的选择非常重要,英文文档实验证明,相对较小的 K 值(100 ~300)就可以取得有效的结果^[7,11,15]。

当潜在语义分析用于分类时,分类文本也通过与产生的新矩阵的降维变换用相同的 K 维表示,其具体数学变换方法如下:

$$d^* = d^T U_k \sum_k^{-1}$$

其中, d 为初始文档向量, d^* 为降维变换后的文档向量。

一旦检索项用 K 维表示出来后,检索项与文档项之间的空间距离就可以通过点积求出,通过点积的大小我们就可以将相关文档以相关度顺序列出。

3 基于LSA 的层次分类系统关键技术

3.1 文档的特征表示

将文本表示为空间向量,首先就要进行文本分词处理。大量的研究工作表明,在文档分类中,与单个字、词相比,用词组来表示文档能够获得更好的分类结果,因此我们用词组为基本词条单位来表示文档的内容语义^[11]。

3.2 文档的向量空间模型

向量空间模型是文档表示的主要方法。与布尔模型相比,该模型考虑到不同词条对文档内容影响程度不同,传统的空间向量方法假设词语语义是相互独立的,每个词语都被看作向量空间中的一个正交基本向量,实际上,词语之间存在很强的关联性,即出现“斜交”现象,影响了文本处理的结果。LSA 利用

这种关联性,通过对文本集中词语的上下文使用模式进行统计转换,获得一个新的低维的语义空间(图 1)。

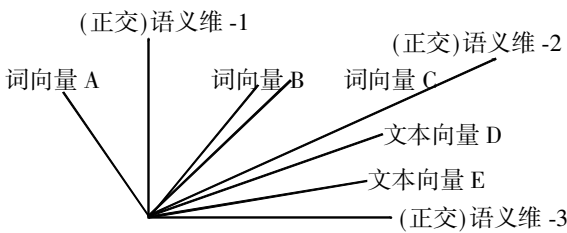


图 1 3 维 - 潜在语义空间示例

与普通分类问题的数据空间相比,文档空间是一个高维,稀疏空间,文档中包含的不同词条数以万计,但对每个具体的文档而言,真正出现在文档中的词条只有上千个。文档中的维数过高会降低分类速度与准确度,因此需要对文档集中的词条进行筛选,从而减少文档向量空间的维度。

在文档预处理中,使用停用词表(Stop List)切除非信息词是提高结果准确度,降低计算冗余的常用方法。非信息词一般是由停用词表定义,其中主要有构成语法的词和一些高频词。大多数文本信息处理系统使用的停用词表含有的停用词基本相同,使用相同的停用词表可以安全地切除文档中的一般冗余词,很少会产生系统的处理精度下降,但也很难显著地提高系统效率。我们使用词频来衡量词条在具体文档集中的区分能力,并由此进行词条选择。考虑一个常见现象:文档集中有相当部分的词条出现频率很低,这些词条的出现很难决定文档的主旨,那么就可以将这些词条略去。在进行一些初步的试验后,我们将阈值设定为 4,在将冗余词条略去后,文档集中有较强区分能力的词条基本被保存下来。

3.3 类重心分类方法

类重心分类方法的基本思想是将文档看成词条的集合,并为类别引入重心矢量的概念。每个类别的重心矢量可以由该类别的训练文档的矢量计算而来,这样文档分类的问题就转换为计算文档矢量与类别重心矢量之间相似度的问题。

在训练阶段,使用公式, $C_k = \sum d / \sum \alpha d \quad D_{Training}^{C_k}$, 计算出每个文档类别 C_k 的类别重心 C_k ; 对于每个待分类的新文档 d_i , 计算文档向量 d_i 与每个文档类别重心 C_k 之间的相似度 $Sim(d_i, C_k)$, 选取相似度最大的一个类别作为新文档的类别。

3.4 扩展潜在语义层次分类方法

类重心分类方法简单易行,在训练文档时可以脱机进行,而分类时计算复杂度与类别个数相关,同样是采用矢量空间模型表示文档的 K 近邻方法在文本分类时,计算复杂度与训练集中的文档个数成正比,这样类重心分类方法在分类速度上明显快于 K 近邻方法。

在处理文档的空间表示时,虽然采用了去除稀有词和使用停用词表处理的方法,词条的维度仍然高达几万维。如此高的维数,不但使得文档空间的存储占用了大量的物理空间,同时其中含有大量的无关数据,影响了文档分类的速度。而且这种分类方法的向量空间模型采用词条集合来表示文档,其中假设文档中的词条之间是线性无关的,文档空间是正交的。但实际上文档空间一些词条的出现有着各种各样的关系,如同义、蕴含、反义、关联。一些相关文档的选词可能有很大的不同,而选词的不同会造成同类文档的分类差异。这样,假定词条间正交,而同时使用夹角余弦来计算文档间相似度缺乏理论依据。解决词条无关性的一个直观想法是构造一个树型或网状结构

的词典^[9], 在词条之间建立联系来反映同义和蕴含等关系, 并在表示文档时进行词条的关系转换。但是由于此方法只能显式地表示一小部分词条之间有限的语义关系, 而且词典的构造和维护都相当困难, 因此在对大规模网络文本进行分类时并不实用。通过对相似词条进行聚类的方法^[5], 用聚类后的词条类来表示文档。然而这种方法在对词条相关性假设进行修正的同时却引入另外一个假设: 文档空间是两两无关的。显然, 这也是不合理的。目前通过对文档进行预处理来加强文档语义, 同时进行降维。其余方法主要有主成分分析 (PCA) 方法^[22]。PCA 采用文档空间变换, 将文档的原始空间向量从 N 维词条空间映射到 R 为主成分空间。对于 $n \times m$ 文档词条矩阵, PCA 使用 $m \times m$ 协方差矩阵来考察词条的相关性, 并由此求出 K 个最大的特征向量。对于 $m \times m$ 协方差矩阵, 其存储代价为 $O(m^2)$, 寻找 K 个最大特征向量的代价为 $O(Km^2)$ 。对于一般文档库, 其中出现的词的数量通常达到数万以上, 这样 PCA 所带来的计算和存储代价是不可接受的。而 LSA 不需要计算 $m \times m$ 协方差矩阵, 所以当 $n < m$ 时, LSA 的计算和存储代价远小于 PCA, 而这个条件对于一般文档库都能达到。

这里我们采用对原始训练文档进行潜在语义分析的方法, 通过分析大量的文本集, 自动生成关键字 - 概念 (语义) 之间映射规则, 合理地表示文档空间, 为分类提供文档准确完整的信息, 较好地表达了文档词条之间的相关性, 同时有效地降低了词条维度。

(1) K 提取与降维

在得到词频后, 我们可以通过对文档矩阵的 SVD 分解对矩阵空间进行主成分分析, 并以此进行空间降维。在 LSA 空间结构中, 文本和词语依据语义上的相关程度组织存放: 分散在不同文本中的同义词空间位置相邻。LSA 方法对语义空间的维度进行约简, 消除语义表达中的“噪音”(词语罕见或者不重要用法含义)。词语含义是词语多种含义的带权平均(如果词语的实际语义偏离这个平均语义很远, LSA 在表达会产生偏颇)。包含不同词语组但主题语义相近的文本位置相邻。文本的含义取决于整体单词的使用模式, 而不是文本中具体包含的单词。文本向量是由各个语义维带权的线性组合表示的。LSA 利用潜在的语义结构表示词条和文本, 将词条和文本映射到同一个 K 维的语义空间内, 均表示为 K 个因子的形式, 向量的含义发生了很大的变化, 它反映的不再是简单的词条出现频率和分布关系, 而是强化的语义关系, 在保持了原始的大部分信息的同时, 克服了传统向量空间表示方法时产生的多义词、同义词和单词依赖的现象。同时, 在新的语义空间中进行相似度分析, 比使用原始的特征向量具有更好的效果, 因为它是基于语义层而不仅是词汇层。

对于原始的 $m \times n$ 词条 - 文本矩阵, 通过 LSA 分析提取出 K 维语义空间, 在保留大部分信息的同时使得 $K < \min(m, n)$, 这样用低维词条、文本向量代替原始的空间向量, 可以有效地处理大规模的文本库。SVD 空间变换所得到的文档向量恰好是原文档向量映射到新语义空间的前 K 个主成分, 其对应的奇异值可以在数学上表示为每个对应的主成分所对应的对整个文档空间的方差贡献。利用主成分方差贡献率递减的性质, 我们可以对 N 维随机向量进行降维, 即在保留 K 个主成分时, 总是保留具有最大方差贡献率的前 K 个主成分, 而将其余的忽略。

(2) 扩展 K 提取与降维

单纯在 SVD 分解后使用 K 值对文档矩阵进行降维, 是目前使用 SVD 对文档集预处理采用的通常方法。但是我们发现: 潜在语义分析将所有文档集进行整合处理, 在平衡文档集的所有语义特征的过程中, 对于大子集的特征, 语义分析之后, 其作为整个语义集的主要特征表现出来, 即作为奇异值高的相应向量, 而那些小语义集的特征, SVD 分解后, 同样在文档矩阵中表现出来, 但是对应于奇异值低的相应向量。在 SVD 分解后使用 K 值对文档矩阵进行降维, 整个语义集的主要特征保留了下来, 即那些大语义集的特征被保留下来, 同时那些小语义集的特征被舍去。在对 SVD 处理后的文档进行分类时, 那些小语义集的语义特征没能得到表现, 分类精度因此被降低。

我们在对 SVD 分解后的文档矩阵进行分析后发现, 在将 SVD 分解后得到的文档矩阵 K 值提高后, 其分类精度仍然继续上升; 对于我们所使用的文档集, 在 K 值提高到 700 时, 其分类精度基本达到最高值; 随后继续加大 K 值, 其分类精度变化产生异常, 甚至有所下降。我们认为, 这是由于在提高 K 的取值并达到一定阶段后, 文档集的代表特征已经被基本提取, 接着提取的是那些表征噪音的向量空间。但是提取太多 K 只是加大矩阵计算量, 加大处理空间。基于以上分析, 我们对 SVD 分解后的文档矩阵采用如下处理: 先取较大的 K 值, 提取文档的较完全特征属性。计算每个文档矩阵的表征特征, 对文档向量中达到某个阈值的语义空间作分类特征处理。我们发现作简单的停用处理就可以达到满意的结果。在具体应用中, 对于我们采用的文档集, 先取文档向量中最大的维为基准, 截取其余值大于其 80% 的维度作为特征维度进行计算。

(3) 层次分类

对于某个文本, 分类器对应每个类别进行匹配, 都要计算出空间向量之间的相似度来指示该文本属于该类别的程度。

在进行特征提取时, 一般做法是将所有文档都视为同一层次, 提取各类在同一层次上的特征向量; 在分类时, 再计算待分类文本向量与各类的相似度, 将其划分到相应的文档类。当文档类较少并且文档类之间区分度较大时, 该匹配策略能够有效地进行文本分类; 但当文档类的数目较多, 且类间区分不明显时, 上述策略就能进行很有效的区分, 并且分类时所需的匹配计算时间较大。我们可以把匹配策略改进为树型结构, 将文档类划分为树状的多个层次, 由内容相近的类组合成一个大类, 内容相近的大类再向上组成根类, 并计算出每个类的类重心; 在进行相似度匹配时, 根据文档层次自上而下进行匹配。采用此种匹配策略, 使得文档的区分特征更为明显, 有助于提高分类准确性, 同时大大减少了相似度计算和匹配的次数, 提高了分类的速度。

(4) 算法过程

在上述分析的基础上, 我们设计了一种基于潜在语义分析的层次类重心文档分类方法。

在训练阶段, 扩展潜在语义层次分类由以下几个步骤构成:

将训练文档中的文档向量经过预处理后得到原始文档矩阵 $A_{m \times n}$ 。

对原始文档矩阵进行 SVD 分解, 得到文档特征矩阵 V_k 。

根据精度需要选取 K 个奇异值 λ_i , 且 $\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_K$ 大以及对应的 K 个奇异向量 $\overline{P_i}$ 。

将训练文档中的每个文档向量 $\overline{d_i}$ 映射到新的文档空间中, 并对每个文档向量进行归一化。

从底层文档类开始, 计算在新空间中文档的类别重心矢量 $\overline{C_k}$, 由层次上升, 从底层类重心得到上层类重心, 直至分类树根节点。

- 在分类阶段, 扩展潜在语义层次分类由以下几个步骤组成:
- (1) 将测试中每个待分类的文档的空间向量 $\overline{d_i}$ 映射到新的文档空间中, 得到 K 维文档向量 $\overline{d_i^*}$ 。
 - (2) 对 K 维文档向量进行扩展 K 处理。
 - (3) 在新的文档空间中, 自上而下在每个类别层次中比较 K 维文档向量与当前层次中每个类别重心矢量 $\overline{C_k}$ 的相似度 $\text{Sim}(\overline{d_i}, \overline{C_k})$, 选取相似度最大的作为当前类别, 再在当前类别的子类别中进行新的类别匹配。

3.5 基于 LSA 的层次分类系统实现(图 2)

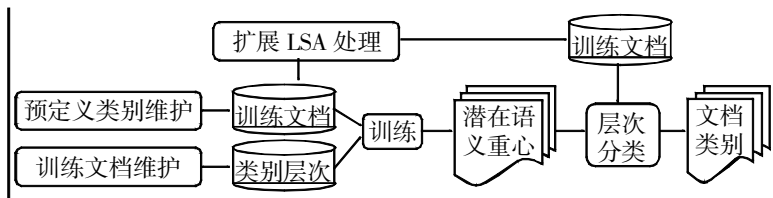


图 2 层次分类系统示意图

4 实验结果与评价

实验中使用的文档库包含近 20MB, 共 1 551 篇计算机领域的中文文档, 主要来源于“计算机世界”、“网络世界”、“微电脑周刊”等 Web 站点。这些文档经过人工分类后, 标记为 34 个文档类, 这些文档类经过逐层向上类别合并为根类别共 5 类: 计算机安全、计算机软件、计算机网络、计算机应用、计算机硬件。实验平台为 P3-650, 256MB 内存。实验中主要测试了原始层次分类方法和扩展潜在语义层次分类方法的分类速度和分类准确度。实验采用 10 - 折交叉确认, 即初始文档集被划分为 10 个互不相交的子集, 每个集大小大致相等, 训练和测试进行 10 次, 再第 i 次迭代, 第 i 个子集作为测试集, 其余子集都用于训练分类法, 依次轮换, 并计算平均值。

对文档库中所有的文档采用无词典抽词方法抽取词条, 并采用矢量空间模型将每个文档表示为 13 543 维的稀疏矢量, 在此基础上使用原始层次分类方法和扩展潜在语义层次分类方法进行文档分类, 记录分类的执行时间和准确度。

错误分类级为原分类级至分类器分类级跨域层次数, 错误级率 = 错误分类级 / 分类文档数。其实验结果比较如表 1 所示。

表 1 实验结果比较				
方法	分类时间 (s)	错误文档数	错误分类级	错误级率
原始 TC3	125.7	128	468	0.301741
LS- TC3	92.7	133	468	0.301741

扩展潜在语义层次分类出现的错误中, 有 78 个分类错误出现在分类树的最上层, 占错误总数的 58. 65%; 同时在树的叶节点出现的分类错误为 44 个, 占错误总数的 33. 08%。在原始层次分类出现的错误中, 有 81 个分类错误出现在分类树的最上层, 占错误总数的 63. 28%; 同时在树的叶节点出现的分类错误为 31 个, 占错误总数的 24. 22%。由于我们选用的文档集特征使得新文档空间较大, 其绝对 CPU 计算量其实较原文档大, 但是由于文档表示较原文档简单, 使得读取时间和内存缓冲时间大大降低, 提高了总的处理速度。不同参数下实验结果比较如表 2 所示。

表 2 不同参数下实验结果比较

方法	分类时间	树型匹配次数	错误文档数	错误分类级	错误级率
K = 220 LS- TC3	62. 5	3894394	241	819	0.528046
K = 700 LS- TC3	133. 4	12214797	84	285	0.183752
K = 400 LS- TC3	92. 7	7069074	133	468	0.301741
MOD- KLS- TC3	74. 9	3756733	122	475	0.305673

我们可以发现, 使用扩展潜在语义层次分类进行特征提取, 其分类精度与 K = 400 维时基本相同, 甚至在错误文档数上有所减少; 树型匹配次数与 K = 220 维树型匹配次数相差无几, 而 CPU 计算时间主要与树型匹配次数相关, 这样扩展潜在语义层次分类在保证精度的同时大大降低了计算时间。

扩展潜在语义层次分类方法采用了基于潜在语义分析的降维技术, 这比原始层次分类方法基于词条频度降维更为有效。LSA 采用了文档空间变换技术, 更好地表示了文档的语义空间, 文档表示的维度是真正两两无关, 因此弥补了原始层次分类方法的理论缺陷。同时文档分类采用了树型层次结构, 有效地实现了现实分类应用结构。

从定义上讲, 所有非标注降维技术都用来进行标注降维, 但是由于数据集的不确定, 很多标注数据集使用非标注降维后将大大降低文本类别的区别性。对于很多数据集, 有些类别的特征变量可能对于整个数据集来说是降维对象, 这主要是由于数据集的不平衡性造成的, SVD 分解后的文档集是从整体上将文本进行平衡, 这样就使得局部特征被略去, 造成分类误差。LSA 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, 有效地获得和表示词汇的语义知识, 以提高后续处理的精度。

文本分类精度的提高主要可以从提高类别之间的区分度入手, 哪种文本特征表示的区分度高, 其分类精度就越高。LSA 只是一种文本的原特征表示方法, 并没有增大文本类别之间的区分度, 所以并没有提高文本分类精度, 但是就语义空间的表示, LSA 分析的效果还是毋庸置疑的。在 LSA 分类文本表示基础上, 采用其他处理方法, 提高文本类别的 Bhattacharyya 距离具有进一步的研究价值。

5 结论

本文探讨了潜在语义分析在文本分类系统中的应用, 提出了扩展潜在语义层次分类算法。扩展潜在语义层次分类方法采用文档空间变换和潜在语义分析技术, 将文档的原始矢量从 N 维词条空间映射到 K 维语义空间, 以便训练与分类展开工作, 理论上弥补了词条无关假设的缺陷, 在保证分类准确度的情况下, 提高了分类速度。最后对 LSA 在分类算法中的应用前景进行了总结。

参考文献:

[1] eerwester, et al. Indexing by Latent Semantic Analysis[J] . Journal of the Society for Information Science, 1990, 41 (6) : 391-407.

[2] Yiming Yang. An Evaluation of Statistical Approach to Text Categorization[J] . Information Retrieval Journal, 1998.

[3] Stuart Weibel. Metadata: The Foundations of Resource Description [J] . D-Lib Magazine, 1995, (7) .

[4] Landauer T K, Dumais S T. A Solution to Plato 's Problem: The Latent Semantic Analysis Theory of the Acquisition, Induction, and Representation of Knowledge[J] . Psychological Review, 1997, 104: 211-240.

[5] 吴立德, 等. 大规模中文文本处理[M] . 上海: 复旦大学出版社, 1997. (下转第 165 页)

线程, 在程序中安排了三种级别(紧急、最高、正常) 的线程。分别对应三种功能类别的程序: 同步线程、收发线程(即数据采集和处理以及数据收发线程) 和显示线程。根据优先级不同, 系统对各线程请求 CPU 时间的响应也不同(最高、高、低)。下面是对线程的定义、启动和线程程序安排的说明:

```
定义线程
UINT SynchThread( LPVOID pParam)
UINT Sync_Send_ReceiveThread( LPVOID pParam)
UINT DisplayThread( LPVOID pParam)
定义线程相关参数
CEvent SynchEvent1( FALSE, TRUE) 同步事件, 被同步线程置为有信号状态, 被线程置为无信号状态
CEvent SynchEvent( FALSE, TRUE) 同步事件, 被同步线程置为有信号状态, 被线程置为无信号状态
int CSWXView: : OnCreate( LPCREATESTRUCT lpCreateStruct)
{ ...
开启线程
设置程序进程为实时
:: SetPriorityClass( :: GetCurrentProcess( ), REALTIME_PRIORITY_CLASS)
设置同步线程为紧急
:: AfxBeginThread( SynchThread, GetSafeHwnd( ), THREAD_PRIORITY_TIME_CRITICAL)
设置收发线程为最高
:: AfxBeginThread( Sync_Send_ReceiveThread, GetSafeHwnd( ), THREAD_PRIORITY_HIGHEST)
设置显示线程为正常
:: AfxBeginThread( DisplayThread, GetSafeHwnd( ), THREAD_PRIORITY_NORMAL)
}
```

以上为线程的定义和开启方法, 下面分别对几个线程功能进行介绍。

同步线程是使其他各线程同步工作的关键, 其他线程都是在等到同步线程发出的同步信号后开始执行对应的程序。在同步线程中, 系统对 COM1 口初始化, 等待 COM1 口 CTS 和 DSR 上的外触发同步信号。当有外触发信号产生时, 同步线程根据信号类型发出相应的同步信号。各线程之间的时序关系如图 1 所示。

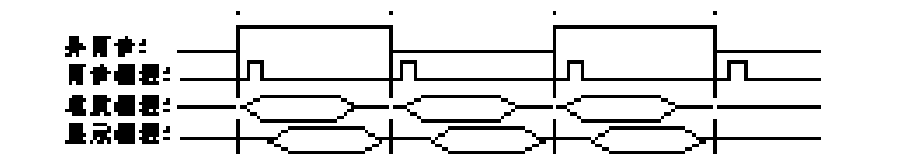


图 1 各线程之间时序关系

收发线程是执行数据收发通信任务以及数据采集和处理线程。该线程优先级别低于同步线程, 当同步线程发出同步信号时, 收发线程开始执行收发通信程序及编、解码程序, 数据采集和数据处理程序。

显示线程用于显示采集到的数据和处理结果。显示线程的优先级别最低。

5 结论

在角变形实时测量系统中, 有很高的同步、逻辑和实时性的要求, 为此本系统在软件设计过程中采用了多线程的编程技术, 充分利用了同步多线程的优势, 在系统中设计了三个不同功能的线程, 满足了系统实时性要求, 完成了舰艇角变形的实时测量任务。多线程设计是当前软件设计过程中必备的一项关键技术, 可以大大提高程序的执行效率, 实现系统实时性的要求。

参考文献:

[1] 张明慧. 基于视频图像的动态实时测量技术的实现[D] . 长春: 长春光学精密机械与物理研究所, 2003 .

[2] 于起峰, 陆宏伟, 刘肖琳. 基于图像的精密测量与运动测量[M] . 北京: 科学出版社, 2002 .

[3] 希望图书创作室. Visual C ++ 技术内幕 6. 0[M] . 北京: 化学工业出版社, 2000 .

[4] 李光明. Visual C ++ 6. 0 经典实例大制作[M] . 北京: 中国人事出版社, 2001 .

[5] 刘金宁, 等, LabWindows/ CIV 下多线程技术在某发控设备测试中的应用[J] . 电子测量与仪器学报, 2002, 16: 710-714 .

[6] 乔立岩, 等. 多线程测控程序设计方法研究[J] . 测试技术学报 , 2002, 16: 1145 -1149 .

作者简介:

张明慧(1974-) , 女, 吉林长春人, 讲师, 在读博士, 主要从事图像处理、光电检测和自动控制仪器的研究工作; 张尧禹(1973-) , 男, 辽宁黑山人, 助理研究员, 博士, 主要从事靶场设备、瞄准系统、跟踪系统的研究工作; 黄廉卿(1942-) , 男, 吉林长春人, 研究员, 博士生导师, 主要从事图像压缩和图像处理方面研究工作。

(上接第 154 页)

[6] 林鸿飞, 姚天顺. 基于潜在语义索引的文本浏览机制[J] . 中文信息学报, 2000, 14(5) : 49-56 .

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

[8] Noriaki Kawamae. Latent Semantic Indexing Based on Factor Analysis [Z] . 2001 .

[9] 邹涛. 基于 WWW 的信息发现技术研究[D] . 南京大学计算机系博士学位论文, 1999 .

[10] 王继成. 基于元数据的 Web 信息检索技术研究[D] . 南京大学计算机系博士学位论文, 2000 .

[11] Furnkranz, T Mitchell, E Riloff. A Case Study in Using Linguistic Phrases for Text Categorization on the WWW[Z] . Working Notes of the 1998 AAAI/ICML Workshop on Learning for Text Categorization .

[12] G Dietterich. Approximate Statistical Tests for Comparing Supervised Classification Learning Algorithms[Z] . 1997 .

[13] intsch E, Steinhart D, Stahl G. Developing Summarization Skills Through the Use of LSA-based Feedback[J] . Interactive Learning Environments, 2000, 8(2) : 87-109 .

[14] Zelikovitz, H Hirsh. Using LSI for Text Classification in the Presence of Background Text[Z] . 2001 .

[15] in Hongl-Fei. The Mechanism of Text Title Classification Based on

Examples[J] . Journal of Computer Research & Development, 2001, 38(9) : 1132-1136 .

[16] 林鸿飞. 基于示例的文本标题分类机制[J] . 计算机研究与发展, 2001, 38(9) : 1132-1136 .

[17] alcolm Slaney, Dulce Ponceleon. Hierarchical Segmentation Using Latent Semantic Indexing in Scale Space[Z] . 2001 .

[18] Kaban, M A Girolami. Fast Extraction of Semantic Features from a Latent Semantic Indexed Text Corpus[Z] . 2002 .

[19] alter Kintsch. On the Notions of Theme and Topic in Psychological Process Models of Text Comprehension[Z] . Thematics: Interdisciplinary Studies, 2002. 157-170 .

[20] W Berry, Z Drmac, E R Jessup. Matrices, Vector Spaces, and Information Retrieval[J] . SIAM Rev. , 1999, 335-362 .

[21] Toutanova, F Chen. Text Classification in a Hierarchical Mixture Model for Small Training Sets[Z] . 2001 .

[22] E Jackson. A User ' s Guide to Principal Components[M] . John Wiley & Sons, Inc. , 1991 .

作者简介:

王怡(1980-) , 男, 硕士研究生, 主要研究方向为智能信息检索; 盖杰(1979-) , 女, 硕士研究生, 主要研究方向为智能信息检索; 武港山(1967-) , 男, 副教授, 博士, 主要研究方向为智能信息检索; 王继成(1973-) , 男, 工程师, 副教授, 博士, 主要研究方向为智能信息检索。