第 35 卷 第 7 期 计 算 机 工 程 2009 年 4 月
Vol.35 No.7 Computer Engineering April 2009
·人工智能及识别技术· 文章编号：1000—3428(2009)07—0183—03 文献标识码：A 中图分类号：TP311

基于贝叶斯信息准则的文本主题数估计

王晓斌，温 春，石昭祥
(电子工程学院网络工程系，合肥 230037)

摘 要：特定领域的主题识别和关键词提取有着广泛的应用，但通过人工指定识别或文本聚类自动生成的主题类别缺乏客观的度量方法。该文结合基于 BIC 准则的模型选择理论和独立分量分析技术对主题的数量进行概率估计，给出主题数量在 BIC 意义下的统计分布。在此基础上实现了文档矩阵的 ICA 分解，并根据分离的独立分量获得主题的关键词及其权重。实验表明，该方法在没有领域知识支持的情况下能估计出反映文本集合的主题数并提取相应的关键词。

关键词：主题识别；关键词提取；独立分量分析；贝叶斯信息准则

Text Topic Number Evaluation Based on Bayes Information Criteria

WANG Xiao-bin, WEN Chun, SHI Zhao-xiang
(Department of Network Engineering, Electronic Engineering Institute, Hefei 230037)

【Abstract】There are many applications that can benefit from topic identification and keyword extraction. The traditional way of choosing the topic number depends on human labeling or automatic clustering which is immeasurable. This paper utilizes the Bayes Information Criteria(BIC) based model selection theory to evaluate the probability of each topic numbers taking. After the topic number is acquired, the paper implements the Independent Component Analysis(ICA) decomposition of term-document, then calculates the weight and extracts the keyword according to the ICA separating matrix. Experiments show this method extracts the keyword in a meaningful way.

【Key words】 topic identification; keyword extraction; Independent Component Analysis(ICA); Bayes Information Criteria(BIC)

1 概述

独立分量分析(Independent Component Analysis, ICA)技术是近年来信号处理领域的研究热点之一。该技术可以在不知道接收信号参数的情况下，仅仅根据输入信号的基本统计特征，由观测信号恢复出源信号。自 1997 年提出快速算法 FastICA[1]以后，该技术进入了实用化阶段。目前，独立分量分析在文本挖掘领域也得到了一定的应用，如文本流的主题识别、文本分类、聚类等，成为继主分量分析 PCA(也称潜在语义索引 LSI)之后，统计信号处理技术在文本挖掘领域的又一新应用。

与经典 ICA 理论相对应，文本挖掘领域的 ICA 模型将文档矩阵中的特征词作为传感器、文档序列作为采样点，在投影后的 LSI 空间中搜索具有非高斯性的投影方向(即分解坐标基)，再将 LSI 空间投影到统计上相互独立的分量上。使用此方法，可以有效地分离文本信息中的高斯噪声，分解后的独立分量表示统计上相互独立的主题。利用分解出的独立分量对词条分量逐个加以分析，还可以提取出主题的关键词。与其他使用启发式信息的统计方法相比，ICA 模型具备完整的框架定义和清晰的优化准则，能同时将主题分类、聚类和关键词提取以及 LSI 固有的同现分析能力统一到一个框架下，因此，颇具吸引力。

在实际应用中，由于 PCA 和 ICA 都是纯数据驱动的方法，因此确定主题数量存在一定困难。本文在深入研究基于 ICA 的主题识别和关键词提取问题的基础上，使用盲信号处理中动态分量的估计技术[2]对文本集合中主题个数进行估计。算法直观地给出主题数的概率分布，体现了模型选择理论中极大似然意义下的主题数估计。

2 基于贝叶斯信息准则的主题数估计

不同主题的文档往往具有不同形式的特征词概率分布，主题以独立随机变量的形式隐含在文档信息中。假设文档集是由多个统计上相互独立的主题文本线性混合而成，那么根据中心极限定理，混合后的词频分布比单个主题下的词频分布要更接近高斯分布。一般的，独立分量分析就是通过计算各种非高斯性判据来确定隐含的独立随机变量。

2.1 基于向量空间模型的独立分量分析

根据向量空间模型，每个文档 $\boldsymbol{X}_i$，i=1,2,…,d 可由 t 维文档特征词权重向量 $\boldsymbol{w}_j$ 来表示，那么包含 d 个文档和 t 个特征词的可观测文档集 $\boldsymbol{X}$ 可以表示成 $t\times d$ 阶的特征词-文档矩阵。基于空间向量模型的 ICA 模型表示为

$$\boldsymbol{X}=\boldsymbol{AS} \tag{1}$$

其中，$\boldsymbol{X}$ 为可观测的文档矩阵；$\boldsymbol{A}$ 为 $d\times c$ 阶混合矩阵(mixing matrix)；$\boldsymbol{S}=(\boldsymbol{s}_1,\boldsymbol{s}_2,\cdots,\boldsymbol{s}_c)^{\mathrm{T}}$ 为 d 个文档的 c 个主题信息构成的向量；c 表示文档中独立分量(Independent Component, IC)的个数。每个独立分量 $\boldsymbol{s}_c$ 定义了具有相同文档主题的一个类别，因此文档集可分为 c 个类别。利用 $\boldsymbol{S}$ 各个分量间的统计独立性假设和观测矩阵 $\boldsymbol{X}$，借助源信号概率分布的先验知识估计混合矩阵 $\boldsymbol{A}$，能够估计出文档的主题信息 $\boldsymbol{S}$。

由于实际的特征词表规模一般超过 1 000 个，因此矩阵 $\boldsymbol{X}$ 很稀疏，对 ICA 来说这是一种病态学习的问题(ill-posed learning problem)。所以，一般先使用奇异值分解(SVD)对文

作者简介：王晓斌(1977－)，男，博士研究生，主研方向：机器学习，Web 挖掘；温 春，博士研究生；石昭祥，教授

收稿日期：2008-07-15 **E-mail**：wxb-77@163.com

档矩阵进行降维，再进行 ICA 处理。经过 SVD 分解的特征词-文档矩阵表示为

$$\underset{t\times d}{\boldsymbol{X}}=\underset{t\times n}{\boldsymbol{T}}\cdot\underset{n\times n}{\boldsymbol{L}}\cdot\underset{n\times d}{\boldsymbol{D}}^{\mathrm{T}} \tag{2}$$

其中，$\boldsymbol{L}=\boldsymbol{Diag}[l_1,l_2,\cdots,l_n]$ 是由奇异值构成的对角矩阵，各 l_i 称为奇异值；$\boldsymbol{T}$ 和 $\boldsymbol{D}$ 分别保存特征词条和文档的特征向量。对保留前 c 个奇异值的矩阵 $\boldsymbol{L}_{c\times c}\cdot\boldsymbol{D}^{\mathrm{T}}_{c\times d}$ 进行 ICA 处理，相当于对式(2)插入 ICA 分解。分解后的矩阵表示为

$$\underset{t\times d}{\boldsymbol{X}}=\underset{t\times d}{\boldsymbol{T}}\cdot\underset{d\times c}{\boldsymbol{A}}\cdot\underset{c\times d}{\boldsymbol{S}}^{\mathrm{T}} \tag{3}$$

其中，c 为 IC 的个数；$\boldsymbol{A}$ 为源信号阵 $\boldsymbol{S}$ 在 LSI 空间上的投影。

由于 ICA 分解要求观察信号的数量等于源信号数量，因此对文档矩阵保留不同奇异值个数的 SVD 截断，也相当于确定了不同 IC 数的分解模型。一般的，对保留的前 c 个奇异值的矩阵 $\boldsymbol{L}_{c\times c}\cdot\boldsymbol{D}^{\mathrm{T}}_{c\times d}$ 进行 ICA 分解也总能找到 c 个独立分量。

2.2 基于贝叶斯信息准则的主题数估计

对不同 IC 数的分解模型集合，可以使用基于贝叶斯信息准则的模型选择理论对 IC 数(即主题数)进行概率估计。令假设集合中，含有 m 个 IC 的假设标记为 m，$m=0,1,\cdots,M$(其中 0 表示空假设，对应于没有显著独立分量的情况)。对观察数据 $\boldsymbol{X}$，选择模型 m_i 的概率表示为 $P(m_i|\boldsymbol{X})$，使用贝叶斯公式可以重写为

$$P(m\mid\boldsymbol{X})=\frac{P(\boldsymbol{X}\mid m)P(m)}{P(\boldsymbol{X})} \tag{4}$$

先验概率 $P(m)$反映在数据到来之前，对某一模型的信任程度。如果没有特定的先验知识则使用均匀分布。一般的，上述概率模型可以定义为参数向量 $\boldsymbol{\theta}$ 的函数，因此，有生成模型(generation model)分布函数 $P(\boldsymbol{X}\mid\boldsymbol{\theta},m)$，该分布由观察模型给出[3]:

$$P(\boldsymbol{X}\mid m)=\int \mathrm{d}\boldsymbol{\theta}P(\boldsymbol{X},\boldsymbol{\theta}\mid m)=\int \mathrm{d}\boldsymbol{\theta}P(X\mid\boldsymbol{\theta},m)P(\boldsymbol{\theta}\mid m) \tag{5}$$

其中，概率分布 $P(\boldsymbol{\theta}\mid m)$ (证据因子)以参数形式反映先验概率 $P(m)$。除了 $\boldsymbol{X}$ 为空的情况下，该证据因子可通过归一化忽略不计。一般而言，式(5)过于复杂很难进行解析计算，这里使用贝叶斯信息准则(BIC)方法近似计算。通过假设最大先验概率参数 $\boldsymbol{\theta}$ 服从高斯分布以简化式(5)，同时加入惩罚因子[2]，式(5)变为

$$P(\boldsymbol{X}\mid m)\approx P(\boldsymbol{X}\mid\hat{\boldsymbol{\theta}},m)P(\hat{\boldsymbol{\theta}},m)T^{-k/2} \tag{6}$$

其中，k 为参数向量 $\boldsymbol{\theta}$ 的维数；T 为训练实例数(即文档数)；$P(\boldsymbol{X}\mid\hat{\boldsymbol{\theta}},m)$ 为最佳匹配似然；因子 $T^{-k/2}$ 为惩罚项(BIC 项)，目的是偏向选择 k 较小的模型。

回到文档矩阵的主题数估计问题，如果混合矩阵 $\boldsymbol{A}$ 已知，则有主题数(即源信号个数)的最佳匹配似然[4]:

$$P(\boldsymbol{X}\mid\boldsymbol{A},m)=\int P(\boldsymbol{X}-\boldsymbol{AS})P(\boldsymbol{S})\mathrm{d}\boldsymbol{S} \tag{7}$$

根据 ICA 的极大似然估计算法，源信号的无参数分布 $P(\boldsymbol{S})$可由下式给出[5]:

$$P(\boldsymbol{S})=\frac{1}{\pi^{dc}}\exp\left(-\sum_{c,d}\ln\cosh(\boldsymbol{S}_{c,d})\right) \tag{8}$$

综合上述 2 项，以变量 m 替代 c，并对 $P(\boldsymbol{X}-\boldsymbol{AS})$近似估计，有

$$P(\boldsymbol{X}\mid\boldsymbol{A},m)=\frac{1}{\pi^{dm}}\left(\frac{1}{\|\boldsymbol{A}\|}\right)^{d}\exp\left(-\sum_{m,d}\ln\cosh(\boldsymbol{S}_{m,d})\right) \tag{9}$$

其中，$\boldsymbol{A}$ 为 $m\times m$ 阶方阵。

3 算法设计与实现

使用.NET 平台的 C#语言实现了 FastICA 算法和主题数参数估计。如前所述，奇异值分解是文档矩阵降维的重要环节。为了尽量减少开销，实现时使用矩阵类库 Mapack(http://www.aisto.com/roeder/dotnet)进行奇异值分解。该类库由纯 C#编写，算法实现参考了为 Intel 处理器优化设计的 Lapack for Java 例程，运算速度较快。ICA 算法的实现部分可参考文献[1]，这里给出主题数估计部分的算法。

在实际的 BIC 估计中，除了要计算式(9)外还应考虑矩阵 $\boldsymbol{T}$ 中未参加 ICA 分解的那部分的似然度，这通过计算 $\boldsymbol{T}\cdot\boldsymbol{L}$ 矩阵剩余部分的均方差似然获得。参数向量 $\boldsymbol{\theta}$ 的维数通过 $1+k^2+\sum_0^k(M+1-i)$ 计算，其中，M 为矩阵 $\boldsymbol{T}$ 的行数。计算时对上述 2 项似然取负对数以简化计算。在求出所有可能 m 的 $P(\boldsymbol{X}\mid m)$负对数似然后，进行归一化就可以得到 IC 数为 m 的概率分布。

begin initialize $\mathrm{M}\leftarrow\mathrm{row(X)},\mathrm{N}\leftarrow\mathrm{column(X)},\mathrm{k}\leftarrow1$

for $\mathrm{k}\leftarrow\mathrm{k}+1$

$\mathrm{X_k}\leftarrow\mathrm{submatrix(X,0,k,0,M)}$

$\mathrm{fastica(X_k)}$

$$\log\mathrm{P[k]}\leftarrow-\ln\left(\frac{1}{\pi^{\mathrm{k\cdot N}}}\left(\frac{1}{\|\mathrm{A}\|}\right)^{\mathrm{N}}\exp\left(-\sum_{\mathrm{k,N}}\ln\cosh(\mathrm{S_{k,N}})\right)\right)$$

if k<M

$$\mathrm{TL_k}=\mathrm{submatrix(T,1,M,k+1,N)}\cdot\mathrm{submatrix(L,k+1,N,k+1,N)}$$

$$\log\mathrm{P[k]}\leftarrow\log\mathrm{P[k]}+\frac{(\mathrm{M-k})*\mathrm{N}\ln(\sum\mathrm{TL_k^2}/(\mathrm{M-k})*\mathrm{N}+1)}{2}$$

endif

$$\dim\leftarrow1+\mathrm{k}^2+\sum_0^{\mathrm{k}}(\mathrm{M}+1-\mathrm{i})$$

$$\log\mathrm{P[k]}\leftarrow\log\mathrm{P[k]}+\mathrm{N}^{-\dim/2}$$

util $\mathrm{k}=\mathrm{K}$

$\mathrm{k}\leftarrow1$

for $\mathrm{k}\leftarrow\mathrm{k}+1$

$$\mathrm{P[k]}\leftarrow\exp(-\log\mathrm{P[k]}-\min(\log\mathrm{P})/\mathrm{N})$$

util $\mathrm{k}=\mathrm{K}$

normalize P[k]

return P

end

在算法中，$logP(k)$, $logP$, P 均为程序变量。完成 ICA 分解后，将 SVD 分解的特征词子空间 $\boldsymbol{TL}$ 矩阵的前 c 列与混和矩阵 $\boldsymbol{A}$ 相乘并规范化权重，一般权重的门限值可取为 0.3[4]。在 $\boldsymbol{T}\cdot\boldsymbol{A}$ 矩阵中搜索权重大于门限值的元素标记其行号，最后通过在特征词表中查找对应行标号的词即为关键词。

4 实验

反复实验表明，基于 BIC 对主题数估计算法能在有效保留分解矩阵信息的基础上反映主题数的概率分布。从文档集中分离的主题具有明显的意义，提取的关键词和人工标注结果也非常相似。这里给出易于实现和对比的 2 个实例。

4.1 主题数估计实验

实验从 SMART 系统的医学摘要数据库 MED 中抽取 5 类共计 124 篇进行测试，文本预处理参考文献[4]。实验设置估计的主题数上限为 10，估计结果如表 1 所示。

表 1 MED 文本数据主题数的概率分布

主题数	概率/(%)
1	0.0
2	0.2
3	32.7
4	67.0
5	0.1
6	0.0
7	0.0
8	0.0
9	0.0
10	0.0

需要说明的是，表 1 对数据进行了舍入，表中主题数的概率从 1~10 概率值都存在。估计结果的界面如图 1 所示。

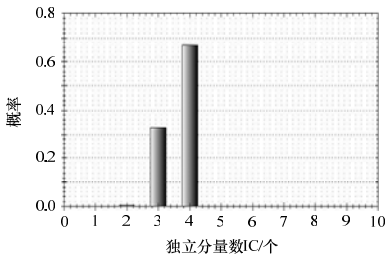

图 1　MED 文本数据的主题数估计

4.2　关键词提取实验

实验从复旦语料库的训练集中选取军事类全部 74 篇文档作为关键词抽取的目标文档集。待抽取文档首先经过中文分词，进行停用词和词性过滤等预处理。为保证有足够数量的词汇，使用简单的文档频率(DF)过滤法，取 DF 大于 2 的词作为抽取的候选词集合，同时使用 CF 规范化权重。

通过 BIC 估计，74 篇文档中共有 11 个子主题，其中有 1 个子主题在设置词权重门限值为 0.3 的情况下没有提取出关键词，其余的结果如表 2 所示。

表 2　关键词提取示例

主题	关键词
主题 1	力量 抵抗 越军
主题 2	中国 日本 海上 自卫队 舰 系统 海军 舰队
主题 3	战士 美国 政府
主题 4	部队 军区 装备 防空 日本 自卫队 旅 师 单位 射程
主题 5	方面 印度 空军 攻击 米格 能力 中国 战斗机 数量 陆军
主题 6	苏联 双方 美国 条约 问题 限制 谈判 削减 战略武器 巡航导弹
主题 7	苏联 战争 增加 庆祝 人民 胜利 卫国
主题 8	军事 美国 古巴 演习
主题 9	武装 反政府 武器
主题 10	战士 军队 建设 工作 转业 干部

从表 1 的结果看，基于 BIC 的主题数估计算法对 MED 数据集的主题数以 67%的概率选择 4，这与文献[4]的结论是一致的，而 BIC 估计使这一问题的解释定量化。另外，从表 2 的应用情况看，文档集合中 11 个子主题分别表示如中日、中越、中印、部队转业安置问题等，识别的主题意义比较明显。

5　结束语

本文采用盲信号处理的动态分量估计技术对文档集合中的主题数进行估计，在尽可能完整保留 SVD 和 ICA 分解矩阵信息的基础上对主题的关键词进行抽取。实验结果表明，基于 BIC 的主题数估计算法能提供主题数的概率分布的有效参考，抽取出的关键词能准确表达主题的内容。

参考文献

[1] Hyvärinen A. Fast and Robust Fixed-point Algorithms for Independent Component Analysis[J]. IEEE Transactions on Neural Networks, 1999, 10(3): 626-634.

[2] Hansen L K, Larsen J, Kolenda T. Blind Detection of Independent Dynamic Component[C]//Proc. of IEEE International Conference on Acoustics, Speech, and Signal Processing. [S. l.]: IEEE Press, 2001: 3197-3200.

[3] Duda R O, Peter E H, David G S. 模式分类[M]. 李宏东, 姚天翔, 译. 2 版. 北京: 机械工业出版社, 2003: 392-394.

[4] Kolenda T, Hansen L K, Sigurdsson S. Independent Components in Text[M]//Girolami M. Advances in Independent Component Analysis. [S. l.]: Springer-Verlag, 2000: 235-256.

[5] Mackay D. Maximum Likelihood and Covariant Algorithms for Independent Component Analysis[R]. Cavendish Laboratory, University of Cambridge, Technical Report Draft 3.7, 1996.

编辑　顾逸斐

（上接第 167 页）

5　结束语

本文提出一种基于改进概率的移动机器人地图创建，通过将距离信任因子引入到超声波传感器模型中，处理由镜面反射引起的测距不准确的问题。通过移动机器人的在线建图的对比实验，说明该方法有比较好的效果，使机器人创建的地图精度得到了提高。虽然本文有效执行高度依赖它的参数，R_{max} 需要根据具体环境情况预先设定，但该参数的引用为进一步研究机器人在未知环境中的高精度地图创建奠定基础。

参考文献

[1] 王　璐, 蔡自兴. 未知环境中移动机器人并发建图与定位(CML)的研究进展[J]. 机器人, 2004, 41(26): 380-384.

[2] 康叶伟, 黄亚楼, 孙凤池, 等. 一种基于 RBUKF 滤波器的 SLAM 算法[J]. 计算机工程, 2008, 34(1): 17-20.

[3] Elfes A, Moravec H P. High Resolution Maps from Wide-angle Sonar[C]//Proc. of IEEE International Conference on Robotics and Automation. [S. 1.]: IEEE Press, 1985: 116-121.

[4] Gasos J, Rosetti A. Uncertainty Representation for Mobile Robots: Perception, Modeling and Navigation in Unknown Environments[J]. Fuzzy Sets and Systems, 1999, 10(1): 1-24.

[5] Thrun S, Fox D, Burgard W. A Probabilistic Approach to Concurrent Mapping and Localization for Mobile Robots[J]. Machine Learning and Autonomous Robots, 1998, 31(1): 29-53.

[6] Smarandache F, Desert J. Advances and Applications of DSmT for Information Fusion[M]. [S. 1.]: American Research Press, 2004: 61-103.

[7] Ribo M, Pinz A. A Comparison of Three Uncertainty Calculi for Building Sonar-based Occupancy Grids[J]. Robotics and Autonomous Systems, 2001, 35(1): 201-209.

[8] Collins T, Collins J, O'Sullivan S. Evaluating Techniques for Resolving Redundant Information and Specularity in Occupancy Grids[C]//Proc. of the 18th Australian Joint Conference on Advances in Artificial Intelligence. Sydney, Australia: [s. n.], 2005: 235-244.

[9] Zou Yi, Ho Y K, Chua Chin Seng. A New Solution for Specular Reflection in the Multi-ultrasonic Sensor Fusion for Mobile Robots[C]//Proc. of IEEE International Conference on Intelligent Robotics and System. [S. 1.]: IEEE Press, 2000: 387-391.

编辑　索书志