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—0175—02 文献标识码：A 中图分类号：TP311

基于独立分量分析的隐蔽 Web 领域聚类

王晓斌，温 春，石昭祥
(电子工程学院网络工程系 602 教研室，合肥 230037)

摘 要：针对隐蔽 Web 主题领域自动识别问题，提出一种基于独立分量分析(ICA)的聚类算法。对查询页面进行页面文本抽取和预处理，利用 TF-IDF 公式计算权重并选择前 *N* 个权重最大的特征词构造文档矩阵，在使用潜在语义索引(LSI)进行特征重构的基础上通过 ICA 分解获得类别信息。利用 LSI 的词共现分析和文本降噪能力提高聚类准确率。实验表明聚类平均准确率达到 90%以上。

关键词：隐蔽 Web；潜在语义；独立分量分析；文本聚类

Hidden Web Domain Clustering Based on Independent Component Analysis

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

【Abstract】 Aiming at organizing hidden Web databases according to their topic domains, this paper proposes an Independent Component Analysis(ICA) based algorithm for hidden Web domain clustering. Text is extracted from search interface pages as common Web pages, and TF-IDF formula is applied to weight terms. After selecting the top *N*-highest weight terms to construct VSM, the algorithm performs a singular value decomposition to implement features reconstruction. It applies ICA decomposition to obtain the cluster information. The main idea is utilizing the co-occurrence analysis and noise eliminating ability of Latent Semantic Index(LSI) to improve cluster performance. Experiment shows that the average precision is higher than 90 percent.

【Key words】 hidden Web; latent semantic; Independent Component Analysis(ICA); text clustering

1 概述

根据 BrightPlanet 公司在 2001 年的统计结果，隐蔽 Web(Hidden Web)数据库包含的数据量是静态页面数据量的 500 倍以上，如何有效利用 Hidden Web 资源已引起研究者的广泛关注。Hidden Web 的查询接口(search interface)以 HTML 表单的形式存在于 Web 页面中，一般具有典型的主题领域(topic domain)特征。为了按主题领域集成 Web 数据库，聚类或分类具有相同主题的数据库显得尤为重要。

目前，对 Hidden Web 领域主题进行识别的主要方法有 pre-query 和 post-query 2 种。其中，post-query 方法通过对数据库提交查询利用返回结果进行主题识别。由于实际的接口抽取准确率和查询匹配率都不可能百分之百正确，因此实现自动查询往往需要人工介入。此外，查询结果只是数据库的部分内容，当数据库的记录具有多个领域的属性信息时，难以保证聚类结果的正确性。post-query 方法都用于接口信息不足的场合，如基于关键字的文本数据库。pre-query 方法依赖于接口表单(forms)的可视特征[1]，即查询表单中的属性标记(attribute labels)和表中其他可利用的文本信息。该方法适合数据库的内容可以完全由表单特征表示的情形，有效性往往依赖于特征文本抽取的准确度。

文献[1]提出把表单的背景文本与表单文本一起参与网页聚类以提高准确性，并将该方法命名为上下文感知的表单聚类方法(context-aware form clustering)。算法使用不同的向量表示表单文本和背景文本，利用 TF-IDF 公式计算权重并选择最有代表性的特征词，相似度通过加权表单文本向量和背景文本向量的相似度计算，聚类算法为 K-means。该方法存在以下 3 个方面的问题：(1)聚类准确度过于依赖特征词的选择；(2)网页中包含的噪声，如公告、导航条，可能严重影响聚类的效果；(3)缺乏对同义词问题的分析和处理，如“car”和“automobile”都暗示数据库领域与汽车有关，但计算机无法识别这种语义关系，而要使计算机具有识别同义词的能力，又会增加软件系统的复杂度。

针对以上问题，本文对上述背景文本辅助聚类的思想进行扩展，在独立分量分析(Independent Component Analysis, ICA)技术框架下实现了 Hidden Web 的主题聚类。算法借助潜在语义索引(Latent Semantic Index, LSI)的词共现分析[2]和降噪功能，从语义处理和文本降噪 2 个方面提高聚类准确度。实验表明，聚类准确度已达到实用标准。此外，算法还继承了 ICA 框架清晰、易于实现的特点，不需要复杂的文本抽取技术和领域知识的支持。

2 独立分量分析与文本聚类

根据向量空间模型，每个文档 X_i, $i = 1,2,\cdots,d$ 可由 t 维文档特征词权重向量 $\boldsymbol{w}_j$ 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。基于向量空间模型的 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 $d\times k$ 阶混合矩阵(mixing matrix)；$\boldsymbol{S}=(\boldsymbol{s}_1,\boldsymbol{s}_2,\cdots,\boldsymbol{s}_k)^{\mathrm{T}}$ 为 d 个文档的 k 个主题信息构成的向量，k 表示文档中独立分量(Independent Component, IC)的

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：2008-09-16 **E-mail：**wxb-77@163.com

个数。每个独立分量 s_k 定义了具有相同文档主题的一个类别，因此，文档集可分为 k 个类别。利用 $\boldsymbol{S}$ 的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 $\boldsymbol{S}$。

用于聚类的特征词表规模通常超过 1 000，因此，矩阵 $\boldsymbol{X}$ 是很稀疏的，对于 ICA 来说，这是一种病态学习问题(ill-posed learning problem)。一般先使用奇异值分解(SVD)对文档矩阵进行降维，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。通过保留前 k 个奇异值，可将 $\boldsymbol{L}\cdot\boldsymbol{D}^{\mathrm{T}}$ 矩阵截断为 $\boldsymbol{L}_{k\times k}\cdot\boldsymbol{D}^{\mathrm{T}}_{k\times d}$ 矩阵。截断矩阵保留了 k 个用于聚类的主分量，而将其余分量当作噪声去除。

对保留 k 个奇异值的 $\boldsymbol{L}\cdot\boldsymbol{D}^{\mathrm{T}}$ 矩阵进行 ICA 处理，相当于对式(2)插入 ICA 分解。分解后的矩阵表示为

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

其中，k 为独立分量 IC 的个数；$\boldsymbol{A}$ 为源信号阵 $\boldsymbol{S}$ 在 LSI 空间上的投影。由于 ICA 算法要求观察信号的数量等于源信号数量，因此保留不同 k 的 SVD 截断就确定了不同 IC 数的分解模型。一般，对保留 k 个奇异值的 $\boldsymbol{L}\cdot\boldsymbol{D}^{\mathrm{T}}$ 矩阵进行 ICA 分解也能找到 k 个 IC。

根据统计自然语言理论，文档集合中 2 个或更多的词经常一起出现(即词的共现现象)不是偶然事件，它表示词之间存在某种语义上的相关[2]，SVD 的另一个功能就是捕捉这种共现关系。例如，“car”和“automobile”可能在同一文档中交替使用，那么通过 SVD 可将它们投影到 LSI 空间的同一个坐标轴上，即 k 个分量上。共现词中的任何一个在文档中的出现都会增加文档向量在该坐标轴上的投影方差，即特征强度。

对矩阵 $\boldsymbol{L}\cdot\boldsymbol{D}^{\mathrm{T}}$ 使用 FastICA 算法[3]可实现 ICA 分解。用式(4)计算矩阵 $\boldsymbol{S}$ 的 Softmax 值[4]可得到由类别概率 $\phi_{i,j}$ 构成的 $k\times d$ 阶矩阵 $\tilde{\boldsymbol{S}}$。使用矩阵 $\tilde{\boldsymbol{S}}$ 的最大值进行 0~1 规范化，设向量 $\tilde{\boldsymbol{S}}_j$ 的最大值为 $\phi_{i,j}$，那么行号 i 即为文档所属类别。

$$\phi_{i,j}=\frac{\exp(\boldsymbol{S}_{i,j})}{\sum_{i=1}^{k}\exp(\boldsymbol{S}_{i,j})} \tag{4}$$

其中，k 表示独立分量数；$\exp(\boldsymbol{S}_{i,j})$ 表示对矩阵 $\boldsymbol{S}$ 中第 i 行的所有元素 s_{ij} 求自然对数后的累加。

3 算法设计与实现

完整的 Hidden Web 聚类处理流程如图 1 所示。处理过程分为数据准备、特征选择和独立分量分析 3 个阶段，共计 6 个步骤。

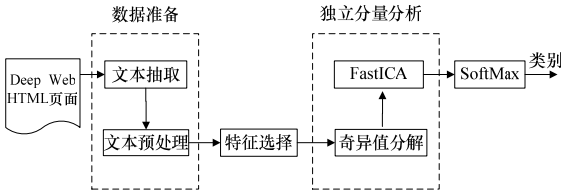

图 1　Hidden Web 接口聚类流程

(1)数据准备

步骤 1 文本抽取。对获得的 HTML 样本页面抽取内容文本。抽取过程不考虑文本降噪问题，将得到的表单的标记词和页面中其他的词构成一篇文档，输出到文本预处理模块。

步骤 2 文本预处理。主要包括词的切分和停止词过滤，英文还需要进行词干还原，中文应过滤与主题无关的虚词。对经过上述处理的文档矩阵滤除文档频率(Document Frequency, DF)小于 2 的词，然后使用 TF-IDF 公式计算词条权重。这样得到的文档矩阵具有很高的维度，为减小 SVD 分解的开销，还需要使用特征选择进行降维。

(2)特征选择

步骤 3 特征选择。使用 Top N 策略对经过 TF-IDF 规范化的词条进行筛选。调节 N 的大小，使文档包含聚类所需的信息。一般情况下，N 值越大则文档矩阵噪声越强，同时保留的聚类信息也越多。为尽量保留聚类信息，可设置较宽松的 N 值，如 N=1 000，而将降噪任务留给下一阶段处理。

(3)独立分量分析

步骤 4 奇异值分解。令 k 为预定义的聚类数，对经过特征选择的文档矩阵进行 SVD 分解，将保留 k 个奇异值的 $\boldsymbol{L}_{k\times k}\cdot\boldsymbol{D}^{\mathrm{T}}_{k\times d}$ 矩阵作为输出，传递给 ICA 算法做进一步处理。

步骤 5 FastICA。使用 FastICA 算法对 $\boldsymbol{L}\cdot\boldsymbol{D}^{\mathrm{T}}$ 矩阵进行 ICA 分解，得到由 k 个表示主题类别的独立分量构成的矩阵 $\boldsymbol{S}$。

步骤 6 Softmax。对矩阵 $\boldsymbol{S}$ 使用 Softmax 算法进行规范化并搜索矩阵 $\tilde{\boldsymbol{S}}$，取 j 列中最大值的行号 i 作为第 j 个接口页面的类别号。

4 实验

笔者在.NET平台下使用C#语言实现了FastICA算法。SVD分解使用矩阵类库Mapack(http://www.aisto.com/roeder/dotnet)实现。该类库用纯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，因此，运算速度　较快。

实验从 UIUC 数据集[5]中随机选取了 200 多个接口页面，共 5 类进行聚类测试。抽取接口页面的文本后，进行英文文本的预处理(使用 Porter 算法进行词干还原)，然后利用 TF-IDF 公式计算权重。

对权重规范化后的文档矩阵进行 N=1 000 的 Top N 特征选择，重组文档矩阵。使用 SVD 分解该矩阵，保留最大的前 5 个奇异值，将奇异值与对应文档分量的矩阵乘积进行 ICA 处理可以获得独立分量矩阵 $\boldsymbol{S}$。$\boldsymbol{S}$ 表示为图 2 所示的灰度直方图(histogram)，其中，Y 轴表示独立分量 IC，X 轴表示文档序列，以不同灰度表示文档在该分量上的值。图中与 Y 轴平行的 4 条黑色线段为文档类别的分界线(Y 轴和 X 轴为序号，没有单位)。

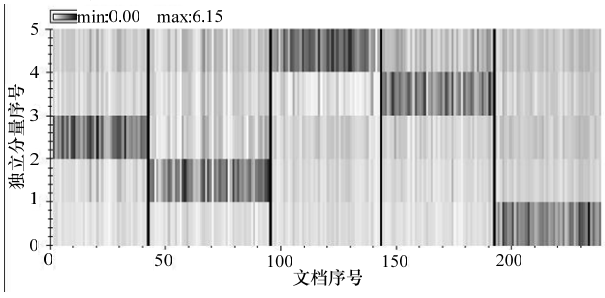

图 2　分离后的独立分量矩阵灰度图

从图 2 中可以看出，经过 ICA 处理的文档在 5 个 IC 分量上表现出明显的区域特征(灰度较大的区域聚集在一个区

（下转第 179 页）