第35卷 第17期 计 算 机 工 程 2009年9月
Vol.35 No.17 Computer Engineering September 2009
·多媒体技术及应用· 文章编号：1000—3428(2009)17—0219—04 文献标识码：A 中图分类号：TP311.52

基于MWH模型的新闻视频语义挖掘

徐新文[1]，李国辉[1]，付畅俭[1,2]
(1. 国防科技大学信息系统与管理学院，长沙 410073；2. 湘潭大学商学院电子商务系，湘潭 411105)

摘 要：根据新闻视频伴随文本及关键帧图像多模异构特征，构建多翼Harmoniums模型，该模型包括多元文本泊松分布和多元颜色直方图高斯分布2个子模型。通过研究隐含主题与观测输入之间的双向依存关系，将其扩展为双层随机场模型，从而对新闻视频进行语义挖掘。在CCTV新闻视频集上进行测试，实验结果验证了该模型的有效性。

关键词：多翼Harmoniums模型；语义挖掘；新闻视频

News Video Semantic Mining Based on Multi-Wing Harmoniums Model

XU Xin-wen[1], LI Guo-hui[1], FU Chang-jian[1,2]
(1. School of Information System & Management, National University of Defense Technology, Changsha 410073;
2. Dept. of E-commerce, Business School, Xiangtan University, Xiangtan 411105)

【Abstract】According to the heterogeneous features of news video accompany text and multimode key frame image, the Multi-Wing Harmoniums(MWH) model is set up, which contains two submodels of multivariate text Poisson distribution and multivariate color histogram Gaussian distribution. By researching the doubleaction dependent relation of implication theme and observation input, this model is extended to two layer random field model, which conducts the semantic mining to news video. It is tested in CCTV news video sets, and the results show this model is effective.

【Key words】Multi-Wing Harmoniums(MWH) model; semantic mining; news video

1 概述

随着处理器速度和网络技术的发展以及海量存储介质的出现，人们对诸如文本、图像、音频和视频等多媒体数据进行建模和挖掘的需求也日益增加。尤其是作为大众媒体的新闻视频，由于其易于获取、信息丰富等特点，因此受到众多研究者的广泛关注。为完成对新闻视频进行有效分类、聚类、检索和关联等数据挖掘任务，必须充分利用其多模式异构特征所提供的丰富信息。因此，对视频各关联数据进行共同建模，探索合适的原始高维特征的低维潜在主题表示成为视频语义挖掘研究一个核心问题。将视频语义挖掘任务和多模式数据(如关键帧图像、音频、伴随文本)融合在一起相互补充，从而获得更好的性能。该思想已广泛应用于许多现有的视频处理方法中。融合策略主要有早期的特征层的融合以及后来的决策级融合。关于哪种融合策略更适合某一任务是个开放性问题，文献[1]对在视频分类中这2种策略进行比较。而本文提出另一种方法，通过对多模式数据低层特征联合建模，生成视频数据潜在语义表示，并用于视频语义挖掘。

2 相关问题

有许多获取视频数据的低维中间层语义表示的方法。主成分分析(Principal Component Analysis, PCA)是最流行的方法之一，它将原始特征映射到一个低维特征空间，同时能保留数据间的差异。独立成分分析(Independent Component Analysis, ICA)和费舍尔线性判别式(Fisher Linear Discriminant, FLD)也是被广泛应用的降维方法。最近出现了对文本和多媒体数据进行潜在语义主题建模的研究。潜在语义索引(Latent Semantic Indexing, LSI)是种根据词条的共现信息探查词条之间内在的语义关联的方法[2]。通过对文档矩阵进行奇异值分解，将矩阵近似地映射到一个低维潜在语义空间，奇异值向量最大限度反映出词条和文档之间的依存关系。经过这样的映射之后，原来不包含或包含很少相同词条信息的文档之间也可能因为词条的共现关系而有较大的相似度。尽管LSI能粗略地获取潜在语义，在自动索引应用中取得较好的效果，但由于其不满足统计学原理，因此存在过拟合现象。后来该思想被扩展为概率潜在语义索引(probabilistic LSI, pLSI)，它将潜在主题建模成词条上的条件概率分布，同样文档也建模成潜在主题上的条件概率分布。由于pLSI是基于概率原理并定义了适当的生成模型，因此它能应用于模型组合，并能进行复杂度控制。潜在狄利克雷分配(Latent Dirichlet Allocation, LDA)是离散数据集(如文本集)的一个生成概率模型。LDA是个层次贝叶斯模型，其中，文档建模成潜在主题集上的一个有限混合，每个主题依次建模成词条概率集上的一个有限混合。在文本建模中，潜在主题概率提供一个清楚的文档表示。LDA后来扩展为高斯混合LDA(Gaussian-Mixture LDA, GM-LDA)和通信LDA(correspondence LDA)[3]，两者都

基金项目：国家教育部博士点基金资助项目(20069998022)；湖南省教育厅科学研究基金资助项目(07C778)

作者简介：徐新文(1974－)，男，博士研究生，主研方向：多媒体信息系统，多媒体数据挖掘；李国辉，教授、博士生导师；付畅俭，副教授、博士

收稿日期：2009-04-15 **E-mail：**xinwen_xu@126.com

可用于对注释数据建模，如带标题的图像或伴随文本的视频。

实际上，上述方法主要是用于将高维的原始特征转换成低维表示，大略地获取数据的潜在语义。但它们主要应用于单模式数据上，不能或不易扩展到多模式异构数据上来。本文将根据新闻视频的多模式异构数据(关键帧图像的颜色直方图和伴随文本关键词)，并基于 Harmony 理论[4]，提出一种多翼 Harmoniums 模型，以进行语义挖掘。

3 潜在语义模型

3.1 基本 Harmoniums 模型

1986 年，Smolensky 在其 Harmony 理论中提出并研究了 Harmoniums 模型，定义一个完整的包含 2 层节点的双边无向图模型，如图 1 所示。其中，$H=\{H_j\}$表示隐含单元集；$X=\{X_i\}$表示输入单元集。

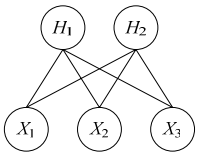

图 1 Harmoniums 模型

Harmoniums 构建如下一个随机场：

$$p(x,h\mid\theta)=\frac{1}{Z(\theta)}\exp\{\sum\theta_i\phi_i(x_i)+\sum\theta_j\phi_j(h_j)+\sum\theta_{ij}\phi_{ij}(x_i,h_j)\} \quad (1)$$

其中，$\phi_e(\cdot)$为模型中定义在一对相连单元上的势函数；θ_e为权重；$Z(\theta)$为配分函数。

Harmoniums 的双向拓扑结构表明如果一层中所有节点都给定，则另一层中的节点是条件独立的。这使基于 2 层间的条件分布函数 $p(x\mid h)$ 和 $p(h\mid x)$ 可以方便地定义 Harmoniums 分布，其中，$p(x\mid h)=\prod_i p(x_i\mid h)$，$p(x\mid h)=\prod_j p(h_j\mid x)$。因此，它在语义上是直接并容易设计的。为了简化，这里考虑的条件概率都采用指数形式：

$$p(x_i\mid h)=\frac{1}{A_i(\{\hat\theta_{ia}\})}\exp\{\sum_a\hat\theta_{ia}f_{ia}(x_i)\} \quad (2)$$

$$p(h_i\mid x)=\frac{1}{B_j(\{\hat\lambda_{jb}\})}\exp\{\sum_a\hat\lambda_{jb}g_{jb}(h_j)\} \quad (3)$$

其中，$\{f_{ia}(\cdot)\}$和$\{g_{jb}(\cdot)\}$分别表示变量 x_i 和 h_j 的充分统计；$A_i(\cdot)$和 $B_j(\cdot)$为配分函数；转换参数 $\hat\theta_{ia}$ 和 $\hat\lambda_{jb}$ 分别定义为 $\hat\theta_{ia}=\theta_{ia}+\sum_{jb}W_{ia}^{jb}g_{jb}(h_j)$ 和 $\hat\lambda_{jb}=\lambda_{jb}+\sum_{ia}W_{ia}^{jb}f_{ia}(x_i)$，由输入和隐含层间所有匹配对产生。局部条件概率正好映射到 Harmoniums 随机场：

$$p(x\mid h)\propto\exp\{\sum_{ia}\theta_{ia}f_{ia}(x_i)+\sum_{jb}\lambda_{jb}g_{jb}(h_j)+\sum_{ijab}W_{ia}^{jb}f_{ia}(x_i)g_{jb}(h_j)\} \quad (4)$$

其中，θ_{ia}，λ_{jb} 和 W_{ia}^{jb} 为与它们对应势函数相关联的参数集。由于存在联合概率的配分函数，因此参数估计很困难，在式中没有用精确的等于符号，而采用比率符号。该模型称为指数簇 Harmoniums(Exponential Family Harmoniums, EFH)。利用这种自下而上的策略从易于理解的局部条件概率来构造特定的 Harmoniums。

在 Harmoniums 模型中，无论是输入变量还是隐含变量都不存在边界独立性。然而，EFH 在隐含变量间拥有条件独立的优点，而它通常在有向图模型中是没有的。这个特性在很大程度上减少了推理复杂度。但由于存在全局的配分函数，因此在参数学习时更加困难。

3.2 新闻视频的多翼 Harmoniums 模型

Harmoniums 模型中的隐含单元和输入单元是对称的，似乎不能解释它们在语义上的因果关系，但如上所述的基于 2 个单元层间局部条件定义，能够提供一种对 Harmoniums 结构双向因果关系的解释手段。从本质上讲，隐含单元 H 可以看作潜在主题，它定义输入，相反地，也可以将 H 视为由输入单元上的一个判别模型所产生的预测器。

在许多多媒体应用中，模型的输入不仅来自单一模式的同构数据，而经常是多模式的异构数据。例如，在典型的多媒体应用中输入包含多种相关的信息，如伴随文本、图像、声音和运动向量等。假设所有输入结合在一起，表现同一中心主题，那么很自然就想到用一个隐含单元集对共享的中心主题进行建模，并将来自各不同类型的观测数据分组成多个输入单元同构排列，每个排列对应一种数据源，这样就构建成一个多翼 Harmoniums 模型，如图 2 所示，它由多个标准的 Harmoniums 模型通过一个共享的隐含单元集组成。从一个标准 Harmoniums 构建多翼 Harmoniums 是简单直接的。

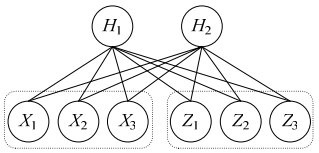

图 2 多翼 Harmoniums 模型

本文基于文本子模型和图像子模型，并采用上述方法对包含有文本和图像信息等多模信息的新闻视频数据进行建模，该模型称为新闻视频多翼 Harmoniums(News Video Multi-Wing Harmoniums, NVMWH)模型。根据传统的文本词袋模型，用文档的词条计数对视频伴随文本建模。假设与每个文档关联的潜在主题特征直接确定一个文档中的每个词条期望比率，为文档中的每个词条的观测计数指定一个泊松分布。该文本模型与多项式模型的关键差异是主题混合直接在由特定主题特征组合的文档所确定的词条比率分布中完成，所以，即使当某词条在文档中仅出现一次或很少几次，主题混合仍是稳定和鲁棒的，而这在新闻视频中是典型的。文本模型如下：对每个词条 $i=\{1,2,\cdots,M\}$，它的比率 x_i 的分布为

$$p(x_i\mid h)=\text{Poisson}(x_i\mid\exp(\alpha_i+\sum_j h_jW_{ij})) \quad (5)$$

其中，泊松率 α_i 的变化由潜在主题特征 h 的加权组合确定。

对于图像输入，本文采用图像的颜色直方图，其模型为典型的高斯分布：

$$p(z_k\mid h)=N(z_k\mid\sigma_k^2(\beta_k+\sum_j h_jU_{kj}),\sigma_k^2) \quad (6)$$

其中，z_k 表示图像直方图的第 k 个柄的值；均值的变化也是由潜在主题特征 h 的加权组合确定。为保证图像输入的条件概率与隐含单元的一致性，采用 σ_k^2 规范 $p(z_k\mid h)$的均值。

最后，对于输入数据的潜在主题特征的隐含单元 h，假定每个特征均为一条件单位方差的高斯分布，均值由观测数据词条计数和颜色直方图所确定。

$$p(h_j\mid x,z)=N(h_j\mid\sum_i x_iW_{ij}+\sum_k z_kU_{kj},1) \quad (7)$$

将以上 3 个模型结合，并对 h_j 积分，得到输入单元的边界分布如下：

$$p(x,z)=\exp\{\sum_i\alpha_ix_i-\sum_i\lg\Gamma(x_i)+\beta_kz_k-\frac{1}{2}\sum_k\frac{z_k^2}{\sigma_k^2}+\frac{1}{2}\sum_j(\sum_i W_{ij}x_i+\sum_k U_{kj}z_k)^2\} \quad (8)$$

将输入的隐含变量的方差定义为 1，以简化参数评估。在 $p(h_j \mid x_i, z_k)$ 中引入协方差矩阵 Σ 能为联合概率提供更大的自由度，但根据边界概率，它并不会导致更一般的表示。

4 基于变分法的模型学习

对一个给定的独立同分布样本 $\chi = \{x_n, z_n\}_{n=1}^{N}$，可以通过梯度上升法在最大或然目标下评估 Harmoniums 的参数。学习规则(如梯度)可以通过对样本的对数或然率进行求导来获取，其模型参数为

$$\delta\alpha_i = \langle x_i \rangle_{\tilde{p}} - \langle x_i \rangle_p, \delta\beta_i = \langle z_k \rangle_{\tilde{p}} - \langle z_k \rangle_p$$

$$\delta(\sigma_k^{-1}) = \left\langle z_k^2 \sigma_k^{-1} \right\rangle_{\tilde{p}} - \left\langle z_k^2 \sigma_k^{-1} \right\rangle_p$$

$$\delta W_{ij} = \left\langle x_i h_j^{'} \right\rangle_{\tilde{p}} - \left\langle x_i h_j^{'} \right\rangle_p, \delta U_{ij} = \left\langle z_k h_j^{'} \right\rangle_{\tilde{p}} - \left\langle z_k h_j^{'} \right\rangle_p$$

其中，$\langle \cdot \rangle_p$ 表示关于分布 p 的数学期望；$\tilde{p}(x) = \sum_n \delta(x - x_n)/N$ 表示经验分布；$p(\cdot)$ 代表模型分布；$h_j^{'}$ 为 $\sum_i W_{ij} x_i + \sum_k U_{kj} z_k$。由于在 p 中存在有配分函数 Z，因此第 2 个数学期望值的计算是很难的。下面介绍变分近似梯度上升学习算法。

对模型分布 p 采用变分近似，采用广义平均场(Generalized Mean Field, GMF)近似处理 Harmoniums 随机场，把分解因子形式作为所有变量的单一边界乘积。实际上，这是个最简单的 GMF，属于标准的平均场模式，但要升级成更好的 GMF 近似是容易的：

$$q(x,z,h) = \prod_i q(x_i \mid v_i) \prod_k q(z_k \mid u_k, \sigma_k) \prod_j q(h_j \mid \gamma_j) \tag{9}$$

其中，$q(x_i \mid v_i)$ 是均值为 v_i 的泊松分布；$q(z_k \mid u_k, \sigma_k)$ 是均值为 u_k 方差为 σ_k 的高斯分布；$q(h_j \mid \gamma_j)$ 是均值为 γ_j 方差为 1 的高斯分布。根据 GMF 法则，可以得到以下的 GMF 近似等式：

$$\begin{aligned} \gamma_j &= \sum_i W_{ij} v_i + \sum_k U_{kj} \mu_k \\ \mu_k &= \sigma_k^2 (\beta_k + \sum_j U_{kj} \gamma_j) \\ v_i &= \exp(\alpha_i + \sum_j W_{ij} \gamma_j) \end{aligned} \tag{10}$$

它最小化了 q 和 p 间的 KL 差分。用 q 代替原始的 Harmoniums 随机场 p，并计算期望梯度进行参数更新。该方案比其他方法更有效，但通常会导致结果精度的降低。尽管变量间的依赖性在式(9)中完全分离，但是 GMF 能提供基于边界的紧密近似。

5 实验

本文采用 CCTV 新闻视频集作为实验数据。视频段被分割成多个镜头，一个镜头被视为一个文档或者一个训练测试例子。

实验共采用 1 256 个视频镜头，分别属于 5 个主题：火灾，洪水，禽流感，朝鲜问题和奥运，每个镜头关联一个类别。其中，主题 1 包括：森林火灾，扑火，火点，大火，消防，扑救，控制，原因。主题 2 包括：大雨，洪水，紧急状态，淹没，气象预报，警戒，灾情，冲毁。主题 3 包括：禽流感，防控，疫苗，农业部，免疫，病毒，家禽，动物。主题 4 包括：朝鲜，美国，六方，会谈，希尔，核设施，朝核，中国。主题 5 包括：奥运，安保，场馆，公安，检测，设施，警戒，危险物品。如图 3~图 7 所示。

(a)场景 1 (b)场景 2 (c)场景 3 (d)场景 4 (e)场景 5

图 3 火灾主题场景

(a)场景 1 (b)场景 2 (c)场景 3 (d)场景 4 (e)场景 5

图 4 洪水主题场景

(a)场景 1 (b)场景 2 (c)场景 3 (d)场景 4 (e)场景 5

图 5 禽流感主题场景

(a)场景 1 (b)场景 2 (c)场景 3 (d)场景 4 (e)场景 5

图 6 朝核主题场景

(a)场景 1 (b)场景 2 (c)场景 3 (d)场景 4 (e)场景 5

图 7 奥运主题场景

对每个镜头，从伴随文本中提取 1 948 个二进制特征和从关键帧中提取 166 维 HSV 颜色空间中的颜色直方图。为平衡 2 种特征分布，对图像特征进行线性规范。

5.1 潜在主题挖掘

NVMWH 模型能自动地从文本和图像中发现有意义的潜在主题，图 3~图 7 列举了通过 NVMWH 学习得到的 18 个主题中的 5 个。

每个主题用与视频镜头相关联的前 8 个关键词和前 5 个关键帧图像进行描述，在潜在主题中它们具有最高的条件概率。前 4 个主题很明显分别与森林火灾场景、洪水场景、禽流感场景和朝核问题场景相对应，它们是基于伴随文本和图像的一个聚类。最后 1 个主题显示 NVMWH 发现的一些特殊模式，这些镜头表现不同主题，因为包含不同的场景，有直升飞机、体育场馆和会议等场景，所以通过查看这些镜头的伴随文本，发现它们语义主题上存在共同的方面。它们都提到了相似的或相同的词语，如“奥运会”、“安保”和“场馆”等。因此，NVMWH 在发现最后这个主题时主要是基于视频伴随文本关键词的相似性。

5.2 分类实验

本文采用分类实验来说明 NVMWH 模型所产生的低维潜在语义特征的预测能力。设置潜在主题维数低于 50 维，原始特征的维数是近 2 000 维文本特征和 166 维图像直方图特征，并评估 NVMWH 模型分类的性能。

为进行比较，本文也执行了 3 个相关的模型：LSI, GM-Mix 和 GM-LDA。GM-Mix 和 GM-LDA 的参数通过 EM 训练得到。用隐含变量的条件概率 $p(h|x,z)$ 获取模型 GM-Mix 的潜在主题，基于主题权重的变分狄利克雷后验概率获取模型 GM-LDA 的潜在主题，详细内容请参阅文献[3]。

对每个算法，用整个数据集进行参数估计，忽略它们的真实类标签。模型一旦学习完成，将每个样本映射到一个低维的潜在主题空间。然后，将数据均匀地分成训练集和测试集，用 SVM 软件包在训练数据上学习支持向量机，并预测测试数据。

图 8 为各种不同模型在不同潜在主题特征维(5 维~50 维)上的分类准确率，可以看出，维数为在 NVMWH 中的隐含单元数。在相同主题维情况下，NVMWH 优于 LSI。这可以证明通过采用 NVMWH 对文本和图像进行建模是很好的设想。GM-Mix 分类性能比较差，这是因为 GM-Mix 建模中受到太多的限制而不能对每个文本图像对获取多个潜在主题。由于后验分布通常在一个潜在主题处到达最大波峰，因此在 GM-Mix 表示中有太多的信息被去除。

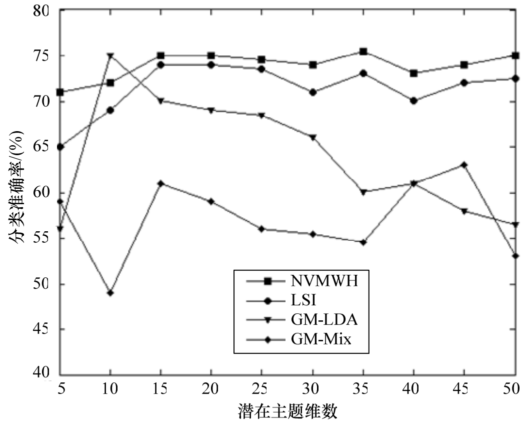

图 8　各模型在不同维主题上的分类准确率

与 GM-Mix 相比，GM-LDA 在文本图像联合建模时提供了更大的灵活性，实际上，它在潜在主题维为 10 维时，还要稍微优于其他方法。但它的准确率曲线随着潜在空间的维度增加而下降很快，可能存在过拟合和低维偏好问题。同时，LSI 和 NVMWH 的性能在整个维数变化范围内相对稳定，这可以反映出它们对潜在主题的表示具有较强的表示能力和鲁棒性。

6　结束语

本文提出一种多翼无向图模型 Harmoniums，并对新闻视频故事进行建模。在该模型中，多元高斯变量表示主题，特定特征的条件分布建模各种类型数据源输入，即用一个多元泊松分布建模文本特征，用多元高斯模型表示关键帧颜色直方图特征，这些概率分布由所有主题共同确定，因此，较好地完成了主题混合。尽管该模型是无向的，但它能从一个双向的生成过程中定义直观的因果关系的潜在语义。同时，模型的概率结构保证了潜在主题的有效推理，主题表示和主题混合机制不同于其他模型，因此，与其他模型相比，能提供不同的用途。

参考文献

[1] Snoek M. Early Versus Late Fusion in Semantic Video Analysis[C]//Proc. of the 13th Annual ACM Int’l Conf. on Multimedia. New York, USA: ACM Press, 2005.

[2] Deerwester S C. Indexing by Latent Semantic Analysis[J]. Journal of the American Society of Information Science, 1990, 41(6): 391-407.

[3] Jordan M. Modeling Annotated Data[C]//Proc. of the 26th Annual Int’l Conf. on Research and Development in Informaion Retrieval. New York, USA: ACM Press, 2003.

[4] Smolensky P. Information Processing in Dynamical System: Foundations of Harmony Theory[M]. Cambridge, USA: MIT Press, 1986.

编辑　陈　文

~~~~~~

（上接第 218 页）

由图 2(a)~图 2(d)可知，本算法能实时检测出移动目标。由图 2(e)、图 2(f)可知，本算法能检测出移动目标(但无法处理移动目标的阴影)，且对移动目标的大小变化具有鲁棒性。

4　结束语

本文算法使用同一个像素的相似度，区分前景和背景像素。采用窗口平均灰度值替代单个像素的相似度，以减少噪声影响。它避免了对背景物体微小运动的误分割，对光线的微小变化具有鲁棒性。因为对背景图像中的所有像素进行实时更新，所以该算法能适应背景光线的突变和缓慢变化等情况。但它不能消除运动目标的阴影，无法自动检测运动目标停止在背景图像中的情况，此类问题有待进一步研究并解决。

参考文献

[1] Paragios N, Deriche R. Geodesic Active Contours and Level Sets for the Detection and Tracking of Moving Objects[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2000, 22(3): 266-280.

[2] 韩鸿哲，王志良．基于自适应背景模型的实时人体检测[J]．北京科技大学学报, 2003, 25(4): 384-386.

[3] Fejes S, Davis L S. Detection of Independent Motion Using Directional Motion Estimation[R]. University of Maryland, Technical Report: CAR-TR-866, CS 3815, 1997.

[4] 林洪文，涂　丹．基于统计背景模型的运动目标检测方法[J]．计算机工程, 2003, 29(9): 97-99.

[5] Kim J B, Kim H J. Efficient Region-based Motion Segmentation for a Video Monitoring System[J]. Pattern Recognition Letters, 2003, 24(1-3): 113-128.

[6] Maadi A E, Maldague X. Outdoor Infrared Video Surveillance: A Novel Dynamic Technique for the Subtraction of a Changing Background of IR Images[J]. Infrared Physics & Technology, 2007, 49(3): 261-265.

编辑　陈　晖
~~~~~~