# PAM概率主题模型研究综述

余淼淼[1] 王俊丽[1] 赵晓东[1] 岳晓冬[2]
(同济大学企业数字化技术教育部工程研究中心 上海 201804)[1]
(上海大学计算机工程与科学学院 上海 200444)[2]

**摘 要** 近年来,主题模型逐渐成为计算机科学领域的一个研究热点,在自然语言处理、文本分类以及信息检索等方面都有很广泛的应用。介绍了概率主题模型的发展后,主要针对PAM模型及其改进的层次PAM和非参PAM进行了分析和比较,层次PAM可以更好地表达主题层次结构;非参PAM则是给定一个基于HDP的非参贝叶斯先验,对复杂结构的模型有更强的表现力。最后对PAM相关主题模型的理论及应用进行了总结,并对未来发展趋势进行了探讨。

**关键词** 主题模型,PAM,文档生成,统计推断
**中图法分类号** TP181 **文献标识码** A

## Research on PAM Probability Topic Model

YU Miao-miao[1] WANG Jun-li[1] ZHAO Xiao-dong[1] YUE Xiao-dong[2]
(Engineering Research Center of Enterprise Digitalization Technology, Tongji University, Shanghai 201804, China)[1]
(School of Computer Engineering and Science, Shanghai University, Shanghai 2000444, China)[2]

**Abstract** Recently, topic model is emerging as a new hotspot of research in computer science, and has a wide range of applications in natural language processing, document classification, information retrieval and so on. The paper mainly analyzed the Pachinko Allocation Model and its improved models. The improved hierarchical PAM is effective at discovering mixtures of topic hierarchies. Nonparametric PAM has more expressive force for complex structures. It has a nonparametric Bayesian prior based on a variant of the hierarchical Dirichlet process. The theory and applications of PAM and its related topic models were summarized, and finally the future directions were discussed.

**Keywords** Topic model, PAM, Document generation, Statistical inference

## 1 引言

随着互联网技术的发展,人们面临着海量、快速增长的数据资源,如何有效利用和挖掘各种信息资源,获取潜在的、有价值的知识成为当前的迫切需求,这促进了概率主题模型的迅速发展及其在网络信息资源自动获取和分析中的广泛应用。

主题模型将一篇文档理解成是由若干隐含主题组合形成的,而这些主题由文本中的特定词汇体现。因此可将隐含主题看作是词的一种概率分布,单个文档则表示为这些隐含主题特定比例的随机混合。给定一个文档集合,主题模型通过参数估计寻找一个低维的多项式分布集合,每个多项式分布称为一个主题,用来捕获词之间的相关信息[1]。因而,主题模型可以在不需要计算机真正理解自然语言的情况下提取可以被人理解的、相对稳定的隐含语义结构,为大规模数据集中的文档寻找一个相对短的描述。

主题模型的起源是隐性语义索引LSI(Latent Semantic Indexing)[2],但LSI并不是主题模型。其工作原理是利用矩阵理论中的奇异值分解技术,将词频矩阵转化为奇异矩阵,通过去除较小的奇异值向量,只保留前$K$个最大的值,将文档向量和查询向量从词空间映射到一个$K$维的语义空间(主题)。这一思想为主题模型的发展奠定了基础。

在LSI的基础上,Hofmann提出了概率隐性语义索引PLSI(Probabilistic Latent Semantic Indexing)[3],该模型被看成是第一个真正意义上的主题模型。通过概率模型来模拟文档中词的产生过程,将文档$d$表示为一个主题混合,文档中每个词作为主题混合中的一个抽样。但PLSI并没有用一个概率模型来模拟文档的产生,只是通过对训练集中的有限文档进行拟合,得到特定文档的主题混合比例。这个过程导致PLSI模型参数随着训练集中文档数目线性增加,出现过度拟合现象,而且对于训练集以外的文档很难分配合适的概率。

针对这些问题,Blei等人在2003年提出LDA(Latent Dirichlet Allocation)[4],它是一个更为完全的概率生成模型,应用非常广泛。它将每个文档表示成主题的混合,而每个主题是单词上的多项式分布。用一个服从狄利克雷分布的$K$维隐含随机变量表示文档的主题混合比例,模拟文档的生成

到稿日期:2012-10-17 返修日期:2013-01-22 本文受国家自然科学基金(61105047),上海市科委(11dz1120702),同济大学中央高校基本科研业务费专项资金,国家科技支撑计划课题(2012BAF10B12)资助。

余淼淼(1989—),女,硕士生,主要研究方向为概率主题模型,E-mail:yumiaomiao_2010@126.com;王俊丽(1978—),女,博士,副研究员,主要研究方向为语义信息获取;赵晓东(1968—),男,高级工程师,主要研究方向为模型可视化;岳晓冬 讲师,主要研究方向为数据挖掘和多媒体。

过程。首先从狄利克雷分布中抽样产生一个文本特定的主题多项式分布，然后对这些主题反复抽样产生文本中的每个词。之后出现了许多基于 LDA 扩展的概率模型[5-9]，LDA 被直接或扩展使用在自然语言处理的众多任务中[8-15]。

LDA 模型中发现的主题可以捕获词之间的相关性，但基于 Dirichlet 分布的抽样假设主题之间相互独立，因此不能够获取主题之间的关系。然而主题的相关性在真实的数据集合中是普遍存在的，忽略这些相关性将限制 LDA 模型对大规模数据集合的表示能力以及对新数据的预测能力[1]。因此，很多学者开始研究更丰富的结构来描述主题之间的相关性。

Blei 在 2006 年提出了 CTM（Correlated Topic Model）[16]，它与 LDA 类似，将每个文档表示成一个主题混合。LDA 模型中主题混合比例是从狄利克雷分布中抽样获得，而 CTM 中主题混合比例是从对数正态分布中抽样获得的。其先验参数包括一个协方差矩阵，用每个主题对之间的协方差描述它们之间的相关性。CTM 只能描述成对主题间的相关性，而且协方差矩阵中参数的数量和主题数目的平方成正比。

基于 CTM 只能描述两两之间相关性的局限性，Li 等人进一步提出了 PAM（Pachinko Allocation Model）[17]，其用一个有向无环图（DAG）表示语义结构，不仅可以描述词之间的相关性，而且可以灵活地描述主题之间的相关性，较 LDA 和 CTM 具有更强的文本表示能力。

同时，近年来随着各种主题模型的不断发展，其在自然语言处理、文本分类以及信息检索等方面都有很广泛的应用。在视频监控数据处理、静态图像内容理解以及统计认知的研究应用等领域也得到了较好的应用[18]。

本文第 2 节介绍概率主题模型的主要内容；第 3 节介绍 PAM 模型的结构以及主要原理；第 4 节介绍在 PAM 模型基础上的改进模型以及实验结果对比；第 5 节介绍目前 PAM 模型的几个应用实例；最后讨论 PAM 改进模型存在的问题及未来进一步的研究方向。

## 2 概率主题模型

当面对一个大规模的文本数据集或者是其他类型的离散数据集合时，为了便于理解，我们总是希望找到这个数据集的一个简短描述和概括来代表或体现出整个数据集的特征信息。对文本数据来说，就是抽取出一个或几个主题概念来描述整个文本数据集。主题模型就是用来发现这种隐含的概括性语义结构，得到主题概念。概率主题模型的主要思想认为文档是若干主题的混合分布，而每个主题又是一个关于单词的概率分布，如图 1 所示。

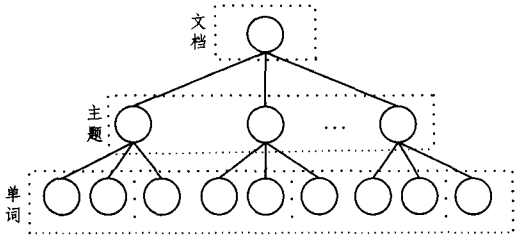


图 1 主题模型思想

### 2.1 概率图模型

图模型是主题模型的一种主要表示方法，它使用图形表达基于概率的相关关系，是概率论和图论的有机结合。图模型结构使得大量基于网络化依赖关系的计算和概率框架得以清晰表达。

概率图模型使用图来表示和操作联合概率分布，通过图论和概率论的结合形成一个多变量统计模拟的形式体系。一个图模型就是一个定义于有向或无向图中的概率分布族。整个模型包括结构元素和参数元素，结构元素就是由边的集合构成的图形模式，参数元素即是由参数化的节点或节点子集构成的参数集合。

关于图模型中的基本表示和基本定理，图 2 中圆圈节点表示随机变量，被填充了的实心圆圈表示可观测变量，空心圆圈表示待估计变量；边则表示概率依赖关系，箭头表示条件概率；图的结构定义了随机变量的条件独立关系。图 2(a)表示的联合概率为 $P(a,b,c)=P(c|a,b)P(b|a)P(a)$。对于更复杂的模型，在图中画出每一个随机变量显然不现实，因而用方框来表示这种重复性结构，图 2(b)就表示了 $y$ 和 $N$ 个 $x$ 的联合概率 $P(y,x_1,\cdots,x_n)=P(y)\prod_{n=1}^{N}P(x_n|y)$。

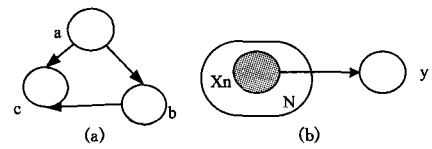


图 2 联合概率分布图模型

### 2.2 文档生成与统计推断

主题模型是文档的一种生成模型，可以根据主题模型所指定的一组概率程序来创建出一个新的文档。首先选择一个文档的主题概率分布，根据这个概率分布，每次随机地从中选出一个主题，再根据这个主题在单词上的概率分布，生成这个文档的一个个单词，这样就可以产生一个新的文档（尽管里面的词可能不具有组成句子或更深层次语义的逻辑顺序）。

该生成模型的逆向操作则可以得到主题信息，即已经有了一些文档的集合，需要反过来推断这个文档集合具体的主题概率分布以及每个主题在词上的概率分布，如图 3 所示。参数估计过程中最重要的两组参数是各主题下的词项概率分布以及各文档的主题概率分布。参数估计可以看成是文档生成过程的逆过程：在已知文档集的情况下，通过参数估计得到参数值，即整个训练过程的输出结果。针对参数估计，我们需要选择最优化的目标函数，即在主题模型中通常是整个语料的概率值 $P(D|\alpha,\beta)$，通过对目标函数进行最大化来估计参数值。

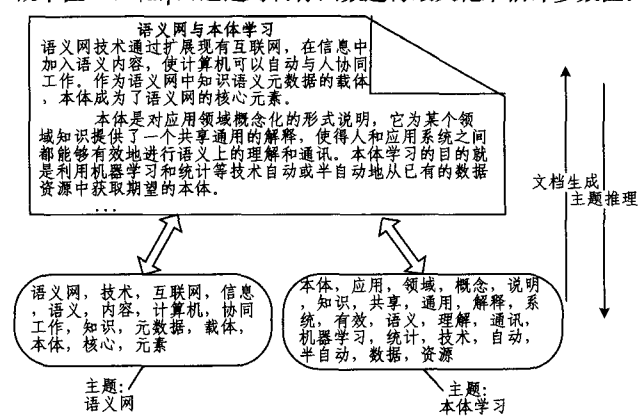


图 3 生成模型

### 2.3 主题推理与模型输入

主题模型的主要输入[18]是文档集合和主题个数 $K$。对于文档集合，多采取词袋（bag of words）的办法，即将每篇文

档视为一个词频向量，其多为词项文档(term-document)矩阵的形式，从而将文本信息转化为了易于建模的数字信息。对于主题个数 $K$ 的大小，可以使用不同的 $K$ 重复实验，当评价指标如困惑度(perplexity)、语料似然值、分类正确率等最优时，认为此时的 $K$ 是模型的最佳选择[19]；或者采用非参数贝叶斯的方法，假设主题个数为无穷多，实际主题个数可以随着语料的规模而变化，训练结束时的主题个数即为 $K$ 的最佳选择[16]。

概率主题模型是非监督机器学习模型，提取的主题具有可解释性，具有清晰的层次结构。LDA 是目前应用较为广泛的一种概率主题模型，它具有全面的文本生成假设。但是，LDA 模型假设主题间是相互独立的，因此不能够获取主题之间的相互关系，限制了其对大规模数据集的表现能力。接下来重点介绍的 PAM 模型则可以灵活表现主题间相关性，具有更好的文本表现力。

## 3 PAM 模型

PAM 模型由 Li 等人在 2006 年提出[17]，该模型是根据日本的一个游戏——“弹珠机”命名的。使用一个 DAG 结构去学习和表现主题相关性，每个叶子节点为词表中的一个词，非叶子内部节点代表一个主题，每个主题是基于它的孩子节点的狄利克雷分布。为了通过该模型产生文档，首先通过每个狄利克雷采样一个多项式，从根节点开始，根据多项式分布对其子节点进行采样，沿着 DAG 的路径采样直到叶子节点产生词为止。这一过程就像“弹珠机”的金属球从顶部进入机器，跌入一组复杂的指针中，碰撞到的指针会改变金属球落下的路径，直到小球落入机器底部。

这种有向无环图结构是非常灵活的，可以是最基本的三层结构，也可以是任意嵌套的，节点间可以是全关联也可以是稀疏关联。PAM 模型中对主题的概念进行了扩展，主题不仅可以是基于词空间的分布，还可以是基于其他主题的分布，内部节点的子节点可以是其他的主题，因此可得到主题之间的关联。如果一个内部节点的所有孩子都是叶子节点，那么可以将其看作一个传统的 LDA 主题。LDA 可以视为 PAM 结构的一种特例，其相应的 DAG 结构只包括顶部的根节点、中间的主题集合以及底部的词集这 3 个层次。

### 3.1 PAM 的模型框架

PAM、LDA、CTM 的模型结构如图 4 所示[17]，每个矩形代表一个单词，每个圆形代表一个主题，主题上的箭头表示对其孩子的分布。(a)LDA：该模型中每个文档对其所有主题服从多项式分布，通过主题产生单词；(b)CTM：低层次的每个主题都是所有单词的多项式分布，每对主题都由另外一个高层主题覆盖；(c)Four-Level PAM：4 个层次包括 1 个总主题、1 组超主题、1 组子主题和词的集合；(d)PAM：任意的有向无环图表达主题相关性，每个非叶子节点代表一个主题，服从狄利克雷分布。

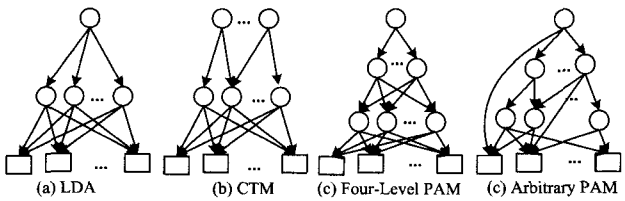


图 4　4 种主题模型的模型结构

PAM 模型中的表示法如表 1 所列。PAM 使用 DAG 结构连接词集 $V$ 和主题集合 $T$，每个主题 $t_i$ 在其孩子节点上服从分布 $g_i$。首先介绍任意 DAG 结构的 PAM 产生过程[17]，每个狄利克雷分布 $g_i$ 以向量 $\alpha_i$ 为参数，向量的维数与主题 $t_i$ 的孩子节点数目相等。

表 1　PAM 模型表示法

| 符号 | 含义 |
|---|---|
| V | 词的集合$\{w_1, w_2, \cdots, w_n\}$ |
| T | 主题集合$\{t_1, t_2, \cdots, t_s\}$ |
| r | 根节点，总主题，主题集合 T 中一个特殊的主题 |
| $g_i(\alpha_i)$ | 主题 $t_i$ 服从的狄利克雷分布 |
| d | 文档 |
| $\theta_{t_i}^{(d)}$ | 主题 $t_i$ 在文档 d 上的多项式分布 |
| $z_{wi}$ | 单词 w 在第 i 个主题上的采样分布 |

产生文档 $d$ 的步骤如下：

1. 通过 $g_1(\alpha_1), g_2(\alpha_2), \cdots, g_s(\alpha_s)$ 选取 $\theta_{t_1}^{(d)}, \theta_{t_2}^{(d)}, \cdots, \theta_{t_s}^{(d)}$，其中 $\theta_{t_s}^{(d)}$ 是主题 $t_i$ 在它的子节点上的多项式分布。

2. 对于文档中的每个单词 $w$：

a)选取一个主题路径 $z_w$ 的长度为 $L_w$：$\langle z_{w1}, z_{w2}, \cdots, z_{wL_w}\rangle$。$z_{w1}$ 是根节点 $r$，$z_{w2}$ 到 $z_{wL_w}$ 是集合 $T$ 中的主题节点。$z_{wi}$ 是 $z_{w(i-1)}$ 的子节点，而且它是根据多项式分布 $\theta_{z_{w(i-1)}}^{(d)}$ 进行选取的。

b)单词 $w$ 是通过 $\theta_{z_{wL_w}}^{(d)}$ 选取的。

通过这个过程，产生文档 $d$，主题分布 $z^{(d)}$ 和多项式分布 $\theta^{(d)}$ 的联合概率是：

$$P(d, z^{(d)}, \theta^{(d)} \mid \alpha) = \prod_{i=1}^{s} P(\theta_{t_i}^{(d)} \mid \alpha_i) \times \prod_{w} \left( \prod_{i=2}^{L_w} P(z_{wi} \mid \theta_{z_{w(i-1)}}^{(d)}) P(w \mid \theta_{z_w L_w}^{(d)}) \right) \tag{1}$$

对 $\theta^{(d)}$ 积分和对 $z^{(d)}$ 求和，我们推算文档的边缘概率是：

$$P(d \mid \alpha) = \int \prod_{i=1}^{s} P(\theta_{t_i}^{(d)} \mid \alpha_i) \times \prod_{w} \sum_{z_w} \left( \prod_{i=2}^{L_w} P(z_{wi} \mid \theta_{z_{w(i-1)}}^{(d)}) P(w \mid \theta_{z_w L_w}^{(d)}) \right) \mathrm{d}\theta^{(d)} \tag{2}$$

最终，整个语料库的概率是通过每个文档的概率产生的：

$$P(D \mid \alpha) = \prod_{d} P(d \mid \alpha) \tag{3}$$

### 3.2 四层 PAM 模型

四层模型是 PAM 模型的一个特例[17]，第一层是总主题 $r$，第二层 $T=\{t_1, t_2, \cdots, t_s\}$ 有 $s$ 个主题，第三层 $T'=\{t_1', t_2', \cdots, t_s'\}$ 有 $s'$ 个主题，最底层是单词。我们把第二层的主题称为超主题(super-topics)，第三层的主题称为子主题(sub-topics)。根部的总主题与所有超主题相关联，每个超主题与所有子主题相关联，子主题与所有单词相关联(模型结构见图 4(c))。

四层 PAM 的 DAG 结构中对主题使用了两种不同的分布。除了一组与总主题 $g_r(\alpha_r)$ 以及超主题 $\{g_i(\alpha_i)\}_{i=1}^{s}$ 相关的狄利克雷复合多项式之外，子主题服从固定的多项式分布 $\{\phi_{t_j'}\}_{j=1}^{s'}$，在整个语料库中通过一个单独的狄利克雷分布 $g(\beta)$ 进行选取。文档 $d$ 的产生过程如下：

1. 根据总主题的分布 $g_r(\alpha_r)$ 选取 $\theta_r^{(d)}$，$\theta_r^{(d)}$ 是超主题服从的多项式分布。

2. 对于每个超主题 $t_i$，根据主题的分布 $g_i(\alpha_i)$ 选取 $\theta_{t_i}^{(d)}$，$\theta_{t_i}^{(d)}$ 是子主题服从的多项式分布。

3. 对于文档中的每个单词 $w$：

(a)通过 $\theta_r^{(d)}$ 选取一个超主题 $z_w$。

(b)通过 $\theta_{z_w}^{(d)}$ 选取一个子主题 $z_w'$。

(c)通过 $\phi_{z_w'}$ 选取单词 $w$。

这种特例模型的构造和产生过程与 LDA 很相似，主要的不同是它另外有一个通过狄利克雷复合多项式构造的超主题层次，其是获取主题相关性的关键部分。四层 PAM 的图模型如图 5 所示。对于每个文档，PAM 通过总主题和超主题的狄利克雷分布选取多项式 $\theta$。对于每个单词 $w$，PAM 通过 $\theta$ 选取一个超主题 $z$ 和一个子主题 $z'$，单词服从的多项式分布 $\phi_{z'}$ 是与 $z'$ 相关联的。对这种结构的另外一种解释是给出了子主题，每一个超主题实质上就是一个独立的 LDA。因此，它可以看作是一系列 LDAs 的组合。

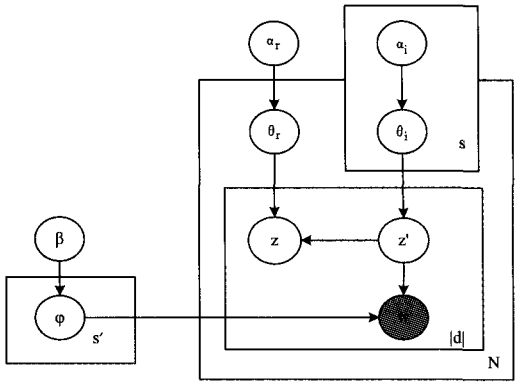


图 5　四层次 PAM 的图模型

根据以上过程产生一个文档 $d$，超主题的分布 $z^{(d)}$、子主题的分布 $z'^{(d)}$ 和多项式分布 $\theta^{(d)}$ 的联合概率是：

$$P(d,z^{(d)},z'^{(d)},\theta^{(d)}|\alpha,\phi)=P(\theta_r^{(d)}|\alpha_r)\prod_{i=1}^{s}P(\theta_i^{(d)}|\alpha_i)\times\prod_{\omega}(P(z_\omega|\theta_r^{(d)})P(z_\omega'|\theta_{z_\omega}^{(d)})P(\omega|\phi_{z_\omega'})) \quad (4)$$

对 $\theta^{(d)}$ 积分和对 $z^{(d)}$、$z'^{(d)}$ 求和，我们推算文档的边缘概率如下：

$$P(d|\alpha,\phi)=\int P(\theta_r^{(d)}|\alpha_r)\prod_{i=1}^{s}P(\theta_i^{(d)}|\alpha_i)\times\prod_{\omega}\sum_{z_\omega z_\omega'}(P(z_\omega|\theta_r^{(d)})P(z_\omega'|\theta_{z_\omega}^{(d)})P(\omega|\phi_{z_\omega'}))\mathrm{d}\theta^{(d)} \quad (5)$$

整个语料库的概率是通过每个文档的概率产生的，对子主题 $\phi$ 的多项式分布进行积分：

$$P(D|\alpha,\beta)=\int\prod_{j=1}^{s'}P(\phi_{t_j'}|\beta)\prod_{d}P(d|\alpha,\phi)\mathrm{d}\phi \quad (6)$$

## 4　PAM 改进模型及其比较

### 4.1　HPAM(Hierarchical PAM)

在 LDA 模型中，可以得到词之间的相互关系，但没有对不同主题之间的相关性进行刻画。Blei 等人在 2004 年提出了 HLDA(Hierarchical LDA)[20]，其采用树结构表示主题的组织构建。每篇文档由树结构中的一条主题路径产生，文档中的每个单词服从于该路径中某层次的一个主题。对主题使用 NCRP 方法进行学习，得到树结构。NCRP 先验需要两个参数：树结构的层数和 $\gamma$ 参数。$\gamma$ 可以看作未采样路径中想象的文档的数量。

明显属于语料库不同子集的文档可以共享一个主题。比如，“专业运动”子层次的各种主题和“电脑游戏”子层次的各种主题可能会使用“游戏”、“选手”、“得分”这些相同的词项。“专业运动”和“电脑游戏”有相同词项的唯一解释是它们有共同的父节点，但这不是语料库最好的组织形式。

PAM 和 HLDA 的一个区别是：前者的主题可能是词项的概率分布，也可能是其他(子)主题的概率分布，而 HLDA 中的每个主题都是词项概率分布。Mimno 等人在 PAM 的基础上提出了 HPAM 模型(Hierarchical PAM)[21]，该模型可以看成是把 HLDA 和 PAM 结合起来，使得 PAM 模型中的非叶子节点也具有词项的概率分布，即 DAG 结构中每个内部节点也服从词库上的多项式分布。HPAM 结合了 HLDA 的主题层次表达和 PAM 混合主题层次多叶子的能力，能够清楚地表达主题层次关系。该模型可以解决得到单一层次的问题，也不像 HLDA 那样要求得到一个准确的树结构，在留存文档似然性、自动发现的主题间互信息以及人工生成分类方面的实验结果都证明其性能得到改进。

HPAM 的图模型如图 6 所示。每篇文档在超主题层 $\theta_0$ 上服从多项式分布，对于每个超主题，在子主题层 $\theta_T$ 上服从多项式分布。DAG 上的每个路径由一个超主题/子主题对 $T_t$ 确定，由多项式 $\zeta_{T_t}$ 产生一个单词。定义 HPAM 在 PAM 的基础上每个节点都是词集上的一个分布，这对于层次主题模型来说是一个极其灵活的结构。

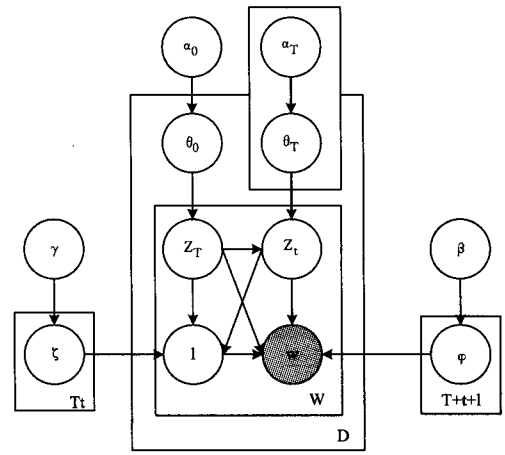


图 6　HPAM 的图模型

Mimno 介绍了 HPAM 的两个变形[21]：变形一，即 HPAM 模型 1，DAG 中每条路径都在路径层次上服从分布 $\zeta_{T_t}$，这些分布被所有的文档共享。模型的产生过程如下：

1. 对每个文档 $d$，选取超主题分布 $\theta_0$ 和每个超主题上子主题的分布 $\theta_T$；

2. 对于每个单词 $w$：

(a)通过 $\theta_0$ 选取一个超主题 $z_T$；

(b)通过 $\theta_{z_T}$ 选取一个子主题 $z_t$；

(c)通过 $\zeta_{z_T z_t}$ 选取一个层次 $l$；

(d)如果 $l=1$，通过 $\phi_0$ 选取一个单词，如果 $l=2$，通过 $\phi_{z_T}$ 选取一个单词，如果 $l=3$，通过 $\phi_{z_t}$ 选取一个单词。

变形二，即 HPAM 模型 2，它与模型 1 相似，但是不包含主题层次上的分布。内部节点的狄利克雷分布有一个额外的维度，比如说，一个有 10 个超主题和 20 个子主题的模型，根节点就有一个 11 维的分布，每个超主题有一个 21 维的分布。这个额外存在的维度相当于内部节点直接关联一个词，并不是通过 DAG 结构的底层关联这个词。模型产生过程如下：

1. 对每个文档 $d$，选取超主题分布 $\theta_0$ 和每个超主题上子主题的分布；

2. 对于每个单词 $w$：

(a)通过 $\theta_0$ 选取一个超主题 $z_T$，如果 $z_T=0$，通过 $\phi_0$ 选取一个单词；

(b)通过 $\theta_{z_T}$ 选取一个子主题 $z_t$，如果 $z_t=0$，通过 $\phi_{z_T}$ 选取一个单词；

(c)通过 $\phi_{z_t}$ 选取单词。

HPAM 不像 HLDA 学习主题的树结构，它通过内部节点分布的狄利克雷多项式参数来表示主题的层次结构，训练这些参数是 HPAM 一个很重要的部分。使用 Wallach (2006)中描述的 Gibbs EM 算法[22]训练根节点在超主题层上的分布以及超主题在子主题层上的分布。允许模型进行多次迭代，然后对每篇文档的每个超主题的词的数量进行周期性的采样。使用 Minka(2000)中提到的固定点迭代方法[23]来估计参数。

### 4.2 NPB PAM(Nonparametric Bayes Pachinko Allocation)

PAM 和之前的 LDA 相比有更强的表现力，四层结构只是 PAM 的一个简化特例，它的主题结构是任意嵌套复杂的，复杂结构有更强的表现力。然而对于复杂结构，选择最优的主题数以及确定合适的主题结构就会变得很困难。基于这个问题，Li 等人在 2007 年提出一个 NPB PAM 模型(Nonparametric Bayes PAM)[24]，针对 PAM 给定一个基于 HDP[25,26]的非参贝叶斯先验，从非结构数据中自动发现主题相关性，并确定不同层次的主题数目，非参方法对于复杂模型更有吸引力。

假设 PAM 的主题被组织为多层次结构，每层都使用一个 HDP 来获取不确定的主题数目。一个标准的 HDP 混合模型中数据有一个预定义的嵌套结构，根据主题分布建立基于动态数据组的 HDPs。NPB PAM 可以视为固定结果 PAM 的一个扩展，每层主题的数目设置为无穷大。为了产生一篇文档，首先从相应的 HDPs 中采样出主题的多项式分布，然后根据这些多项式对主题路径进行反复采样，得到文档中的每个单词。

NPB PAM 描述为 CRP(Chinese Restaurant Process)的一个变形[24]。CRP 过程可以简单描述为：

(1)顾客 $x$ 到达餐厅 $r_j$，选择第 $k$ 个入口 $e_{jk}$，参数为 $\alpha_0$；

(2)如果 $e_{jk}$ 是新的入口，分配一个种类 $c_l$，参数为 $\gamma_0$；

(3)选择种类后，选择桌子 $t_{jln}$，参数为 $\alpha_1$；

(4)顾客坐到一个已存在的桌子旁，会和该桌子的其他客人共用菜单和菜品，否则将给新开的桌子选择菜单 $m_{lp}$，参数为 $\gamma_1$；

(5)顾客使用一个已存在的菜单，则吃该菜单上的菜品，否则为新菜单选择菜品 $d_m$，参数为 $\phi_1$。

根据该过程，使用 PAM 产生语料库文档，文档相当于餐馆，单词相当于顾客；四层 PAM 中，无限的超主题相当于种类，无限的子主题相当于菜品，超、子主题被所有文档共享。产生一个文档中的一个单词，首先去采样一个超主题，超主题的采样过程包含两层 CRPs，分别得到入口和种类。它们作为超主题分布的先验，是一个层次狄利克雷过程(HDP)。然后对已知超主题用同样的方式采样一个子主题。同样地，子主题的分布先验也是一个 HDP，包含三层 CRPs，其分别采样餐桌、菜单和菜品。最后根据单词的多项式分布从子主题中采样得到单词。

### 4.3 PAM 及其改进模型的实验比较

#### 4.3.1 PAM

Li 等人[17]在 2006 年基于 NIPS 数据子集(NIPS00-12)进行实验，分别从人工估计主题和词之间关联性、留存测试数据的似然性以及文档分类的准确性这 3 个方面对 PAM 的性能进行评价。

(1)人工估计主题和词之间的关联性

使用 LDA 产生 100 个主题，使用 PAM 产生 50 个超主题和 100 个子主题。基于相似度产生主题对：对 PAM 中的每个子主题找到 LDA 中和它最相似的主题，标示为一对；同样对于 LDA 中的每个主题，找到 PAM 中最相似的子主题。在去除冗余主题和不相似主题后，提供给 5 个评估者 25 个主题对，评估者在不知情的情况下对每个主题对进行投票。结果如表 2 所列，可见 PAM 的人工评估结果明显优于 LDA，大多数主题对中的 PAM 主题得到过半数的支持。

表 2 人工估计关联性

| | LDA | PAM |
|---|---|---|
| 5 votes | 0 | 5 |
| >4votes | 3 | 8 |
| ≥3votes | 9 | 16 |

(2)留存测试数据的似然性

将分成 75%和 25%的两个数据子集进行实验，对较大数据集进行建模，对较小数据集计算似然值，进行定量测度，将 PAM 和 LDA、CTM、HDP 进行对比分析。PAM 在 50 个超主题时得到最好结果，受子主题数目的影响较大，因此设定超主题的个数为 50，子主题的数目为 20 到 180 之间，采用一个基于 Empirical Likelihood(EL) 的方法计算留存数据的似然值。实验结果如图 7 所示，PAM 子主题数和 LDA 主题数相同时，总是 PAM 的似然值较高，主题数越多，优势越明显。也就是说，PAM 得到的结果始终优于 LDA。当主题数较少时，CTM 的性能最好，在 60 个主题时达到峰值，但仍低于 PAM 在 160 个子主题处的峰值，之后随着主题数的增加，CTM 的似然值开始下降。由于没有预定义的数据结构，HDP 可以自动学习主题的数目，但不能得到任何的主题关联。因此 HDP 的结果不随主题数目的改变而改变，而与 LDA 的最优结果相似，但总是低于 PAM。

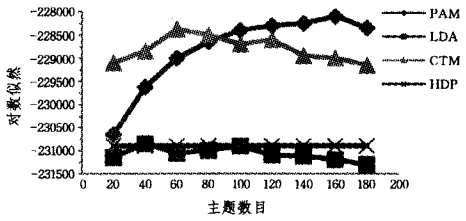


图 7 各模型似然值比较

当改变训练数据的规模时，随着训练数据规模的扩大，所有模型的似然值都持续增加。总体来说，PAM 的性能较好，尤其是训练规模较大时，优于其他所有模型。

(3)文档分类的准确性

在 20 个新闻数据组子集上进行 5 种分类，每类文档都划分为 75%的训练数据和 25%的测试数据。分别使用 LDA 和 PAM 模型对每个类别进行训练，并且计算测试数据的似然值，似然值越高，则测试文档分类的准确度越高。PAM 和 LDA 在文档分类准确性上的表现如表 3 所列，对于所有的文档分类，PAM 的准确度都比 LDA 高 3 到 5 个百分点。

表 3　文档分类准确性

| class | # docs | LDA | PAM |
|---|---|---|---|
| Graphics | 243 | 83.95 | 86.83 |
| Os | 239 | 81.59 | 84.10 |
| Pc | 245 | 83.67 | 88.16 |
| Mac | 239 | 86.61 | 89.54 |
| Windows. x | 243 | 88.07 | 92.20 |
| Total | 1209 | 84.70 | 87.34 |

人工估计主题和词之间的关联性、留存测试数据的似然性以及文档分类的准确性这 3 个方面都证明了 PAM 比 LDA 有更好的表现。

4.3.2　HPAM

Mimno 等人[21]使用美国医学索引数据库(Medline database)中 11 个学术期刊里的 5000 篇文摘作为语料库进行实验，比较 PAM、HLDA 和 HPAM 模型 1 和 2 的性能。

HLDA 不需要一个固定的主题数目。除 HLDA 之外的所有模型，经过多次的运行和交叉验证，选择的主题数目是基本一致的，超主题的数目变化范围是 7 到 13，子主题总的数目范围是 85 到 106。要求模型在更多主题下能够较好地预测隐藏数据，有出色的粒度和可解释性，似然值越高越好。实验结果显示[21]，LDA 在大于 20 个主题时似然值急剧下降，PAM 在更多主题时达到更好的似然性。PAM 和 HPAM1 都是超主题越多性能越好，HPAM2 性能稳定，对超主题的数目不太敏感。HPAM 在主题数目较大时更稳定，在达到 60 个子主题时有些许的下降，但始终比 PAM、HLDA 和 LDA 有更好的经验似然。

4.3.3　NPB PAM

Li 等人[24]分别在 4 个不同的数据集上应用 NPB PAM，并和其他模型的实验结果进行对比分析。

(1)首先将 NPB PAM 应用在结构已知的人工数据集，训练过后，基于相似度将每个已知主题和一个发现的主题相关联，基于主题分布准确度对性能进行评估。在多次实验中，子主题和超主题都有很高的准确度。

(2)将 NPB PAM 应用于真实文本数据，分成 5 部分的 20 个新闻组。将 NPS PAM 发现的主题和 PAM 的 5 个子主题进行对比。结果显示 PAM 的 5 个主题是混杂的，没有和 5 个文档组明确对应，而 NPS PAM 的主题和文档组有明显更强的关系。PAM 不能够自动得到合适的主题数目，而 NPS PAM 通过自动选择主题数目所产生的主题质量较高，这也显示了选择合适主题数目的重要性。

(3)分别对 20 个新闻组分成的 5 个数据子集和 Rexa 数据集进行实验，采用与 PAM 实验中相同的方法对 75% 的较大数据集进行建模，对 25% 的留存测试数据集计算似然值，进行定量测度，实验结果如表 4 所列。

表 4　似然估计

| Models | 20ng | Rexa |
|---|---|---|
| PAM 5-100 | −797350 | |
| PAM 5-200 | −795760 | |
| PAM 20-100 | −792130 | −580373 |
| PAM 20-200 | −785470 | −574964 |
| PAM 50-100 | −789740 | −577450 |
| PAM 50-200 | −784540 | −575086 |
| HDP | −791120 | |
| HLDA | −790312 | −581317 |
| NPB PAM | −783298 | −575466 |

对于 20 个新闻组分成的 5 个数据子集，PAM 在只有 5 个超主题时没有 HDP 表现好，这可能是因为 HDP 使用了额外的文档分组信息，而 PAM 没有。然而，随着超主题数量的增加，PAM 发现特定主题的能力变强，超过 HDP。NPS PAM 能够自动发现数据中的主题结构，和最优设置的 PAM (20-200 和 50-200)表现一样好，比 HDP 和 HLDA 要好很多。对于 Rexa 数据集，得到相似的实验结果。手动选择最优的 PAM 设置是 20 个超主题和 200 个子主题，和 NPS PAM 的结果近似。

## 5　基于 PAM 概率主题模型的发展趋势

从上文的模型介绍可以看出，最初对于 PAM 概率主题模型的研究工作集中在模型的改进和优化方面，包括对参数的扩展和结构的调整等，仍有较多可以入手的研究点。除此之外，主题模型已经不局限于理论研究的阶段，近几年来，PAM 主题模型开始逐渐应用于信息处理各领域，包括文本分类、信息检索、摘要生成和标签处理等。PAM 未来的发展趋势也应主要包括两个方面：主题模型性能的优化以及主题模型在文本处理方面的进一步应用。

首先，在主题模型性能优化方面，需要更高效的训练算法。在 PLSI、LDA、CTM 以及 PAM 这些主题模型中，都是将词项空间变换到主题空间，区别在于主题模型表示上的差别，或者说是在最优化时使用的目标函数不同。由于通常无法求得精确解，因此参数估计问题至关重要，有多种算法可用于估算 PAM 主题模型中的参数，常用的方法有 EM 算法进行变分推理和 Gibbs 抽样。Mimno 等人[27]提出一种基于任意图模型先验的吉布斯采样算法，它能够更好地处理大量文本集合训练推理过程中的复杂关联。Nallapati 等人[28]提出并行的变分 EM 算法来加速训练过程，该算法可应用于 PAM 主题模型的参数求解。其他关注主题模型性能优化的还有文献[29-31]等。

另外一个较为明显的趋势是 PAM 主题模型在信息处理领域的应用越来越广泛。PAM 主题模型本质上是一种对文本的概率建模的方法，因此可以应用在信息处理领域的各个方面。Sethi 等人[32]使用 PAM 模型来进行多文档摘要处理，首先使用主题模型识别多文档集的主题，再通过选择每个主题里最重要的句子生成摘要。先前 LDA 主题模型用于发现语料库表现的主题时，产生的文摘句子之间没有关联，PAM 则用超主题和子主题层次获取主题之间的关联，从而得到更好的文摘。Boulemden 等人[33]在图像检索领域，对图像的局部、全局以及在局部和全局特征的融合中使用 PAM 模型，通过内容图像索引任务中的 SIFT(尺度不变特征转换)技术，对局部图像特征进行提取。Bakalov 等人[34]基于 PAM 提出了两个半监督主题模型 Labeled Pachinko Allocation 和 Labeled Pachinko Allocation-List，其借助含有多标签文档的语料库，根据附加的关键词进行概念的分类。实验显示用户发现主题有益于对分类的解释，大幅增加准确性，能提供更好的信息率，更有效地组织数据和整理资料。Ma 等人[35]提出一个基于标签的四层 PAM 模型(Labeled Four-Level Pachinko Allocation Model，L-F-L-PAM)，应用提出的模型去推理训练数据，获取多标签之间的关系。近几年对于多标签学习问题兴起了一股热潮，目的是有效利用标签间的关系，找到多标签之

间本质的关系。实验结果证明，和其他高性能多标签学习方法相比，该模型在有效性和计算效率方面有很大的优势。Li等人[36]应用PAM的一个扩展版本去进行目标识别，基于PAM方法在一个层次结构里对潜在主题的相关性关系进行建模。在Caltech4和Caltech101数据集上高竞争性的识别结果说明，该方法和大多数现存的目标识别方法相比更具表现力和识别力。

本节所列出的这些工作，在某种意义上说明PAM主题模型在深度和广度上仍在进行着扩展，未来在模型性能优化和模型应用这两个方面都还有很大的发展空间，体现了PAM概率主题模型的生命力。

**结束语** 本文对一些具有代表性的主题模型作了一个比较全面的综述，并分析了几类方法的主要特点，特别集中在PAM概率主题模型及其改进方法的分析和比较。

HPAM模型结合了HLDA的主题层次表达和PAM混合主题层次多叶子的能力来清楚地表达主题层次关系。当主题结构复杂，主题数目过多、分配不均、参数不优时，HPAM仍然可以成功得到主题层次结构，但得到的结果不是最优的。正是由于PAM主题结构的复杂性，很难去估计所有的可能性和选择最优的主题数。NBP PAM针对PAM提出一个非参贝叶斯先验，通过假设一个基于HDP的先验，从非结构数据中自动发现主题相关性，确定不同层次的主题数目，其对复杂模型更具有吸引力。

因此，若能在HPAM的基础上结合NPB PAM中提出的非参贝叶斯先验，则可以解决在主题数目过多、结构复杂时产生的层次结构不优的问题，得到一个对复杂模型依旧灵活且更有吸引力的模型，这将是我们未来努力发展的方向。

## 参 考 文 献


[1] 曹娟，张勇东，李锦涛，等. 一种基于密度的自适应最优LDA模型选择方法[J]. 计算机学报，2008，31(10)：1780-1787

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

[3] Hofmann T. Probabilistic Latent Semantic Indexing[C]// Proceedings of the 22nd Annual International SIGIR Conference. New York: ACM Press, 1999: 50-57

[4] Blei D, Ng A, Jordan M. Latent Dirichlet Allocation[J]. Journal of Machine Learning Research, 2003, 3: 993-1022

[5] 张小平，周雪忠，黄厚宽，等. 一种改进的LDA主题模型[J]. 北京交通大学学报，2010，34(2)：111-114

[6] Wang X, McCallum A. Topics over time: A Non-Markov Continuous-Time model of topical trends[C]// Proceedings of the Conference on Knowledge Discovery and Data Mining (KDD). Philadelphia, USA, 2006: 424-433

[7] Blei D, Lafferty J. Dynamic topic models[C]// Proceedings of the 23rd International Conference on Machine Learning. Pittsburgh, Pennsylvania, USA, 2006: 113-120

[8] 李文波，孙乐，张大鲲. 基于Labeled-LDA模型的文本分类新算法[J]. 计算机学报，2008，31(4)：620-627

[9] 张晨逸，孙建伶，丁轶群. 基于MB-LDA模型的微博主题挖掘[J]. 计算机研究与发展，2011，48(10)：1795-1802

[10] Wang W, Barnaghi P M, Bargiela A. Probabilistic Topic Models for Learning Terminology Ontologies[J]. IEEE Transaction on Knowledge and Data Engineering, 2010, 22(7): 1028-1040

[11] Zavitsanos E, Paliouras G, Petridis S, et al. Learning subsumption hierarchies of ontology concepts from texts[J]. Web Intelligence and Agent Systems, 2010, 8(1): 37-51

[12] Zavitsanos E, Paliouras G, Vouros G A, et al. Discovering Subsumption Hierarchies of Ontology Concepts from Text Corpora [C]// Proceedings of Web Intelligence. 2007: 402-408

[13] 石晶，胡明，石鑫，等. 基于LDA模型的文本分割[J]. 计算机学报，2008，31(10)：1865-1873

[14] 石晶，范猛，李万龙. 基于LDA模型的主题分析[J]. 自动化学报，2009，35(12)：1586-1592

[15] 袁柳，张龙波. 基于概率主题模型的标签预测[J]. 计算机科学，2011，38(7)：175-180

[16] Blei D M, Lafferty J D. Correlated Topic Models[C]// Weiss Y, Scholkopf B, Platt J, eds. Advances in Neural Information Processing Systems 18. Cambridge, MA: MIT Press, 2006

[17] Li W, McCallum A. Pachinko Allocation: DAG-structured mixture models of topic correlations[C]// Proceedings of the International Conference on Machine Learning (ICML). Pittsburgh, Pennsylcania, 2006: 577-584

[18] 许戈，王厚峰. 自然语言处理中主题模型的发展[J]. 计算机学报，2011，34(8)：1423-1436

[19] Griffiths T L, Steyvers M. Finding scientific topics[C]// Proceedings of the National Academy of Sciences. 2004, 101: 5228-5235

[20] Blei D, Griths T, Jordan M, et al. Hierarchical topic models and the nested Chinese restaurant process[C]// Neural Information Processing Systems. 2004

[21] Mimino D, Li W, McCallum A. Mixtures of hierarchical topics with Pachinko Allocation[C]// Proceedings of the ICML. Corvallis, Oregon, USA, 2007: 424-433

[22] Wallach H. Topic modeling: Beyond bag of words [C]// Proceedings of the 23rd International Conference on Machine Learning. 2006

[23] Minka T. Estimating a Dirichlet Distribution[R]. Technical report. M. I. T., 2000

[24] Li W, Blei D, McCallum A. Nonparametric Bayes pachinko allocation[C]// Proceedings of the 23rd Conference on Uncertainty in Artificial Intelligence. Menlo Park, CA: AUAI Press, 2007

[25] Teh Y, Jordan M, Beal M, et al. Hierarchical Dirichlet Processes [J]. Journal of the American Statistical Association, 2006, 101 (476): 1566-1581

[26] 周建英，王飞跃，曾大军. 分层Dirichlet过程及其应用综述[J]. 自动化学报，2011，37(4)：389-407

[27] Mimno D, Wallach H, MaCallum A. Gibbs sampling for logistic normal topic models with graph-based priors[C]// Proceedings of the NIPS Workshop on Analyzing Graphs. Whistler, Canada, 2008

[28] Nallapati R, Cohen W, Lafferty J. Parallelized variational EM for latent dirichlet allocation: An experimental evaluation of speed and scalability[C]// Proceedings of the ICDM Workshop on High Performance Data Mining. Omaha, USA, 2007: 349-354

[29] Asuncion A, Smyth P, Welling M. Asynchronous distributed learning of topic models[C]// Proceedings of the NIPS. Vancouver, Canada, 2008: 81-88


(下转第23页)

其次，计算 Agent $\gamma$ 对与 Agent $\alpha$ 和 Agent $\beta$ 的交易所能获得的利益的评价值，该评价值的计算衡量指标为购买条款和销售预期，具体值见表 2。

表 2　Agent $\alpha$，Agent $\beta$ 的购买条款和 Agent $\gamma$ 的销售预期及权重

| | 价格 | 数量 |
|---|---|---|
| Agent $\alpha$ | 4 | 6 |
| Agent $\beta$ | 5 | 5 |
| 权重 | 0.6 | 0.4 |

计算得：$E(g_\alpha)=4\times0.6+6\times0.4=4.8$，$E(g_\beta)=5\times0.6+5\times0.4=5.0$。

再次，假定 Agent $\alpha$ 和 Agent $\beta$ 同时向 Agent $\gamma$ 提出能导致 Agent $\gamma$ 发生正面情绪变化的劝说提议。其中，Agent $\alpha$ 向 Agent $\gamma$ 提出的劝说策略为支持，Agent $\beta$ 向 Agent $\gamma$ 提出的劝说策略为肯定，拟定这两种劝说的策略值及它们在 Agent $\gamma$ 中的权重，见表 3。

表 3　Agent $\gamma$ 关于 Agent $\alpha$，Agent $\beta$ 所提出劝说的策略值和权重

| | 支持 | 肯定 |
|---|---|---|
| Agent $\alpha$ | 5 | |
| Agent $\beta$ | | 7 |
| 权重 | 0.4 | 0.6 |

同时拟定 Agent $\gamma$ 对 Agent $\alpha$，Agent $\beta$ 各指标的权重值，见表 4。

表 4　Agent $\gamma$ 对 Agent $\alpha$，Agent $\beta$ 的各指标的权重值

| | $\rho_\alpha(\rho_\beta)$ | $\rho_m$ | $\rho_g$ |
|---|---|---|---|
| Agent $\alpha$ | 0.4 | 0.4 | 0.2 |
| Agent $\beta$ | 0.3 | 0.4 | 0.3 |

根据以上数据和权重，计算可得：

$E\{P(\alpha\rightarrow\gamma)\}=0.4\times4.4+0.4\times5\times0.4+0.2\times4.8=3.52$

$E\{P(\beta\rightarrow\gamma)\}=0.3\times4.6+0.4\times7\times0.6+0.3\times5.0=4.56$

$\therefore E\{P(\alpha\rightarrow\gamma)\}<E\{P(\beta\rightarrow\gamma)\}$

综上，Agent $\alpha$ 对 Agent $\gamma$ 劝说成功，完成交易；Agent $\beta$ 对 Agent $\gamma$ 劝说失败，不能交易。

**结束语**　Agent 劝说作为一种较新的自动谈判模式，能较好发挥其人工智能优势，模拟人类思维，使谈判 Agent 通过劝说的辩论方式更好地说服对手，完成交易，这不仅能极大节约成本，还能较好实现谈判双方的共赢。在谈判过程中考虑 Agent 的情绪变化，能更进一步发挥 Agent 在劝说中的人工智能优势，使谈判结果更加趋于理性，谈判双赢局面能更完美实现。

本文在 Agent 情绪变化分类已有研究的基础上，将 Agent 在劝说中的情绪变化分为正面情绪变化、负面情绪变化和情绪无变化 3 类，在建立相应的形式化模型后对正面情绪变化的变化程度进行量化计算，并给出相应的评价方法。与已有研究相比，该分类相对系统、规范，所提出的模型和评价方法也能更好地从量化的角度度量正面情绪变化对 Agent 劝说的影响程度，从而能更好地促使 Agent 更理性地完成谈判，实现合作。

## 参 考 文 献


[1] Huang Ti-yun. Management information system(Revision)[M]. Beijing: Higher Education Press, 2001: 248-250

[2] Fatima S S, Wooldridge M, Jennings N R. A comparative study of game theoretic and evolutionary models of bargaining for software agents[J]. Artificial Intelligence Review, 2005, 23(2): 187-205

[3] Nguyen T D, Jennings N R. A heuristic model of concurrent bilateral negotiations in incomplete information settings[C]//18th Int. Joint Conf. on AI, 2003. Mexico, 2003: 1467-1469

[4] Ramchurn S D. Multi-Agent Negotiation using trust and persuasion[D]. Southampton, England: Faculty of Engineering and Applied Science, School of Electronics and Computer Science, University of Southampton, 2005

[5] 杨佩，高阳，陈兆乾. 一种劝说式 Agent 多议题协商方法[J]. 计算机研究与发展，2006，43(7)：1149-1154

[6] Wu Jing-hua, Jiang Guo-rui, Huang Ti-yun. Using Two Main Arguments in Agent Negotiation[C]//Ninth Pacific Rim International Workshop on Multi-Agents, 2006. China, 2006: 578-583

[7] 伍京华，蒋国瑞，孙华梅，等. 基于 Agent 的辩论谈判过程建模与系统实现[J]. 管理工程学报，2008，22(3)：69-73

[8] Jiang Hong, Vidal J M, Huhns M N. Incorporating Emotions into Automated Negotiation[C]//Proceedings of Agent Construction and Emotions, 2006. Austria, 2006: 123-129

[9] 胡军，史忠植，王茂光，等. 情感智能主体模型[J]. 计算机工程与应用，2007，43(1)：30-38

[10] 陈莉，陈晓云，胡山立，等. 基于情感组织 Agent 的联盟形成研究[J]. 广西师范大学学报：自然科学版，2008，26(1)：146-149

[11] 尹全军，胡记文，冯磊，等. Agent 建模的情感集成研究[J]. 系统仿真学报，2008，20(19)：5152-5157

[12] 张冬蕾，史忠植，潘瑜. 情感主体形式模型[J]. 模式识别与人工智能，2009，22(3)：381-387

[13] Meyer J J C. Reasoning about Emotional Agents [J]. International Journal of Intelligence Systems, 2006, 21(6): 601-619


(上接第 7 页)


[30] Smola A, Narayanamurthy S. An architecture for parallel topic models[C]//Proceedings of the VLDB. Singapore, 2010

[31] Yao L, Mimno D, MaCallum A. Efficient methods for topic model inference on streaming document collections[C]//Proceedings of the KDD. Paris, France, 2009: 937-946

[32] Sethi I, Seetharaman M M. Multi-Document Summarization using Pachinko Allocation Method[D]. Sri Sivasubramania Nadar College of Engineering, Chennai-India 603 110

[33] Boulemden A, Tlili Y. Image Indexing and Retrieval with Pachinko Allocation Model: Application on Local and Global Features[J]. Lecture Notes in Computer Science, 2012, 7457: 140-146

[34] Bakalov A, McCallum A, Wallach H, et al. Topic Models for Taxonomies[C]//Proceedings of the 12th ACM/IEEE-CS Joint Conference on Digital Libraries. 2012: 237-240

[35] Ma Hai-ping, Chen En-hong, Xu Lin-li, et al. Capturing correlations of multiple labels: A generative probabilistic model for multi-label learning[J]. Neurocomputing, 2012, 92: 116

[36] Li Y, Wang W, Gao W. Object Recognition Based on Dependent Pachinko Allocation Model[C]//IEEE ICIP. 2007: 337-340