LDA模型在话题追踪中的应用

张晓艳[1] 王 挺[2] 梁晓波[1]
(国防科技大学人文与社会科学学院 长沙 410074)[1] (国防科技大学计算机学院 长沙 410073)[2]

摘 要 随着对LDA模型的研究越来越深入,文本表示和挖掘能力进一步提高。“话题”是LDA模型中一个非常重要的概念,是特征集合的一个多项式概率分布。话题追踪是根据少数已知相关信息在未知报道流中追踪一个话题,找出与该话题相关的所有报道。把LDA模型用于话题追踪,目的有两个:(一)检验LDA话题对追踪话题的表示能力;(二)检验LDA模型在挖掘训练数据中的追踪话题时,LDA话题和追踪话题之间的关系。实验表明:相对于经典的向量空间模型和一元语言模型,以及专门针对追踪话题提出的事件模型,基于LDA模型的追踪性能更好,但由于粒度不同,LDA模型中的话题和追踪话题并没有直接的一一对应的关系,实现可定制话题的LDA模型是下一步工作的目标。

关键词 LDA模型,话题追踪,话题
中图法分类号 TP301 **文献标识码** A

Use of LDA Model in Topic Tracking

ZHANG Xiao-yan[1] WANG Ting[2] LIANG Xiao-bo[1]
(College of Humanities and Management, National University of Defense Technology, Changsha 410074, China)[1]
(Department of Computer, National University of Defense Technology, Changsha 410073, China)[2]

Abstract As more and more researches are made for the LDA model, its ability of representing and mining has been increased a lot. "Topic" is an important concept in the LDA model, which is represented as a polynomial distribution of the feature set. Topic tracking is monitoring a stream of news stories to find additional stories on a topic identified by several samples. There are two reasons for using the LDA model in topic tracking: one is to show how the performance of the tracking system using the LDA model is; the other is trying to find whether there is some relation between the LDA topic and the tracked topic. The experimental results indicate that the LDA model is better than the vector space model, the unigram language model and the special event model in a topic tracking system. However, since the granularities of two kinds of topics are different, the relation between the LDA topic and the tracked topic is not about bijection. An adjustable LDA model is needed in our future work.

Keywords LDA model, Topic tracking, Topic

1 引言

Latent Dirichlet Allocation(LDA)[1]模型是近年来提出的一种具有文本主题表示能力的非监督学习模型,自2003年由David M. Blei提出以后,已经越来越广泛地被各界人士关注。LDA模型从一个数据集合中挖掘指定个数的潜在话题模型,然后用这些话题模型表示一个文本,从而达到特征降维的目的。LDA模型的同一个“话题”中的特征通常比较相关或相近,存在如高共现度、高相关度等的联系。例如,一个文档同时含有“电脑”和“计算机”两个不同的特征,在LDA模型中通过进一步抽象,则可以把这两个特征映射到同一个话题里去;如果一个文档中包含“地震”和“玉树”(地震发生的地点),经过LDA模型的进一步抽象后,这两个特征也将在一个话题中关联起来。

话题追踪是话题发现与追踪(Topic Detection and Tracking, TDT)[2]的主要研究任务之一,是信息追踪的一种实例,是新知识发现中非常重要的技术之一。该技术帮助人们把分散的相关信息有效地汇集并组织起来,以便从整体上了解一个话题,在较多相关领域有着广泛的应用价值。话题追踪根据已知话题(通过少数报道定义描述)在未知数据流中追踪该话题的发展演化情况,即找出描述该话题的所有报道。这里的“话题”指一个核心事件或活动,以及所有与之直接相关的事件或活动。它与LDA模型中的“话题”相比:前者是有关一个真实存在的具体的事件;后者则更接近于“主题”概念,描述一类事件。

由于LDA模型在相关特征聚类以及主题表示方面的优

本文受国家自然科学基金(60873097,60933005)资助。

张晓艳(1981—),女,博士,讲师,主要研究方向为自然语言处理、话题发现与追踪,E-mail:zhangxiaoyan@nudt.edu.cn;王 挺(1970—),男,博士,教授,博士生导师,主要研究方向为自然语言处理、计算机软件;梁晓波(1969—),男,博士,教授,硕士生导师,主要研究方向为语料库语言学、认知语言学。

势，假定 LDA 模型对话题追踪中的“话题”比向量空间表示模型有更好的表示能力，本文初步试探了 LDA 模型在话题追踪中的应用情况，并与其他表示模型进行了比较。实验结果表明，LDA 模型比基于向量的模型的话题表示能力更强，并仍有进一步改进的空间。

本文第 2 节介绍话题追踪中的相关工作；第 3 节介绍 LDA 模型；第 4 节中介绍 LDA 模型如何用于话题追踪；第 5 节给出实验结果，并进行分析讨论；最后做出总结。

2 相关工作

在已有的话题追踪研究中，很多模型都曾被用于表示追踪话题和测试报道，大致可以分为两类：基于向量的模型和基于概率的模型。

基于向量的模型：以向量空间模型为主，包括多向量模型、词汇链等。有些是模型表示和相似度计算时都基于向量，例如多向量模型[3,4]；有些是在计算相似度时再转换为向量表示形式，例如图表示模型[5,6]。基于这类模型的追踪系统是目前话题追踪中的主流，也是很多研究的基准模型。

基于概率的模型：以一元语言模型为主，包括语言模型[7]和相关模型[8]等。其中后者在概率模型的基础上，还向追踪话题模型中扩充了主题相关知识，所以基于相关模型的追踪系统性能较好一些。这类模型在话题追踪研究中也占据了一定的地位。

在我们看来，基于向量的模型和基于概率的模型在特征选择上是一样的，区别仅在于特征的度量方法以及模型的使用方法上。两者各有优缺点，基于向量的模型受限于独立性假设，基于概率的模型受限于数据稀疏。

本文使用的 LDA 模型属于基于概率的模型，是近年来提出的一种具有文本主题表示能力的非监督学习模型，在数据挖掘和自然语言处理中的效果都较好。本文就是 LDA 模型在话题追踪中应用情况的试探研究。

3 LDA 模型

首先做如下一些符号定义：

- 特征(f)：离散数据的基本单元，通常是词典中的一个条目；
- 报道(s)：有 N 个特征，表示为 $s=\{f_i \mid 1\leqslant i\leqslant N\}$；
- 训练数据(D)：由 M 个报道构成，$D=\{s_1,s_2,\cdots,s_M\}$；

生成概率模型仅考虑特征在报道中的词频，忽略它们出现的先后顺序以及相互关系。假设一篇报道包含 k 个潜在话题 $z_j(1\leqslant j\leqslant k)$：

- 一元模型[7]仅根据特征的比例分布确定概率模型，没有考虑潜在话题，过于简单。式(1)是其文档产生过程。

$$p(s)=\prod_{i=1}^{N}p(f_i) \tag{1}$$

式中，$p(f_i)$是 f_i 的分步，可从训练样本中获取。

混合一元模型[9]首先随机选择一个潜在话题，然后再依次生成报道中的所有特征，缺点在于只允许一篇报道有一个潜在话题。式(2)是其文档产生过程。

$$p(s)=\sum_{j=1}^{k}p(z_j)\prod_{i=1}^{N}p(f_i\mid z_j) \tag{2}$$

式中，$p(z)$是 z 的分布，$p(f\mid z)$是给定 z 时 f 的条件分布，都可从训练样本中获取。

PLSI 模型[10]考虑了多个潜在话题，但容易过拟合，处理新文档的能力较差，而且复杂性随着文档数量增加变大。式(3)是其文档产生过程。

$$p(s,f_i)=p(s)\sum_{j=1}^{k}p(f_i\mid z_j)p(z_j\mid s) \tag{3}$$

式中，s 是训练数据中的文档，$p(s)$是文档 s 的分布，$p(z\mid s)$是给定 s 下的 z 的分布，$p(f\mid z)$是给定 z 时 f 的条件分布。

LDA[11]将一个文档表示为一个特定比例的主题混合，该比例从 Dirichlet 分布中抽样产生，即把文档在 LDA 话题上的概率分布看作是一个符合 Dirichlet 分布 α 的 k 维随机向量，克服了上述模型中的不足。式(4)是 LDA 模型产生一个文档的过程。

$$p(s)=p(\theta\mid\alpha)\prod_{i=1}^{N}p(z_i\mid\theta)p(f_i\mid z_i) \tag{4}$$

式中，θ 是一个 k 维的随机向量，$p(\theta)$是 θ 的分布，它的具体函数形式就是 Dirichlet 分布，这一分布保证 θ 的 k 个分量 $\theta_1,\theta_2,\cdots,\theta_k$ 都取连续的非负值，且 $\theta_1+\theta_2+\cdots+\theta_k=1$；$z_i$ 是离散随机变量，$p(z_i\mid\theta)$是给定 θ 时 z_i 的条件分布，它的具体函数形式很简单，就是把 θ 直接拿来作为概率值：$p(z_i\mid\theta)=\theta_i$，也就是说 z 取第 1,2,…,k 个主题的概率分别是 $\theta_1,\theta_2,\cdots,\theta_k$；$f_i$ 是离散随机变量，$p(f_i\mid z_i)$是给定 z_i 时 f_i 的条件分布，这里每个 z 都是特征集合上的一种分布。

总之，一元模型→混合一元模型→PLSI→LDA 是对文本建模的一步步完善。实际应用中还有这样那样的特征有待考虑，LDA 融合不同的特征产生了形态各异的 LDA 模型，例如考虑了 LDA 话题之间关联性的 Correlated Topic Model[12]，考虑了 LDA 话题动态演化特征的 Dynamic Topic Model[13]。本文属于 LDA 模型在话题追踪中的试探应用，仍然使用最基本的 LDA 模型。

4 追踪

在所有话题相关报道上训练出 LDA 模型之后，每一个报道 s 都表示为 k 个 LDA 话题上的概率分布 $P_s=\langle p_1,p_2,\cdots,p_k\rangle$。本文对追踪话题采用基于质心的表示方法，即根据一个追踪话题的相关报道的 LDA 话题分布$\{\langle p_{i1},p_{i2},\cdots,p_{ik}\rangle\mid 1\leqslant i\leqslant n\}$，取其质心作为该追踪话题的 LDA 话题分布 $P_t=\langle\frac{1}{n}\sum_{i=1}^{n}p_{i1},\frac{1}{n}\sum_{i=1}^{n}p_{i2},\cdots,\frac{1}{n}\sum_{i=1}^{n}p_{ik}\rangle$。由于所有报道都表示为 k 个 LDA 话题上的概率分布，在判断追踪话题和测试报道的相关度时，只需考虑这些概率分布之间的异同，不用再进一步明确 LDA 话题中的具体内容。利用 Clarity 计算两个概率分布的相关度：

$$\begin{aligned}Clarity(T,s)=&-KL(P_t\mid\mid P_s)+KL(P_t\mid\mid GE)-\\&KL(P_s\mid\mid P_t)+KL(P_s\mid\mid GE)\end{aligned} \tag{5}$$

式中，GE 是一个泛化的概率分布，本文从 LDA 模型的训练数据中获得，取所有训练报道的概率分布的质心；$KL(P_1\mid\mid P_2)$是 Kullback-Leibler 差异度量值：

$$KL(P_1\mid\mid P_2)=\sum_x P_1(x)\log\frac{P_1(x)}{P_2(x)} \tag{6}$$

如果一个追踪话题和一个测试报道之间的相关度大于预定义阈值，就认为报道与追踪话题相关，否则无关，相关度也被保留用于获取一个方法的最优性能。

5 实验

5.1 实验数据

为了评估本文的工作，在 TDT4[1] 的中文语料上进行了相关实验。TDT4 由语言数据联盟(Linguistic Data Consortium，LDC)提供，主要包括原始新闻文本和电视或广播新闻节目的转录文本，时间跨度为 2000 年 10 月－2001 年 1 月，专门用于话题发现与追踪研究。使用的中文数据集，共包括 70 个话题的话题描述及其测试报道流，其中每个话题描述通常包括 4 个相关报道，测试流是整个报道流中话题描述时间点之后的报道流。虽然话题追踪具有时序特征，要求顺序判断报道流中每个报道的话题相关性，但本文实验中不同测试报道之间的判断相互独立，因此它也可以看作是一个二值分类问题。LDA 模型基于 70 个话题的已知相关报道进行训练。

所有已知相关报道和测试报道在用于训练 LDA 模型和其他表示模型前都经过预处理。预处理对每篇新闻报道做分词、词性标注和停用词过滤处理，之后获得用于模型表示的词候选特征集合。如果一个词标记了多个词性，则按多个特征处理。报道长度在预处理之后统计得出。本文采用的分词和词性标注器是中科院计算所的汉语词法分析系统 ICTCLAS[2]，使用的停用词表有 507 个。

5.2 评价方法

本文使用 TDT2003[3] 的评价方法及其实现软件评价追踪系统的性能。该方法基于追踪结果中相关报道的丢失率和误判率构造主要评价指标：错误识别代价。其计算公式如下：

$$C_{det}=C_{miss}\times P_{miss}\times P_{target}+C_{fa}\times P_{fa}\times P_{non\text{-}target} \tag{7}$$

式中，C_{det} 是系统的错误识别代价，值越小说明系统追踪性能越好；C_{miss}、C_{fa}、P_{target} 和 $P_{non\text{-}target}$ 都是预定义值，取值分别为 1、0.1、0.02 和 0.98，其中前两个是丢失或误报一个相关报道的代价，后两个是一个报道与待追踪话题是否相关的先验概率；P_{miss} 与 P_{fa} 分别是系统追踪结果中话题相关报道的丢失率和误报率。

C_{det} 标准化之后，最小值仍然为 0，当取值为 1 时表示该追踪系统的性能不强于总是给出肯定判定或者否定判定的系统。标准化公式如下：

$$(C_{det})_{norm}=\frac{C_{det}}{\min(C_{miss}\times P_{target},C_{fa}\times P_{non\text{-}target})} \tag{8}$$

如果评测多个话题追踪的综合性能，P_{miss} 与 P_{fa} 有报道平均和话题平均两种计算方法，通常后者能给出更好的性能估计，因此本文中展示的所有实验结果都基于话题平均。

根据上述评价方法产生的追踪性能包括两种：当前性能和最优性能。前者是对确定决策阈值下的追踪系统的追踪结果进行评价的结果。当决策阈值遍历取值空间时，所有不同决策阈值下的追踪结果的 P_{miss} 与 P_{fa} 构成一个曲线图(称为 DET 曲线)，其中 $(C_{det})_{norm}$ 值最小的点对应的系统性能被认为是最优性能。在实际应用中，一个追踪方法的当前性能和最优性能都应受到关注。

5.3 实验结果

5.3.1 LDA 中 k 值和追踪话题个数的关系

在本节中，基于 LDA 模型的追踪方法与以下几种追踪方法进行了比较。

• 基于单向量表示模型的追踪方法：作为经典组合，单向量模型、经典 $tf*idf$ 权重计算方法和余弦相似度是目前话题追踪研究中使用最为广泛的基准系统之一。

• 基于一元语言模型的追踪方法：除了基于向量方法之外的另一经典方法。本质上来说，它和基于 LDA 模型的追踪方法都属于基于概率的方法，详见文献[15]。

• 基于事件表示模型的追踪方法：事件模型是我们已经提出的一种多向量模型，使用支持向量机模型对事件模型对应的多相似度进行整合获取追踪决策值，详情见文献[3]，它和 LDA 模型的相同之处都是把文本表示看作一个文档层-中间层-候选特征层的一个过程，不同在于如何将候选特征层面抽象到中间层面。如图 1 所示：事件模型的中间层是对特征集合的一次全划分；LDA 模型的中间层和候选特征之间是全关联的关系。

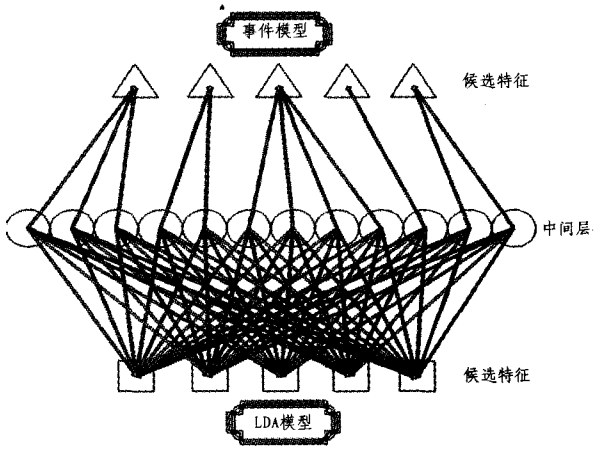

图 1 LDA 模型和事件模型候选特征到中间层抽象示意图

和本文基于 LDA 模型的追踪方法一样，上述方法都是首先计算追踪话题和测试报道之间的相关度，通过与预定义阈值相比较判断测试报道是否描述了追踪话题。不同之处主要在于使用了不同的表示模型。实验结果如表 1 所列。

表 1 各种追踪方法性能

	基于 LDA 模型的方法	基于向量模型的方法	基于概率模型的方法	基于事件模型的方法
		当前性能		
P_{miss}	0.0530	0.0118	*	0.0454
P_{fa}	0.0151	0.0902	*	0.0247
$(C_{det})_{norm}$	0.1268	0.4538	*	0.1665
		最优性能		
P_{miss}	0.0437	0.0998	*	0.0931
P_{fa}	0.0160	0.0108	*	0.0092
$(C_{det})_{norm}$	0.1222	0.1527	0.1334	0.1382

*：原论文中只提供了最小错误识别代价，其他评测值没有提供。

从上表中，可以看到：

• 基于 LDA 模型的追踪方法无论是当前性能还是最优

[1] http://projects.ldc.upenn.edu/TDT4/

[2] 中科院计算所汉语词法分析系统 ICTCLAS3.0 白皮书，http://www.i3s.ac.cn/Manual/

[3] http://www.itl.nist.gov/iaui/894.01/tests/tdt/tdt2003/evalplan.htm

性能在上表中都是最好的，充分说明了 LDA 模型在话题追踪中的适用性，在抓住文本中的话题方面比向量模型、一元概率模型和事件模型都要好。

• 在当前性能和最优性能的差距上：事件模型的优势之一是当前性能接近最优性能，现在这一优势要转移到 LDA 模型上，其差距仅有最小代价的 3.76%。

• 在丢失率和误判率上（主要观察当前性能里的这些评价值）：基于 LDA 模型的追踪方法能较大幅度地降低误判率，原因在于 LDA 模型对概率进一步抽象，即在 LDA 话题上的概率分布，能够进一步减小一些噪音的影响，从而使追踪话题和测试报道的相关度中减少了伪相关的成分，减少了误报。

• 需要提出的是表中的基于事件模型的方法训练数据不局限于已知相关报道，否则会由于训练数据过少而性能急剧下降。表中的性能是该方法的一个正常水平发挥。

需要注意的是：上述 3 种方法不同话题追踪之间相互独立；我们的基于 LDA 模型的话题由于训练数据包含了所有追踪话题的已知相关报道，话题追踪之间有可能会相互影响。这两种方法目前来看是各有利弊的。

5.3.2 LDA 的追踪性能及分析比较

在实验中，我们首先研究 LDA 模型中的 k 值设置是否和训练数据中的话题个数存在关系。对每个训练出来的 LDA 模型，有两种评价方法：

• 困惑度[2]：通过计算 LDA 模型在给定测试集合上的效果来判定，困惑度越低，说明 LDA 模型越好、泛化能力越强。假设给定测试集有 M 的报道 $\{s_i \mid 1<i<M\}$，一个 LDA 模型在该测试集上的困惑度计算如下：

$$\exp\{-\frac{\sum_{i=1}^{M} p(s_i)}{\sum_{i=1}^{M} length(s_i)}\} \tag{9}$$

式中，$length(s_i)$ 是一个文档的长度，$p(s_i)$ 是 LDA 模型在下一个文档上的概率分布。

• 平均相似度[14]：任意两个 LDA 话题的余弦相似度的平均值，不需要借助测试集合，直接评价训练出的 LDA 模型。平均相似度越小，说明 LDA 模型越好。

$$\frac{\sum_{i=1}^{k-1}\sum_{j=i+1}^{k} \cos(z_i, z_j)}{k*(k-1)/2} \tag{10}$$

式中，$\cos(z_i, z_j)$ 是两个 LDA 话题向量表示间的余弦相似度。

从计算量上看，平均相似度是较简单的评价方法，本文采用平均相似度评价 LDA 模型。首先根据文献[14]中的 k 值优化算法，确定了最优 k 值大于 200，但由于 k 值较大时训练 LDA 模型时空复杂性过大，因此本文仅计算了 2～200[4] 范围内抽样 k 值下的 LDA 模型的平均相似度，如图 2 所示，以观察 LDA 话题和追踪话题之间的关系。

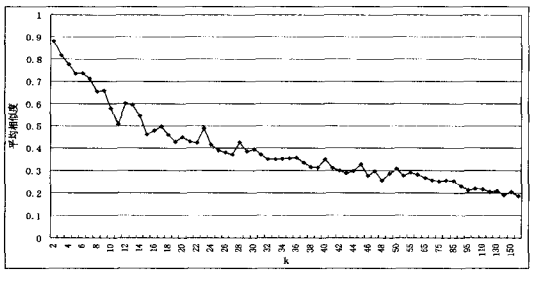

图 2 不同 k 值下的 LDA 模型平均相似度折线图

从图 2 可以看到，k 值从 2 到 200 期间，LDA 模型的平均相似度曲线虽然有些小的起伏，但整个趋势上来看仍然呈下降趋势，即在 200 以下随着 k 值的增加 LDA 模型越来越好。基于这些 LDA 模型的追踪性能也验证了这一点。但我们已经知道训练数据中含有 70 个不同的追踪话题，可见 LDA 模型中的话题和训练数据中的追踪话题并没有一一对应的关系。原因在于：LDA 话题是基于主题设计的，而追踪话题则是具体的粒度更细的事件，两者不属于一个层面。也正是因为这个原因，如何在现有 LDA 模型训练方法的基础上进一步区别更细粒度的追踪话题，是我们下一步解决的问题之一。

在本文接下来的实验中，不采用已有最好的 $k=200$ 的 LDA 模型，而是采用具有一定先验知识的 $k=70$ 的 LDA 模型，检验这种情况下的 LDA 模型是否仍比其他模型的话题表示能力有优势。

除了上面基于所有追踪话题的已知相关报道训练 LDA 模型外，还基于一个追踪话题的已知相关报道训练 LDA 模型，来分析 LDA 模型对单个追踪话题的表示能力，是否能挖掘出描述一个追踪话题的不同方面，即训练一个追踪话题的 LDA 模型时，k 值设为已知相关报道的个数。由训练结果来看，效果并不理想，原因有两个：一是训练数据过少，一个追踪话题的已知相关报道最多只有 4 个，最少是 1 个，过少的训练数据影响了训练模型性能的正常发挥；二是追踪话题的不同侧面是比话题粒度更小的信息块，在挖掘这类信息方面难度更大。

结束语 LDA 模型作为一个主题挖掘模型，已经成功应用到很多文本相关的领域。话题追踪根据少数已知信息从新闻报道流中挖掘指定话题的相关报道，是新知识发现领域中一项非常有意义的研究工作。本文首次把 LDA 模型用于话题追踪，主要解决了两个问题：与传统的向量模型和概率模型以及更专门的事件模型相比，LDA 模型仍然有更好的话题表示能力。下一步工作致力于实现一个可定制的 LDA 模型，使 LDA 话题可以定制，从“主题”到“事件侧面”粒度可调。

参 考 文 献

[1] 张晓艳，王挺. 话题发现与追踪技术研究[J]. 计算机科学与探索，2009，3(4)：347-357

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

[3] Zhang Xiao-yan，Wang Ting. Topic Tracking with Improved Representation Model and Joint Tracking Method[J]. International Journal of Wavelets，Multiresolution and Information Processing，2010，8(6)：913-930

[4] Zhang Xiao-yan，Wang Ting，Chen Huo-wang. Story Link Detection based on Event Model with Uneven SVM[C]//Fourth Asia Information Retrieval Symposium (AIRS'08). Harbin，China，Springer-Verlag，2008：436-441

[5] Eichmann D. Link Detection[R]. Iowa City：School of Library and Information Science，the University of Iowa，2004

[6] Ogilvie P. Extracting and Using Relationships Found in Text for Topic Tracking[R]. Pittsburgh，Pennsylvania，USA，2000

（下转第 152 页）

4) $K>200$ 的 LDA 模型由于机器内存消耗过大，本文没有训练。

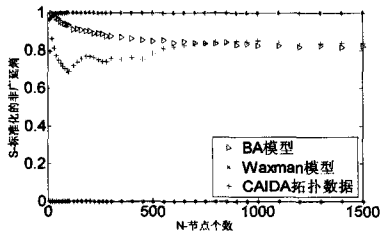

图 1　非广延参数 $q=0.8$

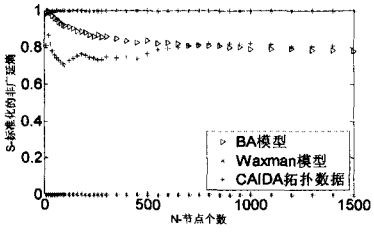

图 2　非广延参数 $q=1$

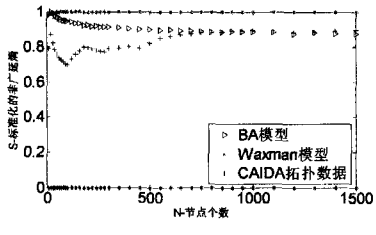

图 3　非广延参数 $q=1.2$

从图中可以看出，Waxman 模型的标准熵趋于 1，说明随机程度越大，节点度分布也越均匀。BA 模型及 CAIDA 拓扑数据得到的标准熵逐步趋于稳定，而且在确定的非广延参数 q 下，N 达到一定数量后，两者的稳定状态十分相近。

为了能够更好地分析这一现象，对不同 q 值在 1500 个节点下进行了不同模型的非广延熵值的计算，得到表 1。

表 1　不同 q 值下的非广延熵

非广延参数 q	CAIDA	BA	Waxman
0.2	0.9078	0.8951	0.9993
0.4	0.8440	0.8203	0.9989
0.6	0.8127	0.7721	0.9989
0.8	0.8168	0.7815	0.9991
1.0	0.8426	0.8223	0.9994
1.2	0.8971	0.8756	0.9997
1.4	0.9382	0.9194	0.9999
1.6	0.9660	0.9600	0.9999
1.8	0.9821	0.9807	1.0000

对表格数据进行拟合得到图 4。

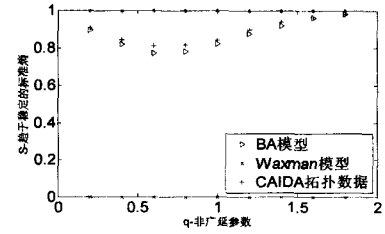

图 4　非广延参数与标准熵的关系

从图 4 中可以看出，在相应的广延参数下得到的稳态、BA 模型及 CAIDA 拓扑标准熵值十分相近。实验结果表明，BA 模型与当前的网络拓扑结构十分相近，对网络拓扑结构的研究可以借用 BA 模型。这从另一个侧面证明了非广延熵适合作为一种新的拓扑度量。但 Waxman 模型与真实网络拓扑还是存在一定的差距。

结束语　本文引入了非广延熵、信息熵及标准熵的定义，通过对 BA 模型、Waxman 模型以及真实拓扑数据进行模拟，用标准化的非广延熵定量地刻画，得出 BA 模型是符合真实网络结构的，基于 BA 模型对网络拓扑的研究是合理的。理论模型是理想化的，在真实的网络中存在路由节点之间的断开与重连操作，这是一个动态的过程，而理论模型忽略了这些因素。为了进一步地模拟真实的网络，下一段的工作需要对理论模型做相应的改进，考虑增加影响网络结构的因素。

参 考 文 献

[1] 汪小帆，李翔，陈关荣. 复杂网络理论及其应用[M]. 北京：清华大学出版社，2006：210-232

[2] 曹克非，王参军. Tsallis 熵与非广延统计力学[J]. 云南大学学报：自然科学版，2005，27(6)：514-520

[3] Faloutsos M, Faloutsos P, Faloutsos C. On power-law relationships of the internet topology[C]// Proc. of ACM SIGCOMM. 1999

[4] Tian Bu, Towsley D. On Distinguishing between Internet Power Law Topology Generators[C]// Proceeding of INFOCOM. New York, 2002, 2: 638-647

[5] 叶中行. 信息论基础(第二版)[M]. 北京：高等教育出版社，2007：32-56

[6] Barabasi A, Albert R. Emergence of scaling in random networks [J]. Science, 1999, 286: 509

[7] Zhou Shi, Mondragon R J. Accurately modeling the Internet topology[J]. Physical Review E, 2008, 70(6): 66-108

[8] 谭跃进，吴俊. 网络结构熵及其在非标度网络中的应用[J]. 系统工程理论与实践，2004(6)：32-36

(上接第 139 页)

[7] Nallapati R. Semantic Language Models for Topic Detection and Tracking[C]// Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology: Proceedings of the HLT-NAACL 2003 Student Research Workshop. Edmonton, Canada, Association for Computational Linguistics, 2003: 1-6

[8] Connell M, Feng A, Kumaran G, et al. UMass at TDT 2004[C]// Proceedings of the TDT2004 Workshop. 2004

[9] Nigam K, McCallum A, Thrun S, et al. . Text classification from labeled and unlabeled documents using EM[J]. Machine Learning, 2000, 39(2/3): 103-134

[10] Hofmann T. Probabilistic latent semantic indexing [C]// Proceedings of the Twenty-Second Annual International SIGIR Conference. 1999

[11] Boyd-Graber J L, Blei D M. Syntactic Topic Models[C]// NIPS 2008. 2008: 185-192

[12] Blei D M, Lafferty J D. Correlated Topic Models [C]// NIPS 2005. 2005

[13] Blei D M, Lafferty J D. Dynamic topic models[C]// ICML 2006. 2006: 113-120

[14] 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

[15] 王会珍，朱靖波，季铎，等. 基于多向量模型的中文话题追踪：自然语言理解与大规模内容计算[D]. 南京：清华大学出版社，2005：669-671