# 基于局部和全局的 LDA 话题演化分析[1]

章 建，李 芳

（上海交通大学 计算机科学与工程系,上海 200240）

**摘 要**：话题演化研究某个话题内容和强度随时间的变化。本文对 LDA 话题演化进行形式化描述，探讨基于全局和局部话题演化的两种建模方式，应用话题相似度和困惑度评测。对房地产话题和奥运会话题实例分析，给出两种不同建模方法在话题演化方面的优缺点。二会报告实验结果表明，全局话题演化能够获得较好的模型参数，方法简单可靠；而局部话题演化则能产生细粒度的话题，反映新话题的产生和旧话题的消亡。

**关键词**：文字信息处理；*狄利特利分布（LDA）*；话题关联；话题演化；

**中图分类号**：TP391 **文献标识码**：A

# LDA topic evolution based on global and local modeling

*ZHANG Jian, LI Fang*

( Dept. of Computer Science and Engineering, Shanghai Jiaotong Univ. Shanghai 200240, China)

**Abstract:** Topic evolution means the changes of contents and strength of a topic over time. The paper first gives the definition of topic evolution, describes two methods of topic evolution based on global and local documents. Two metrics of topic similarity and perplexity are used to evaluate both methods. The evolutions of two topics (the real estate vs. the 2008 Olympic game) are analyzed. Experiments on the recent 5 years of NPC&CPPCC news reports show that topic evolution based on global documents can get good topic model, the evolution method is easy, while topic evolution based on local documents can produce fine topics and show the arise of new topics and the vanish of old topics.

[1]收稿日期：2012-03-30

基金项目：国家自然科学基金项目（60873134）

作者简介：章建(1987－),男,硕士,主要研究领域为自然语言处理,信息检索与信息抽取。

李芳(联系人),女,博士,副教授。电话(Tel.)：021-34205423；E-mail：fli@sjtu.edu.cn

**Keywords**: Text Information Processing; Latent Dirichlet Allocation; Topic Detection and Evolution;

在 TDT 研究中，话题被定义为一个种子事件以及与之相关的所有事件或活动。广义上，话题不仅仅表示具有特定时间、地点、人物的事件以及由此引发的各种事件，还表示某些没有具体时间地点的热点话题，这些话题周期性出现，例如“医疗改革”、“行政体制改革”等出现在“全国两会”的新闻语料中。话题演化反映了某一个话题从它的提出，上升，下降，最后结束，这样一个过程。随着时间的变化，话题强度和内容都会发生变化，即存在话题的迁移[1]。如何自动发现话题以及话题随时间的变化是本文研究的目的。

近年来，LDA（*Latent Dirichlet Allocation*）模型[2]得到了广泛的应用。根据参考文献[3]， LDA 话题演化分为三种方法，其中，对文档集合先离散[4-7]和后离散[8-10]两种方法，尤其受到人们的关注。本文对文档先离散称为基于局部的方法，后离散称为基于全局的方法。本文从文档建模的角度出发，对这两种方法理论描述，然后应用话题平均相似度和困惑度对比，最后通过两个话题实例总结两种方法在话题演化方面不同的特点。

## 1. 话题演化模型

话题定义为一个三元组：$\{t, s, \phi\}$，其中$t$表示话题出现的时间，$s$表示话题在该时间段所具有的强度，$\phi$表示话题在词汇集合$V$上的多项式分布，形式为$\{p(W_1), p(W_2), \cdots, p(W_V)\}$，其中$p(W_v)$表示词汇表中第$v$个词语在话题中出现的概率。*LDA* 话题建模的任务是获取话题的强度$s$以及话题的内容表示$\phi$。

文档$d$表示为$\{t, (w_{d,1}, w_{d,2}, \cdots, w_{d,N_d}), \theta_d\}$，其中$t$表示文档的时间戳，$(w_{d,1}, w_{d,2}, \cdots, w_{d,N_d})$是文档$d$的词语序列，$N_d$为文档$d$的词语数量，$\theta_d$表示文档$d$的话题分布（*LDA* 模型假设一篇文

档有多个话题）。$t$和$(w_{d,1}, w_{d,2}, \cdots, w_{d,N_d})$是话题建模的输入，$\theta_d$是输出。

话题演化定义为：寻找不同时间段上具有相同语义的话题随时间变化的演化路径。它包含两个子任务：话题强度的变化和话题内容的变化。因此，话题演化需要解决两个问题：一是各个时间段话题的内容$\phi$如何表示以及话题强度如何计算，二是不同时间段上具有相同语义的话题如何对应，形成演化路径。

### 1.1 基于全局的话题演化方法

忽略文档的时间戳，对语料进行全局话题建模，然后，根据文档时间戳，划分各个子集，获取各个时间段的局部话题。最后，利用不同时间段上局部话题间的对应关系，产生话题的演化路径。

#### 2.1.1 话题内容和强度的计算

全局 LDA 话题建模后，可以获得所有话题在词汇上的多项式分布$\Phi = \{\phi_1, \phi_2, \cdots, \phi_K\}$，所有文档在话题上的分布$\Theta = \{\theta_1, \theta_2, \cdots, \theta_{|D|}\}$，以及文档$d$中词语的话题类别：$\{z_{d,1}, z_{d,2}, \cdots, z_{d,N_d}\}$。不同时间段的局部话题计算方法：假设整个语料集合由$N$个不同时间段的文档集合$\{D_{t_1}, D_{t_2}, \cdots, D_{t_N}\}$组成，则在时间段$t_n$中的第$k$个话题下词汇$W_v$的概率表示为：

$$p(W_v|t_n, k) = \frac{\sum_{d \in D_{t_n}} \sum_{n=1}^{N_d} 1 * equal(w_{d,n}, W_v) * equal(z_{d,n}, k) + \beta}{\sum_{d \in D_{t_n}} \sum_{w \in d} 1 * equal(z_{d,n}, k) + V * \beta} \quad (1)$$

其中，$w_{d,n}$表示文档$d$中的第$n$个词语，$z_{d,n}$表示文档$d$中第$n$个词语的话题类别，$equal(x, y)$判断$x$和$y$是否相等，若相等取值为 1，否则为 0。$\beta$表示全局建模过程中每个话题在词汇上多项式分布的先验值。因此，全局演化模型中，$t_n$时间段第$k$个话题的内容为：

$$\phi_{t_n,k} = \{p(W_1|t_n, k), p(W_2|t_n, k), \cdots, p(W_V|t_n, k)\} \quad (2)$$

时间段$t_n$第$k$个话题强度则由公式（3）计算，其中$|D_{t_n}|$表示时间段$t_n$中文档的数量，$\theta_d^k$表示文档$d$中第$k$个话题所占的比例。

$$s_{t_n,k}=\frac{1}{|D_{t_n}|}\sum_{d\in D_{t_n}}\theta_d^k \tag{3}$$

### 2.1.2 话题关联和演化

局部话题根据全局话题计算，具有对应关系。理论上，全局话题建模后得到第$k$个话题的内容（在词汇上的多项式分布）为$\phi_k$，则时间段$t_n$下该话题的先验分布表示为：$\phi_{t_n,k}\sim Dirichlet(\phi_k)$。因此，每个时间段上的话题内容，其先验受全局话题的影响，由此形成不同时间段局部话题之间的一一对应，由此产生关联和演化结果。

## 2.2 基于局部话题演化方法

整个文档集按照时间划分，分别局部 *LDA* 建模。不同时间段的话题数量可以相同，也可以不同。通过计算不同时间段话题间的关联度[10]，获取话题在时间上的演化。

### 2.2.1 话题内容和强度的计算

假设时间段$t_n$所有话题在词汇上的多项式分布$\Phi=\{\phi_{t_n,1},\phi_{t_n,2},\cdots,\phi_{t_n,K}\}$，以及话题的多项式分布$\Theta=\{\theta_1,\theta_2,\cdots,\theta_{|D_{t_n}|}\}$。第$k$个话题的内容为：

$$\phi_{t_n,k}=\{\overline{p}(W_1|t_n,k),\overline{p}(W_2|t_n,k),\cdots,\overline{p}(W_V|t_n,k)\} \tag{4}$$

其中，$\overline{p}(W_v|t_n,k)$在局部建模后直接获得。时间段$t_n$第$k$个话题的强度计算采用公式（3）。

### 2.2.2 话题关联和演化

根据[10]，时间段$t_i$上的第$r$个话题和时间段$t_{i+1}$上的第$u$个话题，关联度计算如下：

$$Relate(\phi_{t_i,r},\phi_{t_{i+1},u})=\lambda Sp(\phi_{t_i,r},\phi_{t_{i+1},u})+(1-\lambda)Fp(\phi_{t_i,r},\phi_{t_{i+1},u})+\lambda \tag{5}$$

其中$Sp(\phi_{t_i,r},\phi_{t_{i+1},u})$描述话题间的语义关系，$Fp(\phi_{t_i,r},\phi_{t_{i+1},u})$描述话题间的同一性。为了判断话题$r$和话题$u$是否具有语义关联，设定阈值$\gamma$，当$Relate(\phi_{t_i,r},\phi_{t_{i+1},u})$大于$\gamma$，认为话题$r$和话题$u$在语义上关联。由此得到基于局部演化方法的演化路径。

理论上，时间段$t_n$中的话题$k$，其在词汇上分布的先验表示为：$\phi_{t_n,k}\sim Dirichlet(\phi^0_{t_n,k})$。其中，$\phi^0_{t_n,k}$是时间段$t_n$第$k$个话题在词汇上分布的先验值。对不同时间段，这些参数是预先设定

的常值，没有考虑话题之间的影响。因此，需要确定相邻时间段话题的关联才能找到话题的演化路径。

# 2 实验

## 3.1 模型评测

实验数据为 2007 年至 2011 年的两会报告。先对语料预处理，包括分词，过滤停用词，去除低频词和高频词等。为了更好地对比，全局和局部话题建模将话题个数设为一致，新闻报道数目相同，详细实验数据见表 1。

表 1 模型评测实验数据设置

Tab.1 Experimental data details for model evaluation

| 文集 | 2007 | 2008 | 2009 | 2010 | 2011 |
|---|---|---|---|---|---|
| 文档数目 | 3048 | 3048 | 3048 | 3048 | 3048 |
| 词汇数目 | 23938 | 37783 | 21162 | 17420 | 16464 |
| 全局演化（话题个数） | 160 | 160 | 160 | 160 | 160 |
| 局部演化（话题个数） | 160 | 160 | 160 | 160 | 160 |

### 3.1.1 评测指标

话题间的平均相似度和困惑度（*perplexity*）评测全局和局部话题建模，公式省略。话题平均相似度最小时，LDA 模型最优。困惑度越小则表明模型对未知的数据预测能力越强。

### 3.1.2 评测结果

话题间平均相似度的评测结果见表 2，困惑度的评测结果见表 3。

表 2 平均相似度

Tab.2 Result of the average topics similarities

| | 2007 年 | 2008 年 | 2009 年 | 2010 年 | 2011 年 |
|---|---|---|---|---|---|
| 局部建模 | 0.0592 | 0.0516 | 0.0428 | 0.0403 | 0.0355 |
| 全局建模 | 0.0291 | 0.0271 | 0.0290 | 0.0309 | 0.0305 |

表 3 困惑度

Tab.3 Result of perplexities

| | 2007 年 | 2008 年 | 2009 年 | 2010 年 | 2011 年 |
|---|---|---|---|---|---|

| | | | | | |
|---|---|---|---|---|---|
| 局部建模 | 0.00336 | 0.00208 | 0.00256 | 0.00284 | 0.00304 |
| 全局建模 | 0.00231 | 0.00182 | 0.00168 | 0.00179 | 0.00179 |

根据表 2 和表 3，全局建模优于局部建模。全局建模对整个时间段的文档集合建模，训练数据较多，获得较好的模型参数，话题之间独立性更强，相似度低；而局部话题建模分别对每个时间段文档集合建模，数据少。同样，采用越多数据对模型训练，则模型对未知数据的预测越准确，从而导致全局建模比局部建模具有更低的困惑度。

然而，评价话题演化的优劣，关键是分析演化路径中关联的话题是否在语义上具有一致性。因此，为了更好地分析和对比基于全局和局部话题演化的特点，下面通过房地产和奥运会话题来分析两种方法的结果。房地产话题是在整个时间段上都出现的话题，而奥运会话题是在特定时间段上出现的话题。

## 3.2 实例分析

根据二会报告的实际语料，进行全局和局部建模，具体数据见表 4。

表 4 实例分析实验数据设置

Tab.4 Experimental data details for example analysis

| 文集 | 2007 | 2008 | 2009 | 2010 | 2011 |
|---|---|---|---|---|---|
| 文档数目 | 6127 | 9041 | 4376 | 3031 | 3073 |
| 词汇数目 | 30401 | 49971 | 22287 | 17417 | 16472 |
| 全局演化（话题个数） | 250 | 250 | 250 | 250 | 250 |
| 局部演化（话题个数） | 160 | 250 | 160 | 120 | 120 |

### 3.2.1 话题演化图

图 1 显示了房地产和奥运会话题的演化结果,其中红色部分为全局建模的演化图，黑色部分则为局部建模的演化图，下面依次对两个话题进行分析。

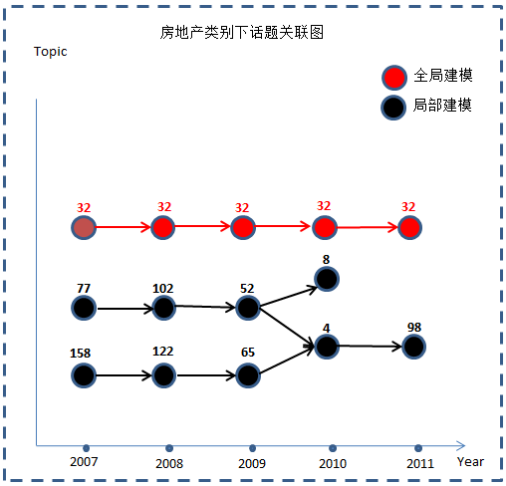


图 1(a) 房地产话题演化图

Fig.1(a) Evolution of the real estate

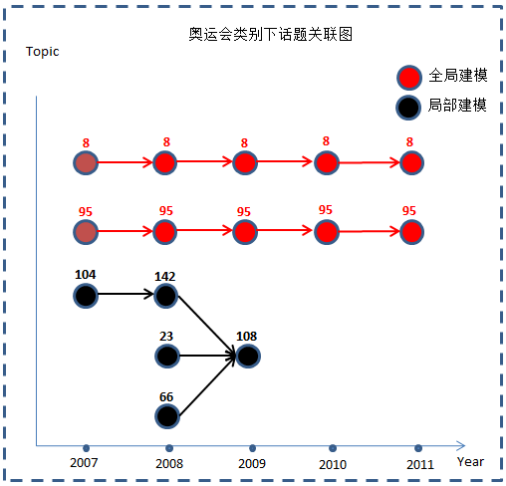


图 1(b) 奥运会话题演化图

Fig.1(b) Evolution of the Olympic game

(a) 房地产话题

表 5 全局话题建模房地产话题的内容演化

Tab.5 Contents Evolution of global modeling for the real estate topic

| 时间 | 话题 | 话题中概率最大的 10 个词语 |
|---|---|---|
| 2007 | 32 | 住房 房价 廉租 房地产 经济适用房 市场 家庭 城市 商品房 房子 |
| 2008 | 32 | 住房 房价 城市 家庭 开发商 适用 公积金 限价 商品房 住宅 |
| 2009 | 32 | 住房 房价 市场 家庭 城市 公积金 适用 建设部 商品房 低收入 |
| 2010 | 32 | 住房 房价 市场 保障性 商品房 城市 供应 房子 适用 买房 |
| 2011 | 32 | 住房 保障性 房价 市场 租赁 调控 房产 城市 需求 房子 |

全局建模得到的话题是一一对应的，例如话题 32，由表 5，房价问题是房地产话题的核心，但每年有各自不同的侧重点，表现为 2007 年廉租房、经济适用房、商品房市场，2008 年住房公积金、城市住房开发、商品房限价，2009 年住房公积金，2010 保障性住房以及商品房供应市场，2011 保障性住房、房屋的租赁和房价的调控。

表 6 局部话题建模房地产话题的内容演化

Tab.6 Contents evolution of local modeling for the real estate topic

| 时间 | 话题 | 话题中概率最大的 10 个词语 |
|---|---|---|
| 2007 | 77 | 房价 房地产 住房 市场 调控 政府 价格 房子 上涨 城市 |
| 2007 | 158 | 住房 政府 廉租 经济适用房 建设 家庭 解决 低收入 合作 保障 |
| 2008 | 102 | 房价 开发商 土地 价格 住房 城市 开发 限价 商品房 房子 |
| 2008 | 122 | 住房 家庭 困难 民生 适用 城市 低收入 收入 公积金 |
| 2009 | 52 | 房价 价格 市场 成本 开发 土地 开发商 行业 调整 销售 |
| 2009 | 65 | 住房 解决 市场 保障 家庭 城乡 困难 城市 适用 公积金 |
| 2010 | 8 | 价格 税费 收费 成品油 我国 市场 产品 成本 资源性 征收 |
| 2010 | 4 | 住房 房价 土地 统计 市场 上涨 城市 政府 地方政府 部长 供应 |

从图 1(a)中可以看到，基于局部建模房地产话题演化得到三条演化路径。其中，{(2007,77),(2008,102),(2009,52),(2010,4),(2011,98)}反映了 2007 年政府对房价的调控，2008 年商品房限价，2009 年住房开发成本，2010 年房价上涨，2011 年房价调控；另一条演化路径{(2007,158),(2008,122),(2009,65),(2010,4),(2011,98)}则反映了 2007 年政府廉租经济适用房出台，2008 年困难家庭民生，2009 年解决住房问题，2010 年房价上涨，2011 年房价调控。前一条路径反映住房价格的形成、调控等，后一条路径则反映了低收入者住房问题。图 1（a）还有一条演化路径{（2007，77），(2008,102),(2009,52)，（2010，8）}前三个话题与房地产相关，而最后一个是成品油的价格和成本，由于房地产的价格成本与成品油的价格成本语义上有一定的近似，所以，形成了两个话题的关联。

(b) 奥运会话题

表 7 全局话题建模奥运会话题的内容演化

Tab.7 Contents evolution of global modeling for the Olympic game topic

| 时间 | 话题 | 话题中概率最大的 10 个词语 |
|---|---|---|
| 2007 | 8 | 开幕式 参加 透露 采取 全世界 晚会 回答 能够 表现 准备 |
| 2008 | 8 | 开幕式 张艺谋 奥运会 北京 透露 参加 导演 表演 全国政协 闭幕式 |
| 2009 | 8 | 阅兵 透露 张艺谋 参加 开幕式 晚会 北京 烟火 全国政协 希望 |
| 2010 | 8 | 兴奋剂 参加 全国政协 北京 透露 崔大林 希望 新华网 阅兵 预案 |
| 2011 | 8 | 北京 希望 全国政协 透露 世界 参加 采取 能够 邰丽华 可能 |
| 2007 | 95 | 北京 体育 运动员 北京市 举办 场馆 奥林匹克 城市 筹办 青岛 |
| 2008 | 95 | 北京 奥运会 火炬 体育 举办 北京市 运动员 传递 圣火 场馆 |
| 2009 | 95 | 北京 奥运会 北京市 体育 蒋效愚 举办 运动员 北京奥组 场馆 邓亚萍 |
| 2010 | 95 | 北京 体育 运动 高尔夫球 奥运会 传递 运动员 北京市 世界 国际 |
| 2011 | 95 | 北京 体育 北京市 举办 吉林 首都 传递 世界 奥运会 场馆 |

根据图 1(b)全局话题演化得到两条有关奥运会的话题演化路径。第一条演化路径（话题 8）反映了：2007 年北京奥运会开幕式的准备情况，2008 年北京奥运会开幕式情况，2009 年张艺谋参加政协并透露 60 周年国庆的执导情况，2010 年崔大林参加政协并提出体改预案，2011 年邰丽华参加政协并提出改善残疾人地位的提案。这些话题都与某项活动（奥运会或者政协会议）的参加或举办有关。在第二条演化路径（话题 95）中，2007 年奥运会的筹办，

2008 奥运会火炬传递，2009 年，2010 年以及 2011 年则是其它体育运动话题。

表 8 局部话题演化下奥运会话题的内容演化

Tab.8 Contents evolution of local modeling for the Olympic game topic

| 时间 | 话题 | 话题中概率最大的 10 个词语 |
|---|---|---|
| 2007 | 104 | 奥运会 北京 体育 举办 世界 北京市 运动员 文明 保安 志愿者 |
| 2008 | 142 | 奥运会 北京 体育 开幕式 张艺谋 运动员 希望 举办 世界 场馆 |
| 2008 | 23 | 北京 北京市 奥运会 交通 措施 吉林 城市 质量 空气 市民 |
| 2008 | 66 | 火炬 北京 传递 圣火 奥运会 仪式 希腊 宁夏 开始 接力 |
| 2009 | 108 | 北京 奥运会 广州 宣传 蒋效愚 亚运会 人员 举办 经验 情况 |

根据图 1(b)局部话题演化的结果，可以得到三条有关奥运会的演化路径：其中，{（2007，104），（2008，142），（2009，108）}反映了 2007 年北京举办奥运会，2008 年开幕式，2009 年事后影响。2008 年北京奥运会分为三个话题：开幕式话题（话题 142），城市交通和空气质量（话题 23）以及 奥运会圣火传递（话题 66），更加精确地描述了 2008 年奥运会的话题，同时该话题在 2010 年、2011 年没有对应话题，反映了话题的消亡。

3.2.2 话题强度变化

图 2 显示了两个话题的强度变化图，红线是全局建模结果，黑线是局部建模结果。

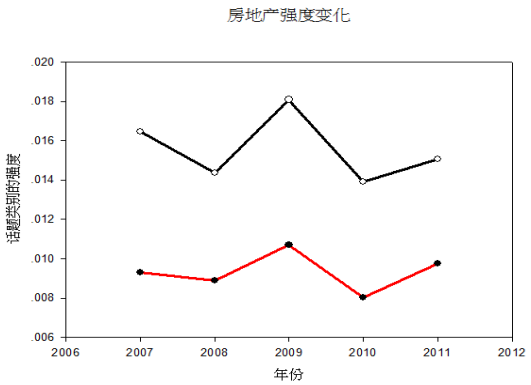


图 2(a) 房地产的强度演化图

Fig.2 (a) topic evolution for real estate

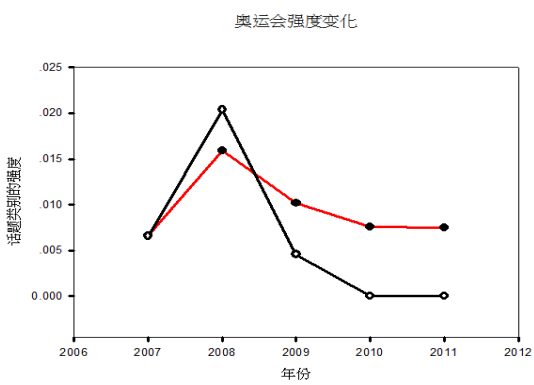


图 2(b) 奥运会的强度演化图

Fig.2(b) topic evolution for the Olympic game

图 2 说明全局和局部话题演化的趋势基本一致。后者趋势变化比前者明显，这是因为全局话题演化认为每个话题在时间上是持续的，不考虑话题的产生以及消亡，导致话题强度变化相对平缓。局部话题演化反映不同时间段新话题产生以及旧话题的消亡，因此，在不同时间段下话题强度的变化比较明显。

### 3.3 分析与总结

表 2，表 3 显示，基于全局建模的话题之间相似度小，模型困惑度小，具有更好的预测能力；局部话题建模可以挖掘更细粒度的话题，例如 2008 年的三个奥运话题。通过对房地产和奥运会话题实例分析，两种建模方法的话题演化各有优缺点，如表 9 所示。

表 9 全局话题演化和局部话题演化对比

Tab.9 Comparison of topic evolution based on global and local modeling

| | 全局话题演化 | 局部话题演化 |
|---|---|---|
| 优点 | (1).利用更多语料建模，话题精确<br>(2).话题对应关系简单，一一对应 | (1).发现新话题的出现和旧话题的消亡<br>(2).描述话题之间多和多的关联关系<br>(3).产生细粒度话题 |
| 缺点 | (1).无法描述新话题的生成和旧话题的消亡<br>(2).当话题消亡时，会将后续语义相似但非同一话题必然地关联起来 | (1).局部话题对新数据的预测能力不如全局话题<br>(2).演化结果随阈值变化 |

## 3 结语

本文对话题演化形式化描述，对基于全局与局部两种话题演化方法进行了指标评测和实例分析。结果显示：基于全局建模的话题演化可以描述持久性话题的演化，通过话题内容和强度的变化，了解该话题的趋势。基于局部建模的话题演化能反映新话题的出现和旧话题的消亡、话题之间多对多的关系，包括话题的分裂或合并，具有更广的应用前景。

目前，话题演化缺乏一个客观的衡量标准，许多话题演化研究都是人为判断,例如基于 ACM 语料学术话题演化研究[11]以及探索事件报道的演化路径[12]。如何衡量话题演化路径的正确性，或者话题在语义上的一致性，需要提出一种客观的话题演化评测标准。局部建模应用更合理的方法来判断不同时间段话题语义的一致性,进一步工作是应用话题的语境改进局部建模话题相似度的计算方法，提高话题演化关系的精度。

**参考文献**:

[1] Juha Makkonen. Investigations on Event Evolution in TDT. Proceedings of HLT-NAACL 2003 student research workshop, Edmonton: 2003: 43-48.

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

[3] 单斌，李芳. 基于 LDA 话题演化研究方法综述[J]. **中文信息学报** 2010,24(6): 43-49

SHAN Bin, LI Fang. A Survey of Topic Evolution based on LDA[J] **Journal of Chinese Information Processing** 2010 24(6) 43-49

[4] David M.Blei, John D.Lafferty. Dynamic Topic Models. Proceedings of the 23rd International Conference on Machine Learning. Pittsburgh, PA, USA. 2006. 113-120

[5] L.Alsumait, D.Barbara, C.Domeniconi. On-line LDA Adaptive Topic Models of Mining Text Streams with Applications to Topic Detection and Tracking. Proceeding of the 8th IEEE International Conference on Data Mining. Washington, DC, USA: IEEE Computer Society, 2008:3-12.

[6] X.Wei, J.Sun, X.Wang. Dynamic Mixture Models for Multiple Time Series. Proceedings of the 20th International Joint Conference on Artificial Intelligent. Hyderabad, India, 2007: 2909-2914

[7] C.Wang, D.Blei, D.Heckerman. Continuous Time Dynamic Topic Models. Proceeding of the 23rd Conference on Uncertainty in Artificial Intelligence, Helsinki, Finland, July 2008. 579-586

[8] D.Hall, D.Jurafsky, C.D.Manning. Studying the History of Ideas Using Topic Models. Proceedings of the Conference on Empirical Methods in Natural Lanaguage Processing. Honolulu, Hawaii, 2008, 363-371.

[9] T. L. Griffiths, and M. Steyvers. Finding Scientific Topics. In: Proceedings of the National Academy of Sciences of the Universitates of America. vol. 101, 2004. Pp. 5228-5235.

[10] 楚克明，李芳. 基于 LDA 话题关联的话题演化[J].**上海交通大学学报** 2010 44(11): 1496-1500

CHU Ke-ming, LI Fang. Topic Evolution based on LDA and topic Association[J] **Journal of Shanghai Jiaotong University**. 2010 44(11): 1496-1500.

[11] Yookyung Jo, et al," "The web of topics: discovering the topology of topic evolution in a corpus" in the proceeding of WWW 2011, Hyderabad, India. March 28-April 1, 2011. 257-266.

[12] Dafna Shahaf, Carlos Guestrin, "Connecting the Dots between news articles" in the Proceedings of the Twenty-Second International Joint Conference on Artificial Intelligence, Barcelona, Catalonia, Spain, 16–22 July 2011 2734-2739