文章编号：1003-0077(2015)02-0179-11

基于上下文的话题演化和话题关系抽取研究

章 建,李 芳

(上海交通大学 计算机科学与工程系,上海 200240)

摘 要：自动挖掘大规模语料中的语义信息以及演化关系近年来已受到广大专家学者的关注。话题被认为是文档集合中的潜在语义信息,话题演化用于研究话题内容随时间的变化。该文提出了一种基于上下文的话题演化和话题关系抽取方法。分析发现,一个话题常和某些其他话题共现在多篇文档中,话题间的这种共现信息被称为话题的上下文。上下文信息可以用于计算同时间段话题间的语义关系以及识别不同时间段中具有相同语义的话题。该文对 2008 年～2012 年两会报告以及 2007 年～2011 年 NIPS 科技文献进行实验,通过人工分析,利用话题的上下文信息,不但可以提高话题演化的正确率,而且还能挖掘话题之间的语义关系,在话题演化的基础上,显示话题关系的演化。

关键词：话题;话题上下文;话题演化;话题关系

中图分类号：TP391　　**文献标识码**：A

Context-based Topic Evolution and Topic Relations Extraction

ZHANG Jian, LI Fang

(Department of Computer Science and Engineering, Shanghai Jiao Tong University, Shanghai 200240, China)

Abstract: Automatic extraction of semantic information and its evolution from large-scale corpus has appealed to many experts and scholars in recent years. Topics are regarded as the latent semantic meanings underlying the document collectionand the topic evolution describes the contents of topics changing over time. This paper proposes a novel extraction method for the topics evolution and the topic relations based on the topic context. Since a topic often co-occurs with other topics in the same document, the co-occurrence information is defined as the context of a topic. Topics with its context are used not only to calculate the semantic relations among topics in the same period, but also to identify the same topics across different time periods. The experiments on NPC&CPPCC news reports from 2008 to 2012 and NIPS scientific literature from 2007 to 2011 have shown that the method has not only improved the results of topic evolution but also mined semantic relations among topics.

Key words: topic; topic context; topic evolution; topic relations evolution

当今社会,信息即是财富,如何高效获取信息以及信息动态的变化趋势,是一个值得关注的问题。信息变化趋势可以反映科技领域的发展、新闻事件的变化以及其他任何人们关注的焦点问题的发展。话题[1-3]被认为是普遍关注的信息焦点,例如,“医疗改革”、“行政体制改革”等出现在“全国两会”新闻语料的典型话题。然而话题在不同时间段内可以具有不同的内容,例如,“住房”话题,2011 年主要体现“房价上涨和土地供应”,2012 年则转变为“保障性住房和房价调控”。因此,话题本身会随着时间发生变化,研究话题随时间的变化具有很重要的现实意义和实际应用背景。如何对话题内容的变化进行描述和分析,是本文话题演化研究的目的。

另一方面,话题之间也存在某种关系,如在 2011 年两会报告中,“财政预算”和“三公支出”会在多篇文档中同时被讨论,“司法”和“违法犯罪”也是如此。话题间的这种关系,可以通过它们在文档中共现信息表现出来。共现越频繁的话题,其语义关系也就越强。本文将某话题与其他话题的共现信息,称为该话题的上下文。利用话题的上下文,可以

收稿日期：2013-03-18　**定稿日期**：2014-04-21

基金项目：国家自然科学基金(60873134)

挖掘出同时间段中不同话题间的语义关系，让读者了解到不同信息之间是如何关联的。同时，话题的上下文还可以改进话题演化的结果。

话题内容不但随着时间变化，而且话题之间的关系也随着时间变化，如2010年两会中“教育”与“学术行政化”“大学生就业”“青少年心理健康”等关系较强，而2011年“教育”则与“人才培养”“高考”“财政预算”等关系较强。话题间的关系随时间的变化，能够让读者从更广泛的视角掌握信息的动态趋势。如何挖掘话题关系随时间的变化，是本文话题关系演化研究需要解决的问题。

本文的组织结构如下：第一部分主要介绍相关工作；第二部分是研究方法的描述；第三部分是试验结果和分析；第四部分是结论和展望。

1 相关工作

基于话题模型的话题演化研究已得到了广泛的应用[4]。对多个时间段的文档集合进行话题演化分析时，主要包括两个步骤，即从每个时间段的文档集合中抽取出话题信息以及将各个时间段具有相同语义的话题进行关联。

话题的抽取，即从文档集合中挖掘潜在的语义信息，常用的方法是采用话题模型。目前，已有多种形式的概率话题模型[1]，如PLSI模型[2]，LDA模型[3]等，它们在建模过程中引入了潜在的随机变量—话题。近年来，考虑到文档集合自身的特点，很多研究工作对LDA模型进行扩展，以便模型更好地描述文档的生成过程，如ATM模型[5]引入文档的作者信息，DTM模型[6]引入文档的时间戳信息，STMS模型[7]同时引入文档的作者和时间信息，JST模型[8]引入情感标签，文献[9]中的模型引入文档间的引用信息等。

话题的关联，即将不同时间段中具有相同语义的话题对应起来。实现话题间的对应关系，一般可以采用两种方法，一是在话题建模时直接考虑话题间的对应关系（即前一时间段的话题影响后一时间段的话题），从而在话题抽取的同时也将话题间的对应关系挖掘出来，如DTM模型[6]，CTDTM模型[10]，文献[11]；二是利用关联函数计算不同时间段中话题间的关联度，当两个话题的关联度满足相关阈值时，认为这两个话题具有相同语义，如文献[12-13]。方法一的缺点是，各个时间段的话题数量必须相同且话题间只具有一一对应的关系，其优点是可以在话题建模的过程中直接挖掘出话题间的对应关系；方法二的缺点是，关联函数的选择以及阈值的确定较为复杂，优点是各时间段的话题数量可以根据文档集合的大小进行调整，话题间允许一对多、多对一以及多对多的复杂关联关系。

除了通过话题模型来抽取话题外，还可以采用其它方式，如文献[14]通过文档的新颖性和重要度来判断是否有新话题的产生。而对于话题的关联，文献[14]则基于话题成员文档集合间的交叉引用数量来判断两个话题是否关联。

上述方法都认为同时间段的话题是互相独立，不存在任何关系。然而，现实世界话题之间是存在关系的，某个话题与其它话题在文档集合中存在共现，该共现信息可以作为话题的上下文。在词义消歧方法中，上下文信息可以识别该词汇的语义信息，解决词汇的一词多义问题；命名实体的指代消歧研究中，上下文信息可以用来识别同一命名实体[15-16]，解决命名实体的指代，信息合并等问题。借鉴上述研究领域的思想，在已有工作基础上[13]，本文提出了话题的上下文信息，既可以加强和识别话题本身的语义，有助于话题演化研究，同时，又能揭示同时间段中话题之间的语义关系。

2 研究方法

话题的上下文信息刻画了话题出现的语义环境。如果两个不同时间段中的话题具有相同语义，那么它们的上下文也应具有一定的相似性；相反，如果两个话题的上下文差异明显，那么它们具有相同语义的可能性较小。另一方面共现越频繁的话题，其语义关系越强。因此，本文提出的话题关联方法不仅仅考虑两个话题本身的内容，而且还依据上下文信息。有些话题在内容上较为接近（即在词汇分布上很相似），但实际上并不具有相同的语义，如表1所示的两个话题。

表1 内容相近的两个话题

时间	话题	话题中概率最大的10个词语
2010	16	公务 出国 行政 接待 开支 官员 公款 书记 财务 成本
2011	6	腐败 贪污 官员 渎职 领导 人大代表制度 公开 干部

2010年话题16涉及“三公支出”，而2011年话题6涉及“官员腐败”，这两个话题本身并不具有相

同的语义。但由于这两个话题使用的词语较为接近，如官员、干部、行政等，仅仅通过计算这两个话题内容（即在词汇上的分布）的距离，容易将这两个话题识别为同义性话题。分析发现，话题 16 的上下文有关财政预算，而话题 6 的上下文是违法犯罪等法律相关的。因此考虑话题的上下文信息，可以有助于判断这两个话题不具有相同语义。表 2 和表 3 分别列出话题 16 和话题 6 的上下文信息。

表 2　2010 年话题 16 的上下文

权重	话题	话题中概率最大的 10 个词语
1.0	2	财政 预算 中央 支出 投入 资金 地方 增加 安排 政策

表 3　2011 年话题 6 的上下文

权重	话题	话题中概率最大的 10 个词语
0.500	21	监督 公开 批评 群众 条件 官员 权力 财产 信息 创造
0.259	53	司法 监督 法院 公正 机关 执法 法官 执行 最高 制度
0.241	28	犯罪 案件 最高 刑法 打击 检察院 调解 机关 纠纷 职务

下面列出本文主要使用的符号（见表 4），概念定义以及方法介绍。

表 4　本文主要使用的符号

符号	符号描述
τ	语料、话题或文档的时间戳信息
T_d	文档 d 的显著性话题
t	*LDA* 模型中的某个话题
φ	话题的内容，即话题在词汇上的多项式分布
D	文档集合
$t_{i,j}$	话题 t_i 上下文中的第 j 个话题
$w_{i,j}$	话题 t_i 上下文中第 j 个话题的权重

2.1 概念定义

话题：话题描述了文档集合中潜在的语义信息。本文话题的表示是三元组 $\{\tau,\varphi,C\}$，其中 τ 表示话题出现的时间，φ 表示话题的内容（表示为在词汇上的多项式分布），C 表示话题的上下文信息。

话题的上下文：对于某话题 t_k 来说，它与同时间段中其他话题在文档集合中的共现信息，称为它的上下文，具体表示为 $\{(w_{k,1},t_{k,1}),(w_{k,2},t_{k,2}),\cdots,(w_{k,m},t_{k,m})\}$。其中，$\{t_{k,1},t_{k,2},\cdots,t_{k,m}\}$ 为与 t_k 存在共现的话题集合，而 $\{w_{k,1},w_{k,2},\cdots,w_{k,m}\}$ 为它们的权重，其中权重越大，表明与话题 t_k 的共现次数越多。

话题演化：将不同时间段中具有相同语义的话题进行关联，进而发现它们内容随时间的变化情况，该过程即被称为话题演化。每个话题演化的结果可以由一条或多条演化路径表示，演化路径的具体形式为 $\{(\tau_1,\varphi_1),(\tau_2,\varphi_2),\cdots,(\tau_n,\varphi_n)\}$。

同义性话题：不同时间段中具有相同语义的话题。如两会报告中 2011 年的“住房”话题与 2011 年的“住房”话题，虽然它们侧重的内容有所不同，但它们本质上都是住房相关的，因此它们具有相同的语义。

2.2　话题建模

LDA 模型是一个生成概率模型，同时也是三层的变参数贝叶斯模型[17]。首先假设词由话题的概率分布混合产生，而每个话题是在词汇表上的一个多项式分布；其次，假设文档是潜在话题的概率分布的混合；最后，针对每篇文档从 Dirichlet 分布[18]中抽样产生该文档中的话题比例，结合话题和词的概率分布生成该文档中的每一个词汇。在进行 LDA 建模时，可以根据文献[19]提出的方法确定话题的个数以及利用 Gibbs Sampling 算法[20]对 LDA 模型的参数进行推导，从而得到每个话题的内容。

首先按照时间划分，然后对不同时间段的文档集合进行 *LDA* 建模。在对不同时间段的文集建模时，话题数量可以设置为相同，也可以不同。进行 LDA 建模后，可以直接得到各个话题的内容，即话题在词汇上的多项式分布，即话题的 φ。

2.3　话题上下文抽取

在话题模型中，每篇文档表示为话题的混合分布，其中那些权重高的话题称为文档的显著性话题。如果两个话题同出现在某篇文档的显著性话题中，则称这两个话题存在一次共现。共现次数越多的话题，可以认为它们的语义关系越强。对于某个话题来说，它与同时间段中其他话题的共现信息，称为它的上下文。抽取话题上下文由以下三个步骤完成。

(a) 计算文档的显著性话题

对于每一篇文档，将话题的权重按降序排列，然后取出权重最大的 3 个话题作为文档的显著性话题。根据实验结果，权重最大的 3 个话题通常比其

他话题的权重明显大,这符合常理。

(b) 计算任意两个话题间的共现次数

对于任意两个话题 t_i 和 t_j,它们的共现次数可以表示为公式(1)。

$$Cooccurrence(t_i,t_j)=\sum_{d\in D}\delta(t_i\in T_d\ and\ t_j\in T_d) \tag{1}$$

而 δ 函数可以表示为

$$\delta(boolVal)=\begin{cases}1, & if\ boolVal\ is\ true\\ 0, & if\ boolVal\ is\ false\end{cases}$$

(c) 计算各个话题的上下文

对于任意话题 t_k,其上下文可以形式化地表示为(其中 m 为 t_k 上下文中话题的个数)

$$Context(t_k)=\{(w_{k,1},t_{k,1}),(w_{k,2},t_{k,2}),\cdots,(w_{k,m},t_{k,m})\}$$

假设 $t_{k,j}\in\{t_{k,1},t_{k,2},\cdots,t_{k,m}\}$,其权重 $w_{k,j}$ 可以采用如下公式(2)计算得到

$$w_{k,j}=\frac{Cooccurrence(t_k,t_{k,j})}{\sum_{t_{k,l}\in\{t_{k,1},t_{k,2},\cdots t_{k,m}\}}Cooccurrence(t_k,t_{k,l})} \tag{2}$$

2.4 话题演化

话题演化需要判断两个不同时间段的话题是否具有相同的语义,即同义性话题。对于两个不同时间段的话题 t_i 和 t_j,其距离计算公式如式(3)所示。

$$Distance(t_i,t_j)=\beta\cdot Distance_T(\varphi_{t_i},\varphi_{t_j})+(1-\beta)\cdot Distance_C(Context(t_i),Context(t_j)) \tag{3}$$

其中,β 为选择因子,$Distance_T$ 为计算两个话题词汇分布差异的距离函数,$Distance_C$ 为计算话题上下文差异的距离函数。预先设定阈值 γ,当 $Distance(t_i,t_j)$ 小于 γ 时,认为话题 t_i 和 t_j 是同义性话题。而上下文的距离采用计算公式(4)。

$$Distance_C(Context(t_i),Context(t_j))=Distance_M(ContextCenter(t_i),ContextCenter(t_j)) \tag{4}$$

其中,$Distance_M$ 为计算两个多项式分布差异的距离函数,而 $ContextCenter(t_k)$ 表示的是 t_k 上下文中的所有话题在词汇上的加权多项式分布,其计算公式如式(5)所示。

$$ContextCenter(t_k)=\sum_{t_{k,j}\in\{t_{k,1},t_{k,2},\cdots,t_{k,m}\}}w_{t,j}\cdot\varphi_{t,j} \tag{5}$$

其中,$\{t_{k,1},t_{k,2},\cdots,t_{k,m}\}$ 表示 t_k 上下文中的话题集合,$w_{t,j}$ 为上下文中对应话题的权重,而 $\varphi_{t,j}$ 则为上下文中对应话题的内容,即在词汇上的多项式分布。

2.5 话题关系的抽取

事实上,同时间段中不同话题间存在一定的语义关系。借鉴词汇共现与语义关系,共现次数越多的话题,其语义关系也就越强。对于任意两个话题 t_i 和 t_j,其语义关系强度表示为式(6)

$$SemanticStrength(t_i,t_j)=\frac{2\cdot w_{i,j}\cdot w_{j,i}}{w_{i,j}+w_{j,i}} \tag{6}$$

即关系强度由 $w_{i,j}$ 和 $w_{j,i}$ 的调和平均值表示,其中 $w_{i,j}$ 和 $w_{j,i}$ 的计算可见公式(2)。

根据公式(6)可以计算出各时间段中话题间的语义关系强度,而根据公式(3)则能获得同义性话题。因此,结合这两者,便可以得到同义性话题与其它话题的关系随时间的变化,即话题关系的演化。

3 实验结果

本文选取的实验数据为两会报告(人大会议和政协会议,2008~2012年)的新闻语料以及NIPS科技文献(2007~2011年)。这是因为两会报告和NIPS科技文献中,很多话题会连续多年被讨论且话题内容随时间变化。选取不同领域、不同特点的语料作为实验对象有利于更全面地对本文提出的话题演化和话题关系抽取方法进行验证。

首先对实验数据进行语料预处理,包括分词,过滤停用词,去除低频词和高频词等,然后再利用LDA模型对各年的两会文集进行话题建模。表5列出实验数据及话题个数设置。实验包括三部分:1)验证话题上下文信息抽取的精度;2)基于话题上下文的演化对比实验;3)话题关系的抽取结果以及分析。

表5 语料信息及话题个数设置

	两会报告					NIPS科技文献				
	2008	2009	2010	2011	2012	2007	2008	2009	2010	2011
文档数目	4 340	4 186	2 999	2 970	3 097	201	216	244	260	286
词汇数目	25 785	22 286	17 417	16 473	19 580	17 395	18 237	19 665	20 890	21 949
话题个数	60	60	60	60	60	40	40	40	40	40

3.1　话题上下文抽取实验

在计算每个话题的上下文之前，需要对建模后的话题进行一定的过滤：首先通过信息熵过滤那些词汇权重分布均匀的话题，这类话题可解释性较差；其次，过滤高频话题，这类话题出现在很多文档中，语义特征不强。表 6 列出的是从两会报告和 NIPS 科技文献中抽取的上下文实验结果。

表 6　话题上下文实验结果

	两会报告					NIPS 科技文献				
	2008	2009	2010	2011	2012	2007	2008	2009	2010	2011
具有上下文的话题个数	37	38	37	36	41	21	21	30	29	28
正确个数	15	13	19	15	14	8	7	13	12	11
部分正确个数	18	20	16	18	20	11	12	14	15	16
错误个数	4	5	2	3	7	2	2	3	2	1

正确的上下文表明话题上下文中的所有话题与该话题都具有明显的语义关系（通过人工分析）；部分正确的上下文则表明话题上下文中存在部分话题与该话题有明显的语义关系，而另一部分则没有；错误的上下文则表明话题上下文中的话题都与该话题没有明显的语义关系。表 7 和表 8 列出的分别是从 2011 年两会报告和从 2011 年 NIPS 科技文献中抽取的正确、部分正确以及错误上下文的实例。

表 7　2011 年两会报告中上下文实例

实例类型	话题	话题中概率最大的 10 个词语
正确	30	教育 孩子 学生 高考 子女 招生 北京 自主 学校 学习
	15	人才 培养 大学 学校 教育 学生 教师 专业 进行 学术
	19	教育 教师 农村 投入 经费 公平 资源 学校 职业 幼儿园
	58	就业 农民工 大学生 创业 解决 毕业生 劳动力 城市 用工
部分正确	9	能源 技术 核电 太阳能 发电 汽车 清洁 产业 董事长 世界
	59	环境 减排 节能 指标 污染 保护 排放 完成 约束性 总量
	8	价格 物价 上涨 稳定 市场 政策 调控 影响 因素 国际
错误	34	建筑 工程 城市 存在 北京 进行 部门 住宅 解决 天津
	7	文化 传统 民族 电影 精神 产业 历史 少数民族 艺术 形象

表 8　2011 年 NIPS 科技文献中上下文实例

实例类型	话题	话题中概率最大的 10 个词语
正确	5	functional brain subjects regions subject fmri voxels connectivity spatial units
	17	neurons signal network neuron neural activity spike figure input signals stimulus
	8	graph graphs network matrix edges nodes distance edge random laplacian vertices
	39	image scene object images objects regions annotation segmentation spatial
部分正确	36	lasso sparse regularization group regression sparsity norm selection groups penalty
	30	matrix matrices rank entries columns random gradient trace-norm error low-rank
	23	policy reward state value agent action function actions features policies
错误	10	kernel kernels motion mkl norm compact svm metric basis human
	8	graph graphs network matrix edges nodes distance edge random laplacian

实验结果表明上下文话题中部分正确的所占比例较大，这主要与显著性话题选取的阈值有关，有些文档显著性话题会少于 3 个，而选取权重最大的前 3 个话题作为显著性话题则会引入误差。但总体上

话题的上下文能够反映出话题之间的语义相关性，例如对于表7中话题9，其上下文中的话题59和话题8的权重分别为0.863和0.137。因此，根据实验结果，结合权重，话题的上下文能够描述与该话题的语义关系。

3.2 话题演化实验

本文在进行话题演化时，计算话题词汇分布差异的距离函数 $Distance_T$ 以及计算话题上下文差异的函数 $Distance_M$ 都采用 KL 距离函数，同时选择因子 β 设为0.7，关联阈值 γ 设为2.0（即两话题的距离小于2.0时，认为具有相同的语义）。

作为本文话题演化对比的方法一（简称基准一），在计算两话题的距离时，仅仅利用两话题间的词汇分布差异，不考虑话题上下文，计算两话题的距离公式为(7)。

$$Distance(t_i,t_j)=Distance_T(\varphi_{t_i},\varphi_{t_j}) \quad (7)$$

同样，基准一中 $Distance_T$ 采用 KL 距离函数，话题间的关联阈值设为2.0。

作为本文话题演化对比的方法二，则为DTM话题模型[6]，该模型以前一时间段的分布参数作为后一时间段正态分布的先验，在建模过程中直接挖掘不同时间段的同义性话题。DTM代码实现来自网页信息①。

表9是基准一方法，DTM方法和本文方法得到的话题演化对比结果。

表9 话题演化实验结果对比

	两会报告			NIPS科技文献		
	基准一	DTM	本文方法	基准一	DTM	本文方法
演化路径总数	81	60	97	45	40	68
正确个数	72	60	92	39	40	61
错误个数	9	0	5	6	0	7
正确率/%	88.9	100	94.8	86.7	100	89.7

实验结果表明，DTM模型得到的演化路径都是正确的，话题一一对应，但不能很好地刻画话题内容随时间的变化（见下文演化实例）。而本文方法比基准一方法不但能找到更多的演化结果，而且提高了精度。例如，对于2010年的话题8{教育 教师 学生 纲要 人才 学校 培养 高考 公平 考试 资源}，通过基准一和本文方法计算2011年中的同义性话题，如表10所示。

表10 2011年中与2010年话题8关联的话题

方法	话题	距离	话题中概率最大的10个词语
基准一方法	19	1.636	教育 教师 农村 投入 经费 公平 资源 学校 职业 幼儿园
本文方法	19	1.513	教育 教师 农村 投入 经费 公平 资源 学校 职业 幼儿园
本文方法	15	1.880	人才 培养 大学 学校 教育 学生 教师 专业 进行 学术
本文方法	30	1.969	教育 孩子 学生 高考 子女 招生 北京 自主 学校 学习

从上面的结果中可以看到，2011年中的话题15和话题30都涉及到教育，因此它们同2010年的话题8是具有相同语义，建立了演化关系。同时，增加上下文信息后，距离公式更加精确，例如，话题19、话题15和话题30距离有所减小（基准一方法中，这三者的距离分别为1.636、2.154、2.236）。因此，引入上下文后，使得同义性话题的计算受不同时间段词汇变化的影响减小，同时受阈值 γ 的影响变小。

图1是分别采用本文方法、基准一方法和DTM模型得到与教育相关的话题演化实例。DTM模型得到的演化路径{47，47，47，47，47}中各话题是一一对应，没有话题分裂和合并；本文方法和基准一方法均获得演化路径{17，20，8，19，38}（学生教育话题的演化），和演化路径{17，22，39，58，44}（学生就业话题的演化）。本文方法还能得到演化路径{17，20，8，30，47}（与学生考试相关）和演化路径{17，20，0，15，16}（与人才培养相关），反映话题在更细粒度上的分裂与合并关系（话题内容见表11与表12）。

图2是分别采用本文方法、基准一方法和DTM模型得到与神经元相关的话题演化实例，表13和表14显示了各个话题的内容。DTM得到一条话题内容非常相似的演化路径，基准一方法同样得到一条演化路径，可以反映话题内容的演化。本文方法采用上下文，计算出话题37（涉及神经元）和话题16具有相同的语义。

① http://code.google.com/p/princeton-statistical-learning/downloads/detail?name=dtm_release-0.8.tgz

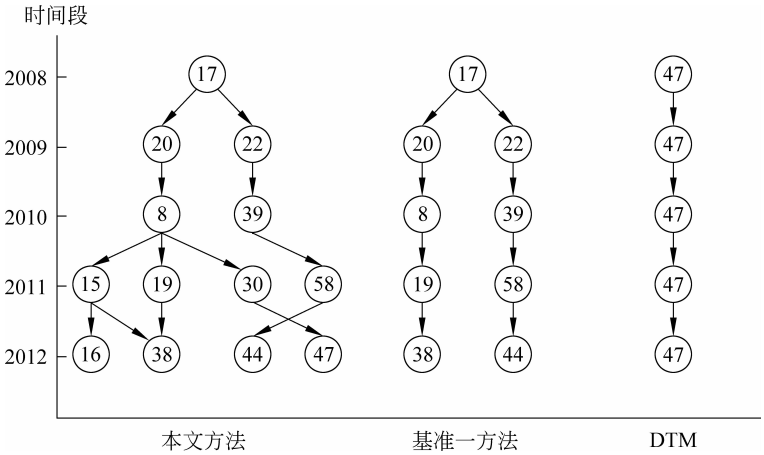

图 1　两会报告演化实例

表 11　图 1 中本文方法和基准一方法各话题的内容

时间	话题	话题中概率最大的 10 个词语
2008	17	学生 学校 教师 人才 素质 大学 培养 职业 农村 大学生
2009	20	教育 学生 教师 农村 学校 孩子 职业 文理 经费 投入
2009	22	大学生 毕业生 人才 创业 培养 大学 教育 就业 学生 岗位
2010	8	教育 教师 学生 纲要 人才 学校 培养 高考 公平 考试
2010	39	就业 培训 毕业生 大学生 创业 职业 人才 企业 岗位 鼓励
2011	15	人才 培养 大学 学校 教育 学生 教师 专业 进行 学术
2011	19	教育 教师 农村 投入 经费 公平 资源 学校 职业 幼儿园
2011	30	教育 孩子 学生 高考 子女 招生 北京 自主 学校 学习 家长
2011	58	就业 农民工 大学生 创业 解决 毕业生 劳动力 城市 用工 培训
2012	16	创新 人才 科技 技术 企业 培养 能力 科研 知识 我国 科学
2012	38	教育 学生 学校 孩子 职业 投入 校长 教师 资源 培养 大学
2012	44	农民工 就业 养老 保险 制度 企业 人员 服务 劳动 职工 生活
2012	47	高考 北京 户籍 公平 子女 参加 异地 政策 教育 招生 城市

表 12　图 1 中 DTM 模型各话题的内容

时间	话题	话题中概率最大的 10 个词语
2008	47	学生 学校 孩子 教师 教育 家庭 培养 儿童 高考 教育部 子女
2009	47	教育 学生 学校 孩子 教师 家庭 培养 儿童 高考 子女 妇女
2010	47	教育 学生 孩子 学校 教师 家庭 培养 儿童 高考 子女 妇女
2011	47	教育 孩子 学生 学校 教师 儿童 培养 家庭 高考 子女 妇女
2012	47	教育 孩子 学生 学校 高考 培养 教师 儿童 家庭 子女 建议

表 13　图 2 中本文方法和基准一方法各话题的内容

时间	话题	话题中概率最大的 10 个词语
2007	16	neurons neuron spike input firing rate network output mean stimulus
2008	7	input neurons neuron learning output synapses rule spike network figure

续表

时间	话题	话题中概率最大的 10 个词语
2008	37	spike neurons stimulus response neural disparity neuron population cells
2009	13	spike cells model neurons network cell information neuron activity figure
2010	4	neurons spike neuron network population neural stimulus input synaptic noise
2011	17	neurons signal network neuron neural activity spike figure input signals

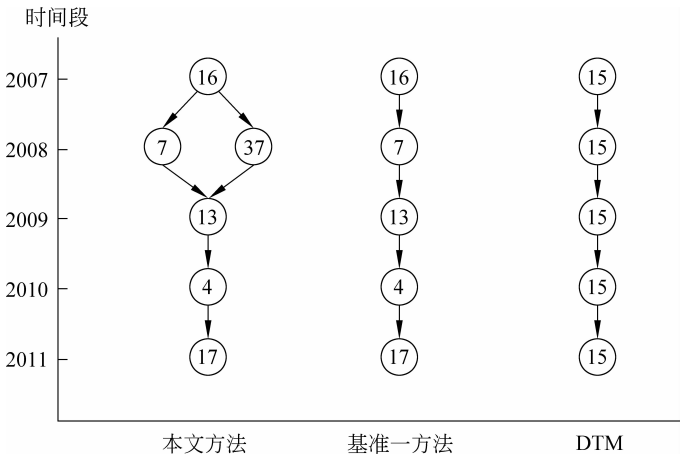

图 2 NIPS 科技文献演化实例

表 14 图 2 中 DTM 模型各话题的内容

时间	话题	话题中概率最大的 10 个词语
2007	15	figure neurons input rate data network spike noise neuron response
2008	15	figure neurons input data spike rate network noise neuron response
2009	15	figure noise network data neurons spike input control response neural
2010	15	figure noise network neurons data neural spike input response stimulus
2011	15	Figure network signal noise neurons data input spike response neural

3.3 话题关系的抽取与演化实验

由公式(6)可知,本文话题关系的抽取依赖于话题上下文计算的结果。对于某个话题来说,如果它的上下文正确,则形成的话题关系也正确;如果上下文部分正确,得到的话题关系也部分正确;如果上下文错误,得到的话题关系也错误。为了更直观地显示话题关系的抽取结果,这里将正确、部分正确和错误的话题关系分别给予数值 1、0.5 和 0。表 15 是话题关系抽取的结果(等价正确个数=正确个数+0.5 * 部分正确个数),话题关系抽取的等价正确率在 60% 以上。

表 15 话题关系抽取实验结果

	两会报告					NIPS 科技文献				
	2008	2009	2010	2011	2012	2007	2008	2009	2010	2011
话题语义关系总个数	37	38	37	36	41	21	21	30	29	28
等价正确个数	24	23	27	24	24	13.5	13	20	19.5	19
等价正确率/%	65	61	73	67	59	64	62	67	67	68

图 3 所示为 2009 年两会报告中与话题 20 有关的语义关系图，话题间连线上的数字代表了关联的强度(公式 6)。根据图 3，话题 20 同话题 22、48 的语义关系最强，与话题 12 ，45，31 语义关系较弱，各话题内容见表 16。

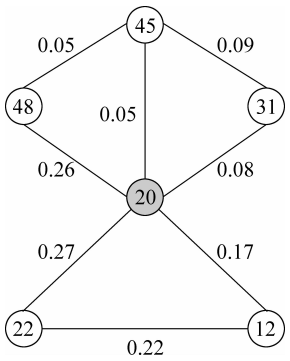

图 3　2009 年话题 20 的关系图

话题之间的关系也会随时间发生变化。选取图 1 中的一条演化路径{17，20，8，19，38}(有关学生教育)，分别计算演化路径中各话题同其他话题的关系，结果见图 4，对应话题的内容见表 17。2008 年，教育话题与话题 50(民族艺术)关系相对较强(强度仅为 0.14)；2009 年，教育话题与话题 22(大学生就业)、话题 48(学术腐败)关系较强；2010 年，教育话题与话题 48(学术行政化)、话题 39(大学生就业)关系强度分别为 0.44 和 0.37，体现了很强的关系。根据表 17 中话题内容可知，2011 年教育话题与话题 15(人才培养)关联，2012 年教育话题和话题 47(异地高考)以及话题 16(创新人才培养)具有比较强的关联。因此，通过演化路径上话题与其他话题的语义关系，反映了目前大学教育众多相关话题随时间的变化，对比话题的一条演化路径({17，20，8，19，38}(话题内容见表 11)，可以传递更多的内容信息。

表 16　图 3 中各话题的内容

话题	话题的内容
20	教育 学生 教师 农村 学校 孩子 职业 文理 经费
22	大学生 毕业生 人才 创业 培养 大学 教育 就业
48	校长 大学 学术 院士 教授 腐败 认为 学生 北大
12	农民工 培训 返乡 城市 创业 失业 解决 劳动力
45	制度 法律 管理 制定 进行 规定 行政 我国 立法
31	保险 养老 保障 制度 退休 单位 人口 改革 职工
22	行政 学术 行政化 级别 教授 大学 学校 校长

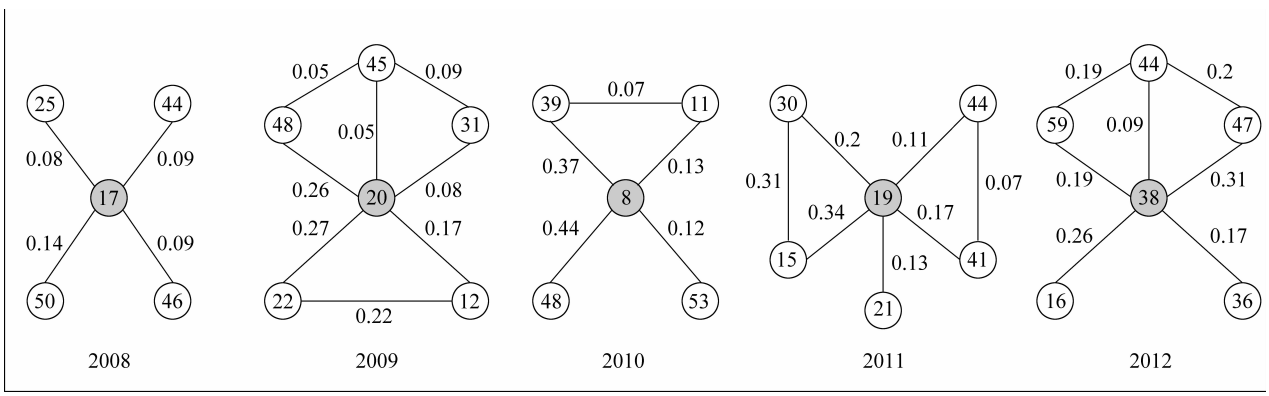

图 4　两会报告话题关系演化实例

表 17　图 4 中部分话题的内容

时间	话题	话题中概率最大的 10 个词语
2008	50	传统 京剧 艺术 民族 作品 作家 精神 语言 演出 创作
2010	48	行政 学术 行政化 级别 教授 大学 学校 校长 权力 认为
2010	39	就业 培训 毕业生 大学生 创业 职业 人才 企业 岗位 鼓励
2011	15	人才 培养 大学 学校 教育 学生 教师 专业 进行 学术
2011	30	教育 孩子 学生 高考 子女 招生 北京 自主 学校 学习
2012	47	高考 北京 户籍 公平 子女 参加 异地 政策 教育 招生 城市
2012	16	创新 人才 科技 技术 企业 培养 能力 科研 知识 我国

同样，在NIPS数据集上，选取图2中的演化路径{16，7，13，4，17}(有关神经元)，分别计算演化路径中各话题同其他话题的关系，可以得到神经元话题同其他话题的语义关系随时间的变化，如图5所示，话题内容见表18。2007年神经元话题同话题34(图像分割)和话题3(脑信号噪声处理)关系较强；2008年，神经元话题同话题37(神经元)和话题11(脑成像)关系较强；2009年，神经元话题同话题31(人类学习记忆)和话题20(神经元模型)关系较强。这些关系体现了有关神经元技术在近几年的发展以及它与图像处理技术，脑信息处理等的关系。

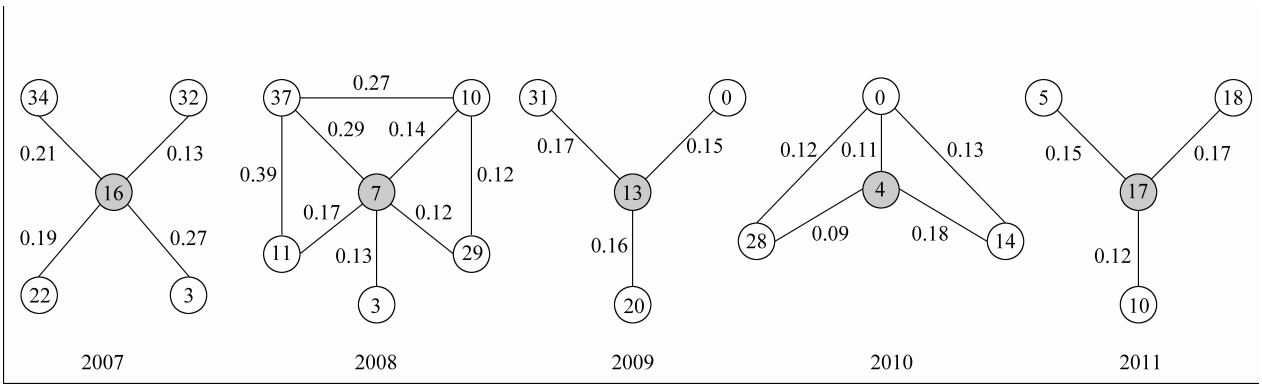

图5 NIPS科技文献话题关系演化实例

表18 图5中部分话题的内容

时间	话题	话题中概率最大的10个词语
2007	34	image segmentation images model ica layer blur gaussian independent local
2007	3	sources source data noise components mixing mixture signals component brain
2008	37	spike neurons stimulus response neural disparity neuron population cells
2008	11	basis data natural brain sparsity images sparse fmri components stimuli
2009	31	model learning task human stimulus figure subjects stimuli memory response
2009	20	distribution spike copula independent model joint neurons data poisson
2010	14	noise model motion contrast phase figure depth visual objects spatial
2011	5	functional brain subjects regions subject fmri voxels connectivity spatial

4 结论和展望

本文提出了一种基于上下文的话题演化和话题关系抽取的方法。首先利用LDA话题模型对各时间段的文档集合进行建模，挖掘潜在的语义信息，即话题。然后通过话题在文档中的共现关系，找到各个话题的上下文。其次，利用上下文信息改进了不同时间段同义性话题的计算，实现话题演化。最后，利用话题的上下文挖掘不同话题间的语义关系，同时结合话题演化的结果，还能得到话题关系在时间上的演化。

本文对两会报告和NIPS科技文献进行实验，结果表明利用上下文信息计算同义性话题，可以获得比基准一方法更多正确的演化结果，同时还能识别因词语使用接近但并非具有相同语义的话题。而与DTM模型相比，采用本文方法进行话题演化，可以得到话题的分裂、合并等复杂的对应关系，且能够较好地反映出话题内容随时间的变化。同时，利用上下文信息还能够挖掘出同时间段中不同话题间的语义关系，在结合话题演化的情况下，还可得到话题关系随时间的演化。本文的主要贡献是：

1) 提出了文档集合话题的上下文概念，并根据话题在文档中的共现，计算话题的上下文；

2) 利用话题的上下文，正确识别不同时间段同义性话题，从而改进了话题演化的结果；

3) 提出了话题之间计算其语义关系强度的公式，挖掘同时间段中话题间的语义关系。

本文提出方法还存在不足，如文档显著性话题个数的选择，如何动态确定某一文档的显著性话题个数，在引入话题的上下文信息的同时，删除其带来的噪音；另一方面，同义性话题计算方法中阈值的确定，如何更合理地权衡话题本身的语义信息与话题的上下文信息在话题演化中的重要性，还需大量的实验结果进行验证。LDA 话题的标签如何自动生成以及演化结果的可视化技术将有助于本文提出方法的广泛应用。

参考文献

[1] Steyvers M, Griffiths T. Probabilistic Topic Models. In: T. Landauer, D. S. McNamara, S. Dennis, W. Kintsch (Eds.), handbook of Latent Semantic Analysys[M]. Hillsdale, NJ. Erlbaum. 2007.

[2] Thomas H. Probabilistic Latent Semantic Indexing// Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. Berkeley, CA, USA, 1999: 50-57.

[3] David M B, Andrew Y N, Michael I J. Latent Dirichlet Allocation. The Journal of Machine Learning Research, 2003, 3: 993-1022.

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

[5] Michal R Z, Thomas G, Mark S, et al. The Author Topic Model for Authors and Documents[C]//Proceedings of the 20th Conference on Uncertainty in Artificial Intelligence. Banff, Canada,2005.

[6] David M B, John D L. Dynamic Topic Models[C]// Proceedings of the 23rd International Conference on Machine Learning, Pittsburgh, Pennsylvania, 2006: 113-120.

[7] Ali D, Li Juanzi, Zhou Lizhu, et al. A Generalized Topic Modeling Approach for Maven Search. APWeb/WAIM 2009, LNCS 5446, 2009: 138-149.

[8] Chenghua Lin, Yulan He. Joint Sentiment/Topic Model for Sentiment Analysis[C]//Proceedings of the CIKM'09, Hong Kong, China, 2009.

[9] R. Nallapati, A Ahmed, E P Xing. Joint Latent Topic Models for Text and Citations[C]//Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. Las Vegas, Nevada, USA, 2008: 542-550.

[10] Chong Wang, David M, David H. Continuous Time Dynamic Topic Models[C]//Proceedings of the 23rd Conference on Uncertainty in Artificial Intelligence, 2008.

[11] Andre G, Alexander H. Topic evolution in a stream of documents[C]//Proceedings of the Ninth SIAM International Conference on Data Mining, 2009: 859-870.

[12] Mei Qiaozhu, Zhai Chengxiang. Discovering Evolutionary Theme Patterns from Text—An Exploration of Temporal Text Mining[C]//Proceedings of the KDD'05, Chicago, Illinois, USA,2005.

[13] 楚克明，李芳. 基于 LDA 话题关联的话题演化. 交大学报,2010:11:1496-1500.

[14] Jo Y Y, John E H, Carl L. The Web of Topics: Discovering the Topology of Evolution in a Corpus[C]// Proceedings of the WWW 2011, Hyderabad, India 2011.

[15] Xianpei Han, Le Sun. A Generative Entity-Mention Model for Linking Entities with Knowledge Base [C]//Proceedings of the ACL 2011. 2011: 945-954.

[16] Xianpei Han, Jun Zhao. Structural Semantic Relatedness: A Knowledge-Based Method to Named Entity Disambiguation[C]//Proceedings of the 48th Annual Meeting of the Association of Computational Linguistics, 2010: 50-59.

[17] Blei D, Jordan M, Ng A. Hierarchical Bayesian Models for Applications in Information Retrieval. In Bayesian Statistics, 2003,7: 25-44.

[18] Antoniak C. Mixtures of Dirichlet Processes with Applications to Bayesian Nonparametric Problems. Annals of Statistics, 1974,2(6): 1152-1174.

[19] Thomas L. G., Mark S. Finding Scientific Topics [C]//Proceedings of the National Academic of Science of United States of America, 2004.

[20] Ian P, David N, Alexander I. Fast Collapsed Gibbs Sampling For Latent Dirichlet Allocation. KDD'08, Las Vegas, Nevada, USA 2008.

章建(1987—)，硕士，主要研究领域为话题探测与话题演化。
E-mail: iamorchid@hotmail.com

李芳(1963—)，博士，副教授，主要研究领域为自然语言处理，信息检索与抽取。
E-mail: fli@sjtu.edu.cn