文章编号: 1006-2467(2010)11-1496-05

基于 LDA 话题关联的话题演化

楚克明, 李 芳
(上海交通大学 电子信息与电气工程学院, 上海 200240)

摘 要: 话题演化可以帮助人们快速获取信息和了解趋势. 提出了一种挖掘话题随时间变化的方法, 通过话题抽取和话题关联实现话题的演化. 对不同时间段的文集进行话题的自动抽取, 话题数目在不同时间段是可变的; 计算相邻时间段中任意 2 个话题的分布距离和话题的特征向量相似度实现话题的关联. 实验结果证明, 该方法不但可以描述同一个话题随时间的强度变化, 还可以描述新话题的产生, 旧话题的消失以及话题内容随时间的演化.

关键词: 话题探测; 话题关联; 话题演化; 潜在狄里特里分配
中图分类号: TP 391 文献标志码: A

Topic Evolution Based on LDA and Topic Association

CHU Ke-Ming, LI Fang
(School of Electronic, Information and Electrical Engineering, Shanghai Jiaotong University, Shanghai 200240, China)

Abstract: Topic evolution will help people to learn information quickly. In this paper, a method was proposed to discover topic's evolution over time by topic detection and relating topics in different time periods. The method applies LDA model on temporal documents to extract topics. The number of topics in different time periods is different. Relating topics in consecutive time periods is based on Jensen-Shannon divergence and features similarity. Experiments show that the method can detect new topics and describe topic's evolution over time effectively. It not only shows that the topics evolve with time, but also that the content of topics change with time.

Key words: topic detection; topic association; topic evolution; latent Dirichlet allocation (LDA)

在传统话题探测与跟踪(Topic Detection and Tracking, TDT)研究中, 新闻话题被定义为一个种子事件引起的若干相关新闻事件的集合. 本文研究的新闻话题不但包括"汶川地震"这样的突发事件, 也包括具有主题的事件, 泛指所有新闻报道涉及的话题. 话题演化反映了某一个话题从它的提出、热议、趋冷、消失的过程. 随着时间的延续, 话题内容的焦点也会发生迁移, 如何描述话题的演化是本文的研究目的.

潜在狄里特里分配(Latent Dirichlet Allocation, LDA)模型[1]是近年来在机器学习领域提出的一个话题模型, 已广泛应用在话题的趋势变化研究中[2-5]. 基于 LDA 的话题演化主要采用 3 种方法: ①时间演化话题(Topic Over Time, TOT)[2], 该模

收稿日期: 2010-05-11
基金项目: 国家自然科学基金资助项目(60873134)
作者简介: 楚克明(1985-), 男, 硕士, 山东菏泽市人. 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, 副教授, 电话(Tel.): 021-34205423; E-mail: fli@sjtu.edu.cn.

型把连续的时间信息引入生成模型中, 表征话题的变化趋势, 但无法对新文档进行扩展, 必须重新建模. ②认为时间是离散的, 在建模前, 根据时间段划分文档. 例如, 动态话题(Dynamic Topic Model, DTM) [4] 和在线话题模型(Online-LDA, OLDA) [5]. ③在整个文集上应用 LDA 模型, 然后, 根据文档的时间戳划分子集, 通过话题在这个子集上的分布来刻画该时间段的话题[3], 也称为 LDA 后离散方法. 上述方法注重话题随时间强度的变化, 并假设话题数目在不同时间段是不变的, 很难反映话题的出现与消失.

实际上, 话题的语义(话题本身的内容) 随时间会发生变化. 本文把话题的演化研究分为 2 部分: 话题强度和话题内容的变化. 本文提出的基于 LDA 话题关联的话题演化研究, 在不同时间段抽取不同的话题, 然后计算话题关联, 显示其演化关系. 该方法不但描述话题强度的变化, 还能够发现话题在内容上的变化.

1　话题演化模型

1.1　话题的定义

在 LDA 中, 话题被定义为一组语义上相关的词并用该词与话题相关的权重来表示. 在时刻 t, 话题 j 可以表示为

$$Z_j^t = \{(v_1, P(v_1 \mid z_j)), (v_2, P(v_2 \mid z_j)), \cdots, (v_n, P(v_n \mid z_j))\}$$

其中, $v_i \in V$, V 为总词汇表; $P(v_i \mid z_j)$ 为话题 j 选择词 v_i 的概率, 通常取 $P(v_i \mid z_j)$ 较大的若干个词作为话题 Z_j^t 的特征词组表示话题 Z_j^t.

1.2　话题的抽取以及话题数目的确定

首先确定一个时间间隔 δ, 根据文档的时间戳和 δ 来划分文档. 然后, 对每个 δ 内的所有文档应用 LDA 模型进行话题抽取, 话题数目 K 必须事先确定. 根据实际情况, 每个时间段文档数目不同, 因此, 不同时间段中话题数目随时间动态变化. 本文根据文献[6] 中方法选取使得模型最优的话题数目.

1.3　话题的关联与演化

话题的演化关系是指在相邻时间的话题中存在着同一和关联关系, 则称这些话题随着时间的变化存在演化关系. 判断话题的关联以及话题的同一性是研究演化关系的关键. 话题的同一性是根据话题特征词组的相似度衡量; 话题和后续话题的关联则通过话题的语义关联度来衡量.

假设相邻时间 t_i 和 t_{i+1} 的文集上经 LDA 模型抽取得到的话题为 $Z_r^{t_i}$、$Z_s^{t_{i+1}}$, q 和 p 分别是 $Z_r^{t_i}$ 和 $Z_s^{t_{i+1}}$ 在 V 上的概率分布, f_r、f_s 分别是话题 $Z_r^{t_i}$、$Z_s^{t_{i+1}}$ 的特征词组, 则使用 p 和 q 的 Jensen-Shannon divergence 公式, 可计算话题 $Z_r^{t_i}$ 和 $Z_s^{t_{i+1}}$ 的语义关联度:

$$S(Z_r^{t_i}, Z_s^{t_{i+1}}) = -\mathrm{JS}(p, q) = -\frac{1}{2}[D(p \| m) + D(q \| m)] \quad (1)$$

式中:

$$m = \frac{1}{2}(p + q), \quad D(p \| q) = \sum_i^{|V|} p_i \log \frac{p_i}{q_i}$$

语义上关联的话题并不一定存在演化关系, 所以需要计算话题的同一性. 由于话题的特征词都是在 V 空间上的, 因此, 特征的相似度使用余弦公式进行计算:

$$F(Z_r^{t_i}, Z_s^{t_{i+1}}) = \frac{\sum_{w_i \in f_r \cup f_s} P(w_i \mid Z_r^t) P(w_i \mid Z_s^{t+1})}{\sqrt{\sum_{w_i \in f_r \cup f_s} P(w_i \mid Z_r^t)^2 \sum_{w_i \in f_r \cup f_s} P(w_i \mid Z_s^{t+1})^2}} \quad (2)$$

话题的语义关联度和特征的相似度对话题的演化具有不同的影响, 本文采用的方法是使用两者的线性组合来识别相邻时间段具有演化关系的话题. 话题 $Z_r^{t_i}$ 和 $Z_s^{t_{i+1}}$ 的演化关联度为

$$\mathrm{Relate}(Z_r^{t_i}, Z_s^{t_{i+1}}) = \lambda S(Z_r^{t_i}, Z_s^{t_{i+1}}) + (1-\lambda) F(Z_r^{t_i}, Z_s^{t_{i+1}}) + \lambda \quad (3)$$

本文设定一个阈值, 如果 2 个话题 $Z_r^{t_i}$、$Z_s^{t_{i+1}}$ 的演化关联度大于该阈值, 则判断这 2 个话题具有演化关系.

2　实验结果与分析

为了验证本文基于 LDA 话题关联的演化方法, 选择了全国人民代表大会与中国人民政治协商会议的报告(简称两会) 和 NIPS 数据, 分别反映社会关注话题和学术话题, 以证明方法的一般性. 设置时间间隔 δ= 1 a, 划分语料, 在每个局部语料上自动抽取话题, 话题数目可变, 模型参数 α、β 分别设置为 50 k(k 为话题数) 和 0.1; 然后, 计算话题之间的演化关联度, 如果关联度大于阈值, 则话题具有演化关系. 本文实验选取 LDA 后离散方法作为基准方法 (Baseline) 进行对比.

2.1　实验数据

实验数据见表 1、2. 对上述语料进行预处理, 包括分词、过滤停用词等, 两会报道的语料过滤了人名; NIPS 会议论文过滤了部分公式中使用的符号.

表 1 两会语料

Tab. 1 NPC & CPPCC corpus

文集	文档数目	文集词表数目
2007 两会语料	6 127	30 402
2008 两会语料	12 755	54 431
2009 两会资料	4 377	26 877

表 2 NIPS 会议文集

Tab. 2 NIPS conference document collection

文集	文档数目	词表数目
2003NIPS	194	18 127
2004NIPS	206	18 497
2005NIPS	205	19 295
2006NIPS	202	20 342
2007NIPS	217	21 360
2008NIPS	245	23 988

2.2 Baseline 方法

本文采用的 Baseline 方法为: 首先将所有语料建立 LDA 模型, 然后对话题进行时间上的离散化. 使用话题在当年文集所有话题中所占比重来描述话题的权重, 第 j 个话题在时段 t 的比重计算式如下:

$$\theta_j = \frac{1}{D_t}\sum_d \hat{\theta}_j^d \qquad (4)$$

式中: D_t 为时段 t 文档的数目; $\hat{\theta}_j^d$ 为时段 t 第 j 个话题在文档 d 中所占比例的后验估计值.

2.3 两会报告实验结果与分析

2.3.1 *话题抽取* 根据文献[6], 分别计算 2007~2009 两会语料话题数目, 如表 3 所示.

表 3 两会语料话题数目

Tab. 3 Topic number for NPC & CPPCC corpus

年份	Baseline	本文方法
2007	250	150
2008	250	180
2009	250	160

相比 Baseline, 本文方法基于局部时间段(年)的文集建模, 可以得到更为精确的描述话题内容的特征词组. 如话题"官员财产申报制度", 其在 2007~2009 的两会中每年都有提及, 但内容稍有差异, 表 4 给出了特征词的对比, 括号中数字为话题编号.

分析表 4 话题词, 官员财产申报制度是在官员反腐监督预防制度的话题之后得到的. 因此, 在不同时间段上建立话题, 可以更好地反应话题的时序性和变化. 本文方法还探测到了具有较强时序性的话题, 如2007两会话题中的物权法话题(物权法, 规定, 草案, 财产, 保护, 法律, 国家), 煤矿安全话题(事故, 安全生产, 煤矿, 没有, 总局, 生产, 死亡), 2009 两会话题中的水价改革话题(水价, 水资源, 实行, 水利, 能力, 改革, 饮水, 控制), 而这些话题并没有被 Baseline 方法探测到.

表 4 话题词的对比

Tab. 4 Comparison of topic words

采用方法	年份(话题号)	话题词
Baseline	2007~2009(173)	官员, 财产, 申报, 制度, 公开, 公务员, 监督, 公布, 纪委
本文方法	2007(98)	官员, 反腐败, 监督, 预防, 制度, 中央, 反腐, 加强, 惩治
本文方法	2008(42)	官员, 财产, 申报, 制度, 公布, 收入, 制定, 全国, 建议
本文方法	2009(95)	官员, 财产, 申报, 制度, 公开, 公务员, 出国, 干部, 公示

2.3.2 *话题关联* 通过对两会报告(2008~2009 年)话题关联的人工检测, 话题关联的精度与查全率如表 5 所示, 对比使用 Jenserr-Shannon divergence 计算方法[7], 在保持相同的查全率下提高了查准率.

表 5 话题关联的查全率与查准率

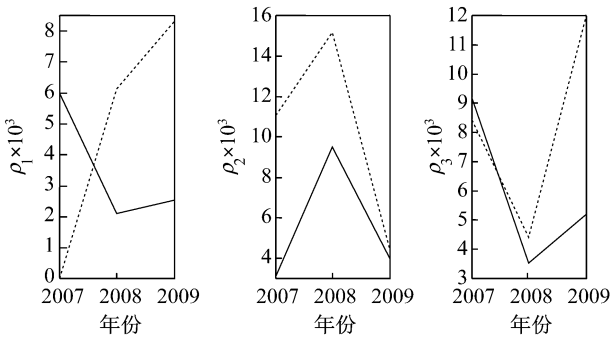
Tab. 5 Recall and precision for relating topics

方法	查全率	查准率
Jenserr-Shannon divergence(阈值 0.5)	0.93	0.52
特征相似度(阈值 0.4)	0.87	0.62
本文方法(λ= 0.6)	0.93	0.63

2.3.3 *话题演化* 分别计算 2007 年与 2008 年以及 2008 年与 2009 年两会话题的关联度, 设置阈值(0.46)判断相邻时间间隔中话题的演化关系. 图 1 选取四川灾后重建、北京奥运会、大学生就业 3 个话题, 表示了话题在该年会议中所占的比重 ρ. 根据实际数据, 在 2009 两会上, 大学生就业话题受到相当程度的关注.

四川灾后重建话题的演化过程, Baseline 方法得到的话题变化过程是不准确的。因为 2007 年底 2008 年初南方发生了雪灾, 四川是雪灾的受灾区, 是 2008 两会讨论的内容; 2008 年四川地震, 是 2009 两会讨论的内容; 而 2007 年两会中并没有讨论四川的受灾和重建情况, 因此, Baseline 方法描述的话题强度变化并不合理.

本文方法还可以得到话题内容上的变化. 例如, 房地产话题在两会报告中的演化关系, 表 6 是自动抽取的关于房地产话题在这 3 年报告中的特征词, 表 7 是应用前述方法计算得到的演化关联度.



(a) 四川灾后重建　(b) 北京奥运会　(c) 学生就业话题

—— Baseline,　——本文方法

图 1　话题强度随时间的演化

Fig. 1　Topic strength evolution over time

表 6　房地产话题的内容变化

Tab. 6　Content change for real estate topic

话题	时间	话题的特征词(前 8 个)
房地产(59)	2007	住房, 政府, 房价, 房地产, 廉租, 经济适用房, 市场, 建设
住房保障(6)	2008	住房, 保障, 政府, 廉租, 建设, 家庭, 解决, 收入, 政策
房价拐点(14)	2008	房价, 房地产, 市场, 拐点, 房子, 开发商, 价格, 调控, 出现
房价成本(17)	2008	成本, 政府, 税费, 房价, 费用, 开发商, 房地产, 开发, 地方
住房保障(48)	2009	住房, 廉租, 建设, 解决, 保障, 市场, 城乡, 家庭, 城市
房价调整(76)	2009	房地产, 房价, 市场, 价格, 成本, 政府, 调整, 开发, 开发商, 政策

表 7　相邻时间段话题(表 6)之间的关联度

Tab. 7　Related degree of topics in Tab. 6

	2008(6)	2008(14)	2008(17)
2007(59)	0.775	0.665	0.521
2009(48)	0.859	0.387	0.326
2009(76)	0.349	0.771	0.617

本文假设关联度大于阈值 0.46 存在演化关系. 由表 7 可见, 房地产话题的内容变化存在演化关系, 如 2007 房地产话题在 2008 年分化为房价成本、房价拐点和住房保障.

设定阈值来判定话题的演化关系也存在一些不足. 分析一些错误结果发现, 对于某些关联度大于阈值的话题对, 不存在演化关系. 例如, 2007 年话题 19(征收, 开征, 物业税, 实施, 代表, 研究, 税收, 认为, 房价, 消费税)与 2008 话题 59(税收, 印花税, 征收, 税率, 交易, 政策, 起征点, 降低, 调整, 收入)其关联度为 0.50, 大于阈值(0.46). 可以看出, 后者并非前者话题的演化, 两者更像是同一主题下的子话题. 又如 2007 话题 104(问题, 解决, 存在, 方面, 进行, 需要, 得到, 没有, 机制, 根本)与 2008 话题 122(问题, 解决, 需要, 提出, 方面, 措施, 采取, 关注, 针对, 存在, 得到)关联度为 0.92, 但这些词并不能作为话题的特征词, 只是语义上耦合的词, 因此, 无法解释的话题又在一定程度上干扰了话题关联的演化研究.

2.4　NIPS 语料实验结果与分析

2.4.1　*话题抽取*　2003~2008NIPS 语料话题数目见表 8, 表 9 列出了聚类话题的对比, 它反映了话题内容的变化.

表 8　NIPS 话题数目

Tab. 8　Topic number for NIPS corpus

年份	Baseline	本文方法	年份	Baseline	本文方法
2003	95	45	2006	95	50
2004	95	50	2007	95	50
2005	95	45	2008	95	60

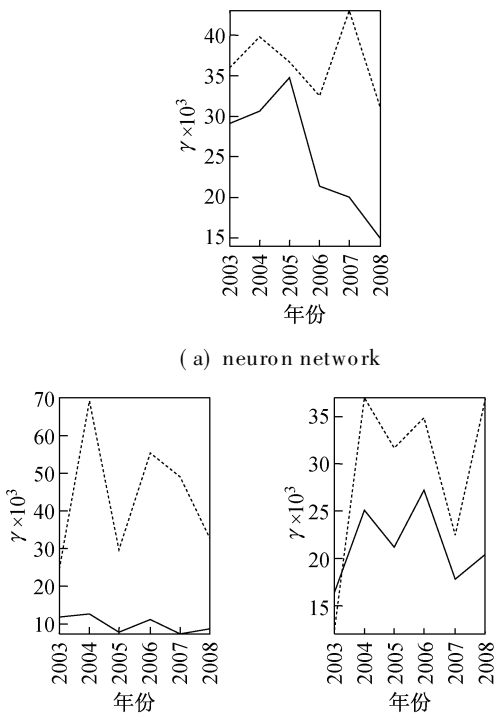
由表 9 可见, 2006NIPS 谱聚类算法(spectral)重新受到人们的关注. 总观本文实验结果可知, 2007 年话题 25 lasso(lasso, sparse, regression, compressed, source, variables, approach, energy), 话题 8 topic model(topic, topics, LDA, document, words, documents, word, dirichlet, code, variational) 反映了近年讨论的学术热点.

表 9　话题 cluster 的对比

Tab. 9　Comparison of topic words for "cluster"

Baseline	2003~2008	clustering, cluster, clusters, data, spectral, algorithm, points, partition, normalized, similarity
本文方法	2003 NIPS	clustering, clusters, cluster, cost, data, partition, features, algorithm, centers, distance
	2004 NIPS	clustering, cluster, clusters, mixture, alignment, data, simples, time, guassian, pairwise
	2005 NIPS	clustering, cluster, clusters, affinity, games, similarity, normalized, graph, propagation, spectral
	2006 NIPS	clustering, cluster, clusters, **spectral**, helicopter, density, margin, maximum, measure, points
	2007 NIPS	clustering, cluster, clusters, matrix, universum, data, K-means, points, trace, problem
	2008 NIPS	clustering, cluster, clusters, target, quality, tasks, load, spectral, dual, measures

2.4.2 *话题关联与演化* 图 2 显示了 neuron network, classification, image 3 个话题随时间的强度变化. 图中, γ 为话题的强度, 表示话题在该年会议中所占的比例. 2 种方法所描述的变化趋势基本一致. 因此, 本文方法同样可以应用在学术话题上.



(a) neuron network

(b) classfication (c) image

图 2 话题强度随时间的演化

Fig. 2 Topic strength evolution over time

根据阈值来判定话题的演化在 NIPS 数据集上也存在一些错误的实例. 如 2007 话题 29(sample, variance, monte, carlo, samples, error, dataset, strati, relative, sampling) 与 2008 话题 5(error, sense, subspace, SVD, sampling, node, dictionary, tree, kernel, query) 的关联度为0.49, 虽然大于阈值, 经人工检验 2 个话题的特征词不具有话题的演化关系. 主要原因是这 2 个话题在语义空间上的距离比较近, 属于关联话题, 但不存在同一性. 虽然调整阈值可以避免发生上述错误, 但有可能遗漏了一些演化关系. 如何设置合适的阈值需要进行进一步研究.

3 结 论

本文提出了基于 LDA 话题关联的话题演化方法, 通过话题抽取和话题关联 2 个步骤完成, 话题的演化关联不仅考虑了话题词的分布距离, 还参考了话题特征的相似度, 实验结果表明:

(1) 基于局部时间段话题抽取, 可以比较精确的描述话题的语义.

(2) 基于局部时间段话题抽取的独立性, 可以探测到新话题的产生, 旧话题的消失.

(3) 基于话题关联的演化可以探测话题内容的变化, 反映了话题之间一对多, 多对多的演化关系.

通过分析实验结果, 进一步的工作是解决如何过滤话题自动抽取中得到的不可解释的话题, 如何过滤存在关联, 但不具有同一性的话题, 正确识别话题的演化关系.

参考文献:

[1] Blei D M, Ng A Y, Jordan M I. Latent dirichlet allocation[J]. **Journal of Machine Learning Research**, 2003(3): 933-1022.

[2] Wang X, McCallum A. Topic over time: A non-markov continuous-time model of topical trends[C]// **ACM SIGKDD 2006**. Philadelphia, USA: [s. n.], 2006: 424-433.

[3] Hall D, Jurafsky D, Manning C D. Studying the history of ideas using topic models[C]// **Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing**. Waikiki, Honolulu, Huawaii: [s. n.], 2008: 363-371.

[4] Blei D M, Lafferty J D. Dynamic topic models[C]// **Proceedings of the 23rd International Conference on Machine Learning**. Pittsburgh, Pernsylvania: [s. n.], 2006: 113-120.

[5] Alsumait L, Barbara D, Domeniconi C. On-line LDA: Adaptive topic models for mining text streams with applications to topic detection and tracking[C]// **In ICDM**. Pisa, Italy: [s. n.], 2008: 3-12.

[6] Griffiths T L, Steyvers M. Finding scientific topics[J]. **Proc Natl Acad Sci USA**, 2004, 101(Suppl 1): 5228-5235.

[7] 楚克明, 李 芳. 基于 LDA 的新闻话题的演化[C]// 第 5 届全国信息检索学术会议. 上海: [s. n.], 2009.