基于自然语言处理的主题模型及其发展分析

加日拉·买买提热衣木 常芙蓉 刘 晨 杨 礼

（喀什大学 计算机科学与技术学院，新疆 喀什 844000）

摘 要：自然语言处理作为计算机应用的重要构成，属于人工智能的范畴，计算机技术以及人工智能技术在信息技术范围之内。主题模型在自然语言处理领域中越来越受到重视，相关学者对于基于自然语言处理的主题模型的研究越来越深入，对此通过对基于自然语言处理的主题模型及其发展进行分析，希望可以为主题模型的发展提供一定建议。

关键词：自然语言处理；主题模型；人工智能

中图分类号：TP391.1 **文献标识码：**A **文章编号：**1003-9767（2017）24-042-03

Theme Model based on Natural Language Processing and Its Development Analysis

Jiarila Maimaitiriyimu, Chang Furong, Liu Chen, Yang Li

（Kashgar University, School of Computer Science and Technology, Kashgar XinJiang 844000, China）

Abstract: As an important part of computer application, natural language processing belongs to the category of AI. Computer technology and AI technology are in the range of information technology. The topic model in natural language processing in the field of more and more attention, the relevant researches on the topic model based on natural language processing more and more in-depth, this theme through the model of natural language processing and its development based on the analysis, hoping to provide some suggestions for development of theme model.

Key words: natural language processing; thematic model; artificial intelligence

在自然语言处理中，主题就是词项的实际概率分布，在此领域中可以将主题作为词项的主要概率分布，通过主题模型利用词项进行文档级的信息抽取，获得一些语义类似与相关的主题集合，同时可以把词项中的相关文档变换为主题空间，并且在文档的低维空间中进行表达。

1 主题模型

1.1 主题模型的输入

主题模型就是文档的集合，因为交换性的架设因素的影响，其相当于词项文档中的矩阵，如表 1 所示。

表 1 词项文档的矩阵

	d1	d2	d3	d4	d5	d6
ship	1	0	1	0	0	0
boat	0	1	0	0	0	0
ocean	1	1	0	0	0	0
voyage	1	0	0	1	1	0
trip	0	0	0	1	0	1

通过对词项文档的矩阵进行分析可以了解到在主题模型中涵盖了 ship、boat、ocean、voyage、trip 五个词项以及六个文档，在文档中相同的词项可以出现多次。同时较为重要的就是输入主题的个数，在常规状况之下必须要在模型训练之前制定主题个数的大小，要基于自己的经验开展，确定数量的简单方式就是对不同的数量进行重复的实验，在评价的指标最优时则可以确定最佳选择，其中重要的就是困惑度、语料似然值以及分类正确率几个因素。同时一些学者也通过非参数贝叶斯的方式确定主题数目，在实践中主题个数可以随着语料的实际规模的变化而不断波动，在训练结束时活动的主题个数就是实际数目的最佳选择。

1.2 主题模型中的基本假设

主题模型中的一个重要假设是词袋假设，即一篇文档内的单词可以交换次序而不影响模型的训练结果。所谓的可交换就是与顺序没有关系，和条件独立同分布等价。在一些

基金项目：青年科学基金项目（项目编号：2016D01B010）；新疆高校科研计划青年项目（项目编号：XJEDU2016S076）。

作者简介：加日拉·买买提热衣木（1987-），男，维吾尔族，新疆喀什人，硕士研究生，助教。研究方向：人工智能、自然语言处理。

LDA 的派生模型中，为了构造对应的模型会打破其可交换。

1.3 表示主题模型

在表示主题模型过程中，可以通过图模型以及生成过程两种方式进行表示，图 1 为图模型方式。

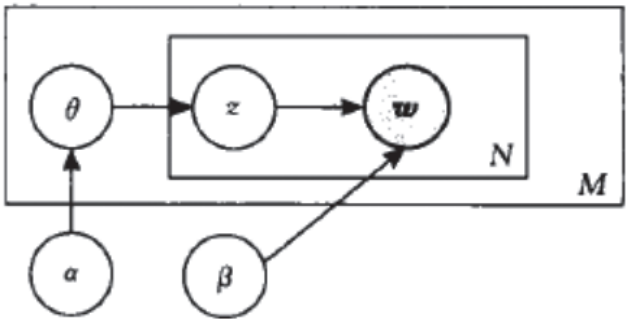

图 1 图模型方式

1.4 参数估计过程

不同主题之下的词项分布概率以及相关文档的主题分布概率是重要的两个参数。其中参数估计就是其生产过程的一个逆过程，例如，在 LDA 模型中，基于其图模型就可以获得其语料的概率数值，可以这样表示：

$$\prod_{d-1}^{M}\int p(\theta_d / a)\left(\prod_{\pi=1}^{Nd}\sum_{zdm}P(Zdn/\theta_d)p(\omega dn/Zdn,\beta)\right)d\theta_d$$

1.5 新样本的推断

在完成主题模型的训练之后，为了进行新样本的推断，可以利用主题模型交换其相关词项空间表达的文档，形成全新的主题空间，这样就会获得其主要的概率分布，新样本的推断在实践中可以实现查询，可以有效应用在信息检索中。

2 LDA 模型

在 2003 年相关学者提出了 LDA（Latent Dirichlet Allocation）模型，这是一种基于 PLSI 获得的主题分布概率。一些学者通过对其进行完善获得了一个较为完整的生产模型。现阶段此种主题模型在一些图像处理以及自然语言处理中得到应用。现阶段的主题模型工作的开展多数都是基于 LDA 模型的完善与修改，也有将整个 LDA 模型作为概率模型重要部件的应用方式。

2.1 LDA 模型算法框架

为了获得主题之间隐藏的信息内容，必须要获得文本之间的相关主题信息内容，然后再利用主题信息获得相关主题之间的演化信息，对此在实践中可以基于以下流程开展。

（1）划分时间窗口，以时间为主要顺序标准，对时间窗口进行划分处理，然后将其对应的时间文本放置到其时间窗口之中。

（2）文本预处理，必须要进行文本的预处理，较为常见的处理方式主要就是分词、过滤噪音、过滤非法字符以及去除停用词等。

（3）利用 LDA 建模获得原始的文本主题。通过对相关时间窗口的文本开展 LDA 建模操作，对不同的时间窗口文本进行模型参数设置以及主题数目的确定。

（4）原始主题的过滤，在原始主题中会存在一些没有实际意义的主题，会降低主题演化的精准性，对此必须要及时过滤去除。

（5）演化分析，在进行分析之前，必须要计算不同主题之间的相似度，关联相关主题的时间窗口，对其进行演化分析。

2.2 文本预处理

在对 LDA 模型进行文本建模操作之前，必须要对文章进行预处理，主要采取分词、去除文本的噪声以及去停用词等方式。

2.3 LDA 建模

在完成预处理之后，必须要对每一个时间窗口中的文本进行 LDA 建模，首先要对 LDA 模型进行推理。

2.4 主题过滤方法

可以通过文本集收集到各种不同的原始主题，多数的原始主题并没有实际的表示意义，在实践中无法对其进行解释。对此在实践中必须要对文本进行过滤处理，如果一个主题在文档中出现的次数较多，就可以将其认定为有意义的主题，可以通过权重计算方式进行主题设置，进而进行过滤处理。

2.5 主题演化分析

2.5.1 主题强度演化

基于文本的实际发布时间，将相关文本离散到对应的时间片中，对其进行 LDA 建模，在通过过滤主题之后，就会获得相关时间片的主题信息，再对其进行主题强度的计算，就会获得在相同时间序列之上不同的主题强度数值，进而了解其变化趋势，为分析提供参考。

2.5.2 主题的内容演化

主题内容演化就是不同主题中特征词序列在不同时间片中存在的差异，这种差异多为语义之上的关联性，在实践中可以通过建立主题关联、主题关联过滤判别主题演化关系等对其进行主题内容演化操作。

3 主题模型发展趋势

主题模型的多数工作都是在其特定的任务中，极少数会应用到参数的扩展以及引入上下文的信息内容，主要就是因为这二者的工作主要就是基于主题模型进行整体的修改，对其进行研究的内容相对较少，通过综合分析，可以了解到在今后的主要研究趋势具体如下。

3.1 主题模型的实践应用

更为重视主题模型的性能，意味着主题模型的实用性得到了人们的肯定，对此在实践中必须要不断探究更为完善的训练方式与算法，相继出现了 EM（Varia-tional EM）算法、基于 LDA 模型以及 HDP 模型的分布式算法以及基于 LDA 模型的在线变分贝叶斯方法等几种方式，这些方式可以有效提升训练的效果与质量。

3.2 主题模型与跨语言的融合

主要就是因为机器翻译本身具有一定的自然语言处理特征，其可以累积海量的跨语言语料，可以为主题模型提供参考。对此相关学者相继提出了 ML-LDA 模型在基于跨语言的语料信息中进行主题的抽取、在每一个主题中获得更多的语言，这样更适应跨语言的网络应用。同时一些学者提出了 Joint LDA 模型，利用西班牙语以及英语的语料进行信息的采样，应用双语词典，在整个模型的训练结束之后不同的主题就会获得不同语言模式的混合主题。此种方式在实践中灵活应用了跨语言信息检索。

同时，一些学者还提出了 CTRF 模型，利用此种模型在实践中可以将单词的特征以及单词中主题的依赖关系进行充分的融合，通过通用性的机器学习方式，不单纯针对具体的任务进行学习，这也是一种全新的发展趋势。

4 结 语

对于主题模型的发展来说，不同的工作之间具有较为密切的关系，虽然因为人类语言的本质因素的影响在学术界没有统一，尚存争议，作为概率生产模型的主题模型自身也有待完善，但是相信在今后的发展中人们会对其进行深入的研究分析，会基于实际状况构建更为完善的主题模型，主题模型也会在各个领域中得到广泛的应用。

参考文献

[1] 徐戈，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.

[2] 茅利锋 . 基于主题模型的主题演化分析及预测 [D]. 南京 : 南京邮电大学 ,2016.

[3] 肖智博 . 排序主题模型及其应用研究 [D]. 大连 : 大连海事大学 ,2014.

[4] 朱佳晖 . 基于深度学习的主题建模方法研究 [D]. 武汉 : 武汉大学 ,2017.

[5]Griffin J E. An adaptive truncation method for inference in Bayesian nonparametric models[J]. Statistics and Computing, 2016, 26(1-2): 423-441.

（上接第 41 页）

数据库性能问题的常用解决方法如下。

（1）监视相关性能数据。

（2）通过分析 SQL 语句的执行找出占用资源最大的事务并优化处理。

（3）定位锁冲突，修改发生锁冲突严重的应用逻辑。

（4）对较大规模的数据或者无法通过常用优化措施解决的锁冲突进行处理，比如利用负载均衡或数据分布式处理等。

除此之外，常见的系统性能调优方法还有如下几方面。

（1）当 CPU、内存利用率较高、资源使用成为系统瓶颈时，可以增加 CPU 和内存的个数；提高 CPU 主频，更换较大内存，将 Web 服务器与数据库服务器分开部署；调整软件的设计与开发。

（2）当系统性能无法满足服务器端应用需求时，调优措施有：检查软件架构设计是否合理；代码开发是否符合规范要求；设置的软件参数是否匹配；检查应用服务器端架构设计是否合理；检查应用服务器和数据库服务器的匹配是否满足系统性能需求。

（3）当网络传输带宽成为系统瓶颈时的调优措施有：增加带宽；压缩传输数据等。

3 结 语

性能测试的目的是验证软件系统是否能够达到用户提出的性能指标，发现软件系统存在的性能瓶颈并加以优化，因此，性能测试应关注高吞吐量、高商业风险及高服务器负载类型的业务。

通常情况下，性能调优是在系统故障定位的前提下实施的，而故障定位的过程，对测试工程师的素质和测试工具的能力是一次严格的考验，因此，系统性能测评和性能调优，能够直接体现一个软件测试机构的技术实力和核心竞争力。

参考文献

[1] 陈松立，杨春晖，戴青云，等 . 一种非插桩采样嵌入式软件性能测试方法 [J]. 软件 ,2014,35(12):1-4.

[2] 李杰 . 一种高性能服务器的设计与性能评估 [J]. 软件 ,2014,35(12):88-93.

[3] 丁小盼，周浩，贺珊，等 . 基于 OpenStack 的云测试平台及其性能分析研究 [J]. 软件 ,2015,36(1):6-11.

[4] 张晓清，龚波，田丽韫，等 . 国产自主可控应用性能优化研究 [J]. 软件 ,2015,36(2):5-9.

[5] 阮晓龙 .HTTP 协议状态检测与性能分析软件的设计与实现 [J]. 软件 ,2015,36(7):136-141.

[6] 刘峰，鄂海红 . 基于海量数据的消息队列的性能对比与优化方案 [J]. 软件 ,2016,37(9):33-37.

[7] 李霄，王常洲，田雅 . 计算机应用系统性能测试技术及应用研究 [J]. 软件 ,2013,34(4):69-73.

[8] 肖扬，于艳华 . 基于 IaaS 云平台的应用性能管理研究与应用 [J]. 软件 ,2013,34(12):241-245.