Contents

Chapter 1

绪论

1.1 研究背景及意义

- 传统的新闻传播方式的弊端，数据量之大，有效信息获取非常困难
- 新技术的出现，是研究内容变得可行
- 如果采用新的技术，人们可以快速有效，直观的了解信息

主题模型在机器学习和自然语言处理等领域是用来在一系列文档中发现抽象主题的一种统计模型。近年来，主题模型已经被广大的学者和研究人员应用在了文本信息挖掘、文本分类、图像分类和理解、情感分析、学术文章挖掘、语义分析、推荐系统等诸多方面 [17][18][19][20][21]。情感分析 (opinion mining, sentiment analysis) 是近年来新出现地一个研究方向，其基本任务就是从用户生成的包含观点和意见的文本中抽取出这些观点和意见，然后生成情感摘要，进行情感分类，自动构建情感词典等等情感分析任务。主题模型在情感分析最主要的任务就是学习出来用户讨论和用户评论中的内容主题。在情感分析中每一个 topic 通常被称为 aspect，在情感分析中，会将词汇区别为情感词汇和主题词汇。学术文章挖掘是主题模型的一个重要应用，通过对学术文章的挖掘来进一步理解研究领域的发展和进化，对于了解之前的科研成果和未来的发展趋势有非常重要的意义。对于学术文章语料集合，LDA 模型并没有考虑到文章的作者，实际上是把所有的作者都看作完全等同的。Rosen-Zvi 等人 (2004)[15] 提出了作者主题模型 (Author-Topic Model, ATM)，该模型假定每一个作者都对应一个主题分布，文档生成过程中，先随机选取一个作者，然后根据该作者的主题分布确定下一个词的主题。ATM 模型假定，不同作者当选定同一主题后，他们写文章时使用的词都是一样的，显然这不符合实际情况。针对这个问题，Kim 等人 (2012)[16] 提出了实体主题模型 ETM(Entity Topic Model)，该模型引入层次的 Dirichlet 先验，将词与主题分布的先验、词与实体分布的先验通过线性组合，作为与主题实体对分布的先验，这样每个词的生成都与主题实体对相关。

现在随着 Twitter, Weibo 等社交媒体的快速发展，对社交媒体的主题挖掘也成了一个热点的研究问题，目前主要应用的模型有 LDA, Author-Topic Model 和 Twitter-LDA, MB-LDA[21] 等等。

1.2 国内外研究现状

- 信息检索技术的发展
- 主题模型等文本挖掘技术，以及扩展模型
- 新闻主题挖掘相关技术比较

主题模型 (Topic Models)[1] 的基本思想是，一个文档是由多个主题混合而成的，而主题是在词库上的一个概率分布。主题模型是一个生成式模型，为了生成一个文档，首先选择一个主题的概率分布，然后对于文档的每个词，根据主题的概率分布随机的选择一个主题，并从该主题中选择一个词。利用统计学的一个方法，我们可以推理出生成该文档集合的主题集合。

主题模型的起源是隐性语义索引 (Latent Semantic Indexing, LSI)[2]。LSI 并不是概率模型，因此也算不上一个主题模型，但是其基本思想为主题模型的发展奠定了基础。在 LSI 基础上，Hofmann(1999; 2001)[3][4] 提出了 pLSI(Probabilistic Latent Semantic Indexing) 模型，但 pLSI 并没有用一个概率模型来模拟文档的产生，只是通过对训练集种的有限文档进行拟合，得到特定文档的主题混合比例。这样就导致了 pLSI 模型参数随着训练集中的文档数目线性增加，出现过拟合现象，而且对于训练集以外的文档很难分配合适的概率。Blei 等人 (2003)[5] 在 pLSI 基础上加以扩展，提出了 LDA(Latent Dirichlet Allocation) 模型。LDA 模型用服从 Drichlet 分布的 K 维隐含随机变量表示文档的主题混合比例，来模拟文档的产生。Dirchlet 分布作为多项分布的共轭先验，很好的简化了统计推理问题。LDA 模型中，主题数 T 通常是根据我们的先验知识随机选取的，Griffiths 和 Steyvers(2004)[6] 通过引入标准 Bayes 方法，将主题数的确定转变成一个模型选择问题，从而确定最佳的主题数，并且作者提出了使用 MCMC 和 Gibbs 采样 [7] 的方法来进行参数评估。

LDA 模型的其中一个局限是，它难以描述主题间的关联关系。而在许多文本语料中，隐主题之间本就有很强的关联性，比如：在一个关于 Science 的语料中，有关基因的文章也有可能是有关健康和疾病的，而不太可能是关于天文学的。产生该局限性的原因是，在主题的 Dirichlet 分布中，我们假设各个分量几乎是独立的。Blei 和 Lafferty(2007)[8] 提出了 CTM(Correlated Topic Model) 模型，该模型通过 Logistic Normal Distribution 来描述主题的分布。相比 LDA，CTM 模型能够更好的表达真实语料中主题的分布，然而模型复杂化之后，计算复杂度也随着增大。不管是 LDA 还是 CTM 都假设语料库中的文档是可交换的，但是在许多实际的语料中，该假设并不合适，如，学术期刊，邮件，新闻等等的内容，都是随着时间不断演化的。为了显示地描述和发现主题的动态变化情况，Blei 和 Lafferty 又提出了 DTM(Dynamic Topic Model)[9]。在 DTM 模型中，作者按时间片对文档集合进行划分，然后分别对每个时间片内地文档用 LDA 模型进行建模，而时间片 t 的主题是从时间片 t-1 的主题进化而来的。和 CTM 模型一样，作者采用 Logistic Normal Distribution 来描述时序主题的不确定性。但是对于一些应用，可能对于事件的粒度要求更高或者很难划分数据，X. Wang(2006)[10] 提出了一个非马尔科夫连续时间模型，该模型除了文本信息以外，将时间标签也作为可见信息，然后通过主题分布信息同时关联起来词汇和事件标签。但是该模型假定主题集合不随时间变化而变化，C. Wang(2008)[11] 进一步松弛了这种假设，提出另一种连续时间主题模型，在这个模型中，主题集合随着时间变化而变化。

LDA 从提出到现在已经被广泛应用于文本挖掘和信息处理领域。并且现在也出现了许多关注主题模型性能的工作，这说明主题模型已经不局限于理论研究阶段，它的实用性得到了认可，因此需要更高效的训练算法。Mariote 等人 (2007)[12] 提出了变分 EM(Variational EM) 算法来对训练过程进行加速，以便应用于多处理器和分布式环境。Asuncion 等人 (2008) 给 LDA 模型和 HDP 模型提出了分布式算法在保证全局正确性的前提下，各个处理单元能够独立进行 Gibbs 采样。Hoffman 等人 (2010)[14] 提出了 LDA 模型的在线变分贝叶斯方法（Online Variational Bayesian）。

1.3 本文的主要研究内容

- 新闻故事线定义
- 时序文档的主题挖掘
- 命名实体的关联分析，以及他们对主题的影响
- 新闻故事线的可视化技术

1.4 本文的组织结构

Chapter 2

主题模型及其应用

2.1 主题模型介绍

2.2 主题模型的应用

Chapter 3

新闻故事线自动生成方法研究

3.1 时序新闻语料时间片划分

3.2 新闻事件主题演化跟踪

3.3 新闻事件人物关系建模

Chapter 4

新闻故事线可视化方法研究

4.1 文档可视化方法研究

4.2 主题演化可视化

4.3 人物关系可视化