# 热点话题发现

姜庆鸿1），刘子宇1），刘宏玉1），张宗毓1）

1）（计算机学院 北京理工大学 北京市 中国 100081）

**摘要：**话题检测与跟踪（Topic Detection and Tracking）将这些分散的信息有效地汇集并组织起来，从而帮助用户发现事件的各种因素之间的相互关系，从整体上了解一个事件的全部细节以及与该事件与其它事件之间的关系。本文介绍了TDT的背景、内容和研究方向。同时，着重介绍了主题模型的相关知识，对于经典的主题模型进行了详细的阐述。最后，对话题发现这一主题的发展进行了分析和总结。

**关键词：**话题发现和跟踪，主题模型

**小组分工：**刘宏玉：负责传统主题模型相关内容；姜庆鸿：负责第一章和LDA模型相关内容；刘子宇：负责BTM模型、PAM模型相关内容和排版、摘要；张宗毓：负责关键问题、难点，以及展望部分。

**Hot Topic Detection**

Jiang Qinghong 1), Liu Ziyu1), Liu Hongyu1), Zhang Zongyu1)

1)（School of Computer Science, Beijing Institute of Technology, Beijing, China, 100081）

**Abstract:** Topic Detection and Tracking effectively brings together and organizes these disparate pieces of information to help users discover the interrelationships between the various elements of the event, and to get an overall picture of the full details of an event, The relationship between this event and other events. This article introduces the background, content and research direction of TDT. At the same time, it mainly introduces the related knowledge of the topic model, and elaborates on the classic topic model in detail. Finally, the topic of the development of the topic was analyzed and summarized.

**Key words:** Topic Detection and Tracking, Topic Model

# 概述

话题检测与跟踪（Topic Detection and Tracking）是一项信息处理技术，这项技术旨在帮助人们应对日益严重的互联网信息爆炸问题，对新闻媒体信息流进行新话题的自动识别和已知话题的持续跟踪。话题检测与跟踪 (Topic Detection and Tracking, 简写为TDT)起源于早期面向事件的检测与跟踪（Event Detection and Tracking，简写为[EDT](https://baike.baidu.com/item/EDT)）是近年提出的一项[信息处理](https://baike.baidu.com/item/%E4%BF%A1%E6%81%AF%E5%A4%84%E7%90%86)技术，这项技术旨在帮助人们应对日益严重的互联网信息爆炸问题，对新闻媒体[信息流](https://baike.baidu.com/item/%E4%BF%A1%E6%81%AF%E6%B5%81/2319006)进行新话题的自动识别和已知话题的持续跟踪。与EDT不同，TDT检测与跟踪的对象从特定时间和地点发生的事件扩展为具备更多相关性外延的话题，相应的理论与应用研究也同时从传统对于事件的识别跨越到包含[突发事件](https://baike.baidu.com/item/%E7%AA%81%E5%8F%91%E4%BA%8B%E4%BB%B6/4402)及其后续相关报道的话题检测与跟踪。由于话题检测与跟踪相对于信息检索、数据挖掘和信息抽取等自然语言处理技术具有很多共性，并且面向具备突发性和延续性规律的新闻语料，因此逐渐成为当前信息处理领域的研究热点。

TDT涉及两类最主要的信息获取问题，即信息的检测与集成、信息的采集与跟踪。两方面的研究课题分别与目前信息检索(IR)和信息过滤(Information Filtering,简写为 IF) 对应的问题非常相似。在IR系统中,用户通过动态地定义需求(Query),从海量信息中检索满足自己当前兴趣的信息,信息以相关度为尺度进行组织、集成与反馈；而IF系统中,用户通过定义静态的用户需求(Profile),从动态变化的信息流中实时地获取相关知识,这种知识的获取方法侧重于跟踪信息的时空进程并将最新的相关信息反馈给用户。基于这些相似点,许多基于IR和IF的信息获取技术都相应地应用于TDT并获得了良好的效果,尤其近期逐渐发展起来的个性化信息检索技术和自适应信息过滤技术,都与TDT研究具有更深层次的共性。但是,TDT在许多方面与IR和IF存在差异,比如对TDT的新事件检测任务(New Event Detection,简称为NED),系统欠缺任何话题的先验知识,TDT系统必须在对话题毫不了解的情况下,自主地进行识别与检测,这一点与具备了背景知识或先验需求IR系统截然不同。同时,话题检测系统通常需要维护固定的存储空间保存曾经发生过的话题线索,从而作为衡量新话题的背景信息。对于话题跟踪而言,话题对应的“Query”是隐含给定的,构成话题的是若干(1～4篇)相关报道样本,这与具备明确需求(Profile)的IF问题也不相同。因此,面向 IR和IF 的相关方法更多地作为TDT 的基础研究 ,而不能完全解决TDT的相关问题。

## TDT应用背景

随着信息传播手段的进步，尤其是互联网这一新媒体的出现，我们已经摆脱了信息贫乏的桎梏，进入一个信息极度丰富的社会。在目前信息爆炸的情况下，信息的来源已不再是问题，而如何快捷准确的获取感兴趣的信息才是人们关注的主要问题。目前的各种信息检索、过滤、提取技术都是围绕这个目的展开的。由于网络信息数量太大，与一个话题相关的信息往往孤立地分散在很多不同的地方并且出现在不同的时间，仅仅通过这些孤立的信息，人们对某些事件难以做到全面的把握。一般的检索工具都是基于关键词的，返回的信息冗余度过高，很多不相关的信息仅仅是因为含有指定的关键词就被作为结果返回了，因此人们迫切地希望拥有一种工具，能够自动把相关话题的信息汇总供人查阅。话题检测与跟踪（Topic Detection and Tracking，以下简称TDT）技术就是在这种情况下应运而生的。通过话题发现与跟踪，人们可以将这些分散的信息有效地汇集并组织起来，从而帮助用户发现事件的各种因素之间的相互关系，从整体上了解一个事件的全部细节以及与该事件与其它事件之间的关系。目前此方面的研究已经引起了人们的普遍兴趣。

TDT技术可以用来监控各种语言信息源，在新话题出现时发出警告，在信息安全、金融证券、行业调研等领域都有广阔的应用前景。此外，它还可以用来跟踪某个话题的来龙去脉，进行历史性质的研究。

## TDT发展历程

话题检测与跟踪是一项旨在依据事件对语言文本信息流进行组织、利用的研究，也是为应对信息过载问题而提出的一项应用研究。

TDT 源于1996 年美国国防高级研究计划委员会（DARPA）提出需要一种能自动确定新闻报道流中话题结构的技术。随后，DARPA、卡内基·梅隆大学、Dragon 系统公司以及马萨诸塞大学的研究人员定义了TDT的相关内容，并检验信息检索中基于主题的技术在TDT 中的应用情况，这些研究及评测被命名为TDT pilot。TDT pilot 从1996 年下半年持续到1997年，把问题以易于处理和评测的方式确定下来，标志着TDT 研究领域正式确立。此后为推动TDT 发展，借鉴信息抽取（MUC）、信息检索（TREC）等研究的成功经验，DARPA 以及美国国家标准技术研究所（NIST）先后资助并主持了TDT 系列评测1。这种以评测为驱动的研究方式在TDT 中同样取得了成功。

由确立至今，TDT 经历了一个研究内容逐渐细化，研究方法逐渐多元化，研究重点逐渐向实际应用靠近的过程。把这段历史划分为以下4 个阶段：

### 1.2.1 1997～1999年发展期

除了TDT 研究内容、评测方法等基本确定，研究方法也得到初步开发。研究重点有两个：检验信息检索中基于主题的方法在TDT 中的有效性；话题追踪对已知话题相关语料数量的敏感度。

### 1.2.2 2000～2002年繁荣期

大量信息检索、信息过滤、文本分类中的技术从各自的角度看待及解决TDT问题，都取得了较好的效果。命名实体抽取技术在新闻文本表示上发挥了较大作用。

### 1.2.3 2003～2004年突破期

由于研究内容和研究对象的特殊性，相邻领域中的技术很难使TDT 性能取得进一步改进。而且，系统在实际应用中的性能下降比较明显。这些都促使TDT 研究需要有所突破，主要表现在表示模型及话题结构的改进上。期间，解决跨语言问题也成为TDT 的重点之一。为此，研究内容、实验语料和评测方法都作了相应的调整。或许考虑到与实际应用越贴近，一些研究条件越难限制或定义，无法在统一的平台上比较不同系统的性能，2004年之后TDT 评测没有再继续。

### 1.2.4 2005年至今

评测虽然结束，但基于话题的研究仍是自然语言处理领域中的热点，一方面基于TDT 语料的话题研究仍在继续，另一方面新应用背景下的话题研究也得到了积极的开展，文献分别探讨了在普通网页、博客、邮件、网络搜索结果中的话题组织和发现问题。随着TDT 研究的发展和深入，一些新的研究动向也为TDT 带来了新的问题和研究思路，例如：数据流挖掘、热点话题发现、基于话题的多文档摘要等。数据流挖掘主要指对实时、动态、按时间排序的信息流进行的一系列相关研究，包括分类、聚类、概念漂移问题解决、时效问题解决等。TDT 的研究对象也是一种动态数据流，两个领域的发展是相互促进的。热点话题发现、基于话题的多文档摘要稍有不同，是对数据流经过TDT相关技术处理后所得结果的进一步整理，以更方便人们对信息的使用。热点话题发现对话题的受关注度进一步分析，进而过滤出热点话题。基于话题的多文档摘要是对描述同一话题的文档集合进行提取，把该话题的发生、发展集中到一个文档中，高效提供简洁全面的话题。国内TDT 研究起步较晚，多在评测之后，主要集中在层次化话题发现、话题追踪和新事件发现几个研究内容上，热点话题发现也是较多关注的焦点。

可以看到，话题检测与跟踪和信息抽取研究一样，其建立与发展是以评测驱动的方式进行的。这种评测研究的方法具有以下一些特点：明确的形式化的研究任务、公开的训练与测试数据、公开的评测比较。它将研究置于公共的研究平台上，使得研究之间的比较更加客观，从而让研究者认清各种技术的优劣，起到正确引导研究发展方向的目的。

## TDT的研究内容

与一般的信息检索或者信息过滤不同，TDT所关心的话题不是一个大的领域（如美国的对华政策）或者某一类事件（如恐怖活动），而是一个很具体的“事件（Event）”，如美国911事件、江泽民访美等等。为了区别于语言学上的概念，TDT评测会议对“话题”进行了定义：所谓话题（Topic），就是一个核心事件或活动以及与之直接相关的事件或活动。而一个事件（Event）通常由某些原因、条件引起，发生在特定时间、地点，涉及某些对象（人或物），并可能伴随某些必然结果。通常情况下，可以简单地认为话题就是若干对某事件相关报道的集合。“话题检测与跟踪”则定义为“在新闻专线（Newswire）和广播新闻等来源的数据流中自动发现主题并把主题相关的内容联系在一起的技术 ” 。例如，“俄克拉荷马城爆炸案”这个主题包括1995年美国联邦大楼被炸、悼念仪式、州和美国联邦政府的一系列调查、对Timothy McVeigh的指控等等。这个定义和其它与话题有关的研究不同，那些研究主要处理信息分类问题，比如任何与爆炸有关的事件。处理分类问题需要专门的分类体系，注解起来效率低而且主观色彩浓厚。TDT与其它研究不同之处还在于它强调新事件的发现，希望找出不在人们意料之中的或没有人知道如何去查询的事件。

NIST为TDT研究设立了五项基础性的研究任务，包括面向新闻广播类报道的切分任务；面向已知话题的跟踪任务；面向未知话题的检测任务；对未知话题首次相关报道的检测任务和报道间相关性的检测任务。

## TDT的话题定义

最初的TD T 研究( TDT Pilot , 1996～1997) 将话题定义为“事件”。事件发生在特定时间和地点的事情。比如“,2001年9月11日针对纽约世贸大厦的恐怖袭击”是一个事件 ,而泛指的恐怖袭击则不是。此外 ,事件包括可预期事件(如“政府选举”)和突发事件(如“飞机失事”) 。从TDT2开始 ,话题的定义有了更加广泛的含义 ,不仅包含了由最初事件引起或导致发生的后续事件 ,同时还包含了与其直接相关的其他事件或活动。直到TDT5 ,话题都一直沿用如下定义。

话题定义: 一个话题由一个种子事件或活动以及与其直接相关的事件或活动组成。

根据话题的定义 ,一篇报道只要论述的事件或活动与一个话题的种子事件有着直接的联系 ,那么这篇报道就与该话题相关 ,比如关于“飞机坠毁”与“坠毁殉难者葬礼”的报道都可以认为与坠毁事件直接相关 ,因此可以作为该话题的一个组成部分。但话题的外延并不是无限的 ,比如关于“联邦航空局通过调查飞机坠毁的原因修改航空条例”的报道与飞机坠毁的话题并不相关。

## TDT任务

### 1.5.1 报道切分任务

报道切分（Story Segmentation Task，简称SST）的主任务是将原始数据流切分成具有完整结构和统一主题的报道。比如，一段新闻广播包括对股市行情、体育赛事和人物明星的分类报道，SST要求系统能够模拟人对新闻报道的识别，将这段新闻广播切分成不同话题的报道。SST面向的数据流主要是新闻广播，因此切分的方式可以分为两类：一类是直接针对音频信号进行切分；另一类则将音频信号翻录为文本形式的信息流进行切分。

### 1.5.2 话题跟踪任务

话题跟踪（Topic Tracking Task，简称TT）的主要任务是跟踪已知话题的后续报道。其中，已知话题没有明确的描述，而是通过若干篇先验的相关报道隐含地给定。通常话题跟踪开始之前，NIST为每一个待测话题提供1至4篇相关报道对其进行描述。同时NIST还为话题提供了相应的训练语料，从而辅助跟踪系统训练和更新话题模型。在此基础上，TTT逐一判断后续数据流中每一篇报道与话题的相关性并收集相关报道，从而实现跟踪功能。

### 1.5.3 话题检测任务

话题检测（Topic Detection Task，简称TD）的主要任务是检测和组织系统预先未知的话题，TD的特点在于系统欠缺话题的先验知识。因此，TD系统必须在对所有话题毫不了解的情况下构造话题的检测模型，并且该模型不能独立于某一个话题特例。换言之，TD系统必须预先设计一个善于检测和识别所有话题的检测模型，并根据这一模型检测陆续到达的报道流，从中鉴别最新的话题；同时还需要根据已经识别到的话题，收集后续与其相关的报道。

### 1.5.4 首次报道检测任务

在话题检测任务中，最新话题的识别都要从检测出该话题的第一篇报道开始，首次报道检测任务（First-Story Detection Task，简称FSD）就是面向这种应用产生的。FSD的主要任务是从具有时间顺序的报道流中自动锁定未知话题出现的第一篇相关报道。大体上，FSD与TD面向的问题基本类似，但是FSD输出的是一篇报道，而TD输出的是一类相关于某一话题的报道集合，此外，FSD与早期TDT Pilot中的在线检测任务（On-line detection）也具备同样的共性。

### 1.5.5 关联检测任务

关联检测（Link Detection Task，简称LDT）的主要任务是裁决两篇报道是否论述同一个话题。与TD类似，对于每一篇报道，不具备事先经过验证的话题作为参照，每对参加关联检测的报道都没有先验知识辅助系统进行评判。因此，LDT系统必须预先设计不独立于特定报道对的检测模型，在没有明确话题作为参照的情况下，自主地分析报道论述的话题，并通过对比报道对的话题模型裁决其相关性。LDT研究可以广泛地作为TDT中其它各项任务的辅助研究，比如TD与TT等等。

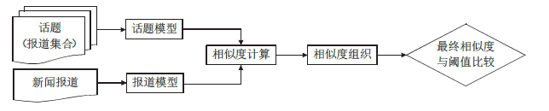


图1 TDT统一研究框架

随着话题检测与跟踪研究的逐步深入与发展，历次NIST举行的TDT评测都对该领域内的各项子课题提出了新的设想与方向，因此相应的评测任务也随之有所更改。比如，TDT2004撤销了报道切分任务（SST）,其原因不仅在于评测语料TDT5中没有包含广播类新闻报道，同时也由于应用中的大部分实例片断本身具备了良好的可区分性。此外，TDT2004将首次报道检测任务（FSD）转换成新事件检测任务（New Events Detection，简称NED）。虽然TDT2004对NED与FSD给与了相同的定义，但本文将这两者定义为目的不同但相互依存的任务。FSD与NED的区别在于前者注重鉴别事件初次报道的时空位置，后者除此之外还需要检测更多相关于事件的报道并进行汇总。此外，TDT2004首次提出了有指导的自适应话题跟踪（Adaptive topic tracking，简称ATT）和层次话题检测（Hierarchical topic detection，简称HTD）概念。

## 统一研究框架

TDT 的多个任务虽然研究目标不同，但仍存在着较多共同点。对其系统结构进行抽象，提出TDT 统一研究框架，如图1所示。

由图1可以看到，TDT 系统首先对话题或报道进行表示，然后确定其话题关联度计算方法（即表示模型间的相似度计算），最后根据实现任务的不同组织相似度与决策阈值比较。

几乎所有TDT 研究任务在一开始都存在一个选择：当前报道与已知报道比较，还是与已知话题比较。一般来说，话题关联识别和新事件发现多采用“报道/报道”比较；话题追踪和话题发现多采用“报道/话题”比较，但并不绝对，当把一个话题看作包含中心事件及其多个侧面的报道集合时，“报道/报道”和“报道/话题”可以互换。然而，随着研究的深入，发现“报道/话题”（或者“话题/话题”）更适合TDT中的各类比较。这是因为：TDT 以话题为粒度，以话题关联为中心，对一篇报道或一个报道集合，重要的是获取其包含的话题，而非某个单独的事件。首先对一篇报道进行话题相关或主题相关的信息扩充，然后建立其话题模型进行比较，系统性能得到了较大改进。

接下来下一章详细分析、比较图中几个主要模块的关键技术。

## TDT的实现方法

构造一个实用化的TDT系统是进行TDT研究的主要目的之一，也是检验现有方法优劣的基础。从参评的数量来看，话题发现和话题跟踪两个子任务最受关注。因此我们介绍的实现方法也以这两个任务为主。总体而言，要实现话题发现与跟踪功能，需要解决以下主要问题：

1）话题/报道的模型化

2）话题－报道相似度的计算

3）聚类策略

4）分类策略（阈值选择策略）

显然，对这种相关性必须做一个界定，不能任由集合无限扩大。为此，TDT会议组织者在构造TDT语料时，对挑选出来的每个话题都定义了相关性判定规则。

## TDT研究现状

TDT作为信息处理领域新颖的研究分支逐步成为国内重要的研究热点。相比于国外以统计概率模型为主体的研究趋势,国内的相关研究更侧重基于TDT本身的特色进行探索。TD T处理的信息是面向真实新闻事件的报道,其语义描述的精确性和可区分性更依赖于实体元素；此外 ,事件的产生和后续发展包含了报道之间的时序关系,其要求TDT系统不能单一基于内容建立相应的话题模型,而是融合时序特性参与检测报道间的关联性和跟踪话题的演化趋势。在此基础上,国内的相关研究也面向建立结构化和层次化的话题模型进行了初步尝试。名实体是描述话题或报道语义的一类特殊语言单位,其对于精确刻画核心内涵和区别不同主题具有重要意义。TDT系统应用名实体改进性能的方法主要包括如下两方面:⑴名实体特征权重的再分配,即希望区别名实体与其他特征对语义描述的能力; ⑵名实体相关性与其他特征相关性的线性组合,即希望通过人工或自动的方式调整名实体在相关性匹配过程中发挥的作用。国内较早将名实体融入TDT系统的研究来自贾自艳,其将文本内的特征标记为人名、地名和主题信息等类别,并预先指定每种特征类别的价值系数,特征的最终权重为词频和其所属类别价值系数的乘积。赵华则通过分析英文写作的习惯,自动识别新闻报道中首字母大写和全部大写的特征 ,其认为该类特征不仅包含名实体,也包含报道需要重点强调的特征 ,并在此基础上采用相关度加权和的方式评估话题与报道的相关性。上述方法在一定程度上改进了 TDT系统的性能,但由于是经验性地分配权重或调整相关性线性比例 ,因此无法保证系统性能的稳定性。张阔基于χ2 分布统计TDT2中各名实体类别(人名类、地名类和机构名类等) 与各话题类别(金融类、自然灾害类和科技类等) 的关联性 ,并将这一关联性的量化指标融入特征权重的再分配 ,其在提高 NED 系统性能的同时确保了这一改进的稳定性。限制名实体在 TDT 领域性能的另一因素是义同形不同的实体无法匹配。针对这一问题 ,宋丹面向地点类名实体建立地理树,匹配过程基于两名实体在地理树中路径的覆盖率进行计算,如“北京”在地理树中的路径“亚洲 —中国 —北京”与实体“北平”的路径基本一致,其高覆盖率可以有效匹配两实体之间的关联性 ,但该方法因无法处理诸如人名类等其他实体而存在局限性。在此基础上,骆卫华基于概念一致性匹配同义的名实体,其通过建立别称表和后缀表识别不同形态的名实体是否隶属于同一概念 ,如通过别称表识别“李光耀”和“李资政”为同一概念; 而基于后缀表识别“江苏省”和“江苏”为同义实体 ,该方法的缺陷在于依赖词典的规模和训练语料的新旧 , 对于报道流中最新出现的名实体依然无法匹配。如前文所述 ,话题起始于种子事件并包含后续相关事件 ,而构成事件描述的一项重要特性是其产生的时间 ,因此话题模型内各相关报道之间往往具备时序关系。国内将时序融入TDT领域的主要策略是将其作为相关性评估的附加元素,通过线性加权的方式调整相关度指标。贾自艳建立了统一时间表述方式的机制,在此基础上将当前报道与话题框架下新近事件的时间取差值 ,并利用该指标削弱基于内容匹配获得的相关度 ,其基本思想是: 报道与事件时序关系越近,则它们相关的概率越大。该方法提高了TDT系统检测与跟踪话题演化趋势的性能,与其相似的工作是赵华面向话题演化边界识别的研究,其训练一项表征话题演化周期的阈值,检测后续报道与话题模型内最新事件的时间差是否高于该阈值,将满足这一条件的报道作为话题演化的边界,该方法同样改进了TDT系统的性能。但由于上述方法或基于经验性的假设 ,或依赖于训练语料的规模 ,因此不能确保系统性能的稳定。有助于解决这一缺陷的研究来自宋丹的时间“覆盖矩阵”,其将相关性匹配双方的时间信息统一为标准格式 ,并分别映射于横纵时间轴上的点 ,基于对角线检测所有同步点及其时间间隔,在此基础上以所有间隔的覆盖率描述匹配双方时序关系的相似性。该方法可获得相对稳定的性能提高。但话题的出现存在跳跃性,即在较长历史时间段内 ,同一话题在一定周期内规律性的出现。这一现象限制了上述假设,即时序关系较近或匹配双方包含较多近似时间信息则相关性较高。因此 ,国内在TDT领域应用时序关系的研究仍有较大可提升的空间。话题模型层次化和结构化是目前TDT领域重要的研究方向。其中,层次化面向将同一话题下的相关报道组织为宏观到具体的层次体系;结构化则侧重挖掘和表征同一话题的不同侧面。国内尝试建立层次化话题模型的研究来自骆卫华和张阔, 前者首先基于时序关系对报道分组,然后进行组内自底向上的层次聚类 ,最后按时间顺序采用单路径聚类策略合并相关类;后者则面向报道全集建立层次化的索引树 ,树中第一层节点对应特定话题,而其子树则描述了该话题的层次体系 ,其建树过程基于输入的报道相对于树中各层次节点是否为新事件进行组织。上述两种策略在改进检测性能的同时也提高了系统效率,但在如何基于层次关系刻画话题语义及其演化趋势方面仍需要更深入的探索。针对结构化话题模型的研究来自赵华和金珠。前者分别尝试基于时序和特征分布密度识别话题演化的边界,在此基础上以演化边界为划分将话题描述为初始质心和当前质心两项子结构,后续报道与话题的相关性取自其与两项质心的相关性最大者。后者则对话题内的相关报道进行聚类,抽取聚类中的特征建立事件框架以描述话题的不同侧面,此外其通过HowNet建立事件内的情态关系和角色框架 ,扶助描述话题不同侧面的倾向性。该两种方法提高了话题跟踪系统的性能,尤其前者对话题演化趋势的识别和描述提高了跟踪系统的实用性。

总体而言,国内相关研究侧重挖掘TDT领域的特性,在方法上注重统计策略和自然语言处理技术相结合,在研究趋势上逐步面向融入数据挖掘、事件抽取和篇章理解等相关技术。此外,国内相关研究也朝着更加细化(如话题演化趋势的识别与跟踪)和更加实用化的方向发展。

# 关键技术分析

## 表示模型

为报道和话题建立表示模型是TDT中最基础的问题。下面分析和比较了TDT中经常使用的表示模型，包括特征选取、特征度量以及模型本身的改进等。

### 2.1.1 向量空间模型

向量空间模型一直是TDT中的主流表示模型。向量特征的选取范围通常是对报道词集进行停用词过滤后的候选集合，特征选取会直接影响使用该模型的系统性能。在TDT 中，向量模型的特征选取是一个日益体现新闻报道特点，逐渐细化的过程。开始时，对所有候选特征不作区分，用一个向量表示所有内容；

然后，为抓取报道的主要内容，抽取实体词进行向量表示，或者用两个向量分别表示实体词和其他名词。根据英文单词的拼写特征和位置特征抽取重要词进行表示；最后，为着重表现事件结构，对时间、地点、人物等特征词用不同向量分别表示；极端情况下，认为特征词的重要性随词性变化，根据词性对候选集合进行划分。特征选择方法的演变说明：人们已经意识到新闻文本的特殊性，TDT的表示模型应以话题为中心。向量模型中的特征度量通常是tf\*idf计算，以及在此基础上的改进方法。

对于新闻报道短小及话题漂移问题，向量空间模型多采用信息扩充的方法解决。例如，使用之前报道流中最新话题相关的报道进行扩充；用背景语料构建实体关联图，对表示模型中的实体词扩充。但也注意到，由于向量空间模型独立性假设的限制，文本到表示模型转换时丢失了所有的关联信息。图模型为解决该问题进行了初步研究，下面会详细介绍。

### 2.1.2 概率模型

概率模型是信息检索的经典模型，比向量模型有更好的理论基础。在特征选择上，概率模型和向量空间模型类似，都朝着与新闻报道特点结合越来越紧密的方向发展，试图细化话题描述的各个角度。例如，从事件词和主题词两处看待话题构成；使用实体词和其他词两种概率分布。实际研究中，向量模型比概率模型尝试了更多的特征选择方式。TDT 主要使用两种概率模型：语言模型和相关模型。面对稀疏问题，语言模型用背景语料进行参数平滑，相关模型用背景语料中的主题相关报道进行信息扩充。因而，相关模型的特征更丰富，概率分布的统计基础更扎实，在TDT中的应用也较好。概率模型对话题漂移关注较少。

从性能来看，概率模型尤其是相关模型能达到与向量模型相当甚至更好的效果。但相关模型和一元语言模型也存在特征独立性假设，n元语言模型中的词概率分布也只是依赖物理距离较近的前n-1个词。常用的解决方法是：分析文档内的依存关系并进行度量，将其用于计算条件概率分布。就用该方法使语言模型得到了更好地发挥。

### 2.1.3 词汇链

词汇链由一系列词义相关的词组成，最初被用于分析文本结构，帮助鉴别文本中心观点。词汇链被用于表示TDT的新闻报道，正是希望其能够捕获新闻

报道包含的话题。目前，TDT中构建词汇链通常借助Wordnet等语言资源中已有的词间关系。但这种关系不依赖具体的新闻报道，不能体现新闻报道内部真实的关联信息。

对词汇链的使用还没跳出向量模型的范畴，只是划分特征集的一种方式，其效果仍在基于向量模型的系统性能范围内。若想进一步改进，有必要获取报道内部特征间的真实关系，在词汇链的关系识别及使用方法上有所突破。

### 2.1.4 图模型

图模型最能体现新闻报道内容和结构，也最难构造。它关键在于含有其他模型经常忽略的关联特征，用结构图而非集合来表示文本，符合“新闻报道以事件为中心并向多个侧面发展”的特点。图模型在TDT的新闻表示中还属试探性研究。

TDT 中有两次图表示尝试：首先为每一句建立实体词距离图，然后合并距离图获取整篇文档图表示；其二首先识别特征词链接关系，然后以特征词为质心构造关系图。这些图中的关联关系都是通过计算结点在文本中的物理距离获得，从某种角度看，属于伪关联特征，可能给表示模型带来不必要的噪音。使用图模型的TDT 系统效果目前还不太理想，但有进一步研究的价值，尤其是图中关联关系的识别有较大的研究空间。

### 2.1.5 其他表示方法

除上述表示模型，还有一种分类器方法，即从报道或话题中抽取一个特征集合，用分类器对该集合分类。例如，用最大熵模型对抽取出的实体词、互指关系和实体词关系等特征进行判断；按信息增益选取特征，然后用决策树模型对其分类。但由于使用历史数据训练出的分类器会由于话题漂移而不够准确，把TDT 看作一种分类问题并不是很适合。

除了从模型本身考虑话题表示，TDT 还用一些独立于模型的技术进一步刻画报道描述的话题。例如：解决话题漂移的无监督反馈、自适应技术，使用已处理的相关报道对话题进行调整；对语料语言、来源、文本类型等进行标准化或分类处理；对话题、报道进行标准化进一步消除语料内部差异等，都有助于TDT 性能的改进。此外，话题漂移和时间关系密切，时间信息一直是报道表示中较受关注的因素。在TDT 研究中：挖掘报道中的时间表达式，通过表达式之间的关系来增强或减弱报道间的话题相似度；研究多种粒度的时间信息在动态数据流中的作用；层次话题发现对数据流首先按时间进行全划分，在每个划分中进行分层聚类，然后再聚成整体分层结构；都取得了较好的结果。

可以看到，在表示模型上，研究者已意识到新闻语料的特殊性，努力根据新闻本身的特点来抽取所描述话题的关键元素并进行合理度量，其中话题漂移更是模型表示不可忽视的问题。话题漂移是流数据挖掘中概念漂移的一个实例。概念漂移是数据流研究的特有问题，是与传统静态数据的主要区别之一，已得到越来越多地关注和研究。传统模型大多建立在静态数据集上，假设文本主题不漂移，这与话题的发展特性不符，TDT 的表示模型应该是动态的。解决方法主要分为两类：对传统表示模型改进，使之适应动态数据流的约束和特征，例如Ensemble Classifiers、CVFDT；或者根据流结构特征，提出新框架解决漂移问题，例如ANNCAD。这些方法都取得了与静态数据下的表示模型近似甚至更好的效果。但TDT 领域针对解决话题漂移还处在初级阶段，大多采用自适应、无监督反馈等主动学习技术，很少对模型本身进行改进，对训练样本的选择策略研究也较少，有待进一步深入。

## 相似度计算

相似度计算和表示模型关系密切，需要结合模型本身特点并充分利用表示模型的内容。TDT 对相似度研究较少，大多使用已有的方法。下面仍按照表示模型分类介绍：

1）向量模型为每个文档都建立等长向量表示。向量间相似度是相同维特征相似度的总和，计算方法有很多种，如Okapi 公式、Clarity、ellinger、Tanimoto、WeightSum等，其中余弦相似度最常用，也最有效。相关实验还表明：Hellinger 和Clarity 两种方法中，前者较适合新事件发现，后者较适合话题关联识别。

2）TDT 中概率模型的相似度计算有两种：语言模型下的生成概率和相关模型下的概率分布差异度量，前者不对称，后者对称。TDT 比较模型间的话题相似度，比较双方位置对等。因此，对称的相似度计算应该更适合，实验也表明相关模型下的分布差异度量表现更好。

3）词汇链的使用仍在向量模型的范围，相似度也基于向量空间模型。

4）比较两个图的相似度通常有两种方法：一是基于图论，例如，文献用Cook-Holder 度量图间距离，但这种方法的时间复杂性通常较高；二是把图转换为其他表示模型后比较，例如，每个质心结点对应一个向量，用向量相似度度量质心间的距离，然后综合构成图相似度。

5）特征集合间的相似度通常由分类器的判断值决定，例如，使用一种基于最大熵模型的多值分类器，但由于数据流中话题个数的不确定以及计算量大的原因不被看好；用决策树解决话题追踪的二值分类问题，由于已知信息少，训练出的模型性能有限；使用了一种自组织神经网络对报道分类。

## 基于相似度的组织方式

组织相似度是TDT 系统的重要组成部分，不仅体现出不同任务间的差异，对系统性能也会有较大影响。相似度组织通常从以下两方面来考虑：

从实现的任务考虑，话题关联识别和话题追踪通常相似度直接与阈值比较，新事件发现和话题发现则多采用KNN 组织方式；从改进系统性能考虑，例如，对新闻报道包含信息的粒度进行划分，采用“主题-话题-事

件”逐步细化的方式，用相似度来分层聚类。还有一种组合方法通常可以改进系统性能，并优于单个组成系统的性能。组合方法提供了一种改进系统性能的框架，独立于其组成部分，组成部分之间应是辅助互补的关系。比较有代表性的组合方法有：分别用支持向量机、投票方法、决策树组合向量模型间的各类相似度；尝试了支持向量机的组合方法；线性组合多个弱分类器分类结果；线性组合两个互补的语言模型系统；则线性组合各类实体词间的相似度；对篇章相似度、句子关联度和地点关联度进行组合。这些方法都较好地改进了系统的整体性能。上述两类相似度组织方式之间并不互相排斥，可以同时使用多种。

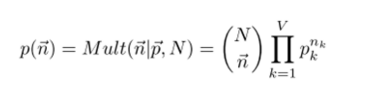
# 主题模型相关知识

## 主题模型简介

LDA主题模型的主要目标就是从一篇文档或多篇文档计算出其主题－词的概率分布情况．LDA主题模型是一种典型的词袋模型，也是一种无监督学习算法，但相对于传统的词袋模型，LDA增加了词之间的关联，在讲LDA模型之前，先介绍下 Unigram Model （词袋模型）和Bayes Unigram Model（贝叶斯词袋模型）：

### 3.1.1 Unigram Model（词袋模型）

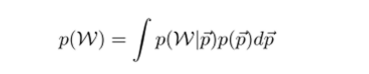
LDA既然是聚类算法，而聚类算法大多数时候，都在寻找两个东西的相似度。最开始，大家想要判断两篇文档是否相似，最简单直接的方法就是看文档里出现的词是否一样，其个数是否相近。于是Unigram Model（词袋模型）就是实现这样的思路设计的。所以为了得到文档集合中，所有文档的共性的规律，词袋模型，假设：一篇文档生成的过程就是 独立的抛一个具有M面的骰子（M是所有词的个数），N次（N是该文档里词的个数），这样文档的生成，刚好可以看作是个多项式分布：



文档集合中，每个词出现的概率就是要求的参数，通过EM算法可以确定下来，这样就得到了模型。

### 3.1.2 Bayes Unigram Model（贝叶斯词袋模型）

在词袋模型中，我们简单的认为文档里每个词出现的概率是个定数（即骰子的每个面的概率），但Bayes学派不这么认为，他们认为这些概率应该是一个随机过程产生的，于是生成一篇文档的过程可以描述为：先随机抽取一个M面的骰子， 再用这个骰子独立抛N次。那么这个模型的分布如下：

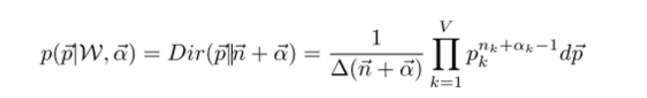


其中后边部分https://images2018.cnblogs.com/blog/1180120/201808/1180120-20180829225943822-1321585911.png，是多项式分布，我们已经知道，为了方便计算我们假设https://images2018.cnblogs.com/blog/1180120/201808/1180120-20180829230015106-1714926802.png为Dirichlet分布，它是多项式分布的共轴分布

简单介绍下 Dirichlet 分布：比如 抛了100次骰子，得到6个面的一个概率，记为一个实验，重复这个实验100次，那么这100次的实验中，这6个面的概率的概率分布，就是Dirichlet分布，它是分布之上的分布。

例如：1点（骰子六个面之一） 在这100次实验（每个实验抛100次） 是 0.15的概率为 0.12，实际我们这么想，100次实验中，有12次，1点在一个实验内出现了15次，可以看作是总共抛10000次，1点出现15×12=180次。这10000次实验，视为一个大的多项式分布，于是可以得出他们有相同的概率分布公式，这就是前面所提到的共轴分布，且有如下性质：

先验的Dirichlet分布+多项式分布 = 后验的Dirichlet分布



上述的例子中，你会发现，它与我们的Bayes Unigram Model（贝叶斯词袋模型）已经很相似了。一个实验里的100次抛骰子，可以看作是先验的Dirichlet分布，也就是模型中确定骰子各个面概率的那个随机过程，而重复这个这个实验100次，可以看作是后面的根据这个骰子确定文档的一个过程。

词袋模型本身有一定的局限性：

1）“词袋”模型仅关注词项出现的频率，而未考虑它们之间的区别，这会导致一词多义问题的存在；

2）“词袋”模型同样没有考虑到不同词项可能会表达同样意思的情况，即“一义多词”的问题，比如“苹果”"iPhone”两个词在电子产品主题里就属于一个意思；

3）使用“词袋”模型来表示文档时，其维度可能会很高。然而，我们知道两篇文档是否相关不仅仅要看它们之间的词语重复度，更重要的影响因素是文本内容中存在的隐含语义关系。

LDA算法，就将文档和词，投射到了一组Topic上，试图通过Topic找出文档与词间，文档与文档间，词于词之间潜在的关系；由于LDA属于无监督算法，每个Topic并不会要求指定条件，但聚类后，通过统计出各个Topic上词的概率分布，那些在该Topic上概率高的词，能非常好的描述该Topic的意义。

从主题模型的生成过程可以看出，一篇文档中词的产生是先通过得到文档中各个主题的概率值，再根据每个主题下各个词的概率来最终确定文档中词的概率分布。可以用这个公式P（词语|文档）= ∑主题P（词语|主题）ｘP（主题|文档）来表示每个词语在生成文档中出现的概率，用矩阵表示就如图2所示：

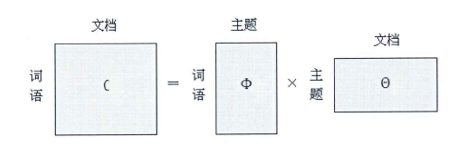


图2 词语在生成文档中出现概率的矩阵表

矩阵C表示文档的词语概率分布，矩阵。表示主题下的词语概率分布，矩阵。表示文档下的主题概率分布。所以，从主题模型的建模过程可以看出:首先，文本的生成过程是通过定义的主题模型来模拟;其次，利用对己知数据的学习来估计出主题模型的各个参数;最后，通过已得到的模型，来估计文本中各个主题的词概率分布和整个文本集中每篇文本的主题概率分布，进而通过算法挖掘出各个文档的语义相似度。

## 传统主题模型

### 3.2.1 LSI模型

LSA（latent semantic analysis）潜在语义分析，也被称为 LSI（latent semantic index），是 Scott Deerwester, Susan T. Dumais 等人在 1990 年提出来的一种新的索引和检索方法。该方法和传统向量空间模型（vector space model）一样使用向量来表示词（terms）和文档（documents），并通过向量间的关系（如夹角）来判断词及文档间的关系；不同的是，LSA 将词和文档映射到潜在语义空间，从而去除了原始向量空间中的一些“噪音”，提高了信息检索的精确度。

传统向量空间模型使用精确的词匹配，即精确匹配用户输入的词与向量空间中存在的词，无法解决一词多义（polysemy）和一义多词（synonymy）的问题。实际上在搜索中，我们实际想要去比较的不是词，而是隐藏在词之后的意义和概念。

LSA 将词和文档映射到潜在语义空间，从而去除了原始向量空间中的一些“噪音”，提高了信息检索的精确度。LSA 的核心思想是将词和文档映射到潜在语义空间，再比较其相似性。LSA 模型就是对词-文档共现矩阵进行SVD，从而得到词和文档映射到抽象出的topic上的向量表示。

引用吴军老师在 “矩阵计算与文本处理中的分类问题” 中的总结如图3：

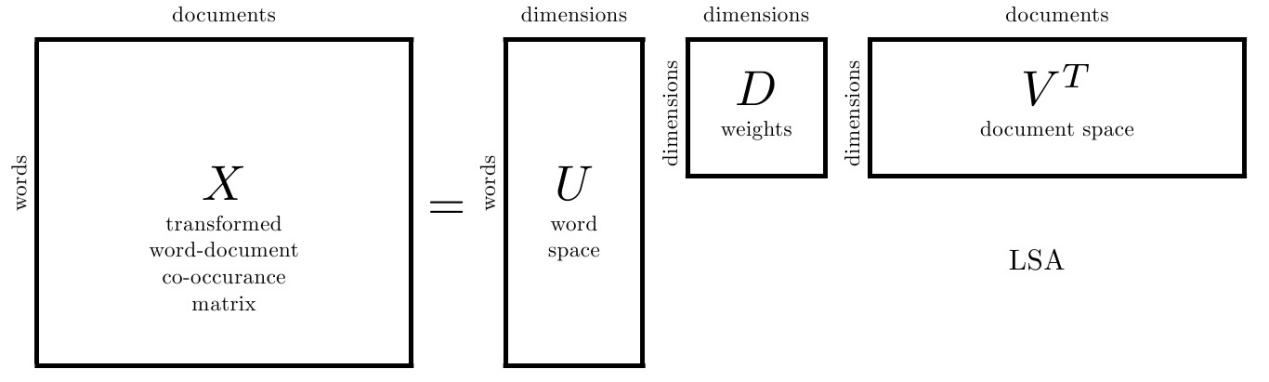
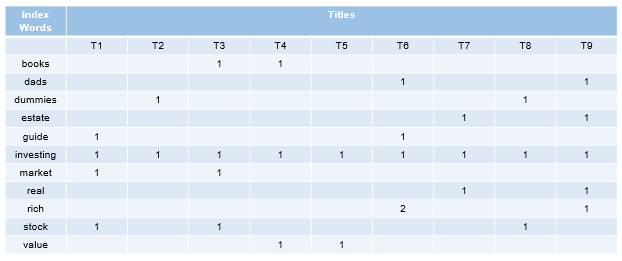
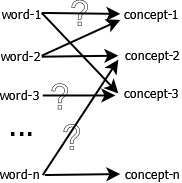


图3 矩阵计算与文本处理中的分类问题

表1 Term-Doucement矩阵



三个矩阵有非常清楚的物理含义。第一个矩阵 U 中的每一行表示意思相关的一类词，其中的每个非零元素表示这类词中每个词的重要性（或者说相关性），数值越大越相关。最后一个矩阵 V 中的每一列表示同一主题一类文章，其中每个元素表示这类文章中每篇文章的相关性。中间的矩阵 D 则表示类词和文章类之间的相关性。因此，我们只要对关联矩阵 X 进行一次奇异值分解，我们就可以同时完成了近义词分类和文章的分类。（同时得到每类文章和每类词的相关性）。传统向量空间模型使用精确的词匹配，即精确匹配用户输入的词与向量空间中存在的词，无法解决一词多义（polysemy）和一义多词（synonymy）的问题。实际上在搜索中，我们实际想要去比较的不是词，而是隐藏在词之后的意义和概念。

可以通过一个简单的例子来看一下整个流程：

Term-Document 矩阵如表1：这里的一行表示一个词在哪些title中出现了，一列表示一个title中出现了哪些词（停词已去除）。比如说T1这个title中就有guide、investing、market、stock四个词，各出现了一次。

对 Term-Document 矩阵做SVD分解如图4：

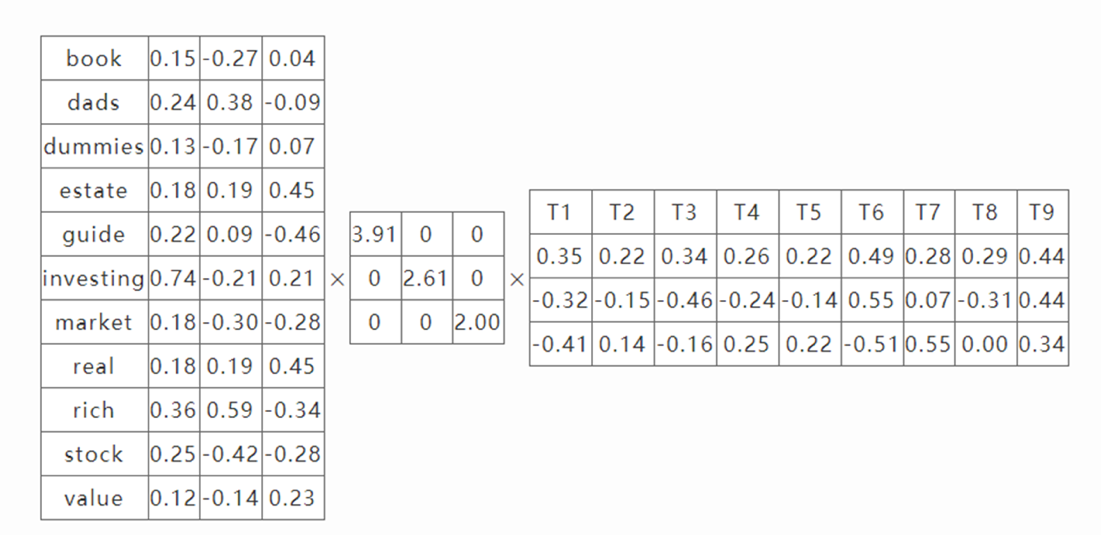


图4 Term-Document矩阵的SVD分解

可以将左奇异向量和右奇异向量都取后2维（之前是3维的矩阵），投影到一个平面上（潜在语义空间），可以得到：

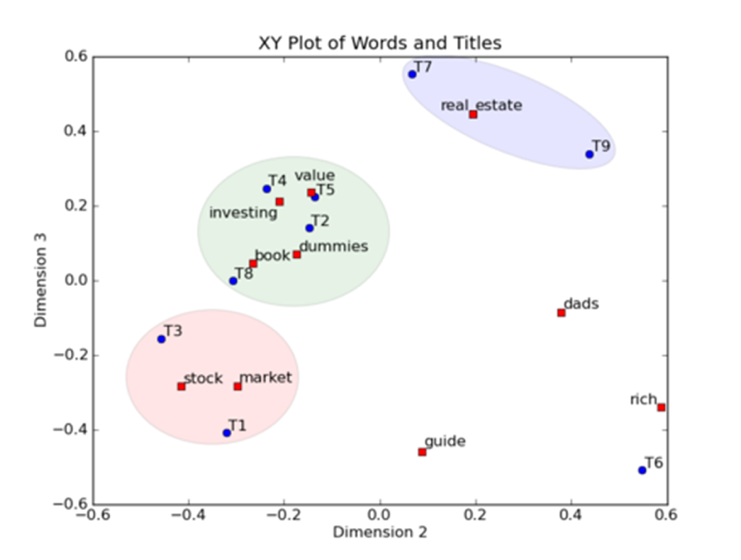


图5 词的分布表示

在图上，每一个红色的点，都表示一个词，每一个蓝色的点，都表示一篇文档，这样我们可以对这些词和文档进行聚类，比如说 stock 和 market 可以放在一类，因为他们老是出现在一起，real 和 estate 可以放在一类，dads，guide 这种词就看起来有点孤立了，我们就不对他们进行合并了。按这样聚类出现的效果，可以提取文档集合中的近义词，这样当用户检索文档的时候，是用语义级别（近义词集合）去检索了，而不是之前的词的级别。这样一是可以减少我们的检索、存储量，因为这样压缩的文档集合和PCA是异曲同工的，二可以提高我们的用户体验，用户输入一个词，我们可以在这个词的近义词的集合中去找，这是传统的索引无法做到的。

LSA 通过将词映射到topic上得到distributional representation（词的分布表示如图5），进而缓解文档检索、文档相似度计算等任务中所面临的同义词（一义多词）问题：比如我搜索“Java 讲义”，如果系统只是用字符匹配来检索的话，是不会返回一篇出现了“Java 课件”但通篇没出现“讲义”这个词的文档的。所以说，单纯地从词-文档共现矩阵取出词向量表示和文档向量表示的向量空间模型，尽管利用了大规模文档集的统计信息，仍然是无法直接从“语义”这个层面上去理解文本的。但是 LSA 这种将词映射到topic上的向量表示，很难去应对一词多义问题：比如“Java”这个词既可能指编程语言，也可能指爪哇岛，即使这两种含义的“Java”都在文档集里出现过，得到的 LSA 模型也无法很好地区分。但是LSA不能解决一词多义（polysemy）问题。LSA将每一个词映射为潜在语义空间中的一个点，也就是说一个词的多个意思在空间中对于的是同一个点，并没有被区分。

另外，SVD的优化目标基于L-2 norm 或者 Frobenius Norm（弗罗贝尼乌斯范数）的，这相当于隐含了对数据的高斯分布假设。而 term 出现的次数是非负的，这明显不符合 Gaussian 假设，而更接近 Multi- nomial（多项） 分布；

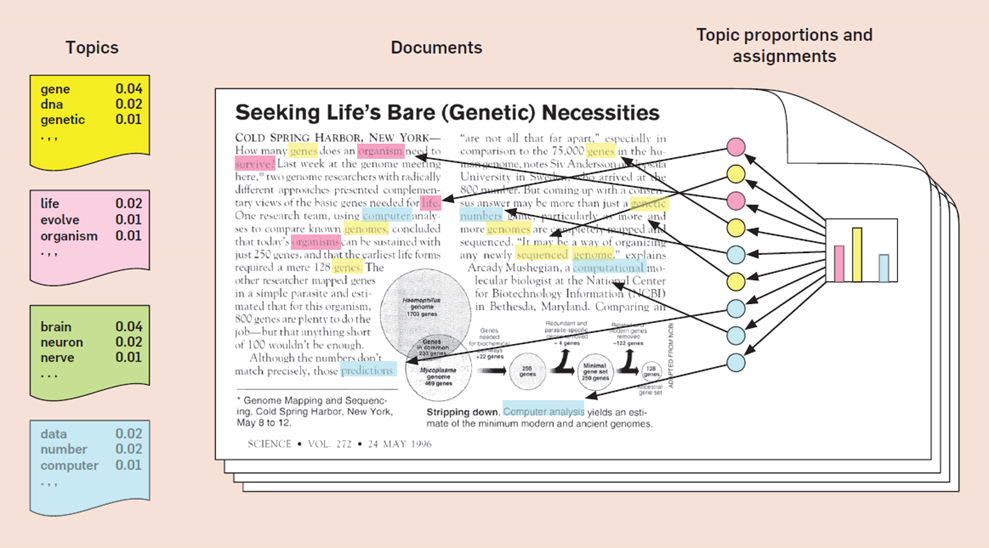


图6 话题生成

特征向量的方向没有对应的物理解释；SVD的计算复杂度很高，而且当有新的文档来到时，要更新模型需重新训练。

### 3.2.2 PLSI主题模型

日常生活中人们是如何构思文章的？如果我们要写一篇文章，往往是先确定要写哪几个主题，比如构思一篇自然语言处理相关的文章，可能40% 会谈论语言学，30% 谈论概率统计，20% 谈论计算机，10% 谈论其他主题：说到语言学，我们容易想到的词包括语法、句子、乔姆斯基、句法分析、主语……我们之所以能马上想到这些词，是因为这些词在对应的主题下出现的概率很高，我们可以很自然的看到，一篇文章通常由多个主题构成、每个主题大概可以用与该主题相关的频率最高的一些词来描述。

LSA可以用来解决一义多词的问题，但缺乏严谨的数理统计基础，同时SVD分解非常耗时，pLSA就是在这样的背景下被提出的。一篇文档（Document）可以由多个主题（Topic）混合而成，每个Topic都是词汇上的概率分布，文章中的每个词都是由一个固定的Topic生成的。

和很多模型一样，pLSA 遵从 bag-of-words 假设，即只考虑一篇文档中单词出现的次数，而忽略单词的先后次序关系，且每个单词的出现都是彼此独立的。



图7 PLSA中生成文本的游戏规则

在PLSA模型中，Hoffman认为上帝是按照如下的游戏规则来生成文本的如图7:

PLSA模型的文档生成过程就可以用图形化的表示为图8：

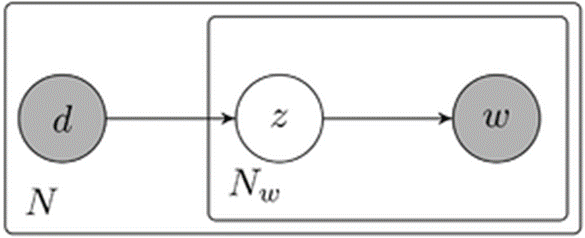
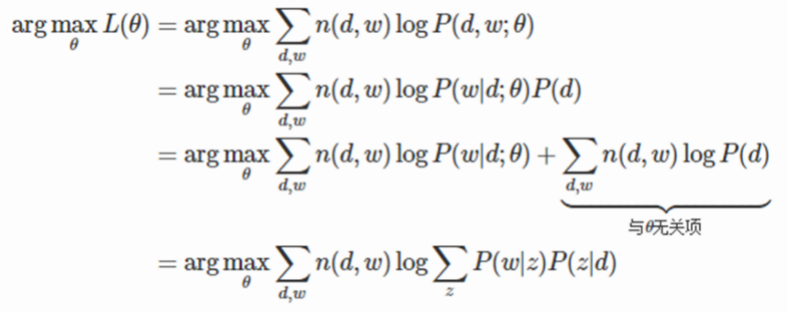


图8 PLSA模型的文档生成过程

想象某个人要写 N篇文档，他需要确定每篇文档里每个位置上的词。假定他一共有 K个可选的主题，有 V个可选的词项，所以，他制作了 K个 V面的 “主题-词项” 骰子，每个骰子对应一个主题，骰子每一面对应要选择的词项。然后，每写一篇文档会再制作一颗 K面的 ”文档-主题“ 骰子；每写一个词，先扔该骰子选择主题；得到主题的结果后，使用和主题结果对应的那颗”主题-词项“骰子，扔该骰子选择要写的词。他不停的重复如上两个扔骰子步骤，最终完成了这篇文档。重复该方法 N次，则写完所有的文档。在这个过程中，我们并未去关注词和词之间的出现顺序，所以PLSA也是一种词袋方法，即只考虑一篇文档中单词出现的次数，而忽略单词的先后次序关系，且每个单词的出现都是彼此独立的；并且我们使用两层概率分布对整个样本空间建模，所以PLSA也是一种混合模型。

而真实的PLSA方法两个“骰子”可能就不是均匀的，因为每个主题的概率不一定一样，主题下每个词的概率也不一定一样。在PLSA模型中，我们需要做的就是如何求出这两个概率。

我们来看下PLSA的核心思想（概率图模型），如图9。

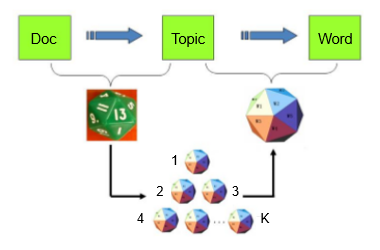


图9 PLSA的核心思想

实心的节点 d 和 w 表示我们能观察到的文档和单词，空心节点 z 表示我们观察不到的隐藏变量，用来表示隐含的主题。

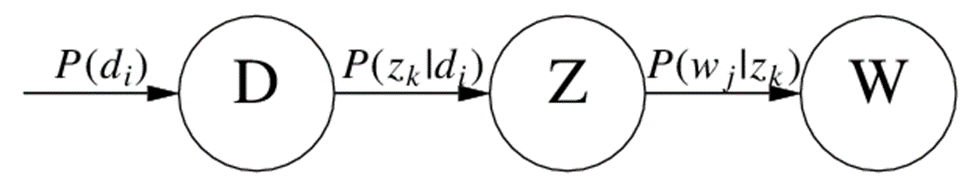
具体来说，该模型假设一组共现（co-occurrence）词项关联着一个隐含的主题类别 ∈{,…, } 。有如下三个相关的概率：

：表示词在文档 中出现的概率；

： 表示某个词 在给定主题 下出现的概率；

：表示某个主题 在给定文档 下出现的概率。

利用这三个概率，可以按照如下方式得到“词-文档”的生成模型：



1）按照概率 选择一篇文档

2）按照概率 选择一个隐含的主题类别

3）按照概率 生成一个词

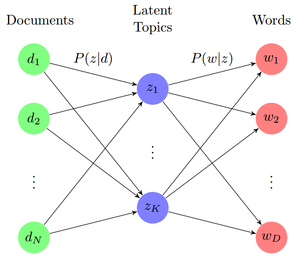
并且每个主题在所有词项上服从 Multinomial （多项）分布，每个文档在所有主题上服从 Multinomial 分布。同时，我们认为在给定 Topic 𝑧 的条件下，单词 𝑤和文档 𝑑之间是条件独立的，也就说：

那么，

模型中的参数包括：和 。可以采用极大似然估计，目标函数为：

其中，表示单词𝑤在文档𝑑中出现的次数。

记 （我们要估计的参数）， 取对数，得到

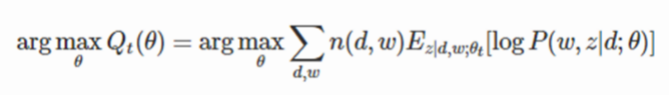


这是一个非凸优化（non-convex optimization）问题，利用EM算法来估计参数。

EM算法的核心思想是：

1.E-step: 假定参数已知，计算此时隐变量的后验概率；

2.M-step：带入隐变量的后验概率，最大化样本分布的对数似然函数，求解相应的参数重复。



利用矩阵分解来解决实际问题的分析方法很多，如PCA（主成分分析）、ICA（独立成分分析）、SVD（奇异值分解）、VQ（矢量量化）等。在所有这些方法中，原始的大矩阵 V 被近似分解为低秩的 V=WH 形式。这些方法的共同特点是，因子 W 和 H 中的元素可为正或负，即使输入的初始矩阵元素是全正的，传统的秩削减算法也不能保证原始数据的非负性。在数学上，分解结果中存在负值是正确的，但负值元素在实际问题中往往是没有意义的。例如图像数据中不可能有负值的像素点；在文档统计中，负值也是无法解释的。

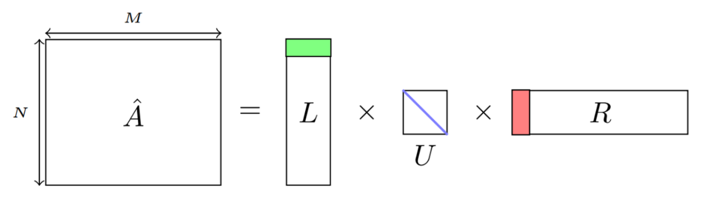


图10 矩阵分解示意图

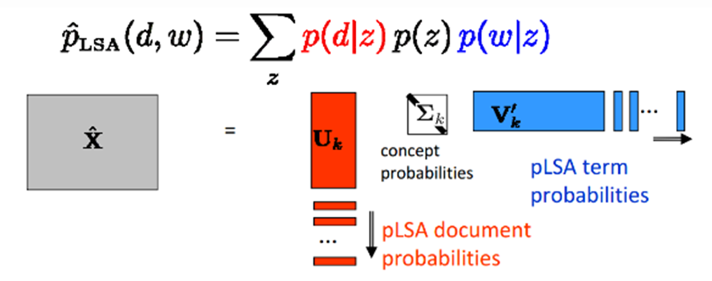
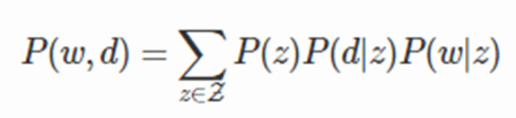


图11 PLSA和LSA对应关系

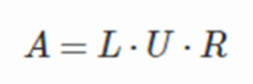
因为我们要得到的是概率，负值是没有意义的，所以这里可以采用NMF（非负矩阵分解）的方式来呈现 PLSA，根据公式



可以考虑利用矩阵分解的方式得到:， ，，形式为

如图10，图中，

矩阵中，N表示的是文档数，M表示单词总数；



L包含文档概率𝑃𝑑|𝑧)；

U是主题的先验概率𝑃𝑧)，是一个对角矩阵；

R表示单词概率𝑃𝑤|𝑧)。

PLSA模型是有向图模型，将主题作为隐变量，构建了一个简单的贝叶斯网，采用EM算法估计模型参数。相比于LSA略显“随意”的SVD，PLSA的统计基础更为牢固。PLSA与LSA之间的对应关系如图11所示。

从图中我们可以看到pLSA与LSA之间的对应关系，其中pz)刻画了Latent Space，也即topic space 的信息； pw|z)刻画了topic space与term space之间的关系，对应着LSA中的正交基V；在文档分类时，这两部分也就是我们在模型训练结束需要保存的信息，当一个新的文档的到来时，我们可以再次利用 EM 算法得到新的文档与主题的对应关系pd|z)，并由此得到文档在topic空间上的表示pz|d)。

PLSA的优势体现在以下四个方面：

1）定义了概率模型，而且每个变量以及相应的概率分布和条件概率分布都有明确的物理解释；

2）相比于LSA隐含的高斯分布假设，pLSA隐含的Multi-nomial分布假设更符合文本特性；

3）pLSA的优化目标是使KL-divergence最小，而不是依赖于最小均方误差等准则；

4）可以利用各种model selection和complexity control准则来确定 topic的维数。

当然，它也存在着一些不足：

1）概率模型不够完备：在document层面上没有提供合适的概率模型，使得pLSA并不是完备的生成式模型；

2）随着document和term个数的增加，pLSA模型复杂度也线性增加，变得越来越庞大。

## LDA主题模型

LDA主题模型由David M.Blei等在2003年提出，它是关于离散数据的概率生成模型，用主题概率来表示一篇文档。LDA模型在自然语言处理领域中有着很广泛的应用。

LSI概率化得到概率主题模型PLSI，但是PLSI不是完整的概率主题模型，因为它在文档层并没有使用概率模型，也就是说文档的主题并没有提供概率生成模型，它会导致模型的参数会随着语料库的规模线性增长，从而使得模型出现过度拟合的问题，还同时会引起在测试数据上不知如何分配文档概率的问题。针对这个问题，LDA在PLSI模型的基础之上又做了一些改进，使主题模型的效果更好。LDA加入了三层贝叶斯模型，使的文档的主题概率分布也使用概率生成模型得到，也就是说每篇文档的词来自于主题的分布，而主题的分布的参数也是随机生成的。贝叶斯参数的估计可以用变分推断方法，也可以使用EM算法。

### 3.3.1 LDA模型表示

假如给定一个主题文档集合M，它含有m个文档，含有K个相互独立的主题，每个主题呈现随机多项式概率分布特点。每个主题下又存在若干个多项式概率分布词汇，这里的多项式概率分布均满足Dirichlet分布．上述文档的表现形式可刻画为LDA主题模型，它的生成过程如下:

1) 从 Dirichlet分布先验知识α中取样生成文档Mm的主题分布θm .

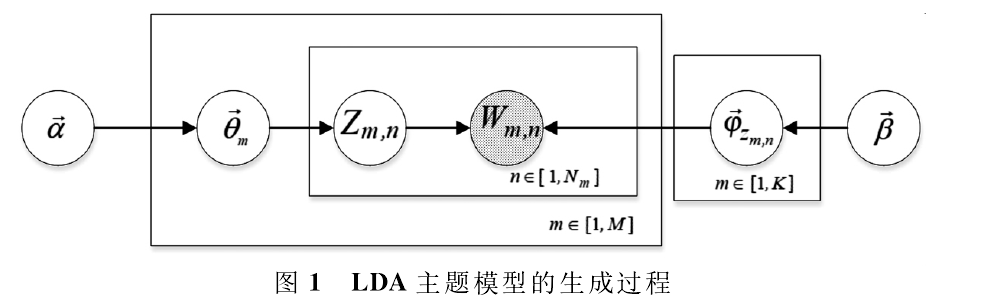
2) 从1)中的主题分布θm中取样生成文档Mm的第n个词汇的主题Zm,n．

3) 从Dirichlet分布先验知识β中重复取样生成与主题Zm,n相对应的词汇分布φZm,n．

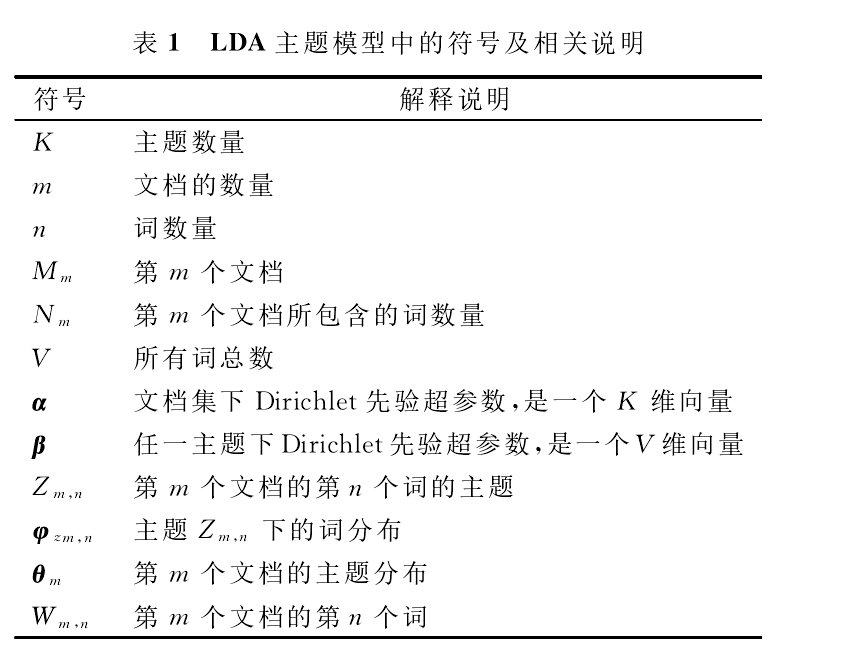
4) 结合主题Zm,n，从词分布φZm,n中采样生成词汇Wm,n．

5) 重复执行以上步骤2)和3), 重复执行以上步骤2)、3)、4)，直至所有词汇被选择．

以上步骤的直观过程如图1 LDA主题模型的生成过程：



LDA模型的符号相关说明见表1：



### 3.3.2 LDA模型参数估计

模型的参数估计可以看成是训练过程的逆过程，就是在已知这结果的情况下通过参数估计方法得到参数值。我们通常选择Gibbs抽样的方法实现参数估计过程，因为这种方法更容易描述和实现。使用Gibbs抽样方法学习LDA主题模型的过程如下:

1）初始化，对语料库中的每篇文档的每个词，随机赋予一个主题topic编号;

2）重新扫描语料库，对每个词w，按照Gibbs抽样公式重新采样语料库的主题，在语料中进行更新;

3）重复以上语料库的重新采样过程直至Gibbs抽样收敛;

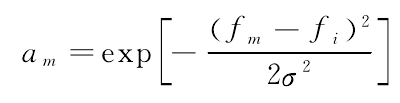
4）统计语料库中的topic-word的共现频率矩阵。

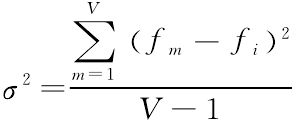
得到主题下词对的概率分布后，通过统计每篇文档中的主题，就可以计算出文档下主题的概率分布。

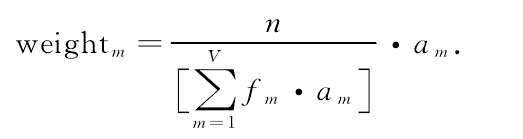
### 3.3.3 改进的W-LDA主题模型的推导

­­­­­­­­­­­­­­­­­­­­­­­如果仅使用传统的LDA模型，高频词会将代表主题的多数词淹没掉，从而使得主题表达能力受到较大影响．为解决这一问题，有研究者通过停用词或者设置阈值剔除掉高频词，但是这些方法在一些应用领域中由于缺乏停用词词库或者词库不全，且阈值大小难以把握，因此效果并不理想．在此背景下，使用高斯函数对特征词作加权处理，以此构建一种改进的 W-LDA主题模型．具体的实现方法及过程如下：

首先运用下式对文档中的词am加权



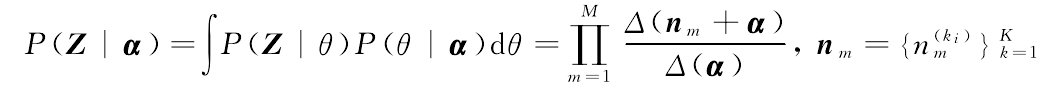
其中方差fm 是第m个词的词频数，fi是词频数中间的第i个词的词频数．为使数据集被加权前后其总词数大体一致，在上式的基础上作均值化处理，处理公式如下式所示：平均加权值 weight m将被应用于主题及其词语分布期望中的计算



在 LDA主题模型推导中，本文采用Gibbs抽样算法，该算法快速高效，可生成马尔科夫链，继而可求得一个复杂的多元分布．主题多项式分布θ由先验参数α 生成，Zm,n主题下的词语多项式分布是由β生成，则有如下所示的联合概率分布：

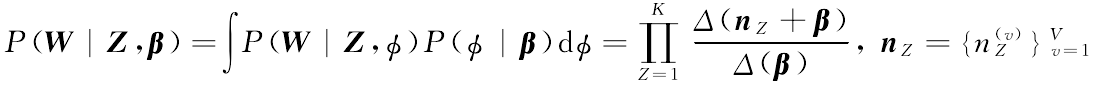


求解式上式分为两个过程：一是根据先验参数α采样主题Z的过程，二是根据主题Z和先验参数β采样词语的过程．式子中的因子 Ｐ（Ｚ｜α）的计算公式为：



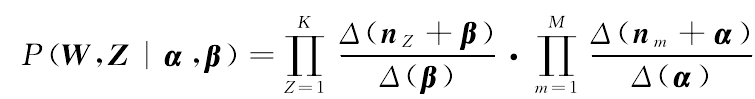
其中，n(ki)m代表第K主题在第m文档中出现的次数

同理，可得到:

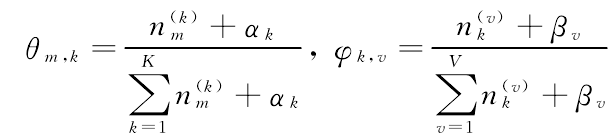


其中，n(v)Z表示在主题Z中词语v出现的次数，K为模型的乘积．

可得到词语和主题的联合分布公式为



在这个公式的基础上，依据变量W下的隐性主题变量Z的条件分布和狄利克雷分布情况，可求得第m文档的主题及主题下的词语分布期望为:



当把主题分布到第m文档时，对n(ki)m变量不是简单地累加1，而是累加平均加权值．

### 3.3.4 基于LDA模型的应用

1）文本分类

文本分类是指根据训练文档集学习得出分类系统，从而判别未知新文本的类别。LDA作为一种无监督文本表示模型，其本身并不能直接判断文本类别，必须集成到合适的分类模型中协同工作。在最初利用LDA模型的分类过程中，LDA对给定训练集中所有文档进行特征降维，将文本训练集表示成主题的概率分布，在得到的主题-文本矩阵上选择一种分类算法进行训练，构造出文本分类器。假设同一类别的文档共享同一类主题，不同类别的主题之间是独立的，因此每个类别的文档都将对应一个LDA子模型。在判断新文本类别时，将训练得到的每一个LDA模型对新文本进行主题推断，计算出此文档的主题概率分布，之后根据分类算法判别新文本属于哪一个类别。新文本的类别其实只有一种，但在分类过程中却存在着对主题的强制分配，因此很多研究人员针对这一缺陷进行了改进，如李文波等人提出通过附加类别标签的Labeled-LDA模型改善分类效率，该模型协同计算新文档在各类别上隐含主题的分配份额，克服了传统LDA模型强制分配主题的缺陷，提高了LDA分类性能。Zhao等人提出主题文本分类的gI,DA新算法，通过在原始LDA中添加主题，类别分布参数，模型中的文本可划分为不同种类，每个类别共享一组主题，使得文本是由相关的类别所生成，极大避免了主题-词的错误分配。

2）文本聚类

文本聚类作为一种无监督的学习方法，将相似度较高的相关文本划分为相同类别集合，低相似度的文本则分布在不同的类别集合中，不同类别的集合对象彼此间差异较大。传统的聚类方法完全依靠词频等基本特征作为聚类分析依据，忽略了文本语义信息，聚类效果并不明显。人们开始考虑引入主题模型对文本建模，按照语义关系对文本进行聚类分析，每一聚簇内文本共享同一主题，使聚类更符合文本语义特征，提高了文本聚类的准确性。如Lee等人提出了一种基于I,DA技术的交互可视分析系统iVisClustering，将LDA应用于自动聚类过程中，通过具有代表性的关键字为每一个子聚簇赋予一个概括总结的同时，生成聚类结果，并允许用户实现对聚类过程的指导。Xie等人基于文本聚类和主题建模互相作用的思想上，提出一种MUCTM（multi-grain clustering topic model）模型，使文本聚类和文本建模可以同时发生。通过估计全局主题分布参数和局部主题分布参数，可以得出更多对聚类分析有用的文本特征，从而提高聚类效果。

3）摘要抽取

摘要抽取是指从文本中抽取出能概括出文本核心内容的总结性句子，包括单文档和多文档摘要抽取。LDA模型可建立多文档之间的语义联系，排除文档间的冗余成分，提高了多文档摘要抽取的准确性。如Arora等人假设文档中的每条语句对应同一个主题，利用L DA技术计算出每个主题下的单词权重，而每条语句定义为各词语权重的向量，然后利用奇异值分解算法计算出最能表达这些主题含义的句子作为文档摘要，这种面向主题的方法利用LDA和SVD混合模型降低了最终文摘中的冗余成分。Petinot等人将文档层次看成是一棵具有代表类别的内部结点和代表文档的叶子结点，提出hLLDA（hierarchical Labeled Latent Dirichlet Alloca-tion）模型，利用多层次标签建立每个标签与主题之间的对应关系，通过类别分层自动提取文档摘要。

4）情感分析

情感分析是指通过挖掘带有个人观点的文本，分析并整理出用户所表达的情感倾向信息。最初的J情感分析主要针对文章整体的情感倾向，而在一篇文章中通常可能会涉及多个话题的讨论，用户对不同的话题通常会持不同的情感倾向，同一个情感描述词在不同文本主题下所表达的情感倾向也可能完全不同，因此I,i等人考虑到情感表达与文本所描述主题间的对应关系，提出基于主题的情感分类模型Sentiment-LDA，该模型在LDA基础上引入了一个情感变量，假设文本中每个单词是由主题和情感变量的联合分布共同生成，实现了对一篇文本所包含的所有主题进行相应的情感分析。Fu等人在LDA基础上依赖Hownet词典对博客文本进行多方面话题情感分析，通过训练LDA模型得出多个话题子模型，采用KL-divergence算法判断文本各段落的话题，利用基于Hownet的情感词标注方法计算每个博客话题所包含句子的情感定位，从而总结出博客文本中所有话题的情感定位。

## BTM主题模型

我们当前正处于社交媒体逐渐取代传统媒体的信息革命时代,在线的社交媒体平台,如Facebook、Twitter,以及国内的新浪、腾讯等公司,它们创建了“普通用户,即是发布者也是传播者”的信息传播模式,使得信息的产生与传播都比以前更加即时、方便、快捷。随着移动智能终端的普及,越来越多的用户愿意随时随地发表自己的意见,分享自己的经历,甚至表达自己的政治观点等,社交媒体已经成为了一个关于公共事件、舆情研究,甚至个人情绪观点的有价值的数据源泉。互联网普通用户每天通过社交媒体产生海量（TB级）的短文本数据,其中的信息含量也是传统媒体所无法比拟的。

因此,研究如何从这些海量社交媒体短文本中发现有用的信息是当前一项重要的挑战,也已经成为国际学术界的一个研究热点。主题模型被用来进行文本挖掘已经被证明是一个非常有效的手段,随着即时通讯的发展,挖掘这些海量短文本的主题特征也变得越来越重要。但是,因为短文本的稀疏性,把传统的主题模型（如LDA）直接作用在短文本上,挖掘的主题特征效果及性能都不是很令人满意。本文在研究LDA主题模型和BTM（Biterm Topic Model）主题模型的基础上,提出一种针对于海量短文本的主题模型。在此主题模型中,通过对每篇短文本获取“词对（biterm）"共现的方式来对文本建模,代替传统主题模型中的“词”共现的方式,这种方式可以在一定程度上缓解数据稀疏的问题,提高了模型对主题挖掘的效果。同时,又保留了一些传统主题模型的优秀特性。

近年来，关于使用主题模型对海量短文本进行主题挖掘的研究受到了业界广泛的关注，各种模型、方法也不断涌现。

在国外，为了改善LDA直接应用于短文本而产生的语义稀疏问题，Xuan-hieu Phan等使用外部背景资料与短文本进行结合，来丰富短文本的语义信息，从而解决文本数据稀疏问题。D.Bollegala等通过搜索引擎的片段来扩展短文本的上下文，使用这些搜索片段来尝试解决短文本的数据稀疏性。Hotho A等引入WordNet从而添加短文本的特征扩展。 Grube:等人假设同一个短文本的词语都出之于同一个主题。 Michal Rosen-Zvi等人将每条tweets的发布者信息

都作为该tweets的附加信息。 Sahami Me等人将领域相关的搜索片段作为短文本的背景扩展来增加上下文语义。Daniel Ramag等人将推文的HashTag作为tweets标签，对其直接使用LDA也有不错的效果，前提是每条tweets都必须有Hashtag。

在国内，Yan XY等人将整个语料库打碎成不同的biterm，通过对biterm训练学习，来获取主题分布和词分布。 Weng J等人将同一个作者发布tweets聚成一篇长文档，再用LDA训练，当作长文档来处理。 Hong L等人将包含相同词语的微博都聚合成同一篇文档，且其效果确实比直接用LDA更好。Zhao W X等人直接在tweets上用混合主题模型。 Yue Lu在优化公式的基础上的改进，增加先验信息来区别不同的主题。Xuerui Wang等在传统模型上添加了词与词之间的顺序的考虑。Qiaozhu Mei等引入了一些验证和关联信息，添加了规则化因子。Mei Q等人将主题分为局部主题和全局主题，前者用于特征后者用于实体。Cheng X等人考虑主题数目对模型的影响，将但层次的主题模型扩展为多粒度的主题。 Xiaojun Quan等人综合考虑了语句片段和文本的主题分布，用来提高聚类效果。J Xu等人将短文本的主题特征与文本的tags标签相结合，来提高分类效果。张晨逸等综合考虑了微博的联系人之间的关联关系和微博文本之间的关联关系。

与传统文本挖掘只是利用文档中词汇的重复度来衡量文档的相似度不同，主题模型可以挖掘文字的潜在含义，因此更符合人类语言的习惯。LDA主题模型被广泛应用于文本信息挖掘，但是它是基于文档级别的，LDA模型生成文档的过程为:先通过文档一主题分布生成一个主题，然后根据主题一词汇多项分布抽取一个词汇，如此反复操作以生成一个完整的文档。LDA在长文本领域取得了不错的成绩，但是对于微博、新闻标题等这些不同于一般文档的短文本，由于它包含的词汇较少，缺乏上下文信息，所以LDA在短文本中的效果并不太理想。

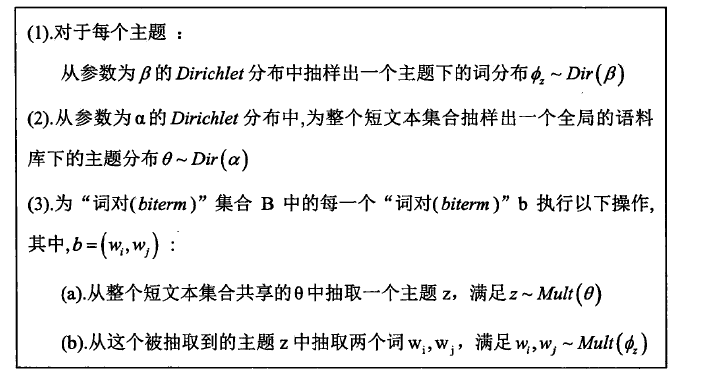
为了解决LDA不适用于短文本的情况，晏小辉教授等人在2013年5月的World Wide Web会议上提出了一种针对短文本来进行主题建模的BTM模型，该模型并不像LDA一样基于文档级别，而是针对整个语料库的信息进行建模的，这样就能减少某些文档因为字数少而对建模效果产生的影响。

BTM模型的建模过程是基于词对（biterm）来进行的，对于词对，BTM模型的作者给出的解释是文本片段中无序的两个不一样的词。通过词共现的方式，可以丰富短文本的信息，而且由于短文本本身词汇就比较少，因此其之间的词汇的联系通常也更加紧密，由于BTM的建模去除了文档的界限而基于整个语料库来建模，因此在推导文档的主题分布的时候不能直接得到，必须再经过一轮推导过程才能得到。

### BTM模型表示

BTM模型的目的是解决单一文本信息量不足导致的主题建模效果不理想的问题。它也借鉴了一元混合模型的思想，一元混合模型认为一篇文章只有一个主题，BTM在建模的过程中采用词对为单位进行，词对的构建采用滑动窗口的形式进行，通过在整个语料库中滑动窗口来构建词对集合，这种方式可以在一定程度上解决短文本数据稀疏的问题。

“共现词对”（biterm）可以理解为一个文本片段中的各个词汇不同的组合方式，由于短文本中词汇之间的联系通常比较紧密，比如“海洋”这个词会经常与“海水”或“鱼”一起出现;而“天气”则会经常与“晴天”或“暴雨”等词一起出现。基于BTM模型的文档生成过程可以描述如下：



其中，θ是整个语料库词对集的主题多项式概率分布，φ为一个主题下的词对的多项式概率分布，Z为通过语料库主题多项式概率分布下得到的一个主题，wi/wj为一个词对下的两个无序的词对。将以上过程转换为BTM的概率模型图来描述，如图13所示。

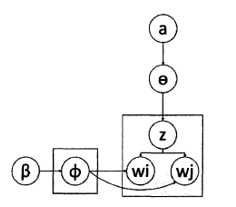


图13 BTM的概率模型图

图中，θ是BTM语料库中的主题概率分布，φ是主题一词对的分布，z为词对的主题标号，wi/ wj是词对中的两个词，IBI是BTM语料库中词对的数目。BTM模型并不是直接从语料库建模，而是先生成BTM语料库后，再通过BTM语料库进行建模。

## PAM模型

PAM模型是由Li等人在2006年提出，该模型是根据日本的一个游戏——“弹珠机”命名的。使用一个DAG结构去学习和表现主题相关性，每个叶子结点为此表中的一个词，非叶子内部结点代表一个主题，没个主题是基于他的孩子节点的狄利克雷分布。为了通过该模型产生文档，首先通过每个狄利克雷分布采样一个多项式，从根节点开始，根据多项式分布对其叶子结点进行采样，沿着DAG的卢金采样知道叶子结点产生次为止，这一过程就像“弹珠机”的金属球从顶部进入机器，跌入一组复杂的指针中，碰撞到的指针会改变金属球落下的路径，知道小球落入机器底部。

这种有向无环图结构是非常灵活的!可以是最基本的三层结构，也可以是任意嵌套的，节点间可以是全关联也可以是稀疏关联PAM模型中对主题的概念进 扩展，主题不仅可以是基于词空间的分布，还可以是基于其他主题的分布，内部节点的子节点可以是其他的主题，因此可得到主题之间的关联。如果一个内部节点的所有孩子都是叶子节点。那么可以将其看作一个传统的LDA主题，。LDA可以视为PAM结构的一种特例，其相应的DAG结构只包括顶部的根节点，中 间的主题集合以及底部的词集这3个层次。

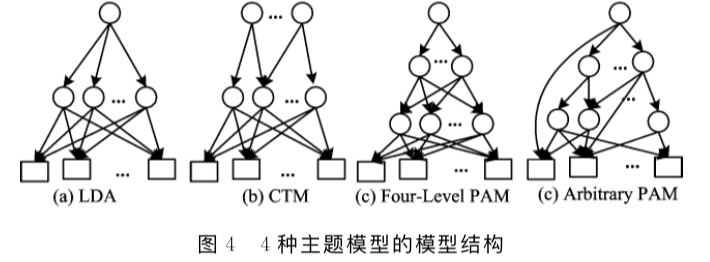


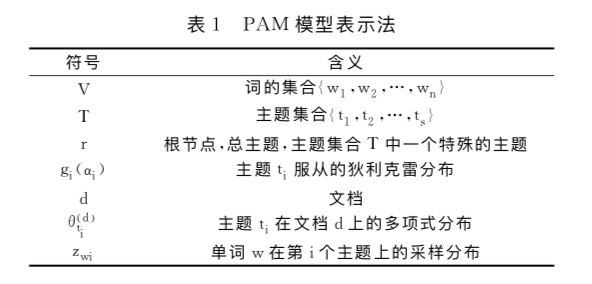
图14 4种主题模型的模型结构

### PAM模型表示

PAM 、LDA、CTM的模型结构如图14所示，每个矩形代表一个单词，每个圆形代表一个主题，主题上的箭头表示对其孩子的分布。（a）LDA：该模型中 每个文档对其所有主题服从多项式分布，通过主题产生单词；（b）CTM：低层次的每个主题都是所有单词的多项式分布，每对主题都由另外一个高层主题覆盖 （c）Four-Level PAM：4个层次包括1个总主题、1组超主题、1组子主题和词的集合；（d）PAM：任意的有向无环图表达主题相关性，每个非叶子节点代表 一个主题，服从狄利克雷分布。

PAM模型中的表示法如表2所示：

表2 PAM模型表示法



产生文档d的步骤如下：

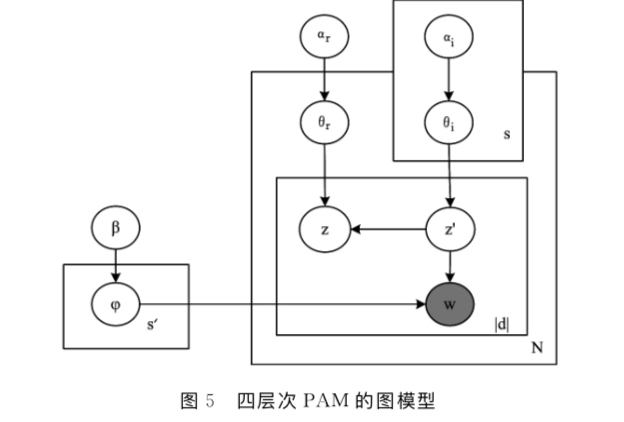
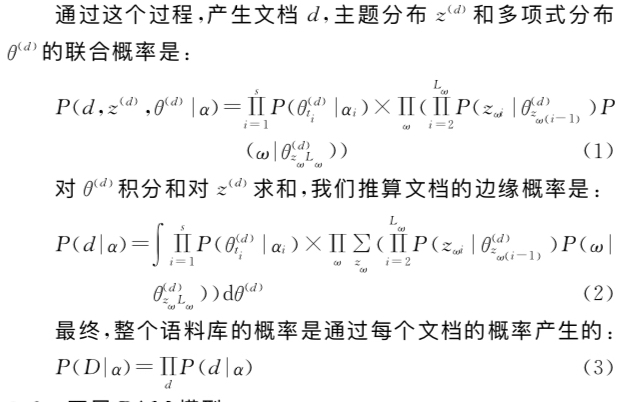
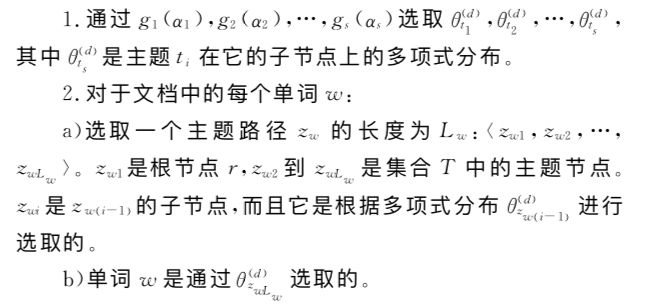


图15 四层次PAM的图模型

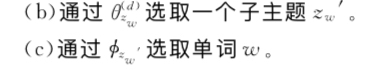


## 四层PAM模型表示

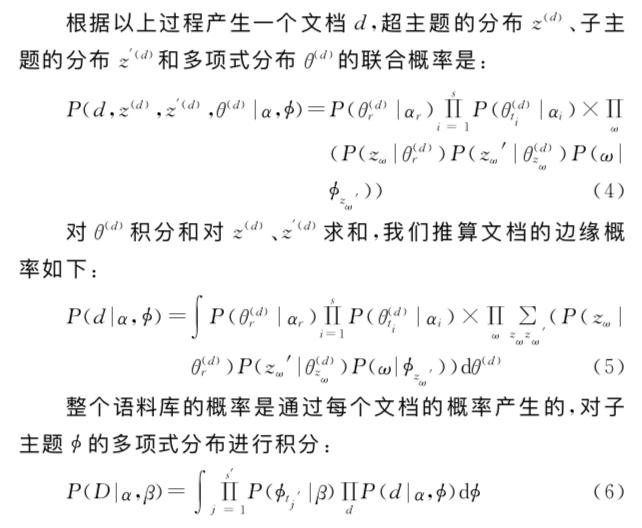
四层模型是PAM模型的一个特例。第一层是总主题r，第二层*T={t1 , t2 ,t3 ,··· ， ts}*有s个主题，第三层*T‘={t1‘ ,t2 ’, t3‘ ,···，ts’}* 有*s’*个主题，最底层是单词。 我们把第二层的主题称为超主题（super-topics），第三层的主题称为子主题（sub-top-ics）。 根部的总主题与所有超主题相关联。每个超主题与所有主题相关联。 主题与所有单词相关联。

四层PAM的DAG结构中对主题使用了两种不同的分布。文档d产生

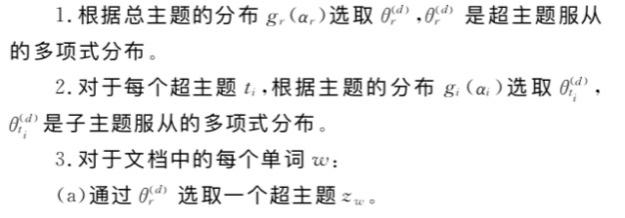
过程如下：



这种特例模型的构造和产生过程与LDA很相似，主要的不同是它另外有一个通过狄利克雷复合多项式构造的超主题层次，其是获取主题相关性的关键部分。四层PAM的图模型如图5所示。对于每个文档，PAM通过总主题和超主题的 狄利克雷分布选取多项式，对于每个单词w，PAM通过多项式选取一个超主题z和一个子主题z’，单词服从的多项式分布是与z’相关联的。对这种结构的另外一种解释是给出子主题，每一个超主题实质上就是一个独立的LDA。因此，它可以看作是一系列LDAs的组合。



## PAM改进模型



### HPAM模型表示

在LDA模型中，可以得到词之间的相互关系，但没有对不同主题之间的相关性进刻画。Blei等人在2004提出了HLDA（Hierarchical LDA），其采用树结构表示主题的组织构建。每篇文档由树结构中的一条主题路径产生，文档中的每个单词服从于该路径中某层次的一个主题。对主题使NCRP方法进行学习，得到树结构。NCRP先验需要两个参数：树结构的层数和y参数。y可以看作未采样 径中想象的文档的数量。

明显属于语料库不同子集的文档可以共享一个主题。比如，“专业运动”子层次的各种主题和“电脑游戏”子层次的各种主题可能会使用“游戏”、“选手”、“得分”这些相同的词项。“专业运动”和“电脑游戏”有相同词项的唯一解释是它们有共同的父节点，但这不是语料库最好的组织形式。

PAM和HLDA的一个区别是：前者的主题可能是词项的概率分布，也可能是其他（子）主题的概率分布，而HLDA中的每个主题都是词项概率分布。Mimno等人在PAM的基础上提出了HPAM模型（Hierarchical PAM）。 该模型可以看成是把HLDA和PAM结合起来，使得PAM模型中的非叶子节点也具有词项的概率分布，即DAG结构中每个内部节点也服从词库上的多项式分布。HPAM结 合HLDA的主题层次表达和PAM混合主题层次多叶子的能力，能够清楚地表达主题层次关系。 该模型可以解决得到单一层次的问题， 也不像HLDA那样要求得到一个准确的树结构，在留存文档似然性、自动发现的主题间互信息以及 成分类方面的实验结果都证明其性能得到改进。

HPAM的图模型如图16所示：

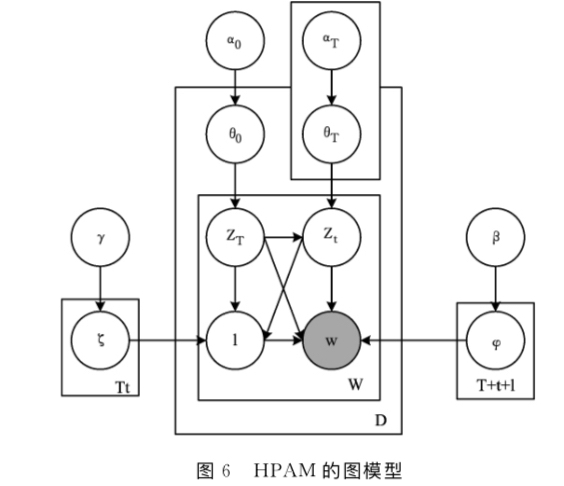
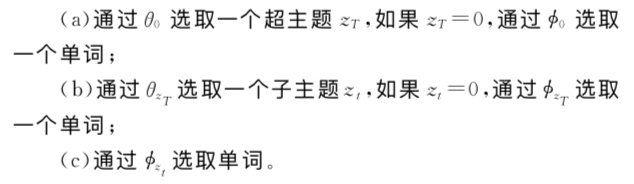
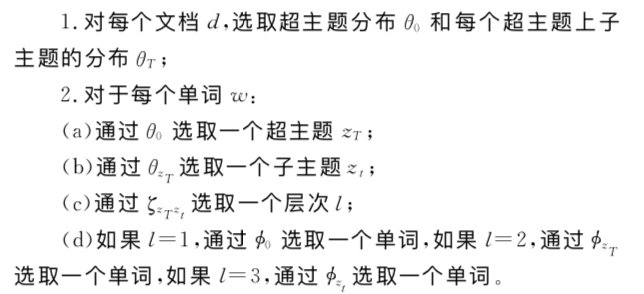
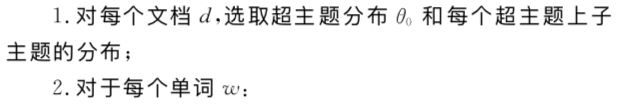


图16 HPAM的图模型



Mimmo 介绍了HPAM的两个变形：变形一，即HPAM模型1，DAG中每条路径都在路径层次上服从分布，这些分布被所有的文档共享。模型的产生过程如下：





变形二，即HPAM模型2，它与HPAM模型1相似，但是不包含主题层次上的分布。内部节点的狄利克雷分布有一个额外的维度，比如说，一个有10个超主题和20个字主题的模型，根节点就有一个11维的分布，每个超主题有一个21维的分布。这个额外存在的维度相当于内部节点直接关联一个词，并不是通过DAG结构的底层关联这个词。模型产生过程如图17。

HPAM不像 HLDA学习主题的树结构，它通过内部节点分布的狄克多项式参数来表示主题的层次结构，训练这些参数是HPAM一个很重要的部分。 使 用 Wallach（2006）中描述的Gibbs EM算法训练根节点在超主题层上的分布以及超主题在子主题层上的分布。允许模型进行多次迭代，然后对每篇文档的每个超主题的词的数量进行周期性的采样。使用 Minka（2000）中提到的固定点迭代 法来估计参数。

### NPB PAM模型表示

PAM和之前的LDA相比有更强的表现，四层结构只是PAM的一个简化特例，它的主题结构是任意嵌套复杂的，复杂结构有更强的表现力。然而对于复杂结构，选择最优的主题数以及确定合适的主题结构就会变得很困难，基于这个问题。Li等人在2007年提出一个NPB PAM模型（Nonparametric Bayes PAM），针对PAM给定一个基于HDP的非参贝叶斯先验，从非结构数据中自动发现主题相关性，并确定不同层次的主题数目，非参方法对于复杂模型更有吸引力。

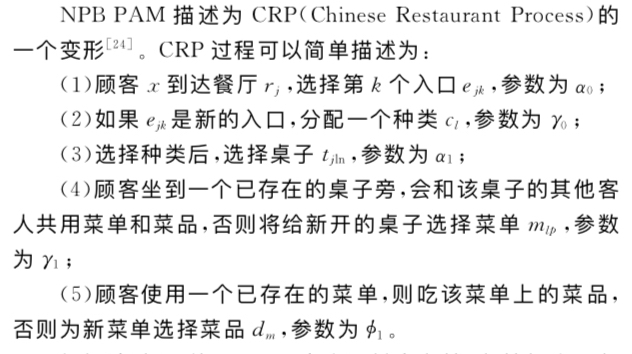


图17 NPB PAM过程

假设PAM的主题被组织为多层次结构，每层都使用一个HDP来获取不确定的主题数目。 一个标准的HDP混合模型中数据有一个预定义的嵌套结构，根据主题分布建立基于动态数据组的HDPs。 NPB PAM可以视为固定结果PAM的 一个扩展，每层主题的数目设置为无穷大。为了产生一篇文档，首先从相应的 HDPs中采样出主题的多项式分布，然后根据这些多项式对主题路径进行反复采样，得到文档中的每个单词。

根据该过程，使用PAM产生语料库文档，文档相当于餐馆，单词相当于顾客；四层PAM中，无限的超主题相当于种类，无限的子主题相当于菜品，超、子主题被所有文档共享。产生一个文档中的一个单词， 首先去采样一个超主题，超主题的采样过程包含两层CRPs，分别得到入口和种类。它们作为超主题分布的先验，是一个层次狄利克雷过程（HDP）。然后对已知超主题用同样的方式采样一个子主题。同样地，子主题的分布先验也是一个HDP，包含三层CRPs，其分别采样餐桌、菜单和菜品，最后根据单词的多项式分布从子主题中采样得到单词。

### 基于PAM模型的发展趋势

从上文的模型介绍可以看出，最初对于PAM概率主题模型的研究工作集中在模型的改进和优化方面，包括对参数的扩展和结构的调整等，仍有较多可以入手的研究点。除此之外，主题模型已经不局限于理论研究的阶段，近几年来，PAM主题模型开始逐渐应用于信息处各领域，包括文本分类、信息检索、摘要生成和标签处理等。PAM未来的发展趋势也应主要包括两个方面：主题模型性能的优化以及主题模型在文本处理的进一步应用。

首先，在主题模型性能优化方面，需要更高效的训练算法。在PLSI、LDA、CTM以及PAM这些主题模型中，都是将词项空间变换到主题空间，区别在于主题模型表示上的差别，或者说是在最优化时使的目标函数不同。由于通常无法求得精确解，因此参数估计问题至关重要，有多种算法可用于估算PAM主题模型中的参数，常用的方法有EM算法进行变分推理和Gibbs抽样。Mimno等人提出 种基于任意图模型先验的吉布斯采样算法，它能够更好地处理大量文本集合训练 推理过程中的复杂关联。 Nallapati等人提出并行的变分EM算法来加速训练过程，该算法可应用于PAM主题模型的参数求解。

另外一个较为明显的趋势PAM主题模型在信息处理领域的应用越来越广泛。PAM主题模型本质上是一种对文本的概率建模的方法，因此可以应用在信息处 理领域的各个方面。Sethi等人使用PAM模型来进行多文档摘要处理，首先使 用主题模型识别多文档集的主题，再通过选择每个主题里最重要的句子生成摘要。先前LDA主题模型用于发现语料库表现的主题时，产生的文摘句子之间没有关联，PAM则用超主题和子主题层次获取主题之间的关联，从而得到更好的文摘。实验结果证明，和其他高性能多标签学习方法相比，该模型在有效性和计算效率方面有很大的优势。Li等人应用PAM的一个扩展版本去进行目标识别，基于PAM法在一个层次结构里对潜在主题的相关性关系进行建模。在Caltech4和Caltech101数据集上高竞争性的识别结果说明，该方法和大多数现存的目标识别 法相比更具表现力和识别力。

# 关键问题及难点

前面的章节对研究框架中主要模块的关键技术作了详细的分析，想要实现各个模块时，必须注意解决关键问题即模型质量问题：

1. 模型质量较差，话题出来的无效词较多且较难清洗干净。
2. 话题之间，区别不够显著，效果不佳。
3. 话题内，词和词的关联性很低。
4. 反映不出场景，笔者最开始希望得到的是一个话题，里面有场景词+用户态度、情绪、事件词，构成一个比较完善的系统，但是比较那以实现。
5. 话题命名是个难点，基本词语如果效果差了，话题画像也很难了。

为解决上述问题，我们小组搜集了许多解决的方法被使用并取得了一定的成效。总结如下：

1. 如果要训练一个主题模型用于预测，数据量要足够大。
2. 理论上讲，词汇长度越长，表达的主题越明确，这需要一个优秀的词库
3. 如果想要主题划分的更细或突出专业主题，需要专业的词典。
4. LDA的参数alpha对计算效率和模型结果影响非常大，选择合适的alpha可以提高效率和模型可靠性。
5. 主题数的确定没有特别突出的方法，更多需要经验。
6. 根据时间轴探测热点话题和话题趋势，主题模型是一个不错的选择。

# 展望

基于统计方法的机器翻译技术作为最热门的话题从1999年开始，进入了一个飞跃上升的阶段。出现这个变化的原因，就是在1999年出现了一个机器翻译的热潮，其最主要的特征是基于统计的方法在这一领域开始占据主导地位，机器翻译的质量出现了一个跨越式的提高。这股热潮持续至今，仍未现衰减之势。同时，基于统计的句法分析的强度也随着这股热潮不断提升。而情感分析在2000年前一直都是比较冷门的话题，但现今研究者对它的青睐不断增加。

统计技术的兴起对这些热门话题的强度上升起了很大的推动作用；另一方面，冷门话题的下降趋势也有不同的表现形式，有的是缓慢下降，有的是急速下降。

通过对面向LDA主题模型的文本分类研究进展与趋势的分析,可以发现,应用LDA主题模型于文本分类,有利于处理大规模文本,不仅合理地降低了特征词矩阵的维度,还能保持原数据集的全面性,不影响分类器性能,解决了传统文本分类中相似性度量和主题单一性问题。然而,尽管LDA主题模型得到进一步改进和完善,但还尚有一定缺陷和不足。LDA是非监督学习模型,不能直接用于文本分类,因此必须嵌入到合适的分类算法中。传统的LDA主题模型存在分类过程中将文档强制在单个类别上分配隐含主题的缺陷；并且由于实际情况中大规模的数据,可能会出现主题范围过大,不能对主题单词的潜在语义进行准确定位,限制了模型的鲁棒性和有效性；没有考虑词序问题,是典型的词袋模型等。

另外在分类器设计方面,朴素的贝叶斯分类法假定词与词之间是独立的,这在实际情况中很难保证,因此当假设条件不满足时,会严重影响分类的准确率和性能。基于支持向量机的分类器在处理具体分类问题时无法选取正确有效的核函数,另外,训练时间与数据集规模大小有关,训练时计算量通常比较大,这也会影响分类器的效率。k-最近邻法计算量大,时间复杂度较高,而且训练样本质量对分类器性能有着很大影响。神经网络法分类效果比kNN和SVM较差,而且时间复杂度比较大。针对这些问题,多种方法的融合、改进可以改善分类效果。特征选择和特征重构是降维操作的关键技术,二者融合有助于改善降维效果。例如把互信息和聚类融合,通过互信息最大化从原始特征空间中选择次优特征子集,借助特征空间的聚类来剔除冗余特征,从而实现特征空间的再次降维。把多种分类算法相融合,利用它们的优点,剔除缺点,从而可以改善分类性能。例如LDA分别与卡方统计、互信息和信息增益进行结合,利用改进后的特征提取方法提取特征词,实验表明结合后的方法比原来的方法分类效果好;另外随着特征词个数的增多,每一种方法的分类性能也有提高。

我国每年突发事件频发,交通事故、火灾等不计其数。如何对这些突发事件进行监测,并实施有效的救援,这是一个难题。现在网络技术发达,其实可以把网络信息进行详细分类,针对网络中出现的信息进行分析、挖掘,过滤出敏感词汇,如地震,失火,车辆相撞,追尾等,分析出可能发生的隐患事件和对已经发生的事件进行追踪,从而实施有效的预防和救援措施,保障人们的生命财产安全。应用于公安系统可以预防犯罪发生和快速破案。也可应用于军队,对我国领海、领土、领空进行监测,一旦发现事故发生或是外部入侵,可以及时采取有效措施,保障我国国民和领域安全。

我国人口众多,排队看病是一个难题,病人流量太大,医院环境嘈杂,可能会影响病人描述病情和医生进行更有效诊断。因此,将文本分类和信息抽取应用于医疗健康系统,将用户输入的咨询信息进行分类和整理,提取出用户的病症信息,然后根据处理后的病症内容进行分类,诊断出可能的病症名称,然后推送给不同的科室医生进行在线回复,还可以根据分析出的病情推荐合理的看病科室。将文本分类应用于医疗健康后,病人可以更方便的对自己的病情进行咨询和就诊,医生也可以根据这些信息对病人病情进行更好、更快捷的诊断。这不仅对病人、医生,还是医院都提供了有利的条件,因此可以在这方面进行更深一步的研究。

就研究对象而言，话题检测和热点检测的目标都是为人们分析信息、了解社会问题提供服务。相关的领域，例如文本分类、事件识别，以及情感计算、意见抽取、信息检索等，都是针对相似问题进行的，他们的研究具有协同性。就媒体种类而言，本文一直在围绕新闻报道展开，因为新闻报道是话题检测与跟踪的经典研究领域，至今仍然有研究和实用意义。今后，研究可以由新闻扩展到博客、微博、论坛等其他媒体领域。

就舆情分析系统的工程而言，有两方面值得深入研究。一方面，系统分析结果的好坏受各个相关组件共同影响。例如更准确的分词、词性标注等预处理可以使得聚类、热点分析的结果更准确。爬虫的好坏决定了语料的丰富的程度。任何一个相关的技术的进步都会影响整个系统的实用性。另一个方面，系统的定位包括产品和实验平台两个方面，他们在易用性和灵活性方面有着互相矛盾的不同需求。

虽然话题检测与跟踪研究已经开展很多年，暗示由于互联网数据来源的多样性与特征抽取的不确定性带来的困难，目前话题检测主要集中在新闻数据上，神会网络上的话题检测到的是相对较少。随着社会网络的兴起，特别是论坛、微博等的广泛应用，面向突发事件等特定需求和面向社会网络数据的新型应用的话题检测具有越来越总要的意义。随着面向社会网络数据的特征选择方法以及用户行为与文本内容惯量挖掘方法的综合婴童，话题检测与跟踪技术的研究与应用会得到进一步的发展。 总之，话题检测、热点检测领域仍然有大量的研究需要完成。

参考文献

1. Allan J. Topic detection and tracking：Event-based information organization[M]. [S.l.]：Kluwer Academic Publishers，2002.
2. Wayne C L. Topic detection and tracking（TDT）：Over-view& perspective[C]//Proceedings of the Broadcast News Transcription and Understanding Workshop， Lansdowne， Virginia，1998.
3. Li Baoli，Yu Shiwen. Research on topic detection and tracking[J]. Journal of Computer Engineering and Applications，2003，39（17）：7-10.
4. Hong Yu，Zhang Yu，Liu Ting，et al. Topic detection and tracking review[J]. Journal of Chinese Information Processing，2007，21（6）：71-87.
5. Allan J，Carbonell J，Doddington G，et al. Topic detection and tracking pilot study final report[C]//Proceedings of the DARPA Broadcasting News Transcript and Understanding Workshop，1998：194-218.
6. Bun K K，Ishizuka M. Emerging topic tracking system in WWW[J]. Journal of Knowledge-Based System，2006，19（3）：164-171.
7. Sekiguchi Y，Kawashima H，Okuda H，et al. Topic detection from blog documents using users’interests[C]//MDM，2006.
8. Cselle G，Albrecht K，Wattenhofer R. BuzzTrack：Topic detection and tracking in email[C]//Intelligent User Interfaces 354 2007，Honolulu，2007：190-197.
9. Mori M，Miura T，Shioya I. Topic detection and tracking for news Web pages[C]//Web Intelligence 2006，Hong Kong，2006：338-342.
10. Ye Huimin，Cheng Wei，Di Guanzhong. Design and implementation of on-line hot topic discovery mode[J]. Wuhan University Journal of Natural Sciences，2006，11（1）：21-26.
11. Pons-Porrata A，Berlanga Llavori R，Ruiz-Shulcloper J. Topic discovery based on text mining techniques[J]. Journal of Information Processing and Management，2007，43（3）：752-768.
12. Yu Manquan，Luo Weihua，Xu Hongbo，et al. Research on hierarchical topic detection in topic detection and tracking[J].Jouranl of Computer Research and Development，2006，43（3）：489-495.
13. Qiu Likun，Long Zhiyi. Hierarchical topic detection and tracking and implementation of system [J]. Journal of Guangxi Normal University：Natural Science Edition，2007，25（2）：157-160.
14. Zhao Hua，Zhao Tiejun，Yu Hao，et al. English topic tracking research based on query vector[J]. Journal of Computer Research and Development，2007，44（8）：1412-1417.
15. Zhang Kuo，LI Juanzi，WU Gang，et al. A new event detection model based on term reweighting[J]. Journal of Software，2008，19（4）：817-828.
16. Zhang Kuo，Li Juanzi，Wu Gang. New event detection based on indexing-tree and named entity[C]//SIGIR，2007：215-222.
17. Zhang Xiaoyan，Wang Ting，Chen Huowang. Story link detection based on dynamic information extending[C]//IJCNLP，Hyderabad，2008：40-47.
18. Kumaran G，Allan J. Using names and topics for new event detection [C]//Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing，2005：121-128.
19. Makkonen J，Ahonen-Myka H，Salmenkivi M. Topic detection and tracking with spatio-temporal evidence[C]//ECIR，Pisa，2003：251-265.
20. Makkonen J，Ahonen-Myka H，Salmenkivi M. Simple semantics in topic detection and tracking[J]. Information Retrieval，2004，7（3/4）：347-368.
21. Hoogma N. The modules and methods of topic detection and tracking[C]//Proceedings of the 2nd Twente Student Conference on IT，Enschede，2005.
22. Shah C，Croft W B，Jensen D. TDT-representing documents with named entities for story link detection [C]//Proceedings of the 15th ACM International Conference on Information and Knowledge Management，Arlington，Virginia，USA，2006：868-869.
23. Lee Changki，Lee G G，Jang M G. Dependency structure language model for topic detection and tracking[J]. Journal of Information Processing and Management ，2007 ，43（5）：1249-1259.
24. Carthy J. Lexical chains versus keywords for topic tracking[C]//Proceedings of the 5th International Conference on Intelligent Text Processing and Computational Linguistics，Seoul，Korea，2004：507-510.
25. Ogilvie P. Extracting and using relationships found in text for topic tracking，CIIR Technical Report IR-209[R]. Undergraduate Honors Thesis，2000.
26. Larkey L S. Language-specific model in multilingual topic tracking[C]//SIGIR，2004：402-409.
27. Elsayed T，Oard，D W，Doermann D，et al. TDT-2004：Adaptive topic tracking at Maryland[C]//TDT-2004 Workshop，Gaithersburg，2004.
28. Spitters M，Kraaij W. A language modeling approach to tracking news events[C]//TDT-2000 Workshop，2000.
29. Wayne C. Multilingual topic detection and tracking：Successful research enabled by corpora and evaluation[C]//LREC Workshop，2000：1487-1494.
30. Rajaraman K，Tan A H. Topic detection，tracking，and trend analysis using self-organizing neural networks[C]//PAKDD，2001：102-107.
31. Dai Shangxue. A study on constructing a news event detection and tracking system with multi-document summarization[D].Taiwan：Institute of Information Management，National Yunlin University of Science and Technology，2003.
32. Yan X, Guo J, Lan Y, et al. A biterm topic model for short texts[C]//Proceedings of the 22nd international conference on World Wide.International Wide Web Conferences Steering Committee, 2013: 1445-1456
33. Blei D M, Ng A Y, Jordan M I. dirichlet allocation[J].the Journal of machine Learning research, 2003, 2003(3): 993-1022.
34. Cheng X, Yan X Lan Y, et al. BTM:Modeling over Short Knowledge and Data Engineering, lEEE Transactions on, 2014, 26(12):2928-2941.
35. 徐戈，王厚峰.自然语言处理中主题模型的发展闭.计算机学报，2011 34(8):1423-1436. DOI:10.3724/SP.J.1016.2011.01423.
36. 张晓艳，王挺.话题发现与追踪技术研究[J]. Journal of Frontiers of Computer Science and Technology 1673-9418/2009/03（04）-0347-11 DOI：10.3778/j.issn.1673-9418.2009.04.002
37. 余淼淼，王俊丽.PAM 概率主题模型研究综述[J].计算机科学2013,40（5）