# 基于领域知识的跨媒体微博突发事件抽取

## 摘要

随着互联网技术的蓬勃发展和普及，网络上的信息呈指数级增长，信息采集、传播的速度、规模达到空前水平，人们对信息的需求与日俱增。中国互联网信息中心（CNNIC）2014年1月公布的“第33次中国互联网络发展状况统计报告”显示，截至2013年12月，我国网民规模达6.18亿，全年共计新增网民5358万人。互联网普及率为45.8%，较2012年底提升3.7个百分点。微信数据作为一种新的社交媒体有着更加迅猛的发展速度，而且据最新的统计显示截至到2014年微信每日活跃账户已经超过一亿了。

目前基于微信公共账号的订阅模式的信息分发方式已经成为了普遍网名获取信息的重要方式，我们会发现不同的人所感兴趣的知识是不一样的，有很多专题文章会被聚集在一起，这些所谓的专题文章在很大程度上是因为他们拥有某些共同的关键词，假设有人对所有与习近平相关的文章很感兴趣，有人会对天津爆炸的所有事件都感兴趣，但是包含某些关键词的文章之间是会存在很大的区别的，如果要细致的分析这些专题的信息内容，我们必须对这些文章进行更加细致的分类，将真正同一个主题的文章集合在一起，例如有关习近平的文章，有一部分是有关股市的，有一部分是关于反腐的，有一部分是有关南海问题的，我们如果能够将这些差异的信息分割开来，那么就能够是的阅读者能够更加清晰地了解这部分知识的机构和布局。

我们称上面的在某个专题内部找出差异信息主体的做法为子话题分析，这部分工作的主要内容就是某些拥有大的共同背景的文章中找出属于文章之间的差异性，每一个不同的部分形成一个子话题，找出每个子话题的代表关键词。真是由于这些文章之间有一个公共的背景，所以子话题分析与话题分析的工作存在实质上的区别，这里举一个例子进行说明，例如使用LDA主题模型进行子话题分析，在这里由于所有的文章都存在一个相似的背景所有使用LDA主题模型不能够彻底的将信息进行细致的分割，有些不同的子话题的文章有可能会因为有共同的背景知识，而使得他们的差异性被淹没，从而没有区分出这些不同的主题，而是将它们纳入同一个主题下。

针对这一挑战，本文提出了基于背景化的LDA子话题模型，这种方法是基于LDA主题模型的一种改进，先抽取所有文档的公共背景，在迭代的过程中加入每个词的产生方式，它有可能来自公共的背景知识，也可能来自子话题主题模型。同时本文也试验了基于背景的PLSA主题模型，它与LDA子话题模型的方法类似，具体细节可以参考论文部分。本文还提出了背景化的KMEANS模型，使用这种文档聚类方法也能达到子话题分析的效果。最后本文正对微信数据本身的特点提出了一种基于微信公众号的先验知识的快速子话题分析方法。

本文在结合微信公众号发表的文章特点的基础上，通过去背景化的策略将各种主题分析和文档聚类方法引入子话题分析模型中，同时利用微信公共账号的先验知识实现了子话题的快速分类：

1. 本文提出了背景化的LDA子话题分析模型，这种模型非常适用于专题内部的子话题分析，与原始的LDA相比使用背景化的LDA子话题模型能够更加准确的分析出专题文章内部的不同子话题之间的信息差异。
2. 本文也设计并且实现了去背景化的KMEANS模型，使用这种去背景化的KMEANS模型我们也可以更加准确地找出不同子话题之间信息差异。
3. 本文最后提出了一种快速子话题分类策略，该方法充分的利用了微信公共账号发表文章的主题专一性，结合微信公共账号之前发表的文章的主题来初步的断定一篇文章的主题，但是这种做法存在一定的猜测性，存在一定的误判风险。

关键词：突发公共事件，事件表示模型，事件抽取，跨媒体，图像分类

## Abstract

Emergent public events threaten the public's life and property safety, and have a significant impact on people's production and life. The prevention and rapid treatment of public emergencies has become the key areas of national science and technology development. In the Internet environment, the network space and the real space to form a mapping relationship, real life events will occur through the network of different channels in different media from different sides to reflect and express. By analyzing the different media data in the network space, the event in real life is perceived in a timely manner, which has become an important technical means to deal with unexpected events.

This paper focuses on the analysis of the micro blog data in real time. Compared to the news data, large amount of micro blog data, timeliness strong, major public emergencies are often on the micro-blog platform first, through the mechanism of the reproduced spread rapidly, have great impact on society, so the micro-blogs for emergency extraction is of great significance. However, the amount of data is large, the text is short, the expression ability is weak, and more popular words, but also to the Weibo event extraction brings unique challenges.

To solve this challenge, this paper proposes the technology of micro blog cross media burst event extraction based on domain knowledge. Many Weibo users preferences figure to express the views of a certain event, image data than the text expressed more abundant semantic, media text and image data fusion cross media event detection, on the one hand intends to fill as short text to express the weakness of ability is insufficient, on the other hand to incident detection of an open a new way of thinking.

Based on the characteristics of public security, this paper analyzes the text and image data of the micro blog, and then realizes the identification of the event:

(1) in this paper, we propose an event extraction method which is based on the fusion of micro blog text and image. To extract image color and gradient features to calculate the similarity, and then according to the theme of the picture to supplement the short text, to a certain extent, the expansion of the short text semantic, improve the effect of event extraction.

(2) this paper focuses on the short text event extraction technology based on domain knowledge. Aiming at the problem of micro blog topic diversity in the open field, the method of filtering and re checking (classification) is adopted in the field of micro blog. The domain model of the candidate text filtered out in the field of the field is re classified, which improves the speed and precision of event extraction.

(3) in view of the candidate text filtered out in the field, this paper proposes a domain event representation model. The domain event representation model effectively describes the temporal and spatial characteristics of the emergency, the subject category and the trigger relationship between the disaster factors and the main body, and further improve the accuracy of the classification.

(4) aiming at the problem that the micro blog is not allowed to use new words and popular words, the paper puts forward a new method based on statistics. Such as Ebola, MAS, representing, Yong Wen line "neologism is key feature to describe the event, the event extraction is of great significance. It can improve the accuracy of word segmentation and guarantee the accuracy of classification. On the other hand, it can expand the domain knowledge, join the domain dictionary, and improve the extendibility of event extraction.

Keywords: sudden public events, event representation model, event extraction, cross media, image classification

## 目录

摘要

Abstract

图目录

表目录

1. 绪论

1.1 课题背景

1.2 研究目的和内容

1.3 本文组织结构

1.4 本章小结

1. 国内外研究现状

2.1 文档建模

2.2 事件表示模型

2.3 文本事件检测

2.4 图像事件检测

2.5 本章小结

1. 去背景化的LDA子话题模型

3.1 去背景化的思想

3.2 LDA子话题模型中的去背景化思想

3.3 LDA子话题模型去背景化的的实现

3.4 公式推导与实验结果分析

3.5 本章小结

1. 去背景化的KMEANS子话题模型

4.1 KMEANS子话题模型的去背景化思想

4.2 实验方法与实验结果

4.5 本章小结

1. 基于微信公共号先验知识的快速子话题分类模型

5.1 微信公共号发表文章的主题一致性

5.2 微型公共账号历史发表文章主题研究

5.3 建立主题模型分类体系框架

5.4 快速的子话题分类策略

5.5 实验结果

5.6 本章小结

1. 总结与展望

6.1 全文总结

6.2 未来研究方向

参考文献

## 绪论

### 1.1 课题背景

随着互联网技术的蓬勃发展和普及，网络上的信息呈指数级增长，信息采集、传播的速度、规模达到空前水平，人们对信息的需求与日俱增。中国互联网信息中心（CNNIC）2014年1月公布的“第33次中国互联网络发展状况统计报告”显示，截至2013年12月，我国网民规模达6.18亿，全年共计新增网民5358万人。互联网普及率为45.8%，较2012年底提升3.7个百分点。微信数据作为一种新的社交媒体有着更加迅猛的发展速度，而且据最新的统计显示截至到2014年微信每日活跃账户已经超过一亿了。

目前基于微信公共账号的订阅模式的信息分发方式已经成为了普遍网名获取信息的重要方式，我们会发现不同的人所感兴趣的知识是不一样的，有很多专题文章会被聚集在一起，这些所谓的专题文章在很大程度上是因为他们拥有某些共同的关键词，假设有人对所有与习近平相关的文章很感兴趣，有人会对天津爆炸的所有事件都感兴趣，但是包含某些关键词的文章之间是会存在很大的区别的，如果要细致的分析这些专题的信息内容，我们必须对这些文章进行更加细致的分类，将真正同一个主题的文章集合在一起，例如有关习近平的文章，有一部分是有关股市的，有一部分是关于反腐的，有一部分是有关南海问题的，我们如果能够将这些差异的信息分割开来，那么就能够使得阅读者能够更加清晰地了解这部分知识的机构和布局。

我们称上面的在某个专题内部找出差异信息主体的做法为子话题分析，这部分工作的主要内容就是在某些拥有大的共同背景的文章中找出属于文章之间的差异性，每一个不同的部分形成一个子话题，找出每个子话题的代表关键词。真是由于这些文章之间有一个公共的背景，所以子话题分析与话题分析的工作存在实质上的区别，这里举一个例子进行说明，例如使用LDA主题模型进行子话题分析，在这里由于所有的文章都存在一个相似的背景所有使用LDA主题模型不能够彻底的将信息进行细致的分割，有些不同的子话题的文章有可能会因为有共同的背景知识，而使得他们的差异性被淹没，从而没有区分出这些不同的主题，而是将它们纳入同一个主题下；使用KMEANS算法也同样会存在同样的问题，有时候两篇不同主题的文章有可能由于相同的主题背景词语存在比较多公共的相同的背景知识，有可能会被归纳到同一个主题下，通过调节主题的个数也许能够解决相关的问题，但是却需要经过很多次尝试，这样子的做法显然不能被很多人接受。

针对上面存在的问题，我们急需寻找一种能够在专题文章内部寻找到存在差异性的子话题的方法，并且这种方法能够适用于专题文章数量很大的时候，也就是该方法能够适用于海量数据，目前学术界还没有人针对该问题做出深入的研究，相关的文献也很少，所以本文针对上面的问题提出了自己的解决方案。

微信公共账号的发表数量非常的大，如果直接进行当线程的迭代，算法的性能并不好，我们做过相关的统计，当文档数据达到10万篇以上的时候，算法的运行时间将会变得非常的长，这在实际的系统中是无法接受的，所以，本文将针对去背景化的LDA算法进行并行化的处理，可以是多核的并行化，也可以是多机的并行化算法。

此外，本文还针对微信公共账号的属性进行了相关的统计和研究，研究发现很大一部分（可以转换为比例）公共账号的发表文章的专业性非常的强。图1-1是一个账号名称为军事的公共账号的相关介绍。

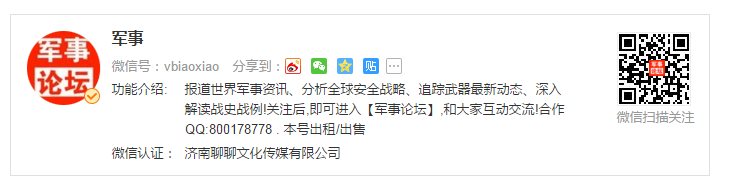


图1-1 账号军事的相关介绍

从图中我们可以很清楚的看到，这是一个与军事报道关联性非常强的微信公共账号，该账号的所发表的文章应该很大部分上都是与军事相关的。通过观察军事账号的相关内容，如图1-2所示，我们可以发现确实如此，“军事”账号发表的文章绝大部分都是与军事主题极度相关。



图1-2 “军事”账号的相关发表内容

我们在举一个例子，我们观察账号“全球体育赛事”的简要描述，以及它的发表文章的内容。

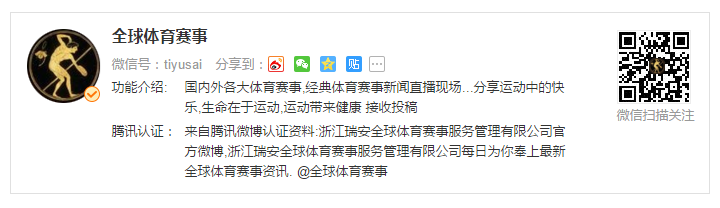




图1-3 账号“全球体育赛事”及其相关的文章代表

这部分信息绝对不是没有意义的，本论文将从另外一个角度来实现主题的挖掘，在挖掘的过程中充分的利用账号本身的属性信息，例如上面的两个账号，我们可以看到它们发表的内容具有极强的账号黏性，或者说是极强的专业性。

本论将充分学些账号的历史黏性以及充分的利用这部分信息来提高主题分析的准确性和加快主题挖掘的速度。

### 1.2研究目的和内容

本文的主要研究内容是如何针对大量的拥有共同的背景知识的文章进行子话题挖掘，原本的一些话题挖掘分析方法在这种专题文章数据下运行会存在话题内部信息分布不集中的问题，主要是因为语聊中的文章中存在公共的背景知识，他们之间有很多相同的词汇量，普通的方法可能会因为这些公共的背景知识而无法区找出它们之间的差异。而本文的工作目的就是针对上面存在的问题，设计并且实现一些能够在专题文章内部寻找到存在差异性的子话题的方法，并且由于微信数据量非常的大，我们希望这些方法能够适用于海量的数据。同时本文还将从微信公共账号数据本身的特性出发，考虑如何学习微信公共账号发表文章的转有特性，并且充分利用这部分数据来提高主题学习的性能和速度。

### 1.3 本文组织结构

本文分为六个大章，第一章为绪论主要说明本文的选题背景和研究目的和内容。第二章是讲解相关领域的国内外研究现状，其中主要讲解话题挖掘算法的研究现状以及分布式话题挖掘算法的研究情况。同时也有一些关于子话题研究算法的现状描述。第三章讲解了去背景化的LDA子话题挖掘算法的设计思路，并且讲解了去背景化的思想及其原始考虑，最后实现了该算法并且给出了实验结果分析。第四章讲解了去背景化的KMEANS算法的设计与实现，并且给出了实验结果分析。第五章描述了基于微信公共号的特点提出的一种快速的经验性的子话题挖掘算法，这种算法很有特色，并且也是基于微信公共号的先验知识提出的一种策略。第六章给出了全文的总结和未来的展望。

## 国内外研究现状

### 2.1 话题挖掘算法研究

话题发现的相关研究主要是起步于上世纪 90 年代中期，研究主要集中在关联检测、话题检测、话题跟踪、跨语言TDT 等研究领域。自从1996年以来,该领域进行了多次大规模评测，为信息识别、采集和组织等相关技术提供了新的测试平台。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。

在现有的研究中，话题发现的常用算法包括聚类、矩阵分解、话题模型、神经网络等。其中，应用的最为广泛的方法是基于文本聚类的方法，该方法主要包括特征选取，文本表示模型和文本聚类三个步骤。也有人使用SVD分解的方法来实现隐语意挖掘，从而形成话题挖掘的算法，Huffman在2003年提出了PLSA的主题模型，从此也开启了学者和数据挖掘工程师使用主题模型来对文本的主题挖掘的模式，后面又提出了LDA主题模型，由于它的理论基础和实践意义都得到了很好的证明所以现在LDA主题模型被广泛的应用于文档的主题模型挖掘。

### 2.2 基于微博和新闻的话题挖掘算法

之前有很多的学者针对微博和微信数据提出了很多的话题话题挖掘算法，多数的学者使用的是聚类算法，也有学者提出使用隐主题的思想进行文本的软聚类也得到了不错的效果。

聚类分析是一种无监督学习方法，是一个将对象集合划分成多个子集的过程。在话题检测中可以利用聚类分析，将主题相关的文档归于一类，形成话题。常用的聚类算法有层次聚类算法，基于密度的DBSCAN算法等，基于划分的KMeans聚类算法等。在早期为了处理线上数据，有人提出了在线话题检测技术。一种之前比较流行的算法Single-Pass聚类算法就是大家使用比较多的在线话题发现算法。这种算法是一种思想很简单的方法，初始的时候可以对小文档集合进行一个简要的KMeans聚类，然后形成一个初始的聚类团，然后流进来的文章就会与初始的簇进行相似度计算，如果相似度符合我们初始设定的阈值，就直接将文章分配相关的簇中，并且重新更新该簇的信息，反之，如果文章没有超过初始设定的阈值，就会将该文章当做是一个新的聚簇。这个算法最初是由J. Allan等人提出的，后面出现了很多改进的版本。由于这个算法简单易懂而且能够达到一定的效果，并且非常适合大数据量的在线流式数据处理，所以得到了很多人的认可和接受。

层次聚类算法（HAC）可以分为两类：自顶向下的分裂的方式和自底向上凝聚的方式[KR1990]。层次聚类算法的聚类效果较好，而且可以设定阈值以获得不同粒度的结果，满足不同的需求。但是算法的复杂度相对来说较高，而且由于其处理过程不可逆性，导致该算法一般只用于回顾式话题发现，不适用于在线话题发现。[YPC1998]提出了组平均聚类（Group Average Clustering）算法以改进层次聚类算法，进行回顾式话题发现。该算法采用基于分治的Bucket策略，使算法的复杂度降低一个数量级，同时由于有效地运用文档时间聚集性的特征，一定程度提高了话题检测结果的正确率。层次聚类算法可以很好的控制话题的个数，并且结果是可以预测的，所以在很多时候层次聚类算法的想过都相对较优，这也是为什么现在还有人在使用层次聚类算法的原因。

基于密度的DBSCAN聚类算法利用类的密度的连通性可以快速发现任意形状的类。其基本思想是：对于一个类中的每个对象，在其给定半径的领域中包含的对象不能少于某一给定的最小数目。[Cha 2011]研究了利用DBSCAN算法对Chatter流行话题进行检测，但是实验显示，检测的话题效果并不是很理想。

KMeans算法是最早由[Mac1967]提出的一种基于划分的聚类算法，并对其收敛性等做了理论上的分析与证明。该算法是一种基于形心的划分技术，形心可以有多种定义，例如用分配给该簇对象的均值或中心点定义。对象与该簇代表之差用欧氏距离度量。其中簇可用簇内均方误差度量，它是中所有对象和形心之间误差的平方和，定义为，目标试图让生成的结果簇之间尽可能紧凑和独立。该算法实现简单，效率比较高。但是受K值和初始中心点选取以及迭代过程中心点选取等因素影响较大。

高斯混合模型（GMM）[McL1988]是一种聚类算法，每个component就是一个聚类中心。即在只有样本点，不知道样本分类（含有隐含变量）的情况下，计算出模型参数。GMM认为数据是从几个单高斯分布模型GSM中生成出来的。

多变量GSM的概率密度函数表示为：

(2.1)

GMM的概率密度函数为：

(2.2)

其中模型参数（，和）可以通过EM算法求得。

利用训练好的模型就可以判断一个点所属的分类，方法是：

Step1：随机选择K个component中的一个（被选中的概率是k）；

Step2：把样本代入刚选好的component，判断是否属于这个类别，如果不属于则回到Step1。

GMM每一步迭代的计算量比较大，大于KMeans。而且GMM的求解办法基于EM算法，因此有可能陷入局部极值。

上面这些方法都是基于聚类的算法，也有学者使用软聚类的算法进行话题的学习。

LDA 是一种非监督的学习方法，这种非监督的学习方法非常适用于分析未知数据集。但是LDA更适合于检测关键词级的话题，比如热词新闻等，而VSM模型更适合于表示文档级别的话题。

[孙 2010] 提出将动态主题模型加入LDA模型中，以改善LDA模型对文档输入顺序无关的假设性，同时考虑时间片对文本语料的影响，实验表明该方法在新闻语料中的话题检测一定程度上提升了效果。

Yiming Yang提出了采用分类计数将所有相关的报道分类道不同的组中的方法，他只选择每组中相似度最大的相关报道来描述一个话题，其结果在性能上得到了很大的提升。

使用文本之间的的余弦相似度来报道不同话题之间的相似度是最开始的方法，后面M. Spitters等人提出一个机遇语言模型的相似度函数。

上面的所有模型基本上都是基于词袋模型来完成的，基于词袋的统计模型存在一个公认的问题，那就是在进行文本相似度计算的过程中会只关注词袋，忽略了其他的一些特征。例如文档的发表时间这个特性就是会被比较容易忽略的特质，举一个例子，例如有关“恐怖袭击”这个话题，如果有一堆有关恐怖袭击的文档，它们有的是与9.11事件有关的报道，有些是有关叙利亚恐怖袭击的新闻，也有一些是有关云南昆明火车站的恐怖袭击时间，但是由于这些实践中都涉及了“恐怖”，“死亡”，“凄惨”，“重伤”等词汇，可能这些文档都会被聚集在一起，但是我们会发现这些事件其实根本就不在同一个时间段，如果将它们的时间特征加入到应用中，可以极大的改进聚类的效果。引入时间窗口可以很好的解决上面的问题。

也有学者提出了在另辟蹊径，提出了与传统的向量空间模型不一样的策略，大连理工的宋丹提出了按照语义将特征词划分为4个组（任务、时间、地点、内容）并形成4个向量空间，每个空间进行独立的权重计算和相似度计算，并提出了构造地理树的方法，计算地点的相似度。

还有的学者提出了基于多策略聚类的话题发现算法，他认为众多的文档可以提前进行预划分，也就是，提前将文档划分成多个没有交集的组，然后分别在组内进行聚类，在获得了组内的簇的信息之后，然后再对不同组之间的类簇进行聚类，这种思想可以参考层次聚类的做法，可以得到不错的结果。

也有的学者针对话题漂移的问题提出了一些新的解决方案，东北大学的往会珍认为使用特征权重调整的自适应方法可以帮助在话题跟踪中提高其性能。最后实验结果表明她的想法是可行的。

而另外一个学者针对主题漂移的问题提出了自己的解决方案，潘渊采用在增量聚类过程中不断修正主题重心的策略。哈尔滨工业大学的王巍提出了基于多中心模型的网络特点话题发现算法，利用话题重心判别文档流中的文档是否属于已有网络热点话题。也有的人提出了加入情感打分的分值做为聚类的特征。

#### 2.2.1话题聚类算法的深入研究

话题挖掘算法以及开始了被很多人都行了相对深入的研究，他们大都针对某些点上相处了一些新的问题并且给出了一个一些自己的见解和想法，他们的主要工作可以总结为以下两个类别，第一类是直接针对现有的聚类算法进行改进，或者提出一些新的点，或者针对以前的模型中不合理的部分进行改进。第二类是寻找新的话题特征。

KMeans算法由于其简单快速的特性，并且聚类效果也较好，是一种常用于话题发现的常用聚类算法。在TDT初期，Dragon System小组用KMeans聚类算法进行回顾式话题检测，文档表示形式先后采用了一元、二元语言模型，相似度度量通过Kullback-Leibler距离公式计算。[雷2006]提出了改进的增量KMeans方法，该方法采用密度函数法初始化类中心，可用于检测回顾式和在线式两种话题。

但由于依赖初始簇中心的随机选择，因此不能保证收敛于全局最优解，从而导致算法不稳定。于是，人们提出了一些初始中心点选取的优化算法，[秦 2006] 提出了一种改进算法。该算法先探测数据集中的相对密集区域，再利用密集区域生成初始类中心点。该方法能够很好地排除类边缘点和噪声点的影响，并且能够适应数据集中各个实际类别密度分布不平衡的情况。但是该方案对于生成密集区域的计算量较大，并且没有很好的区分方法，从而会引发类间多簇的现象。KMeans++算法提出了一种基于权重随机函数的初始中心点选取方法，提高了算法的收敛速度。[牟2008]克服了KMeans算法对初始聚类中心过分依赖的缺点，文章提出采用竞争神经网络和密度思想对经典KMeans算法进行预处理，从而改变经典KMeans算法对初始聚类中心的随机选择。但该方法预处理过程时间复杂度较大。[魏 2012]利用最大边缘相关(Maximal marginal relevance，MMR)的方法，改进KMeans初始类中心的选择，实验表明，该策略可以明显的减少KMeans算法的迭代次数，并且提高聚类的正确率。这些算法主要针对KMeans初始中心点问题进行了优化。

近年来，寻找新的话题特征的努力，主要表现在以下几个方面：话题的时间聚集性，话题的特征词，话题的生命演变特征以及话题的命名实体等。

1）话题的时间聚集性

[APL1998]主要是使用单遍法聚类算法和一个新的阈值控制模型，来进行在线新事件的检测。该阈值控制策略的基本思想是：相距较远的两个报道必须具有较大的相似性才能将其划分为同一个事件，而相距较近的两个文档则需要较小的相似性将其归为同一个事件。

2）话题的特征词

[CLC2007]采用老化理论提取话题的关键词，采用多维关键词向量方式来表示文档，采用层次聚类算法实现话题发现，实验室表明，这种方式比纯粹地采用TF\*IDF的方式，在覆盖率和正确率方面都有明显的提升。

3）话题的生命演变特征

[CCC2009]等人提出的生命特征识别方法认为事件的发展具有一定的特定模式，即产生、发展、壮大和消亡，因此可以通过隐马尔科夫模型训练已知的若干种具有不同生命特征的事件演变方式，然后再对每个新事件的行为模式进行预测。通过对不同的事件演变阶段赋以不同的检测阈值，即通过动态话题阈值策略来改进已有话题检测方法的检测效果。

4）话题的命名实体

[KA2004]在2004年提出了通过文本分类和命名实体来改进新事件检测的效果。该论文通过对文本进行分类，对不同类别给予不同的相似性阈值，通过文本的多重表示方法，即将一个文本表示为三个部分：由所有的特征词构成的表示、由命名实体构成的表示、由非命名实体构成的表示，改进文本内容相似性计算的效果。

[MAS2004]在2004年提出了一种将文本中的特征词进行分类的方法，即将关键词分为地点、名字、时间和一般特征词等，然后在各自的类别上进行文本内容的比较。

### 2.3 分布式系统数据切分

传统分布式数据划分方法主要关注减少系统各个节点之间的通信量，最大化地促进系统并发，以及有效均衡各节点负载等问题。但是在分布式话题检测在数据划分中，由于其具体聚类算法的特性，数据挖掘过程往往依赖输入数据的一些特征属性，比如数据时间，数据来源和文本自身相似度等。因此在数据划分时必须考虑到这些因素，以便能够利用这些属性提高话题挖掘的准确率。

目前，这方面的代表性工作有：

以提高并发度为目标的数据划分则需要结合具体的应用背景实施。[赵葆华 等 2000]提出在快速局域网的多数据库环境下，通过合理的数据划分，使查询能够最大限度地并非执行，从而提高查询的相应时间的方案。[杨小虎 等 2008] 则更多地从负载均衡的角度考虑制定分布式数据库的数据划分策略。[Zhang et al 2006] 在研究并行 K-Means算法时，采用按记录随机切分的数据划分，以达到均衡负载，提高并发度的目的。

以上方案都是着重考虑分布式系统的高并发性，小通信量，高效负载均衡，应用到实际的分布式文本聚类系统中，以上方案对保证聚类结构质量没有保证，会引起数据碎片

[王孝宗 2010]在数据划分中，引入Web 数据时间局部性概念，假设产生时间接近的Web 数据比产生时间间隔大的Web 数据更倾向于属于同一个话题。并以此作为依据进行数据的初始化分。该方案假设限定性过强，对网络文本的抽取时间依赖性过高，可以对于不同通道进行更多细粒度的假定；而且该方案没有考虑输入数据本身的内容属性特征。

### 2.4分布式话题聚类与优化方法

在分布式话题检测方案中，主要涉及局部聚类算法的选择、优化和后期全局话题合并。而全局话题合并中又包括垃圾话题过滤和全局话题排序两部分。以下为相应方向国内外研究现状。 聚类是一种无指导的基于内容相似性的分类方法。[Rijsbergen 1979]对聚类有过历史性的记述。大部分的聚类算法都可以用于话题挖掘，其中K-Means算法是一种应用非常广泛的算法，根据初始化类中心策略的不同，计算文本与类相似度的不同，以及更新类中心的不同等，发展出了很多变种。[MacQueen 1967]最早提出了 K-Means 算法的思想并对其收敛性等做了理论上的分析与证明。算法的基本思想如下：

步骤1. 随机从数据集中选择 K个点作为初始聚类中心；

步骤2. 计算各个样本到聚类中心的距离，把样本加入与它最近的那个聚类中心所属

的类；

步骤3. 计算每个聚类数据对象的平均值作为新的类中心；

步骤4. 重复 步骤2，步骤3直到达到某种退出条件。

但K-Means算法的局限性在于初始簇心的随机选择性对最终结果质量的影响。

[秦 钰 等 2006] 提出了一种K-means算法的改进方案。算法先探测数据集中的相对密集区域，再利用密集区域生成初始类中心点。该方法能够很好地排除类边缘点和噪声点的影响，并且能够适应数据集中各个实际类别密度分布不平衡的情况。但是该方案对于生成密集区域的计算量较大，并且没有很好的区分方法，从而会引发类间多簇的现象。

[牟颖 等 2008]克服了K-Means算法对初始聚类中心过分依赖的缺点，文章提出采用竞争神经网络和密度思想对经典k-means算法进行预处理，从而改变经典K-means算法对初始聚类中心的随机选择。但该方法预处理过程时间复杂度较大。

Yi Wang等人实现了两种针对LDA的并行化策略，这个就是很出名的PLDA算法，他们分别使用了两种分布式框架，MPI和MapReduce，他们的策略都是基于AD-LDA来完成的。

AD-LDA算法的思想就是充分的利用数据水平切分的优势，同时合理的将子模型融合，从而得到最终的子话题的结果。首先将数据集分成P份，分别放到不同的处理器上，然后在各自的内部分别运行LDA的采样，并且在内部各自重新计算文档-主题分布矩阵，经过一个周期之后再将文档中词汇的采样结果分别同步到主运行机上。在主运行机上，等到收到所有的节点的同步信息之后重新计算主题-词语分布矩阵，并且将它们重新分配给各个子节点，重复上面这个过程直到运行结果收敛为止。

ZHIYUAN LIU 等人总结分析了当前LDA并行化算法中存在的问题，他们认为之前的LDA并行化算法可以总结为两种类型：（1）松耦合的分布式计算机并行化算法，（2）紧耦合的类似多核计算机和GPU的分布式算法。他们认为Dirichlet Compound Multinomial LDA[Mimno and McCallum 2007]，Approximate Distributed LDA(AD-LDA)[Newman et al, 2007]是属于前者，它们的策略是在不同的非共享内存的机器上进行Gibbs Sampling。但是这种方法可能会消耗很大一部分时间的机器之间的交互上，这些都是严重影响分布式系统加速的操作。第二类的方法使用了带共享内存的多核CPU或者GPU方法，这种方法能够一定程度上减少机器之间的信息交流，但是当存储和内存被设定之后，这个框架就变得不灵活了。ZHIYUAN LIU等人提出了一种新的做法，它们的方法可以很大程度上的较少机器之间的信息交互，他们的方法包括四个阶段。

(1) *Data placement*. Data placement aims to separate CPU-bound tasks and

communication-bound tasks onto two sets of processors. Data placement enables

us to employ a pipeline scheme (discussed next), to mask communication by computation.

(2) *Pipeline processing*. To ensure that a CPU-bound processor is not blocked by communication,

PLDA+ conducts Gibbs sampling for a *word bundle* while performing

intercomputer communication on the background. Suppose Gibbs sampling is performed

on the words “foo” and “bar”. PLDA+ fetches the metadata for the word

“bar” while performing Gibbs sampling on the word “foo”. The communication time

for fetching the metadata of “bar” is masked by the computation time for sampling

“foo”.

(3) *Word bundling*. In order to ensure that communication time can be effectively

masked, the CPU time must be long enough. Revisiting the example of sampling

“foo” and “bar”, the CPU time for sampling the word “foo” should be longer than

the communication time for the word “bar” in order to mask the communication

time. Suppose we performed Gibbs sampling according to the order of words in

documents, each Gibbs sampling time unit would be too short to mask the required

communication time. Since LDA treats a document as a bag of words and entirely

ignores word order, we can flexibly process words on a processor in any order

without considering document boundaries. Word bundling combines words into

large computation units.

*(4) Priority-based scheduling*. *Data placement* and *word bundling* are static allocation

strategies for improving pipeline performance. However, runtime factors would almost

always affect the effectiveness of a static allocation scheme. Therefore, PLDA+

employs a priority-based scheduling scheme to smooth out runtime bottlenecks.

### 2.5 子话题研究

前面说了很多有关话题挖掘的算法，但是有关子话题的算法研究确实目前还是比较不成熟的，有很多有关搜索引擎的查询结果的子话题挖掘的算法的研究。对搜索结果进行子话题挖掘可以提供更加便利的搜索结果展示，提供层次化的结果展示，是一件非常有意义的事情，有人提出了使用层级聚类的算法对搜索结果进行子话题挖掘，得到了不错的效果。这种方法针对比较少的关键词背景的时候可以得到不错的效果，并且搜索引擎搜素结果的并不需要包含所有的搜索者中的关键词，因此与本文提出的专题内部的子话题挖掘还是存在一定的区别。

### 2.6子话题关键词抽取技术

话题的关键词词组对于话题的展示是非常重要的工作，但是针对这一方面的工作却相对较少。Popescul和Ungar(2000)通过慈祥的卡方信息机文档集频率的组合获得了较好的结果。Glover等人(2002b)shiyong 信息增益来标识Web网页所生成的簇。Stein和Su Eissen提出了基于本体的方法。（Glover et al. 2002a）和（Treeratpituk and Callan , 2006）对层次结构下更复杂的标签问题进行了处理。据我们所知，目前还没有一项完整的工作可以对这种基于标签的聚类算法的结果质量进行比较。理论上说，多文档摘要额也可以用于簇标签生成。

## 去背景化的LDA子话题模型

### 3.1 去背景化的思想

专题文章有一个特点就是它们之间拥有一些公共的背景知识，这些公共的背景知识也许会给话题的学习带来一些不必要的麻烦，因为共同的背景知识会使得不同的子话题之间的关联性增大，从而使得普通的话题分析方法可能无法将这些子话题区分开来，下面我们举个例子来说明公共背景给子话题分析带来的可能的问题。

* 经常性工作都有新的套路、积累了新的经验。在党中央的坚强领导下，各级党组织不仅提高了对加强作风建设的思想认识，增强>了落实管党治党的主体责任，而且明确了新形势下作风建设的思路和方法，丰富了管党治党经验，着力把依规治党与以德治党紧>密结合起来，把经常性教育与集中教育紧密结合起来。&nbsp;&nbsp;&nbsp;&nbsp;习近平总书记在十八届中央纪委六次全会的重>要讲话中，明确了2016年党风廉政建设和反腐败斗争的五大工作要点。一是尊崇党章，严格执行准则和条例；二是坚持坚持再坚>持，把作风建设抓到底；三是实现不敢腐，坚决遏制腐败现象滋生蔓延势头；四是推动全面从严治党向基层延伸；五是标本兼治>，净化政治生态。党的十八大以来，以习近平同志为总书记的党中央以强烈的历史责任感、深沉的使命忧患感，总揽全局、运筹>帷幄，统筹中国特色社会主义伟大事业和党的建设新的伟大工程，把全面从严治党纳入“四个全面”战略布局，把“四个全面”……
* 习近平对食品安全工作作出重要指示习近平对食品安全>工作作出重要指示提交我的留言加载中已留言习近平对食品安全工作作出重要指示2016-02-18中国茶叶流通协会中国茶叶流通协>会中国茶叶流通协会微信号sinictea功能介绍发挥行业中介职能，服务中国茶叶事业。1月28日央视新闻联播报道，中共中央总书记、国家主席、中央军委主席习近平日前对食品安全工作作出重要指示强调，确保食品安全是民生工程、民心工程，是各级党委>、政府义不容辞之责。近年来，各相关部门做了大量工作，取得了积极成效。当前，我国食品安全形势依然严峻，人民群众热切>期盼吃得更放心、吃得更健康。2016年是“十三五”开局之年，要牢固树立以人民为中心的发展理念，坚持党政同责、标本兼治加强统筹协调，加快完善统一权威的监管体制和制度，落实“四个最严”的要求，切实保障人民群众“舌尖上的安全”。
* 今日离开台湾。《李毅台湾观选记》，已经发表了十篇。这篇总结，分为五个部分。第一，简介一下台湾问题的现状与这次选举的结果。第二，检验一下我自己过去的研究，>也就是那三篇文章的结论。第三，探讨一下社会研究方法。第四，对习近平，说几句话。第五，对蔡英文，说几句话。出于对习>近平和蔡英文的尊重，把第四、第五部分，提前放在卷首。本文的主干观点非常明确：和平统一已无可能。恭请海内外台湾问题>研究专家指教，特别恭请两岸的台海问题专家指教。&nbsp;致习近平1.&nbsp;总述：和平统一，已无可能。现在不解决台湾问题>，台湾问题必将无限期拖延下去，且日益恶化。现在，在军事、经济、政治、外交、党心、士气、民气、台情等等方方面面，大>陆武力统一台湾的时机都已经完全成熟。
* 坚持党的领导，思想政治共识有新提升&nbsp;&nbsp;&nbsp;&nbsp;通过常委会议、主席会议、政协党组会议等形式，深入学习中共十八大以及十八届三中、四中、五中全会精神，深入学习习近平>总书记系列重要讲话精神，深入学习市委十二届八次、九次全体会议精神，推动各族各界委员坚定理想信念、把握大局大势、增>进政治认同。召开专题学习会议、座谈交流会议，集中学习姜异康同志、徐景颜同志在省委、市委政协工作会议上的重要讲话，>在人民政协形成学讲话、抓落实、促工作的良好氛围，进一步增强了与中共聊城市委在思想上同心同德、行动上同心同向、工作>上同心同力的责任感和使命感。

上面这几篇文章都是与习近平主席相关的一些文章，它们中有与反腐廉政相关的，也有与台湾问题相关的，还有与食品安全相关的，也有与政府工作相关的，由于篇幅有限，这里我们有没有给出太多的文章举例，知识找出了一部分比较有代表性的文章作为说明，给出这几篇文章的目的是告诉大家这些文章之间有很大的主题差异，但是内容中又有很多的公共的背景知识，例如十八大，十三五，习近平主席等等公共的背景知识，我们现在来观察一下直接使用原始的LDA算法，这里的LDA工具是GibbsLDA++。我们观察一下最后的主题-词汇信息结果，我们取了3487二月底到三月初的与习近平相关的微信公共账号发表的文章，然后设定主题数量为20个，迭代400次的结果，最后观察每个主题的关键词列表。

表3-1 基于GibbsLDA++的非去背景化的运行结果话题

|  |  |
| --- | --- |
| 主题1 | Nbsp 工作 中 中国 发展 新 经济 新闻 改革 说留言建设 做 舆论 学习 习近平 |
| 主题2 | nbsp，发展，中国，工作，中，经济，新，说，做  年 ，建设，改革 ，学习，国家，教育，社会 ，党 |
| 主题3 | nbsp 中国 中 发展 新 经济 工作 国家 美国 说  教育 留言 年 做 改革 建设 关注 世界 |
| 主题4 | nbsp 发展 中国 中 经济 中医药 新 年 企业  工作 服务 改革 建设 签证 创新 国家 社会 健康 |
| 主题5 | nbsp 中国 发展 中 工作 新 经济 说 党 年 建设国家 改革 关注 留言 做 社会 新闻 政策 美国 |
| 主题6 | nbsp中国 发展 经济 中 工作 新 改革 说 年  企业 国家 政策 党 建设 市场 教育 社会 服务 |
| 主题7 | 中医药 中医 发展 服务 健康 nbsp 医疗 中药  体系 机构 建立 中药材 建设 文化 医院 中国 |
| 主题8 | nbsp 发展 中 中国 工作 经济 党 新 做 中医药  社会 服务 建设 留言 教育 改革 创新 企业 |
| 主题9 | nbsp中国 中 发展 工作 新 经济 说 建设 国家  改革 年 做 美国 政策 企业 社会 党 留言 关注 |
| 主题10 | 中国 日本 蒙古国 Nbsp 中 美国 安倍 发展 政府  蒙古 世界 经济 工作 美 新 国家 国 罕见 建设 |
| 主题11 | 举报 Nbsp 工作 农村 电话 发展 户口 新 建设  中国 教育 改革 经济 留言 新闻 农民 国家 领导 |
| 主题12 | Nbsp 发展 中国 中 工作 新 经济 建设 改革 年  国家 留言 指挥 说 企业 关注 政策 战区 分 做 |

由于主题数目太多，这里面我们没有列出所有的话题信息，我们可以发现其中有很多的主题，它们之间的相似程度非常的高，几乎每一个主题中都包含了“中国”，“经济”，“发展”等字样，基本上的主题信息都没有特别的清晰，偶尔能够看到一两个主题信息比较有清晰地主题。

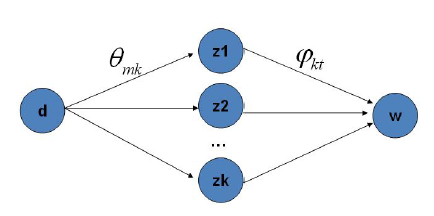
如果把主题之间的Jaccard-Distance当做是主题差异指标的话，这些主题关键词的差异指标将会非常的小。

TODO列出Jaccard-Distance的计算方法

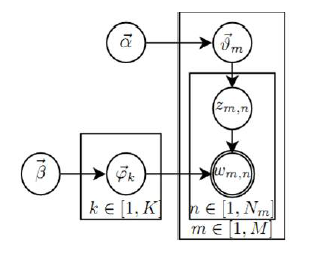
因此为了能够的进行专题文章的子话题挖掘，本文提出了去背景化的话题学习方法，去背景化的思想就是考虑文章的词有可能来自背景也有可能来自除背景知识之外的主题，算法在学习的过程中需要学习的知识包括背景知识和主题知识，最后背景知识是可以被丢弃的部分，为了学习到子话题之间的差异性，我们只需要考虑主题模型的知识就好了。对于去背景化的思想可以有多种做法，本文提出了两种基本的去背景化的策略。第一种就是考虑将背景知识带入算法迭代的过程中，认为背景知识是学习中的一个固定存在，它是语料生成的一部分。第二种思想就是将背景知识先提前剔除，然后在剔除了背景知识之后的文章中学习这些子话题信息。前一种策略被应用到了去背景化的LDA子话题模型中，第二种策略被应用到了去背景化的KMEANS算法中，这两种方法都很好的实践了去背景化的思想。

### 3.2 LDA子话题模型中的去背景化思想

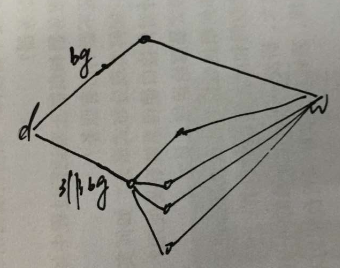
LDA最初始被应用于文档的主题模型的分析，LDA是一个产生式模型，作者认为每一篇文档都是由多个主题构成的，并且每个词语的生成过程都是先生成主题然后再根据主题曲生成这个词语，假设一篇文档有k个主题，那么一个词语的生成过程可能就有k条路径，首先随机的在这K个主题中选取一个主题，然后根据这个主题的词分布矩阵来生成这个词语。原始的LDA认为文档中的主体信息分布符合多项式分布，同时主体中的词语分布也是符合多项式分布的，为了减少过拟合的可能性，LDA给这两个分布都添加了一个共轭先验，他们都是Dirichlet分布，具体的可以参考图3-2 LDA图模型表示。那么去背景化的LDA与原始的LDA算法存在的一个重要的区别就是去背景化的LDA认为一个词语有可能来自背景语料也可能来自于差异化的主题模型中，背景语料在主题模型的迭代过程中不会发生变化，背景语料可以提前通过都整体语料的统计而计算出来，而差异化的主题模型的词语的概率分布则需要在后面的更新迭代的过程中计算出来。去背景化的LDA的一个词语生成的过程应该是如图3-3所示。从图3-3中可以看出每一个词语的生成路径都有了多一种新的选择，第一种是直接从背景中获得，第二种是从主题模型中获得，这里注意区分背景不是一个主题，假设是一个主题的话那么给去背景化的主题模型就相当于在原始的模型上添加了一个主题而已，显然这种做法是肯定不正确的。



文档-主题-词的生成路径



LDA图模型表示



去背景化的LDA的词语生成过程示意图

统计背景语料生成某个词的概率的方法可以有多种，最直接的方法就是根据词语频率进行统计，可以考虑一个词语在一篇文档中出现多次按多次计算，也可以考虑一个词语在一篇文档中出现多次按照一次计算。在LDA的迭代过程中，背景语料生成词语的概率是不变的，不需要进行更新。这就将背景语料的概率信息与差异化的主题信息区分开了，而不会将背景语料当做是一个主题来学习。固定背景预料的信息是去背景化LDA中至关重要的点。这个必须在之后的设计与实现的过程中牢记。

由于背景语料的知识是提前学习的，并且在LDA算法迭代的过程中不会进行更新，所以这部分知识对于调节后面的差异化的主题信息比较重要，固定的背景知识是所有子话题的公共背景知识。在算法的迭代过程中，每次进行主题采样的过程中，不仅要考虑差异化的主体信息，同样需要进行对背景知识的可能性进行采样。如果一个词语在背景语料中频繁的出现，那么在对这个词进行主题采样的时候就可能将它认定为来自背景知识，而非差异化的主题之中。具体的采样细节我们可以参考3.3节的阐述。

### 3.3 LDA子话题去背景化的实现

#### 3.3.1 LDA介绍

LDA模型是对PLSA模型的一个改进，作为频率学派的人一定会认为PLSA已经是模型中的佼佼者了，但是作为贝叶斯学派的学者则不这么认为。贝叶斯学派认为一切的参数都不是一个固定的数字，而是一个分布，理所当然的贝叶斯学派的学者要给PLSA添加一个贝叶斯框架。

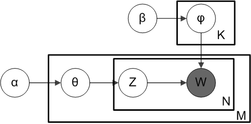


图3-3-1

图3-3-1是LDA模型的Plate

* 是唯一的可视变量，而其他的参数都是隐变量，很多时候，我们都会使用基本的LDA的都会被要求变得更加平滑，从而可以获得更好地结果。上面的图称为概率图，图中的参数解释可以从下面的表3-1 LDA相关变量解释说明表中得到。
* 是一个*K\*V（V*是词汇量的维度*）*的矩阵，每一行都是一个主题的单词分布情况。

LDA主题模型是一个产生式模型，它的生成过程描述如下。每一篇文档都是随机的由多个主题构成的，并且每一个主题又被描述成一个词汇的分布矩阵。LDA模型的产生过程是这样的，语料D是一个由M篇文档组成的集合，并且每一篇文档的单词个数为N_i。

1. Choose  \theta_i \, \sim \, \mathrm{Dir}(\alpha) , where  i \in \{ 1,\dots,M \}  and  \mathrm{Dir}(\alpha)  is the [Dirichlet distribution](https://en.wikipedia.org/wiki/Dirichlet_distribution" \o "Dirichlet distribution) for parameter \alpha

2. Choose  \varphi_k \, \sim \, \mathrm{Dir}(\beta) , where  k \in \{ 1,\dots,K \} 

3. For each of the word positions i, j, where  j \in \{ 1,\dots,N_i \} , and  i \in \{ 1,\dots,M \} 

(a) Choose a topic z_{i,j} \,\sim\, \mathrm{Multinomial}(\theta_i). 

(b) Choose a word w_{i,j} \,\sim\, \mathrm{Multinomial}( \varphi_{z_{i,j}}) .

The lengths N_i are treated as independent of all the other data generating variables (w and z). The subscript is often dropped, as in the plate diagrams shown here.

Mathematical definition[[edit](https://en.wikipedia.org/w/index.php?title=Latent_Dirichlet_allocation&action=edit&section=3" \o "Edit section: Mathematical definition)]

A formal description of smoothed LDA is as follows:

|  |  |  |
| --- | --- | --- |
| **变量名称** | **类型** | **解释** |
| K | integer | 主题个数（例如50个主题） |
| V | integer | 词汇量的大小 (例如 50,000 或者1,000,000) |
| M | integer | 文档集合的大小，就是文章的总数量 |
| N_{d=1 \dots M} | integer | 文档d中的词语的个数 |
| N | integer | 所有文章中的单词的总数量; 也就是 N_d 的求和大小,  N = \sum_{d=1}^{M} N_d |
| \alpha_{k=1 \dots K} | positive real | 主题k的先验值，通常是一个正的实数; 通常每一个主题的先验值得大小都是一致的; 通常是一个小于1的实数, 例如. 0.1, |
| \boldsymbol\alpha | *K*-dimension vector of positive reals | 主题k的先验值得向量表现形式 |
| \beta_{w=1 \dots V} | positive real | 单词w的先验值; 通常每个单词的先验值都是一致的;通常是一个小于1的实数, 例如. 0.001, 通常在文档中单词非常系数的时候会很有用，可以帮助更好地平滑数据 |
| \boldsymbol\beta | *V*-dimension vector of positive reals | 单词w的先验值的向量表现形式 |
| \varphi_{k=1 \dots K,w=1 \dots V} | probability (real number between 0 and 1) | 主题k下单词w出现的概率 |
| \boldsymbol\varphi_{k=1 \dots K} | *V*-dimension vector of probabilities, which must sum to 1 | 主题k的单词概率分布 |
| \theta_{d=1 \dots M,k=1 \dots K} | probability (real number between 0 and 1) | 文档d中主题k概率值 |
| \boldsymbol\theta_{d=1 \dots M} | *K*-dimension vector of probabilities, which must sum to 1 | 文档d的主题分布 |
| z_{d=1 \dots M,w=1 \dots N_d} | integer between 1 and *K* | 文档d中的词语w的主题值 |
| \mathbf{Z} | *N*-dimension vector of integers between 1 and *K* | 文档d的所有词的主题值得向量形式 |
| w_{d=1 \dots M,w=1 \dots N_d} | integer between 1 and *V* | identity of word *w* in document *d* |
| \mathbf{W} | *N*-dimension vector of integers between 1 and *V* | identity of all words in all documents |

We can then mathematically describe the random variables as follows:


\begin{array}{lcl}
\boldsymbol\varphi_{k=1 \dots K} &\sim& \operatorname{Dirichlet}_V(\boldsymbol\beta) \\
\boldsymbol\theta_{d=1 \dots M} &\sim& \operatorname{Dirichlet}_K(\boldsymbol\alpha) \\
z_{d=1 \dots M,w=1 \dots N_d} &\sim& \operatorname{Categorical}_K(\boldsymbol\theta_d) \\
w_{d=1 \dots M,w=1 \dots N_d} &\sim& \operatorname{Categorical}_V(\boldsymbol\varphi_{z_{dw}}) \\
\end{array}


**LDA模型的参数学习**

学习LDA的参数的方法有很多种，其中涉及的变量有一些列的主题信息，它们的单词概率分布，每一篇文档的主题分布，主要使用的方法是 [Bayesian inference](https://en.wikipedia.org/wiki/Bayesian_inference)。最初是的论文使用的是[variational Bayes](https://en.wikipedia.org/wiki/Variational_Bayes)策略来学习后验概率的近似分布。另外还有一种替换式的策略是使用Gibbs Sampling或者Expectation Propagation。

Following is the derivation of the equations for [collapsed Gibbs sampling](https://en.wikipedia.org/wiki/Collapsed_Gibbs_sampling), which means \varphis and \thetas will be integrated out. For simplicity, in this derivation the documents are all assumed to have the same length N_{}. The derivation is equally valid if the document lengths vary.

According to the model, the total probability of the model is:

 P(\boldsymbol{W}, \boldsymbol{Z}, \boldsymbol{\theta},
\boldsymbol{\varphi};\alpha,\beta) = \prod_{i=1}^K
P(\varphi_i;\beta) \prod_{j=1}^M P(\theta_j;\alpha) \prod_{t=1}^N
P(Z_{j,t}|\theta_j)P(W_{j,t}|\varphi_{Z_{j,t}}) ,

First of all, \boldsymbol{\varphi} and \boldsymbol{\theta} need to be integrated out.


\begin{align}
&P(\boldsymbol{Z}, \boldsymbol{W};\alpha,\beta)  =  \int_{\boldsymbol{\theta}} \int_{\boldsymbol{\varphi}} P(\boldsymbol{W}, \boldsymbol{Z}, \boldsymbol{\theta}, \boldsymbol{\varphi};\alpha,\beta) \, d\boldsymbol{\varphi} \, d\boldsymbol{\theta} \\
 = & \int_{\boldsymbol{\varphi}} \prod_{i=1}^K P(\varphi_i;\beta) \prod_{j=1}^M \prod_{t=1}^N P(W_{j,t}\mid\varphi_{Z_{j,t}}) \, d\boldsymbol{\varphi} \int_{\boldsymbol{\theta}} \prod_{j=1}^M P(\theta_j;\alpha) \prod_{t=1}^N P(Z_{j,t}\mid\theta_j) \, d\boldsymbol{\theta}.
\end{align}


The goal of Gibbs Sampling here is to approximate the distribution of P(\boldsymbol{Z}\mid\boldsymbol{W};\alpha,\beta). Since P(\boldsymbol{W};\alpha,\beta) is invariable for any of Z, Gibbs Sampling equations can be derived from P(\boldsymbol{Z}, \boldsymbol{W};\alpha,\beta) directly. The key point is to derive the following conditional probability:

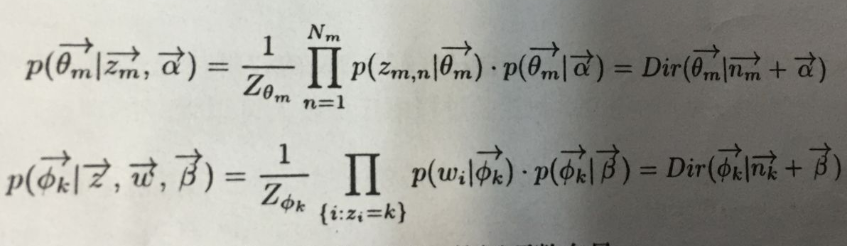
 P(Z_{(m,n)}\mid\boldsymbol{Z_{-(m,n)}},
\boldsymbol{W};\alpha,\beta)=\frac{P(Z_{(m,n)},
\boldsymbol{Z_{-(m,n)}},\boldsymbol{W};\alpha,\beta)}
{P(\boldsymbol{Z_{-(m,n)}}, \boldsymbol{W};\alpha,\beta)} ,

where Z_{(m,n)} denotes the Z hidden variable of the n^{th} word token in the m^{th} document. And further we assume that the word symbol of it is the v^{th} word in the vocabulary. \boldsymbol{Z_{-(m,n)}} denotes all the Zs but Z_{(m,n)}. Note that Gibbs Sampling needs only to sample a value for Z_{(m,n)}, according to the above probability, we do not need the exact value of P(Z_{m,n}\mid\boldsymbol{Z_{-(m,n)}},
\boldsymbol{W};\alpha,\beta) but the ratios among the probabilities that Z_{(m,n)}can take value. So, the above equation can be simplified as:


\begin{align}
& P(Z_{(m,n)}=k\mid\boldsymbol{Z_{-(m,n)}}, \boldsymbol{W};\alpha,\beta) \\
\propto &
P(Z_{(m,n)}=k,\boldsymbol{Z_{-(m,n)}},\boldsymbol{W};\alpha,\beta) \\
= & \left(\frac{\Gamma\left(\sum_{i=1}^K \alpha_i
\right)}{\prod_{i=1}^K \Gamma(\alpha_i)}\right)^M \prod_{j\neq m}
\frac{\prod_{i=1}^K
\Gamma(n_{j,(\cdot)}^i+\alpha_i)}{\Gamma\bigl(\sum_{i=1}^K
n_{j,(\cdot)}^i+\alpha_i \bigr)} \\
& \times \left( \frac{\Gamma\bigl(\sum_{r=1}^V \beta_r
\bigr)}{\prod_{r=1}^V \Gamma(\beta_r)}\right)^K \prod_{i=1}^K
\prod_{r\neq v}
\Gamma(n_{(\cdot),r}^i+\beta_r) \\
& \times  \frac{\prod_{i=1}^K
\Gamma(n_{m,(\cdot)}^i+\alpha_i)}{\Gamma\bigl(\sum_{i=1}^K
n_{m,(\cdot)}^i+\alpha_i \bigr)}  \prod_{i=1}^K \frac{
\Gamma(n_{(\cdot),v}^i+\beta_v)}{\Gamma\bigl(\sum_{r=1}^V
n_{(\cdot),r}^i+\beta_r \bigr)} \\
\propto & \frac{\prod_{i=1}^K
\Gamma(n_{m,(\cdot)}^i+\alpha_i)}{\Gamma\bigl(\sum_{i=1}^K
n_{m,(\cdot)}^i+\alpha_i \bigr)}  \prod_{i=1}^K \frac{
\Gamma(n_{(\cdot),v}^i+\beta_v)}{\Gamma\bigl(\sum_{r=1}^V
n_{(\cdot),r}^i+\beta_r \bigr)}\\
\propto & \prod_{i=1}^K
\Gamma(n_{m,(\cdot)}^i+\alpha_i)  \prod_{i=1}^K \frac{
\Gamma(n_{(\cdot),v}^i+\beta_v)}{\Gamma\bigl(\sum_{r=1}^V
n_{(\cdot),r}^i+\beta_r \bigr)}.
.
\end{align}

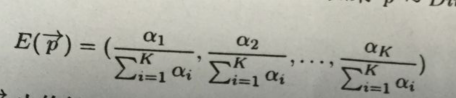

最后一步就是根据Markov链的状态获取主题分布的参数和次分布的参数。

根据贝叶斯法则和Dirichlet先验，以及上文中得到的p(w|z,)和p(z|)各自被分解成两部分成绩和结果，可以计算得到每个文档上Topic的后验分布和每个Topic下的词的后验分布分别如下。

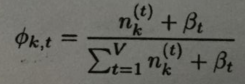


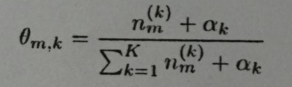
其中，是构成文档m的主题数向量，是构成主题k的词向数向量。

Dirichlet函数的一个性质就是各位的期望值可以按照下面的方式计算：

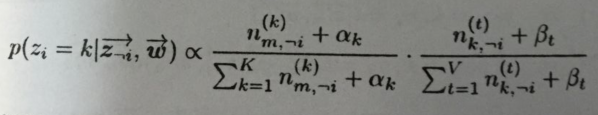


可以看出，超参数ak的直观意义就是时间鲜艳的伪计数，所以最终求解的Dirichlet分布期望为：



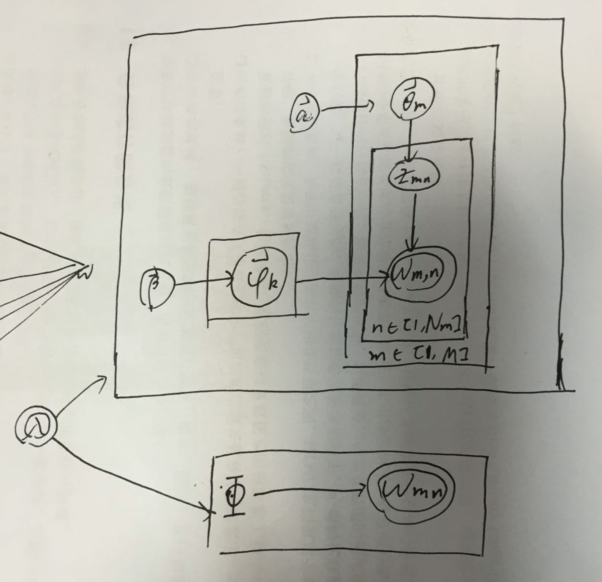


然后将和的结果带入之前得到的的结果中，可得：



仔细观察上述结果，可以发现，式子的右半部分便是P(topic|doc).p(word|topic)。这个概率的值对应着doc->topic->word的路径概率。如此，K个Topic对应着K条路径，Gibbs Sampling便在这K条路径中进行采样。

#### 3.3.2 去背景化的LDA算法模型



去背景化的LDA的关键就是认定文档的内容有可能来自背景语料，也可能来自差异化的话题之中。重新设定产生词语的过程：

1 设定一个概率认定这个词来自于背景还是差异化的主题信息中

2 分别计算这个词语来自背景语料的生成概率和这个词语来自差异化的子话题之中采样后的结果的生成概率。

3 设定规则，如果这个词语来自背景语料的生成概率超过某个阈值那么直接认为这个词语来自背景语料，那么这个词将在后续计算文档的主题概率和主题的词语概率的过程中被忽略。此外，如果背景语料生成这个词的概率不超过某个阈值，就将背景生成路径也加入采样路径。

具体的步骤请流程图。

**流程图**

同时我们也要更新初始化词语的初始主题的过程，随机化的初始过程，考虑加入这个词语可能来自于背景的可能性，先随机生成一个0到1的小数，如果这个数小于背景的概率就认为它的主题信息为-1，否则再次随机生成一个1到K的数设定为它的主题。

#### 3.3.3 实验结果

实验数据取了从2月25日到2月28日的所有的包含习近平等相关词汇的文章，作为我们的训练语料，将实验数据人工标注为20个类别，总共包含3487篇文档，每个类别的细节信息可以参考表3-1 训练数据类别信息表。

|  |  |  |
| --- | --- | --- |
| 类别标号 | 类别名称 | 类别大小 |
| 1 | 中美货币&天合积分 |  |
| 2 | 文化艺术 |  |
| 3 | 考试&高考 |  |
| 4 | 企业成功学&任正非&华为 |  |
| 5 | 教育工作 |  |
| 6 | 心里鸡汤 |  |
| 7 | 经济发展&城镇建设 |  |
| 8 | 干部会议&国防建设 |  |
| 9 | 蒙古&俄罗斯 |  |
| 10 | 企业改革&投资 |  |
| 11 | 农村工作&土地&户口 |  |
| 12 | 医疗健康 |  |
| 13 | 金融经济&市场改革 |  |
| 14 | 中美南海外交 |  |
| 15 | 党员工作与责任 |  |
| 16 | 新闻媒体&舆论 |  |
| 17 | 旅游交通 |  |
| 18 | 中日韩等外交问题 |  |
| 19 | 任志强之党员问题 |  |
| 20 | 行业改革之互联网 |  |

本文针对针对该语料将分别进行普通的LDA算法和去背景化的LDA算法，观察这两种算法在这个数据集合上的聚类结果和关键词抽取的结果，并且提供客观的评价指标，比较两种算法的聚类结果。

20个主题

Noback 400次迭代结果位0.308

Back 500次迭代结果为0.676

Back 600次迭代结果为0.75

从两个方面去分析实验的结果，第一个是从实验的主题关键词的挖掘结果进行评判，第二个是从聚类结果进行实验结果的评判。

为了对主题关键词进行评判，实验组寻找人工评判标准，对于主题关键词的评判结果，我们评判的标准是主题召回率，每次试验我们不同的模型都会给每个主题风别给出20个关键词，当做是该主题的代表，然后去评判能够召回多少原始的主体信息。

第二个主题关键词的评判标准是，不同主题关键词之间的差异化信息量，这个也是我们使用去背景化思想的最主要目的，就是我们希望我们的模型能够很好地抽取出有差异化的主题信息。这个指标比较不好测试，但是还是有一些粗糙的指标。我们这里设定一个指标，就是关键词的平均相似度信息，而关键词词组相似度信息的方法就是通过计算它们之间的同现词的比例。

从聚类的结果来评判聚类结果是聚类的最直接的评价指标，purity方法是极为简单的一种聚类评价方法，只需计算正确聚类的文档数占总文档数的比例：

http://hi.csdn.net/attachment/201003/5/1179789_1267789323l0kf.jpg

其中Ω = {ω1,ω2, . . . ,ωK}是聚类的集合ωK表示第k个聚类的集合。C = {c1, c2, . . . , cJ}是文档集合，cJ表示第J个文档，N表示文档总数。

优点是方便计算，值在0～1之间，完全错误的聚类方法值为0，完全正确的方法值为1。同时，purity方法的缺点也很明显它无法对退化的聚类方法给出正确的评价，设想如果聚类算法把每篇文档单独聚成一类，那么算法认为所有文档都被正确分类，那么purity值为1。而这显然不是想要的结果，因此本文将不使用该指标。

一种能在聚类质量和簇数目之间维持均衡的指标是NMI，它的定义如下：

其中,I是互信息（参见第13章）：

其中，和都是一篇文档属于，和的概率，而第二个等式是对前半部分进行了MLE估计后的结果。

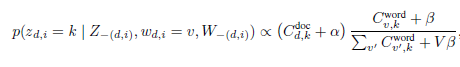
这种评价指标能够综合聚类结果的类别数和类别质量。

与上面的评价指标不同的另外一个评价指标可以将聚类看成是一系列的决策过程，即对文档集上所有的个文档对进行决策。当且仅当两篇文档相似是，我们将它们归入同一簇中，TP决策将两篇相似文档归入一个簇，而TN决策将两篇不相似的文档归入两个不同的簇，在此决策过程中会犯两类的错误：FP决策会将两篇不相似的文档归入同一个簇，而FN决策会将两篇相似的文档归入不同簇。RI计算的是正确决策的比率。

本文将从AC准确率，NMI和RI系数这三个指标来验证去背景化的LDA算法在实际的聚类结果与原始的LDA算法的聚类结果更加合理。

### 3.4 并行化设计

Griffiths和Steyvers提出使用蒙特卡洛方法去学习LDA的参数，使用吉普斯采样法来进行每个词语的主题选取，在迭代的过程中使用对文档中的每一个词语都进行主题采样。如公式（1）所示。

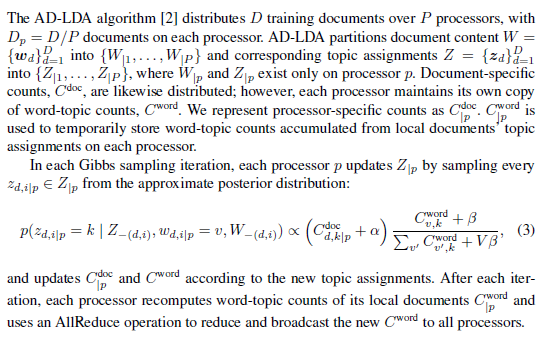


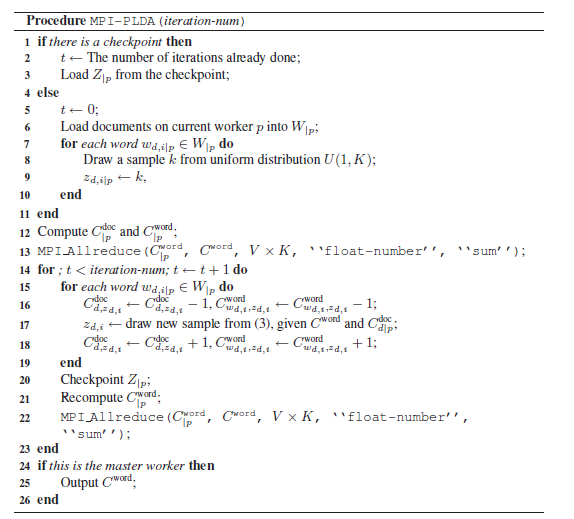
其中k [1;K]，k是主题，v [1;V]是词汇量，wdi是文档d中的第i个单词，W¡(d;i)代表去除wdi后的语料，Z¡(d;i)是W¡(d;i) 对应的主题分布信息。单词v被赋予主题k的总次数，代表的是文档d中有多少单词被赋予了主题k，当迭代了足够多的次数之后，系统就会去更新两个矩阵，一个是文档主题矩阵，一个主题单词矩阵。

单机版本的LDA吉普斯采样算法的时间复杂度是K乘以文档中的总单词数量。之前也有很多的工作用来提高LDA算法的运行速度，而这些方法对于去背景化的LDA算法也同样使用。其中最值得重复使用的是并行化的LDA算法，使用多台机器来运行LDA算法，并且减少算法的运行时间。

LDA算法的并行化思想最重要的就是进行文档分割，将D篇文档分割到P个处理器上，这样的话每一个处理器上需要处理的文档的总的数量就是*Dp* = *D/P,*

每个处理器保存本地的文档数据Cdoc，这个矩阵保存了，每篇文档中包含的不同的主题词汇的数量，同时每个处理器还会保存一份全局的Cword的拷贝，这份拷贝刚开始在每一台机器上都是相同的，但是随着每一台处理器自己更新得过程不会通知其他的机器，随着迭代的进行，每一个处理器上得这部分信息就会出现差异，但是这部分信息的差异对于整体的结果影响不大，每次当处理器处理完本地数据的迭代之后，就会使用All\_reduce函数进行数据的汇总。





## 去背景化的KMEANS模型

### 4.1 KMEANS子话题模型的去背景化思想

#### 4.1.1 KMEANS算法介绍

Given a set of observations (**x**1, **x**2, …, **x***n*), where each observation is a *d*-dimensional real vector, *k*-means clustering aims to partition the *n*observations into *k* (≤ *n*) sets **S** = {*S*1, *S*2, …, *Sk*} so as to minimize the within-cluster sum of squares (WCSS) (sum of distance functions of each point in the cluster to the K center). In other words, its objective is to find:

\underset{\mathbf{S}} {\operatorname{arg\,min}}  \sum_{i=1}^{k} \sum_{\mathbf x \in S_i} \left\| \mathbf x - \boldsymbol\mu_i \right\|^2 

where ***μ****i* is the mean of points in *Si*.

 聚类属于无监督学习，以往的回归、朴素贝叶斯、SVM等都是有类别标签y的，也就是说样例中已经给出了样例的分类。而聚类的样本中却没有给定y，只有特征x，比如假设宇宙中的星星可以表示成三维空间中的点集[clip_image002[10]](http://images.cnblogs.com/cnblogs_com/jerrylead/201104/201104061601448600.png)。聚类的目的是找到每个样本x潜在的类别y，并将同类别y的样本x放在一起。比如上面的星星，聚类后结果是一个个星团，星团里面的点相互距离比较近，星团间的星星距离就比较远了。在聚类问题中，给我们的训练样本是[clip_image004](http://images.cnblogs.com/cnblogs_com/jerrylead/201104/201104061601448982.png)，每个[clip_image006](http://images.cnblogs.com/cnblogs_com/jerrylead/201104/201104061601453159.png)，没有了y。

K-means算法是将样本聚类成k个簇（cluster），具体算法描述如下：

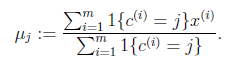
1、 随机选取k个聚类质心点（cluster centroids）为[clip_image008[6]](http://images.cnblogs.com/cnblogs_com/jerrylead/201104/201104061601454064.png)。

2、 重复下面过程直到收敛 {

               对于每一个样例i，计算其应该属于的类

[clip_image009](http://images.cnblogs.com/cnblogs_com/jerrylead/201104/201104061601464654.png)

               对于每一个类j，重新计算该类的质心

[](http://images.cnblogs.com/cnblogs_com/jerrylead/201104/201104061601468308.png)

}

K是我们事先给定的聚类数，[clip_image012[6]](http://images.cnblogs.com/cnblogs_com/jerrylead/201104/201104061601473390.png)代表样例i与k个类中距离最近的那个类，[clip_image012[7]](http://images.cnblogs.com/cnblogs_com/jerrylead/201104/201104061601481537.png)的值是1到k中的一个。质心[clip_image014[6]](http://images.cnblogs.com/cnblogs_com/jerrylead/201104/201104061601496064.png)代表我们对属于同一个类的样本中心点的猜测，拿星团模型来解释就是要将所有的星星聚成k个星团，首先随机选取k个宇宙中的点（或者k个星星）作为k个星团的质心，然后第一步对于每一个星星计算其到k个质心中每一个的距离，然后选取距离最近的那个星团作为[clip_image012[8]](http://images.cnblogs.com/cnblogs_com/jerrylead/201104/201104061601504211.png)，这样经过第一步每一个星星都有了所属的星团；第二步对于每一个星团，重新计算它的质心[clip_image014[7]](http://images.cnblogs.com/cnblogs_com/jerrylead/201104/201104061601515182.png)（对里面所有的星星坐标求平均）。重复迭代第一步和第二步直到质心不变或者变化很小。

KMEANS算法可以采用多种终止条件。

* 当迭代一个固定次数后停止。该条件能够限制聚类算法的运行时间，但是有些情况下，由于迭代次数不足，聚类结果的质量并不高。
* 当文档到簇的分配结果不再改变后停止。除了某些情况下会使算法陷入局部最优外，该停止条件同城会产生比较好的聚类结果，但是运行时间不宜太久。
* 当质心向量不再改变后停止。

**4.1.2 KMEANS的去背景化**

假设我们有初始化的背景词汇{c1,c2,c3….cn}，最基本的去背景化的做法就是将这些背景词汇都去除，这种做法最直接的，但是它的效果往往依赖于背景词汇{c1,c2,c3…cn}的信息的完整性。有时候由于背景词汇非常的稀少，这时候如果之去除这些背景知识仍然不能把两个句子区分开。下面我们可以举一个例子进行说明。假设有如下两个句子A和B。假设背景词汇包括{习近平}

A：习近平主席出访美国。

B：习近平主席出访德国。

这两个句子在原本上的相似度非常的高，使用一般的方法，我们将背景词汇去掉之后，他们的相似度会有所降低，但是相似度也可能非常的高，但是如果我们能够将背景词汇进行扩展，将与背景词汇的同义词或者同现概率很高的词汇也添加到背景词汇中，并且也进行删除，那么再重新计算它们的相似度，相似度的值也会相应的减小。

其实上面的例子是为了说明，首先相同的背景语料对于计算文本的相似度有一定的影响，如果去除相同的背景语料，可以更加准确的还原文本语料之间最真实的相似度，如果能够计算这些背景语料知识。并且，我们假设当背景词汇的信息量不够充分的时候，通过计算的手段收集并且扩充更多的背景语料知识，再重新去计算文本之间的相似度，能够得到更加合理的易于聚类的相似度。

本章的接下去的环节将讲述如何进行背景语料的扩展，并且将它应用到KMEANS算法中。

### 4.2 KMEANS去背景化的设计与实现

在4.1.2中我们讲了在KMEANS算法中如何使用去背景化的思想，其实在4.1.2中描述的方法适用于任何其他聚类算法，4.1.2中描述的方法是一种预处理的去背景化的思想。

#### 4.2.1 直接去背景词汇的策略实现

#### 4.2.2 背景词汇扩展策略实现

1 互信息

2 同义词词典

3 Word2vec的扩展策略

## 微信账号分析

微信公共账号是一种具有先验知识的账号，一个公共账号的先验知识是非常强烈的，很大一部分账号都是有自己发表文章的一个专业领域，而且这个专业领域将大部分时间内不随时间而改变。所以，本文认为学习文章的主题信息可以充分的利用账号的先验知识，帮助我们去认知文章的主题知识，并且当我们充分的利用了账号的主体信息之后，能够带来主体信息分类速度上的极大提升，这是一个非常诱人的信息点。

为了能够更好地学习账号的主题信息，我们需要对账号信息的先验知识进行比较准确地学习，我们将充分利用公共账号的历史文章信息，依据用户的历史数据，我们将对用户贴上一个标签，然后针对那些专业性很强的账号，系统将直接使用该公共账号的历史标签对文中进行打标签。

针对用户的历史数据，可以得出一批数据很强的用户，这些用户的历史数据的主题分布很集中，例如账号“一点易教育”它的历史数据基本上集中为教育，“生活可以更美一点”它的历史数据基本上集中为健康。我们从所有的账号中随机的采样了10000个账号，分析它们的历史主题分布比例，记录它们历史主题分布中概率最高的主题的概率值，我们这里称之为峰值主题概率。最后的统计结果如下：

|  |  |
| --- | --- |
| 概率范围 | 用户数 |
| 0.9~1.0 |  |
| 0.8~0.9 |  |
| 0.7~0.8 |  |

从上面的统计我们可以发现用？？%用户的峰值主题概率都能够达到0.6以上，有？？%的用户的峰值主题概率能够达到0.7以上，有？？%的用户的峰值主题概率能够达到0.8以上，有？？%的用户的峰值主题概率能够达到0.9以上。

上面的统计提供了足够的证据证明某些专业用户确实在不断地发送自己的专业领域内知识。此外，我们还统计了这部分专业领域比较强的用户的专业持续能力。我们从10000个用户中采样1000个用户进行持续性统计，将历史时间段进行切分，以一天或者10天或者一个月单位，分别统计他们在每个时间段内的峰值主题，然后统计峰值主题的分布情况，看一下是不是很多时候这些用户

也许你会问这么做有什么好处呢？

随着系统数据的增大，要实时的对文档进行主题监控，但是这时候要进行像上面的LDA聚类或者其他的主题模型算法，往往都需要进行很多次的迭代，。这时候我们这种利用用户专业性的做法就能够得到彰显其用处了，系统可以直接利用账户的信息建立类别体系，也可以成为主题信息体系。具体的展示可以参考图。

**添加系统展示图。**

本系统统计了所有用户的历史类别信息，对不同的专业形用户做了不同的处理，我们分别取峰值主题概率超过一定阈值的用户作为免处理用户，然后将他们发表的文章直接标记为峰值主题，得到的结果如下。

|  |  |  |
| --- | --- | --- |
| 峰值主题概率阈值 | 准确率 | 速率 |
|  |  |  |
|  |  |  |
|  |  |  |

从上面的表格中可以看到，