1. **绪论**
   1. 研究背景

近年来，互联网技术与应用高速发展。互联网已经成为人们获取信息和资源的最重要的场所。CNNIC统计数据表明：截至2014年6月，中国网民规模达6.32亿，较2013年底增加1442万人，互联网普及率为46.9%。互联网发展重心从“广泛”向“深入”转换，各项网络应用深刻改变网民生活。

然而在互联网中信息资源极大的丰富的同时，美国科学家A.L.Barabasi等人通过研究发现，网民访问互联网时具有明显的“群聚”现象，网民上网产生的大多数的流量都流向少数的几个网站，A.L.Barabasi等人提出了“无尺度(Scale Free)现象”这一概念来描述此现象。无尺度现象主要表现在互联网已经由服从泊松分布的随机网络演化成了服从幂律分布的无尺度网络。无尺度现象导致大量冗余信息在网络中重复传播，造成网络拥塞，形成网络性能瓶颈。于此同时，用户通过关键字检索等方式在传统搜索引擎查找信息时，往往其搜索行为缺乏明确目的性，从而无法及时高效获取真正符合自己需要和兴趣的内容信息，容易迷失在互联网的海量信息资源中。

为了应对当前互联网中信息资源的快速增长与过度冗余，李幼平院士等提出了通过网络实施“深度去冗”，在传统互联网主结构上增加以异步广播通信为基础的次结构，形成“双结构”的播存新型网络（以下简称“播存环境”），并通过“统一内容标签”（UCL，Uniform Content Label），为次结构添加交互性，为用户提供更具个性化的服务。播存环境如图1-1所示，其主要由广播服务器端B，边缘服务器S，用户终端C构成。播存环境的核心流程为：广播服务器端B通过网络爬虫定时爬取网络中热门资源和信息，对资源与信息进行UCL标引，产生UCL。通过广播方式向边缘服务器端S广播内容以及对应UCL，向用户终端C广播UCL。用户终端C通过用户历史行为以及自身配置对用户的兴趣进行建模，根据用户的兴趣模型对收到的UCL进行适配，向边缘服务器S请求感兴趣的内容。同时，边缘服务器S向用户终端C推荐用户可能感兴趣的UCL，用户收到UCL后决定是否向边缘服务器S请求内容全文。当用户请求的内容已经不存在于边缘服务器S时，边缘服务器S可以通过传统互联网方式向网络其他的服务器请求相应内容。



图1-1 播存环境基本结构

播存环境综合了传统互联网和广播网络的优势，并且更适用于当前终端设备存储能力和计算能力有限的移动互联网的情景。在缓解网络中冗余数据问题的同时，为用户提供更具个性化的内容服务。其中，UCL是播存环境的关键概念。由播存环境进行广播分发的内容数据都需要进行UCL标引。它是信息资源的“元数据”，目的是帮助理顺当今繁杂的无结构网络、解决网络中信息资源的发现，查找，识别和管理问题，沟通作者和读者的意向，以便于信息的理解和主动服务，实现从“大数据”到“有限需求”之间的直接去冗。

其中，“话题”字段是UCL重要组成部分。“话题”是对互联网网页内容的高层次提炼，它与“关键词”的主要区别在于：“关键词”是一篇网页文档内容与其他网页文档内容产生区别的重要词汇，关键词没有任何标准，它可以是任何与其他网页产生区别的词语。而“话题”则是从大量网页语料库中提取出的，经过一定规范化处理的“标准”。网页内容可以依据这个“标准”进行归类映射。“话题”可以看做“关键词”的进一步提炼和收束。一方面，广播服务器可以通过“话题”字段对信息内容进行聚类，将相关内容聚敛在一起。另一方面，用户可以通过“话题”字段订阅自己感兴趣的内容，享受更为个性化的服务。

由于当前网络中网页内容过于庞大，且缺乏统一的话题分类标准，同时，面对如此大量的资源采用人工标注的方法不切实际。由于当前网页内容具有很强的时效性，热门话题实时动态变化，所以通过一定的算法机制从采集的热门信息资源中实时动态高效地挖掘网页动态话题是播存环境重要的一环。

另一方面，从用户角度，通过用户的历史浏览阅读行为，对用户从话题角度进行收集、分析、建模以便向用户以便从大量的热门信息中，合理过滤适配出符合用户个性化需求的信息资源，将直接决定用户的满意程度。

话题挖掘、追踪和检测的研究领域目前已经有了一些成熟的模型，比如PLSA模型、LDA模型等等，但是对实时话题挖掘进行的研究比较少，很少有研究突出话题挖掘的时效性。而时效性在播存环境中则至关重要。个性化推荐领域一直是当前的热门研究领域，基于协同过滤的算法、基于内容的算法已经普遍应用在商业网站中。但不论是协同过滤算法还是基于内容的推荐算法，都有自身的缺点，难以应用到当前的播存环境中。

因此，本文希望借鉴传统的话题挖掘模型，并在利用播存环境特有的UCL信息的同时，添加时间因素的反馈，以达到对当前网络中热门网页信息资源中蕴含的话题进行实时挖掘的目的，形成动态话题映射表，在播存环境中进行分发，达到UCL自动生成和基于话题维度的用户个性化推荐的效果。

* 1. 研究目标及内容

本文希望在借鉴传统机器学习领域文本话题挖掘的基础上，设计一种满足播存环境需求，从海量热门新闻网页资源中实时挖掘流行话题的实时话题挖掘模型，通过此模型，实时地动态生成UCL的“话题”字段，形成符合当前网络中话题的分类映射表并向接入服务器进行分发。同时，通过用户端的历史行为记录，挖掘用户“话题”兴趣维度，建立用户兴趣模型，对用户进行个性化推荐。

本文的主要研究内容如下：

（1）网页实时话题挖掘模型

基于机器学习和人工智能领域传统算法，进行改进，克服传统算法中无法实时进行挖掘的缺点，设计一种从海量新闻文本资源中高效、合理地进行话题挖掘、发现新浮现话题的算法，形成反应当前互联网热门趋势的话题分类，形成当前网络话题的分类映射表进行分发。

（2）基于话题的个性化推荐技术

利用用户历史阅读行为、和接入服务器交互行为，根据服务器端分发的话题建立用户话题维度的兴趣模型。不同于传统的协同过滤方法，基于话题模型的个性化推荐无需利用其余用户的信息，计算效率高、响应时间短。并且通过挖掘内容的隐含“话题”，比传统的基于内容的个性化推荐算法更能精确反应用户的兴趣所在。

（3）播存环境中的应用

将上述实时网页实时话题挖掘模型以及基于话题的个性化推荐技术集成中当前的播存环境原型实验系统中。在原型系统中达到预计的网页实时挖掘的效果，形成话题的分类映射表。并且生成原型实验系统的用户的话题维度兴趣模型，进行主动推荐和适配。

* 1. 论文组织结构

本文共分6章，各章的主要内容如下：

第一章为绪论。主要介绍了本文的研究背景、研究意义、研究目标和、研究内容以及论文的组织结构

第二章为相关研究工作介绍。主要介绍了话题挖掘相关领域的工作、个性化推荐领域的相关工作，以及本文工作的研究基础。

第三章为播存环境话题挖掘模型研究与设计。首先分析了传统话题挖掘模型在播存环境中不适用的原因，然后详细介绍了本文设计的应用在播存环境中的话题挖掘模型。

第四章为播存环境中基于话题的个性化推荐。详细介绍了将话题挖掘运用在个性化推荐中的思路，介绍了算法的设计，并且比较了和传统基于用户和基于物品的协同过滤的不同。

第五章为实验部分。首先介绍了播存环境的原型系统的相关工作已经本文涉及的主要模块。然后介绍了第三章和第四章设计的模型在原型系统中的实验效果。

第六章总结了本文的工作并且展望了未来的工作。

1. **话题挖掘及个性化推荐相关研究**
   1. 相关研究技术概述

目前，对于播存环境的研究主要集中在其基本体系结构和UCL标引的定义。对于播存环境中广播服务器端的实时话题挖掘和性化推荐技术的研究较少。由于播存环境承载传输的主要是互联网中热门新闻网页资源等文本数据，所以播存环境中的信息实时话题挖掘对应的关键技术为文本话题实时挖掘以及文本分类。在传统的机器学习和人工智能领域，这些问题已经有了一定的研究。而传统互联网中个性化推荐技术已经得到了充分研究，其中基于协同过滤、基于内容、基于标签推荐等技术已经有了广泛的应用。

根据上一章分析的播存环境的特点并且结合本文的研究内容与目标，本文将对以下两个方面的相关研究工作进行详细分析。

1. 在大量网页信息资源中进行话题挖掘的算法模型。播存环境的广播源点将通过网络爬虫爬取大量的热门网页资源进行分发，回顾传统模型是如何进行话题挖掘和提取的，有利于找到播存环境的不同点，从而对传统模型进行改进。
2. 个性化推荐的相关技术基础。本文将提出一种基于话题的个性化推荐技术，回顾现有的个性化推荐技术的关键思想和优缺点，有利于分析基于话题的个性化推荐技术，判断其是否更加符合播存环境的情景。
   1. 话题挖掘技术相关研究
      1. Unigram模型

Unigram是最简单的文本建模方式。在Unigram模型中，假设一篇文章中的词全部来自一个词典，每次按照一定概率，从词典中选取词语形成文章。Unigram模型假设文章中的词语都是无顺序可交换的而且文档之间也是独立可交换的。所以这种模型也被称为词袋模型。Unigram模型的概率图模型如下：



Unigram生成文档的过程描述如下：

假设词典共有个词，分别为 ，其中选取词的概率为。则每次选取词的过程可以看做一次贝努利实验，服从多项式分布

文章中的每个词可以看做独立同分布的随机变量。则对于一篇文档，生成文档的概率如公式所示：



语料库中的所有文章也可以看做独立同分布，则形成语料库的概率如公式所示



其中，为第个词出现的次数。通过最大似然估计最大化可得参数的最大似然估计为：。

求得后，即完成了对模型的求解。

* + 1. PLSA模型

文献提出了一种基于概率统计技术的文本分析及话题挖掘算法：PLSA（Probabilistic latent semantic analysis）。PLSA引入了隐含话题的概念，克服了Unigram的一些缺点。其在自然语言处理和文本分析方面有着广泛的应用。其基本思路是：每篇文档有某些话题构成，而每个话题下又有许多属于此话题的词。因此每篇文档可以看做在若干话题上的多项分布，而每个话题又可看做是若干词语的多项分布。基于这个思路，PLSA的概率图模型如图2所示：



图2 PLSA概率图模型

其中， 表示单词出现在文档 中的概率， 表示文档中出现主题下的单词的概率，给定主题出现单词 的单词。每个文档在所有主题上服从多项分布，而每个主题在所有单词上也服从多项分布。所以，单词和文档的联合概率分布如公式（1）所示：

 （1）

得到此模型后，文献[6]中采用传统机器学习中的EM算法，对模型的参数进行估计。

EM算法分为两步：

1. E Step计算欲求的隐含变量的后验概率
2. M Step 通过最大化后验概率来更新参数。重新进行迭代直到两次迭代间的误差小于一定的阈值，即收敛

根据上述公式，在E Step中可以得到的后验概率为：



则M Step需要最大化的公式有：



根据上述公式，通过EM算法进行迭代，可以得到相应的模型参数，即：每篇文档的主题分布和每个话题下的词语的分布。得到模型参数后，即有了对整个语料库的在话题维度的估计。当一篇新出现的文档时，分析此文档中出现的词语，可以根据计算得到的模型参数对新出现的文档进行话题分布估计。

* + 1. LDA模型

虽然PLSA在文本挖掘等领域的应用取得了一些成功，但其仍然面临一些问题，比如每个文档在所有主题上服从多项分布的参数和每个主题在所有单词上服从多项分布的参数，在PLSA模型中被视为固定的参数，这样忽略了这些参数的先验分布。为了克服PLSA模型这个缺点，文献[7]提出了LDA（Latent Dirichlet allocation），目前在文本挖掘领域包括文本主题识别、文本分类以及文本相似度计算方面都有应用。

与PLSA模型忽略先验概率不同，LDA算法引入了Dirichlet分布作为先验分布。将Dirichlet分布作为先验分布引入的原因是，Dirichlet分布和多项式分布共轭，使得计算后验分布时仍然符合Dirichlet分布的形式，便于计算。Dirichlet和多项式分布共轭的形式化描述如公式所示：



LDA的概率图模型如图3所示：



图3 LDA模型概率图模型

在LDA模型看来，一篇文档的生成过程如下，它和PLSA一样，仍然属于词袋模型：

* + - 1. 从狄利克雷分布中取样生成文档的主题分布
      2. 从主题的多项式分布中取样生成文档第个词的主题
      3. 从狄利克雷分布中取样生成第个主题的词语分布
      4. 从词语的多项式分布中采样最终生成词语

因此整个模型的所有可见变量和隐藏变量的联合概率分布如公式(2)所示：

 （2）

文献[7]中采用机器学习算法中传统的EM算法中对模型进行求解，得到模型参数。文献[8]采用在综合分析了LDA模型的基础上，设计出了更为简单和易于理解的吉布斯采样(Gibbs Sampling)算法对LDA进行了求解。

吉布斯采样是基于马尔科夫链的蒙特卡洛方法的一个特例。在LDA中运用吉布斯采用的基本过程如图所示：



图 吉布斯采样基本过程

在吉布斯采样开始时，给语料库中的每个词随机分配主题，然后统计每个主题下出现词语和每篇文档中出现的属于主题的词语的数量。在一次迭代中，根据吉布斯采样规则，计算概率，它的含义为：在排除当前的词所属主题的前提下，根据语料库中所有文档的词语的所属的主题，则当前词属于各个主题的概率。在得到这个概率分配之后，可以进行采样，给当前词分配新的主题。以此类推，不断地进行迭代。直到每个文档的主题分布和每个话题下的词语的分布收敛时，迭代停止。则每个档次的主题 也可以得到。

文献[9]提出了一种快速Gibbs采样方法对LDA模型进行求解，文献[10]分析了LDA模型在标签推荐中的应用。文献[11]提出了一种应用于大规模数据分析的并行LDA模型。

* 1. 个性化推荐技术相关研究

个性化推荐技术在最近几年是互联网技术研究中的重点。个性化推荐的形式化表示如所示：



其中 C 是所有用户的集合，S 是待推荐的项目集合，效用函数 u 计算项目 s 对用户 c 的兴趣适配度，即 u: C×S→R，R 是一定范围内全序的非负实数，个性化推荐要研究的问题就是找到适配度 R 最大的那些信息 s\*

个性化推荐的基本思路大致包括协同过滤、基于内容的推荐、以及近几年在社交网络中应用广泛的个性化标签推荐技术。

* + 1. 协同过滤推荐技术（对冷启动、稀疏问题详细）

协同过滤策略的基本思路是：不用物品本身的内容来衡量物品之间的相似度，而是通过同时喜欢两个物品的用户数来衡量物品之间的相似度。因为对很多物品来说，抽取物品特征，建立物品特征向量非常困难。通过协同过滤的方式，可以更简单地衡量物品之间的相似度。从而向用户推荐与用户购买过或者喜爱的物品相似度最大的物品。协同过滤最为常用的用户模型为用户-项目评价矩阵，用户-项目评价矩阵如表所示。其中表示用户 对物品 的评分,则协同过滤可以看做利用已知评分，来计算用户相似度，最后得出未知评分的过程。

一般根据余弦相似性、相关相似性、修正余弦函数相似性等方法来计算用户相似度。

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| 物品  用户 | 项目1 | 项目2 | …… | 项目k | 项目n |
| 用户1 |  |  |  |  |  |
| 用户2 |  |  |  |  |  |
| 用户K |  |  |  |  |  |
| 用户N |  |  |  |  |  |

因此，协同过滤策略的优点通过用户自身的历史数据和其他用户的历史数据来发现用户可能感兴趣的新的内容，而不必对比如音乐，视频等非结构化数据进行内容特征提取，协同过滤的主要有两个不易解决的缺点和问题

1. 冷启动问题：这是由于新加入的物品或者信息由于没有被用户浏览或者反馈喜欢过，系统对它无法进行推荐，而新注册的用户由于没有浏览或者购买记录，无法计算相似用户，因此也难以为他推荐合适的资源信息。
2. 稀疏性问题：当系统中用户不活跃，对物品的评分较少，导致评价矩阵过于稀疏，从而使得推荐引擎缺少足够多的信息进行推荐。导致推荐结果较差。

针对这两种问题，协同过滤经常采用取长补短的方法来进行解决。和其他策略相结合，比如和基于内容的推广方法相结合来缓解这两种问题。还可以采取降维的方法，提高数据的密度以达到提高相似度计算精度的目的。文献[12]介绍了冷启动问题并且介绍了一种解决方案。文献[13]介绍了协同过滤算法在亚马逊网站中的运用。文献[14]介绍了一种与基于用户的协同过滤算法不同的基于物品的协同过滤算法。

* + 1. 基于内容的推荐技术

基于内容的推荐技术的基本思路是抽取用户的行为特征，建立起用户的特征向量。对于欲推荐的物品，也建立其特征向量。通过计算两个特征向量的距离，来衡量用户和此物品之间的相似度，以决定是否向用户推荐此物品。比较常见的计算距离和相似度的方法有余弦相似度、改进余弦相似度、雅克比距离、相对熵等。基于内容的个性化推荐基本过程如图所示。

文献[15]介绍了基于内容的推荐的系统的基本思路和方法。文献[16]介绍了基于内容的推荐技术在书籍推荐和文本分类方面的运用。文献[17]介绍了基于内容的推荐技术在音频推荐中的应用。

 基于内容的个性化推荐的关键步骤主要有2点：用户兴趣特征提取技术和项目特征向量提取。

1. 用户兴趣特征提取。通过对用户的历史行为的收集和分析，提取出用户的兴趣特征。收集用户历史行为的方法主要有显式收集对用户在网站上对物品的评分、用户的自然属性（性别、年龄、学历），隐式收集用户的浏览行为。在收集用户行为的同时如何确保用户的隐私不被侵犯也是当前的研究热点之一。在收集到用户的行为后，提取用户兴趣特征的方法主要有：贝叶斯分类器、决策树方法、聚类算法和用户-评分矩阵方法。
2. 项目特征向量提取。对于结构化的文本或者新闻网页进行特征提取较简单，通常可以根据词语来进行处理。基本方法有基于词频的tf-idf方法、关键词提取方法、话题提取方法。而对于非结构化的信息比如音频视频来说，进行特征抽取非常困难，表现往往较差。通常只能利用其相关的背景信息比如作者、歌手和网站对其进行的分类来进行处理。

基于内容的推荐技术，其优点在于运算比较高效，运算量比协同过滤要小且不存在协同过滤中的冷启动问题。但这种推荐技术只考虑用户的个体性没有考虑用户行为的群体性，单单根据用户历史行为进行推荐，无法发现用户的新兴趣，无法带给用户惊喜

* + 1. 个性化标签推荐技术

标签和标签系统作为一种特殊的元数据，在当今网络的内容整合、管理、表征用户兴趣等方面起着越来越重要的作用。如今采用标签系统的社交网站有：照片分享网站Flickr,书签分享网站del.icio.us,音乐爱好分享网站last.fm。在这些网站中，标签系统使得用户可以创建和管理标签以对内容资源进行标注和分类。用户不仅可以对自己的内容资源标注标签，也可以对其他用户的资源内容进行标注[18]。

标签系统的主要功能有两点：

1. 标签帮助用户组织和管理自己的内容。

2. 标签帮助用户寻找由其他用户分享的相关内容。

由于UCL带有很强的标签属性，所以个性化标签推荐技术对于利用UCL属性有很强的借鉴意义。

标签系统的核心过程为[19]：一旦用户进行登录，他就可以向系统上传资源并对资源进行标注，打上标签。他个人的标签行为集合就成为个人标签集合。而所有个人标签集合就形成了社会化标签集合。一个用户可以对个人标签集合和社会标签集合进行多维度的探索：用户可以浏览他自己上传的资源以及他自己赋予资源的标签。他也可以浏览其他用户所上传的资源并且了解其他用户是如何对这个资源赋予标签的。当点击标签时，他可以知道是谁赋予的资源这个标签。通过这些标签，用户可以寻找和发现自己或者其他用户的自己感兴趣或者相关的资源。随着标签标注的进行，整个系统成为一个有相关用户，相关资源和相关标签构成的复杂网络。系统将有着相同兴趣和爱好的用户连接在一起，随着用户对资源的标签标注，系统可以追踪用户的兴趣特征，这是数据挖掘算法的强有力工具。

基于以上过程，个性化标签推荐系统索要解决的问题是帮助和支持用户对资源进行标签标注，并向用户推荐系统认为合适的标签社会化标签集合建模成一个元组[20]：,其中 是有穷集合，分别代表用户集合、标签集合、资源集合。，是用户、标签、资源组成的三元关系，代表数据集中所有的标注行为。定义，即代表用户对资源所标注的所有标签。

则个性化标签的推荐系统所完成的任务形式化为：对一个指定的用户和一个指定的资源，且同时，推荐一个标签集合，多数情况下先依据一定的标准和相关性产生候选集，选取候选集中按标准评分前几项，产生推荐集。

* + 1. 个性化推荐技术评价指标

文献[21][22][23]对当前个性化推荐技术涉及的评价指标进行了综述。其中的主要的指标有准确度、覆盖率以及多样性。

准确度是最能直观反映推荐系统性能的一个指标，所以它也是推荐系统中使用最多且最看重的一个指标。衡量准确度最典型的指标是平均绝对误差（Mean Absolute Error，MAE）和平均平方误差（Mean Squared Error，MSE）以及标准平均误差（Normalized Mean Absolute Error，NMAE）。它们的计算形式如式（1）和（2）以及（3）所示。

 （3）

 （4）

 （5）

其中为系统中用户打分信息条目的个数，和分别为预测打分和实际打分。 为系统中用户-信息对的个数。和分别为用户打分区间的最小值和最大值。

在常见的TOP-N情景中，推荐系统需要根据用户的历史行为，向每一个用户推荐N个推荐系统评分最高的项目。所以在TOP-N问题情景中，常用公式衡量准确率和召回率



其中表示对用户 返回的推荐项目的集合。是在测试集中用户真正满意的项目的集合。

覆盖率是指算法向用户推荐的商品能覆盖全部商品的比例，如果一个推荐系统的覆盖率比较低，那么这个系统很可能会由于其推荐范围的局限性而降低用户的满意度，因为低的覆盖率意味着用户可选择的商品很少。覆盖率尤其适用于那些需要为用户找出所有感兴趣的商品的系统。其常用的计算形式如4所示：

 （6）

表示所有用户推荐列表中出现过的不相同的商品的个数。 表示所有物品的总数。推荐覆盖率越高，系统给用户推荐的商品种类就越多，推荐多样新颖的可能性就越大。如果一个推荐算法总是推荐给用户流行的商品，那么它的覆盖率往往很低，通常也是多样性和新颖性都很低的推荐。

此外，对个性化推荐系统的评价指标还有：新颖性和多样性。通常，在衡量一个推荐系统性能时，需要根据面对的问题情景的不同、对个性化推荐算法的要求不同选择不同的评价指标进行评判。

* 1. 本章小结

本章首先介绍了现有的传统的话题模型：Unigram、PLSA和LDA模型，分析了它们的理论基础和优缺点，介绍了这些模型的不同的求解方法：最大似然估计法、EM算法以及吉布斯采样方法。。然后介绍了当前应用较广泛的个性化推荐技术：协同过滤个性化推荐、基于内容的个性化推荐技术以及当前在社交网站中非常流行的个性化标签推荐技术，分析了它们的关键思想和核心技术，以及优缺点。最后介绍了个性化推荐的相关评价指标。

1. **播存环境话题挖掘模型研究与设计**
2. 主要问题与研究思路

播存环境的体系结构与传统的互联网体系结构有很大的不同，这因为播存环境这些新的特点造成播存环境面临许多新的问题情景，从而对话题挖掘模型有许多新的需求。下文将首先深入分析因为播存环境自身结构特点而产生的需求。

1. 播存环境实时从互联网中采集全新的热点内容。传统的话题挖掘模型，往往采取离线计算的方式，针对特定的离线语料库进行离线计算，从而进行话题挖掘。而播存环境对实时性要求高、不断爬取新内容的这一特点，要求话题挖掘必须是在线的、流式的，针对不断进入播存环境的新内容可以持续话题挖掘，而不能采取传统话题挖掘模型的离线计算的方式。进行在线的、流式计算的一个关键点就是如何在播存环境的话题挖掘中添加时间这一维度，从而利用时间反馈，使得话题挖掘可以反映互联网中话题的演进。
2. 播存环境中存在UCL这一元数据。传统的话题挖掘模型单纯直接地从语料库中挖掘话题，没有任何元数据来帮助理解语料库中的内容，在进行话题挖掘之前，对于语料库中的内容的分类等信息毫无了解。而播存环境本身对于每一条内容都有与之对应的UCL来提供相关的元数据。因此，播存环境中的话题挖掘有必要利用好UCL这一元数据来提高话题挖掘的准确度。
3. 播存环境中的话题挖掘需要检测到互联网中新出现的话题。由于传统的话题挖掘模型针对离线的语料库进行话题挖掘，是一个一次性的过程，没有不断捕捉和检测新话题的需要。而根据上文的分析，播存环境中的话题挖掘是一个在线的、流式的模型，因而需要在挖掘的过程中检测到互联网中新出现的话题并且呈现出来，即不断更新播存环境中的“话题表”，使其始终能反映当前互联网中最新的话题。
4. 播存环境需要将话题分布映射到UCL的话题字段。UCL中话题字段需要通过话题挖掘模型来填充。而传统的话题挖掘模型产生的最终结果是关于正文对于各个话题的一个概率分布。这种形式的结果无法填充到UCL定长的话题字段中。所以播存环境中的话题挖掘模型必须在形成正文对于话题的概率分布的同时，采取一种机制将这种概率分布映射到UCL定长的话题字段。

针对以上播存环境的特点和需求，经过深入的分析，分别提出了以下的研究方案

1. 传统话题挖掘模型超参数为一个预设值或者固定值，每次挖掘后不会进行更新，从而无法体现时间这一因素。因此可以利用贝叶斯理论的思想，将其改造成流式的模型，根据时间对超参数进行迭代更新。使其能更好的反映当前的话题现状。
2. 传统的话题挖掘模型超参数起始值的设定往往是根据经验或者实验结果来设置，而播存环境中由于有UCL这一元数据，提供了更多的先验信息，因此播存环境中的话题挖掘可以利用UCL中的部分信息来更好的提供超参数的起始值。
3. 播存环境中的话题挖掘模型引入了“话题距离”的概念来衡量一个话题的演进程度，当超过一定的阈值后即可判断这一话题已经变成一个新的话题，从而可以更新话题表，检测到新的话题。
4. 针对从概率分布映射成一个固定值的问题，本文设计了一种综合考虑了概率和词语权重的方法，将播存环境中话题挖掘形成的正文的话题概率分布，映射成其对应的UCL的固定长度字段的值。
5. 播存环境话题挖掘模型框架总览

图为播存环境中话题挖掘模型框架的总览。话题挖掘模块需要和热门信息采集模块、UCL标引模块以及广播分发模块进行交互。



热门信息采集模块从互联网中采集当前热门的信息存储到信息存储的数据库中。话题挖掘模块连续不断地从存储信息的数据库中获得当前的网页内容。话题挖掘模块主要分为两个子模块。

网页预处理模块主要对网页进行相关处理。由于话题挖掘模型面对的是正文文本内容，而前端采集到的是html等撰写的网页。所以需要对网页进行清洗，去除html相关的标签，去除噪音提取出正文。在英文文章中，单词直接可以利用空格天然进行分词以便进行话题提取，分词较为简单。而对于中文文章，则需要通过一定的算法对正文进行分句以及分词，才能进行话题提取。所以在网页预处理模块中设计一个网页分句与分词模块来完成这一工作。同时，当分词后，网页中往往会含有大量的没有任何信息量的词语，比如：的、是的、好的、因为、而且等。这些词语会对话题挖掘造成干扰。所以在网页预处理模块中设计网页停用词去除模块，去除网页中的没有任何信息量的词，进一步提高话题挖掘的准确度。

挖掘算法模块主要对经过预处理后，已经提取出正文的网页进行话题挖掘。在线LDA模块主要按照上文中的分析对网页进行话题挖掘并且更新相关参数。和传统的离线式计算不同，在线LDA模块，按照一定的时间间隔对当前批次的内容进行话题挖掘后迭代更新超参数，当获取下一批次的内容后，用更新后的超参数进行计算。新话题检测模块，主要用来衡量每次话题挖掘后各话题“距离”的变化，当距离超过一定的阈值之后，则判断其为新话题并更新话题表。

话题挖掘模块同时要和UCL标引模块交互作用。当进行话题挖掘时，UCL的部分字段比如大类、栏目等信息已经填充，这些先验信息可以用来帮助话题挖掘模型计算更加准确的超参数的预设值。而此时，UCL的话题字段仍然等待话题挖掘模型的填充。当话题挖掘模型运行结束后，进行UCL话题字段的填充。当正文以及其相应的UCL、话题表形成后，就可以进行广播分发。分到到边缘服务器和用户终端。

下面各小节将对各个模块进行详细的介绍和阐述。

1. 网页预处理模块

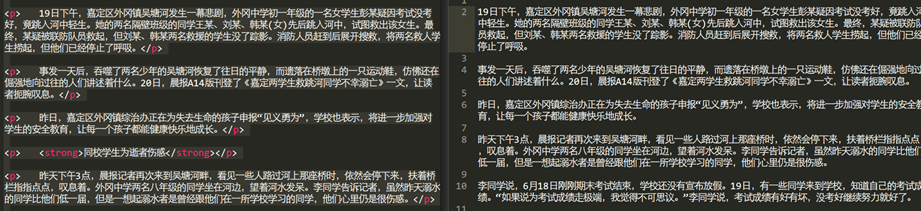
根据前文所述，网页预处理模块的功能是对网页进行清洗，提取网页正文后，对网页进行分句分词，去停用词后，产生便于话题挖掘模型进行处理的内容。由于这部分不是本文研究的重点，所以这部分大多数采用较为成熟的开源实现。

1. 网页清洗模块

网页清洗也被称为正文提取。其作用是从html编写的网页中去除相关html、css、javascript的标签，提取出真正有价值的网页正文。目前网页清洗和正文提取采用的主要算法有：

1. 基于标签用途和固定格式的正文提取算法。根据标签的固定作用和含义比如title,h1,h2等来提取正文。由于互联网中大多数网页都不规范且各个网站规范也不一致。一般需要针对一个特点的信息源来撰写规则。这一工作量较大。
2. 基于标签密度判定的正文提取算法。由于正文部分html标签的密度较低，则可以确定一个阈值，当标签密度小于此阈值时，可以认为此段文字为正文。这种方法实现起来较为简单
3. 基于统计的网页正文提取方法。这种方法将网页先规范化，然后转换成基于XML表示的DOM树形式，再统计相应节点的信息，利用这些信息将DOM树中含有噪声的节点去除，只保留相应的正文节点。

由于网页清洗和正文提取不是本文的研究重点。因此本文直接使用的是开源的网页清洗库Goose，其采用的为第(3)种算法，在各个国家的语言的网页上都有较好的性能。图展示了网页清洗的结果。



1. 网页分句与分词模块

由网页清洗模块得到正文内容后，还需要对正文进行进一步的处理。由于英语的各个单词之间直接有空格作为分隔，所以不需要特定的分词算法就可以得到各个单词然后进行话题挖掘。而对于中文来说，各个词汇直接没有明显的分隔界限，无法直观的进行词语间的划分，所以需要利用专门的算法来进行分词。

与分词相比，本文的分句采用的方法较为粗糙，直接根据正文中的句号、问好等标点，利用其直观的语义进行分句。从实验结果来看，这种分句方法对于网页正文已经有良好的效果，满足了基本的需要。

当前，分词技术主要有以下三种方法。

1. 基于字典和词库的分词方法，这种方法利用已有的字典和词库，来匹配当前的正文，以寻找和当前正文的最大匹配。这种方法较为直观，较容易实现。但由于汉语词汇较多，很多时候字典和词库不能满足当前的需求。
2. 基于词频统计的方法。它首先切分出与词表匹配的所有可能的词,运用统计语言模型和决策算法决定最优的切分结果。它不需要利用庞大的字典和词库，且可以将新词提取出来。
3. 利用马尔科夫链和动态规划的思想，根据概率统计进行分词。其效果优于前两者，摆脱了对于大型词库和字典的依赖，计算量较低，目前是分词技术的主流算法。

由于分词算法不是本文研究的重点，所以本文使用了开源的中文分词库jieba作为本文的分词实现。Jieba分词库基于前缀词典实现高效的词图扫描，生成句子中汉字所有可能成词情况所构成的有向无环图 (DAG)，采用了动态规划查找最大概率路径, 找出基于词频的最大切分组合，对于未登录词，采用了基于汉字成词能力的 HMM 模型，使用了 Viterbi 算法。

1. 去停用词模块

停用词是指为了提高搜索效率和节省存储空间，搜索引擎在处理搜索请求和索引页面时会自动忽略的某些字和词。通常停用词可以分为两种(1)使用十分广泛的词。如“我”、“就”等词 (2)信息量较小的词。比如一些连词和副词：“和”、“接着”、“然后”。

本文采用的去停用词方法是，搜集停用词表，将正文分词后的每个词与停用词表进行匹配，在停用词的表中的词就去除，不在的词语就继续保留。

通过去除这些停用词可以使组成正文的词语更有信息量，使得提取出的话题更有特殊性，而不必收到没有信息量的词语的影响。同时，可以大大减少正文中的词语数量，显著降低话题挖掘的计算量。

1. 播存环境话题挖掘算法模块