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，并提出了“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，用户通过传统搜索引擎查找信息时，往往缺乏明确目的性，无法及时高效获取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。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广播内容以及对应UCL，向用户终端广播UCL。用户终端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，向边缘服务器请求感兴趣的内容。同时，边缘服务器向用户终端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时，边缘服务器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。由播存系统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，以便于信息的理解和主动服务。

“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的重要和与其他网页文档内容产生区别的词汇。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，应用在播存环境系统中进行分发，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。