

分 类 号 TP391

密级

# 基于用户模型的个性化广告推荐技术研究

研 究 生 姓 名： 朱道杰

指导教师姓名、职称： 朱艳辉 教授

学 科 专 业： 计算机技术

研 究 方 向： 智能信息处理

湖 南 工 业 大 学

二〇一七年六月五日

分类号 TP391

密级                     


基于用户模型的个性化广告推荐技术研究  
Research on Personalized Advertising  
Recommendation Based on User Model

研究生姓名： 朱道杰

指导教师姓名、职称： 朱艳辉 教授

学 科 专 业： 计算机技术

研 究 方 向： 智能信息处理

论文答辩日期 2017.6.6 答辩委员会主席 

湖 南 工 业 大 学

二〇一七年三月三十日

## 湖南工业大学学位论文原创性声明

本人郑重声明：所呈交的学位论文，是本人在导师的指导下，独立进行研究工作所取得的成果。除文中已经注明引用的内容外，本论文不含任何其他个人或集体已经发表或撰写过的作品成果。对本文的研究做出重要贡献的个人和集体，均已在文中以明确方式标明。本人完全意识到本声明的法律结果由本人承担。

作者签名：朱道杰 日期：2017年06月12日

## 湖南工业大学论文版权使用授权书

本人了解湖南工业大学有关保留、使用学位论文的规定，即：学校有权保留学位论文，允许学位论文被查阅和借阅；学校可以公布学位论文的全部或部分内容，可以采用复印、缩印或其他手段保存学位论文；学校可根据国家或湖南省有关部门规定送交学位论文。

作者签名：朱道杰 导师签名：朱松林 日期：2017年06月12日

## 摘 要

随着互联网的快速普及和智能携带产品的推广,越来越多的人投入到互联网中,互联网成为现代广告业务的新载体。互联网广告投放具有效益高、覆盖面广等特点,因此受到前所未有的关注。传统的互联网广告投放充满着随机性和不确定性,导致用户在上网的时候面临着铺天盖地的广告,降低了用户的上网体验,造成了广告位的转化率低等问题。为了提升用户体验,研究者提出了个性化广告推荐技术,并成为了近年来的研究热点。个性化广告推荐技术的核心是用户模型,个性化广告服务的质量与用户模型的精确性有直接相关性。本文提出了基于显隐式信息结合的用户模型的个性化广告推荐技术。开展了如下研究工作:

(1)提出一种改进的隐式建模方法。传统的隐式建模方法将用户所有的日志信息作为建模信息。改进的隐式建模方法对用户上网日志进行分析,将查询词与历史文档进行相似性对比,过滤掉相似性值较小的文档,提高用户兴趣模型精确度。

(2)提出了显隐式信息相结合的用户建模技术。显隐式信息结合的用户建模首先是根据用户提交的用户信息初始化用户模型,然后通过对用户上网历史信息进行分析,构建隐式用户模型,对初始用户模型进行更新。

(3)提出了一种基于用户模型的协同过滤广告推荐算法。协同过滤技术是根据用户-项目评分矩阵,挖掘出用户相似集合,通过近邻集合中兴趣相近的用户给目标用户推荐信息。本文将上面提出的用户兴趣模型应用到协同过滤算法中,利用用户模型矩阵替代评分矩阵,实验结果表明基于用户模型的协同过滤推荐算法能够提升广告推荐的精确度。

**关键词:** 互联网, 用户兴趣模型, 个性化广告推荐

## ABSTRACT

With the rapid popularization of Internet and the promotion of smart carrying products, more and more people are investing in the Internet, and the Internet has become a new carrier of modern advertising business. Internet advertising has high efficiency, wide coverage and so on, so it has received unprecedented attention. The traditional Internet advertising is full of randomness and uncertainty, lead users faced with overwhelming advertising in the Internet, reducing the user's online experience, resulting in low conversion rate of advertising. In order to improve the user experience, researchers have proposed personalized advertising recommendation technology, and become the research focus in recent years. The core of personalized advertising recommendation technology is the user model, and the quality of personalized advertising service is directly related to the accuracy of the user model. This paper proposes a personalized advertisement recommendation technique based on explicit implicit information model. The following research work has been carried out:

(1)An improved implicit modeling method is proposed. The traditional modeling method of implicit user will log all information as the modeling information, but the user's interest is the real time change of the implicit improved modeling method adopting the method of historical information classification, only related to the query history information as the user modeling document set.

(2)A user modeling technique combining explicit and implicit information is proposed. Explicit user modeling with implicit information according to the user information to initialize the user model submitted by the user, and then through the Internet history information for users is analyzed, the construction of implicit user model, the initial user model update.

(3)A collaborative filtering advertising recommendation algorithm based on user model is proposed. Collaborative filtering technology is based on the user project scoring matrix to mine the user similar set, and recommend the information to the target user through the users of similar interest in the

nearest neighbor set. This paper will use the user interest model presented above to collaborative filtering algorithm, using the user model matrix instead of score matrix, the experimental results show that collaborative filtering algorithm can improve the accuracy of user model based on advertising recommendation.

Key Words: Internet; User Interest Model; Personalization; Personalized ad recommendations

# 目 录

摘 要.....	I
ABSTRACT.....	II
第一章 绪论.....	1
1.1 研究背景.....	1
1.2 研究目的及意义.....	1
1.3 国内外研究现状.....	2
1.3.1 用户兴趣模型的研究现状.....	2
1.3.2 个性化广告推荐技术的研究现状.....	3
1.4 论文主要研究内容及创新点.....	3
1.5 本文组织结构.....	4
第二章 相关研究.....	5
2.1 用户兴趣建模流程.....	5
2.2 用户兴趣建模技术分析.....	5
2.2.1 用户采集信息来源.....	6
2.2.2 用户信息获取方式.....	7
2.2.3 用户模型的数据结构.....	7
2.2.4 模型构建的技术分类.....	9
2.3 推荐技术.....	10
2.3.1 个性化推荐系统原理.....	11
2.3.2 推荐技术的常用方法.....	12
2.3.3 广告推荐技术.....	14
2.4 本章小结.....	15
第三章 基于显隐式信息结合的用户兴趣模型构建.....	16
3.1 基于显式信息的用户兴趣建模.....	16
3.2 改进的隐式建模方法.....	19
3.2.1 传统隐式信息建模方法.....	20
3.2.2 改进的隐式建模方法.....	22
3.3 显隐式信息结合的用户兴趣建模方法.....	24
3.3.1 显隐式信息结合的兴趣建模框架与流程.....	25
3.3.2 显隐式结合的用户兴趣模型建立.....	27
3.4 实验结果与分析.....	28

3.4.1 实验目的.....	28
3.4.2 实验结果分析.....	30
3.5 本章小结.....	33
第四章 基于用户模型的协同过滤个性化广告推荐技术.....	34
4.1 协同过滤推荐技术.....	34
4.2 基于用户模型的协同过滤个性化广告推荐技术.....	37
4.2.1 个性化广告推荐系统框架设计.....	37
4.2.2 基于用户兴趣模型的个性化广告推荐算法.....	38
4.3 实验设计与分析.....	39
4.3.1 MovieLens 数据集介绍.....	40
4.3.2 数据集的划分.....	41
4.3.3 评价方法.....	41
4.3.5 实验结果分析.....	42
4.4 本章小结.....	43
第五章 总结与展望.....	44
5.1 总结.....	44
5.2 展望.....	44
参考文献.....	46
攻读学位期间主要的研究成果.....	50
致    谢.....	51



## 第一章 绪论

### 1.1 研究背景

随着中国经济实力提升,互联网技术有了跨越式的发展,以阿里巴巴为代表的电子商务系统不仅提升了自身的业务发展,并且带动了很多其它行业发展,例如物流行业、移动支付、互联网广告等行业。尤其是广告行业<sup>[1]</sup>,随着电子商务的发展,广告需求量呈指数式增长。互联网媒体广告相对于传统媒体广告有以下优点,(1)投放快捷,(2)范围广泛,(3)目标具体,(4)效果明显,(5)成本低廉,导致各种各样的广告充斥着互联网,比如,一个不爱游戏的人打开网页却发现网页推送各种各样的游戏广告<sup>[2]</sup>。这种情况不仅造成资源浪费,同时也降低了网站的吸引力。刚刚兴起的互联网广告投放毫无逻辑可言,给用户投放的广告是大众化的,无差异性可言。

在网络快速发展的今天,互联网用户的角色不再是被动接受者,早已转变成为了信息的制造者、传播者。用户对信息有了多种选择,对那些影响用户体验或则与自己兴趣毫无相关的产品推荐丢入到“黑名单”中。只有真正了解用户兴趣需求,才能正确为用户推荐所需要的信息,才能在新媒体营销中立于不败之地。能够挖掘分析出用户的真实兴趣,成为了当下研究主要方向,为基于个性化广告推荐实现提供了参考。只有挖掘出用户兴趣爱好,基于挖掘出的结果并针对性的对用户推荐不同的广告,才能获得更好广告价值。

### 1.2 研究目的及意义

目前,互联网的世界里充满了各种各样的广告,广告运营商之间展开激烈的竞争。随着 Web2.0 的到来,电子商务快速发展,为广告主进行产品展示提供了广阔空间,能够使产品在短时间内被推广到世界各地,被用户所了解。网络用户能快速找到自己感兴趣的信息。这是传统广告无法比拟的。

通过分析用户上网日志,用一种算法或则是模型,表示出用户的喜好,根据不同用户的喜好,向用户推广不同的广告,将是研究者努力奋斗的目标。这样既可以增加广告的点击率,又可以降低运营成本。基于用户个性化的需求,谁的推荐结果与用户的兴趣最为匹配,谁就会获得更高的点击率,获得广告运营市场<sup>[3]</sup>。互联网广告推荐系统具有以下优势<sup>[4]</sup>:

(1) 投放简单快捷,无需传统广告一样排版、印刷,只需要简单的代码就可以了,在数小时内就可以遍布全球。

(2) 投放的范围广。与传统广告相比,只要有网络的地方,就可以接收到互联网广告。所有的网民都能够成为互联网广告的收益人群。互联网广告大大增加了广告的覆盖率。

(3) 投放目标更具体。系统收集用户上网信息,分析用户兴趣爱好,针对不同的用户兴趣爱好投放不同的广告。增加广告点击率,增加广告商的收益。

(4) 效益明显。广告能够在较短时间推荐给用户,同时,用户也能在在最短时间内得到广告信息,并针对自己的兴趣爱好选择自己感兴趣的广告。广告商也能快速的获益。

总体上,基于用户兴趣的个性化广告能够使互联网用户快速找到自己所需要的信息,同时也能够增加广告商的收益。

### 1.3 国内外研究现状

#### 1.3.1 用户兴趣模型的研究现状

随着用户个性化需求的增加,用户兴趣模型的研究取得长足进步,研究人员获得大量的研究成果,如有的用户模型基于标签库,有的用户模型根据用户上网行为,还有就是将向量空间模型运用到用户兴趣模型建立中。

目前的研究大部分是对用户浏览过的网页信息进行分析挖掘,没有从用户的兴趣主题这一角度进行深入地、系统地研究。

Ma 等<sup>[5]</sup>提出在进行用户兴趣建模时,不能从单一的数据源对用户进行兴趣挖掘。单一的数据源具有局限性,数据信息不能全面体现用户兴趣,比如,酒仙网的数据源只能分析出用户喜欢的酒,不能分析出用户喜欢的衣服类型。多数据源的信息相融合,通过互斥理论,用户的兴趣集合更加全面。Wenger 等<sup>[6]</sup>提出了基于标签的用户兴趣模型。Wu 等提出了一种基于微博关键词的用户兴趣模型<sup>[7]</sup>,通过分析用户的微博数据,提取微博中的关键词表示用户喜好。Huang He<sup>[8]</sup>提出示例文档用户兴趣建模技术。Pazzani 等<sup>[9]</sup>通过对浏览页面的标注获取用户兴趣模型。

总的来说,国外在用户兴趣模型领域研究起步比较早,拥有很多研究成果。但是在数据获取、模型表示统一化和评价方面存在不足之处。

国内,林鸿飞等<sup>[10]</sup>研究了示例文档建模技术。田萱<sup>[11]</sup>论述基于向量的用户模型构建。王平等<sup>[12]</sup>提出基于 RSS 信息源的用户模型,该方法针对 RSS 标准新闻源,根据用户浏览信息,通过文本聚类创建用户模型。

总的来说,用户兴趣模型研究在国内是一个新兴的领域<sup>[13]</sup>,国内的用户模型研究工作起步晚。但是也做出了大量的研究成果。

### 1.3.2 个性化广告推荐技术的研究现状

个性化广告是指在投放广告时因人而异，区别对待，不同喜好的人推送的内容和形式应当加以区分。互联网的快速发展，使网络营销成为了一个新的广告投放市场，如何基于用户兴趣来投放具有差异化的广告，并且所推荐的广告信息正是用户所需要的信息，成为了研究者们面临的问题。

google 广告最初是基于用户的行为来分析用户兴趣。由于用户的行为信息会在用户浏览器中保存，尤其是鼠标点击信息。因此，google 广告平台通过模拟用户心理活动，推测出用户感兴趣的广告类型，此方法比较简单，不能得到精准化广告投放效果。

之后，google 提供一些兴趣词让用户选择，根据用户提交的信息进行个性化广告推荐。虽然此模式提高了广告推送的精度，但是降低了用户体验，因为很多用户不愿意提交兴趣词。

广告模式可以基于用户兴趣模型，也可基于用户行为。两者之间的差别在于，基于兴趣模型的广告模式，只会推荐用户感兴趣的广告；基于用户行为的广告模式会大量推荐和用户行为相关的广告。google 机构调查表明，基于用户兴趣模型和基于用户行为的广告模式相比，google 广告平台推荐给用户广告数量减少了，但是用户对广告的点击率有所提升。

Facebook 这种社交平台对于每个用户来讲，个性化已经很鲜明，数据更具有用户代表性，这种数据更加有利于深度挖掘。因为 Facebook 平台既有用户提交的基本信息，还有用户上网浏览信息，数据类型也比较广泛。因此 Facebook 凭借自己的优势，可以推荐给用户相关度更高的广告<sup>[4]</sup>。

在国内，随着社交平台的发展，比如新浪微博等，在很短时间就有上亿的用户群体。新浪平台一开始引入了原生的广告，推出便引起网友反感，因为当用户在浏览微博的时候，看到与自己毫不相关的广告信息。这种广告平台没有对用户进行区分，同质化对待每一个用户，导致用户体验下降，最后失败而终。

## 1.4 论文主要研究内容及创新点

为了能够在用户浏览信息时为其提供个性化广告推荐服务且不需要用户主动参与，我们希望通过用户的兴趣和广告推荐集合的匹配，建立一种广告推荐系统，使其推荐给用户的广告具有个性化特点。本文主要从用户模型和个性化推荐两个方面做了深入研究。用户兴趣模型是个性化推荐关键技术点，模型的好坏关系到个性化广告系统的推荐质量。本文研究内容主要有以下几点：

(1) 提出一种改进的隐式建模方法。传统的隐式建模方法将用户所有的日志信息作为建模信息，但是用户的兴趣是实时变化的，改进的隐式建模方法采取历史信息分类的方法，只将与本次查询相关的历史信息作为用户建模的文档集合。

(2) 提出了显隐式信息相结合的用户建模技术。显隐式信息结合的用户建模首先是根据用户提交的用户信息初始化用户模型，然后通过对用户上网历史信息进行分析，构建隐式用户模型，对初始用户模型进行更新。

(3) 提出了一种基于用户模型的协同过滤广告推荐算法。协同过滤技术是根据用户-项目评分矩阵，挖掘出用户相似集合，通过近邻集合中兴趣相近的用户给目标用户推荐信息。本文将上面提出的用户兴趣模型应用到协同过滤算法中，利用用户模型矩阵替代评分矩阵，实验结果表明基于用户模型的协同过滤推荐算法能够提升广告推荐的精确度。

## 1.5 本文组织结构

本文基本结构如下所示：

第一章，绪论，主要阐述个性化广告推荐系统产生的背景、用户兴趣模型和广告推荐系统的国内外发展现状。然后对本文研究主要内容作了描述，最后阐述了本文的组织结构。

第二章，相关研究。对本章涉及的相关知识做了简单的描述。对用户兴趣模型建模做了详细表述，详细介绍了常用的推荐技术以及个性化广告技术，为后文的研究奠定基础。

第三章，基于显隐式信息结合的用户兴趣模型构建。分析显式建模和隐式建模的弊端，提出基于显隐式结合的用户兴趣模型，并对原有的隐式用户建模做了改进。给出模型总体框架与流程，设计相关实验进行验证。

第四章，基于用户模型的协同过滤推荐技术。将第三章的用户兴趣模型结合到协同过滤算法中，通过与传统的协同过滤算法比较验证模型的有效性和准确性。

第五章，总结与展望。总结本文研究，根据研究中遇到的问题，提出下一步工作计划。

## 第二章 相关研究

本章内容主要是介绍基于用户模型的个性化广告推荐所需要的相关技术，首先描述了兴趣模型建模相关技术，然后介绍分析了常用的推荐技术。

### 2.1 用户兴趣建模流程

用户模型建立过程中涉及到 Web 文本分析、网络爬虫等相关技术，用户模型的创建过程需要借助其它相关技术来实现。图 2-1 表示用户兴趣模型创建的流程

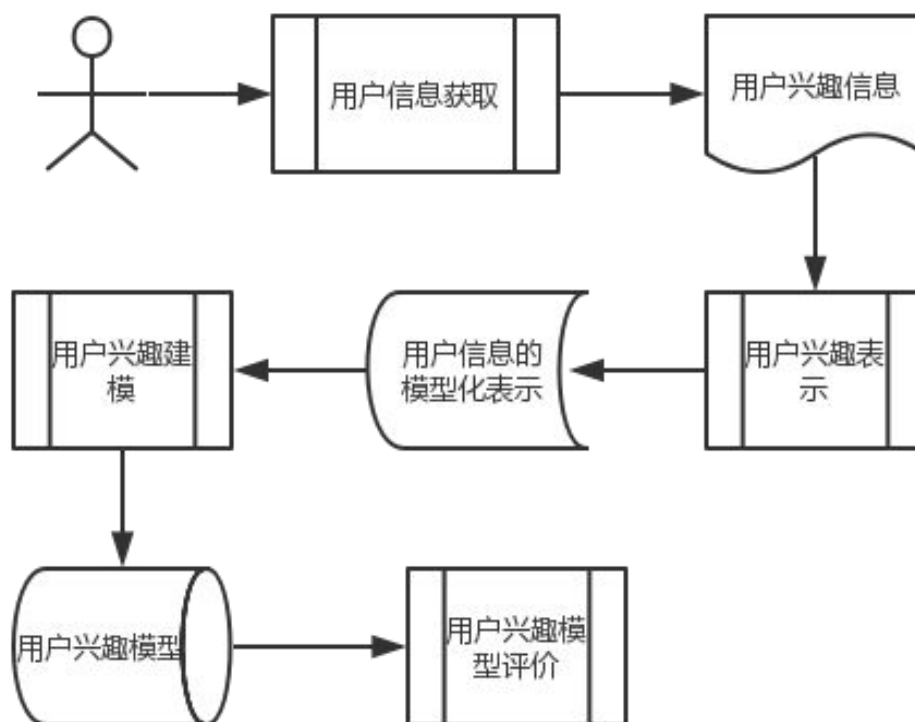


图 2-1 用户兴趣建模流程图

### 2.2 用户兴趣建模技术分析

本文的重点之处在于用户兴趣模型的建立，用户兴趣模型创建的好坏，直接影响向用户推荐的广告准确性，才能将用户所需要的广告推荐给用户。本文将对建模技术中涉及的关键技术详细描述。

### 2.2.1 用户采集信息来源

一个高质量的用户兴趣模型是建立在用户全面、真实的信息之上，因此信息来源至关重要。用户访问网络过程中有多种信息均可以反映出用户兴趣，这些信息都可以作为用户建模数据。基于错误的或则不全面的用户信息建模，用户模型质量会打折扣。所以用户信息选择相当重要，不然的话，构建出的用户模型质量太差<sup>[15]</sup>。大体上，我们会通过以下途径获取用户信息。

#### (1) 用户检索的关键词

用户在使用浏览器键入的查询词能够代表用户兴趣，但是由于查询关键词连贯性不强，并且数量较少，不能全面代表用户兴趣。查询关键字只是用户上网的开始，是用户的部分信息，通过关键词来表示用户兴趣的话就会以偏概全，得不到真正的用户兴趣模型。

#### (2) 浏览器的书签集合

用户的兴趣爱好也能够从浏览器保存的书签中获取到。通常，用户会把重要的或则需要重复查看的页面添加书签。以这种信息构建的用户模型也仅能提供少量的兴趣集。相比于用户浏览信息，书签集中的页面相当少，Polanco<sup>[16]</sup>研究发现用户保存的页面远少于感兴趣页面的三分之一。通过书签集合创建的模型由于训练样本少，不能全面表示用户兴趣。但可以将此信息结合其它的信息丰富用户模型。

#### (3) 用户的上网行为

用户上网行为包括用户在某个页面上执行的动作，以及在页面上花费的时间等。用户上网行为能够反应兴趣方向，例如用户长时间打开某个网页，代表用户喜欢页面内容。用户的浏览行为可以自动获取，无需用户提交。用户上网行为是在查看网页时发生的，因此用户行为必须和页面结合才能建模。

#### (4) 用户上网日志

用户登录信息、浏览页面以及用户访问行为都会记录在服务器的日志文件中，包括用户的唯一 ID，访问 URL、以及访问时间等信息<sup>[17]</sup>。通过分析服务器的日志文件便能收集用户浏览过的页面集合。互联网上收集用户信息的服务器分为三种，网页所在服务器、搜索引擎服务器和代理服务器。网站服务器只有对本站的访问日志，没有用户访问其它站点的日志记录，基于这种信息创建的用户模型不能全面反映用户的兴趣爱好。代理服务器是用户访问网络的中转站，记录了用户所有访问信息。收集用户信息最为全面的是搜索引擎服务器，信息包括查询关键字、浏览网页，浏览时间等。

#### (5) 用户主动提供信息

用户注册某个网站或者系统时，在页面中设置了兴趣选项，需要用户注册时填写。用户注册以后，系统自动将这些信息进行标记、聚类，然后作为用户兴趣爱好方向。这种方式有很大缺点，这种方式必须有用户的参与，牺牲了用户上网体验。有时候需

要提交的信息过于私密，会激起用户的反感，转而提交一些假的数据，基于这些假数据的用户兴趣模型与用户是完全不匹配的，基于错误的用户模型推送的广告信息，是完全不靠谱的。

综上所述，用户的上网轨迹多数都能代表用户喜好。其中，用户浏览的页面和浏览行为是用户建模的最佳信息来源；服务器日志也能代表用户喜好；用户标签虽然较少，但是都是用户喜欢程度较高的信息；用户查询关键词简明扼要，不能单独作为建模信息；用户主动提供的兴趣主题是建模重要信息<sup>[18]</sup>。

## 2.2.2 用户信息获取方式

结合上节所讲的信息来源，本文将信息获取方式分为显式收集和隐式收集方式。

显式收集方式：用户在注册某一网站时，提前指出自己兴趣爱好，比如提交感兴趣的特征词，喜爱的网页内容，提交对网页评价信息等，这种方式需要用户直接参与。事实上，文献<sup>[19]</sup>研究发现，很多用户不愿主动提交数据，担心个人信息被泄露，并且这种方式获取的信息相对较少，用户模型质量较低。显式提交信息时，容易涉及用户隐私，如年龄、性别等。这些信息只是用户兴趣信息的冰山一角，基于这些信息构建出的用户模型是不能代表用户的。

隐式收集：隐式信息收集是在用户上网浏览的过程中的信息，包括网页 URL、鼠标行为以及耗费的时间等。这种信息收集方式，通过用户历史数据挖掘而得到用户兴趣信息。文献<sup>[18]</sup>中的用户兴趣就是基于用户上网轨迹挖掘分析而出的；Webpersonlizer<sup>[20]</sup>就是基于隐式用户信息构建用户模型。

显式信息直接从用户反馈中得到，当用户兴趣发生变化时，需要用户主动更改，旧的用户模型无法体现用户新的兴趣爱好。相对而言，隐式收集方式无需用户主动提交，通过对用户上网的历史数据分析得到，能够根据用户兴趣变化而更新用户兴趣模型。网页数据挖掘分析技术，是以在海量数据为基础，根据用户浏览行为挖掘出用户兴趣信息。面对巨大的数据量，需要强大的人力物力支持和复杂的技术作为支撑<sup>[21]</sup>。

## 2.2.3 用户模型的数据结构

个性化广告推荐是根据用户喜好推荐广告，能否成功构建具有代表用户兴趣的用户兴趣模型，直接关系到个性化推荐的结果。用户兴趣模型是用户兴趣爱好的数学表达，这种数据结构需要很强的代表性和计算能力<sup>[22]</sup>。本文将详细介绍几种常用的表示方法：

### (1) 主题表示法

模型的主题表示法<sup>[23-24]</sup>就是将用户的信息分为我们常见的标签类型。例如某用户对手机类、文化类、体育类、汽车类感兴趣，则模型就可以表示为{手机，文化，体



育, 汽车}。主题表示方法依赖于领域信息, 例如, 雅虎的 MyYahoo 站点就是用户兴趣主题表示。如果用户定制了汽车和手机栏目, MyYahoo 就会将二者保存, 作为用户的兴趣模型。等用户再次访问 MyYahoo 时, MyYahoo 就会把与汽车和手机有关的信息推荐给用户。

### (2) VSM 表示法

VSM 表示<sup>[25]</sup>法主要是将向量空间中的维度信息作为用户模型的特征项向量空间中的点作为用户兴趣项的权重值。通过 VSM 表示的用户兴趣模型是一个  $n$  维特征向量  $\{(t_1, w_1), (t_2, w_2), \dots, (t_n, w_n)\}$  每一个维度代表用户一种兴趣喜好, 其中  $t_i (1 \leq i \leq n)$  表示特征项, 它一般是由字、词、短语等。 $w_i (1 \leq i \leq n)$  表示特征项在模型中的权重, 称为兴趣度<sup>[26]</sup>。该方法可以通过兴趣项的权重值的不同反映出用户不同的喜好, 方便项目后期资源匹配计算等。用户模型会随着兴趣文档增加而变大, 这种情况下系统需要大量的开销。

### (3) Keyword 列表法

Keyword (关键词) 列表法是将用户关键词作为集合表示用户兴趣模型<sup>[27-29]</sup>。假如一个用户特别喜欢汽车, 那么该用户的关键词列表可能为 {奥迪, 大众, 奔驰, 奇瑞}。关键词可以通过用户提交或机器学习得到, 通过机器学习关键词本质上都是通过训练文档得出一个特征集合。

### (4) 基于兴趣粒度表示法

兴趣粒度一般分为粗粒度和细粒度, 粗粒度表示法中兴趣项仅有喜欢与不喜欢之分, 分类比较极端。细粒度表示法对用户兴趣主题表现的更加详细, 算法复杂。基于粗粒度用户模型表示法, 简单实用, 实现起来速度快, 比细粒度表示法使用广泛。但是在推荐质量和精度上无法与细粒度表示法相提并论。表 2-1 是用户兴趣的不同粒度表示示例。

### (5) 基于用户-项目评分矩阵表示

该矩阵是一个  $m \times n$  维的矩阵列  $A_{m \times n}$ ,  $m$  代表用户数量,  $n$  代表项目的数量。用户的兴趣与  $A_{m \times n}$  的每一行相对应,  $A_{i \times j}$  代表第  $i$  个用户对第  $j$  个项目的评价, 若  $A_{i \times j}$  为空, 表明该  $User_i$  对该  $item_j$  未评价, 若有值, 代表  $User_i$  已经评价这个  $item_j$ ,  $User_i$  对  $item_j$  的喜欢程度取决于  $A_{i \times j}$  的大小。

### (6) 基于本体的表示法

本体是哲学学科的词语, 本意指的是存在的系统化表述, 具有规范性、可靠性、可重用性。到了计算机领域, 研究者从新定义了本体概念, 认为本体是对抽象化概念的表示和描述<sup>[30]</sup>, 并将本体的三大特性体现到信息推荐领域。用本体领域知识描述文档和用户兴趣模型, 能够提高推荐系统的准确性<sup>[31]</sup>。文献<sup>[32]</sup>将某一领域的用户模型表示为一个多维度集合:  $UserModel = (Personal\_I, Personal\_O, Personal\_R)$ 。



上述表达式中  $Personal\_I$  为用户基本信息，包含用户姓名、年龄、性别以及兴趣爱好、专业等； $Personal\_O$  为用户信息个性化领域本体； $Personal\_R$  为用户的个性化需求。

本体作为计算机领域概念化模型，描述了该领域中的术语、术语的含义以及术语之间的语义等基本信息。基于本体的用户模式依赖这些语义关系，构建能够提供用户感兴趣领域的抽象视图。但是本体的设计依赖性太强，设计的有效性受到约束<sup>[33]</sup>。

表 2-1 不同粒度示例表

粒度类型		兴趣词
粗粒度		汽车、篮球、滑雪、KTV、周星驰、天天向上、计算机、游戏、黄山、动物、NBA
细粒度类型	兴趣一	旅行、黄山、泰山、泰国、海角天涯、丽江
	兴趣二	体育、NBA、篮球、科比、姚明、中国足球
	兴趣三	计算机、电脑、pad、Java、c++、Mysql

综上所述，利用关键字表述用户的兴趣爱好是最常用的技术手段<sup>[34]</sup>。关键字对用户兴趣爱好的表示还是有很高准确性的，因此要想掌握用户的兴趣爱好，就需要理解关键字。只有了解了用户的兴趣偏好，才能做出高质量的推荐系统。

#### 2.2.4 模型构建的技术分类

根据用户是否参与模型创建过程，本文将用户建模技术分为手工建模和自动用户建模技术。

##### (1) 用户手工建模

用户手工建模是指将用户上传的信息作为用户兴趣模型数据来源，比如用户上传感兴趣的关键词，或是选择系统提供的感兴趣的栏目。用户手工建模技术是最简单的建模方法，在早期的个性化服务中，用户手工建模是主流建模技术。

早期的雅虎网站就是采用的手工定制建模。雅虎站点包含繁多的信息，然而用户感兴趣的信息比较少，为了增加效率，雅虎推出个性化产品 MyYahoo。用户注册 MyYahoo 账号以后，系统会让用户从种类繁多的栏目中选择感兴趣的栏目<sup>[35]</sup>。

手工建模技术简单，效果比较明显，但是也有不足之处。首先手工建模关键词完全有用户提交，降低用户使用的积极性。有研究表明，用户不情愿参与对系统的训练，虽然使用系统训练会给自己带来好处；其次用户提交的关键词不具有代表性，基于这些关键词的用户模型对用户的兴趣表述不够准确。这种模型的系统，关键词基本上都

是设计者自己理解组织的；再者用户的兴趣随着时间会发生变化，这时用户就要修改兴趣信息，也就是说用户模型一旦生成，模型就不会改变。

## (2) 隐式建模

隐式建模也叫自动化建模，隐式建模是根据用户上网历史自动创建用户模型，过程中无需用户干预。卡内基梅隆大学研制的 Personal WebWatcher、麻省理工学院的 Letizia、德国研究中心的 ELFI 等<sup>[36]</sup>都是采用隐式信息构建用户兴趣模型。

Personal WebWatcher 是卡内基·梅隆大学提出个性化推荐系统。用户上网时，该系统会记录用户的上网日志，通过分析上网日志，将日志进行分类，并作为训练集合的数据源，从训练集合中获取用户的兴趣词集合，通过兴趣词集合创建用户的兴趣模型。

麻省理工学院研究出的 Letizia，通过用户上网行为推断用户兴趣喜好<sup>[37]</sup>。比如，用户将某个页面添加到书签，这判定用户喜欢这个页面；如果用户忽略了页面中推荐的信息，则推断出用户对推荐的信息不感兴趣。

隐式用户建模方法从本质上讲是显式建模的一种升级，改变了用户信息获取方式，解放了用户。虽然提高用户上网体验，但是会放大噪声，构建的用户模型质量就会下降。但是总的来说，隐式建模无需用户参与，不会影响用户的上网体验。是用户建模技术的发展趋势。

## 2.3 推荐技术

针对不同的用户推荐不同的项目（本文指的是广告）或则针对同一个用户在不同的时间段推送不同的项目，这就是个性化推荐。不同的用户有着不同的兴趣爱好这就必然导致推荐结果的差异性。合格的推荐系统不仅知道当前的用户兴趣，还能帮助用户发现其它的兴趣爱好。下面我们将介绍个性化推荐原理，并详细介绍目前最为常用的推荐算法--协同过滤推荐算法。

本文将从四个部分来阐述个性化推荐原理，用户部分，项目部分，推荐部分和过滤部分，如图(2-2)所示。每个部分各司其职，共同实现个性化推荐。

现在的推荐系统越来越重视推荐信息的个性化，本文给出个性化广告推荐的定义如下：首先，收集用户历史信息，其次利用已经收集的信息对用户兴趣进行挖掘，最后一步是推送，当用户再次上网时，系统就会根据之前训练好的用户兴趣模型向用户推荐广告<sup>[38]</sup>。个性化广告推荐就是实现广告和用户个性化需求相匹配的过程。根据个性化广告推荐，一个好的个性化推荐系统设计需解决以下问题。

### 2.3.1 个性化推荐系统原理

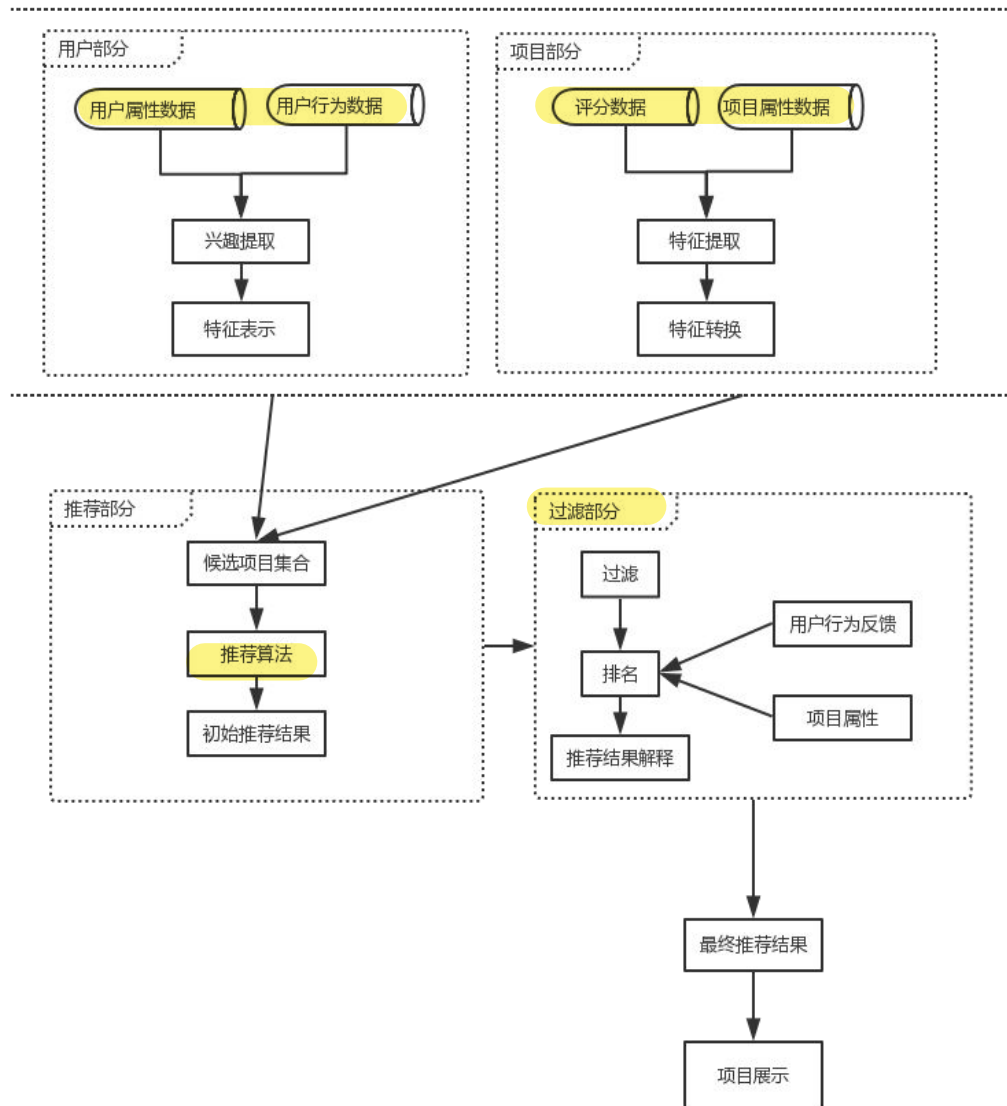


图 2-2 个性化推荐系统原理图

(1) 创建用户兴趣模型。基于用户的显式或则隐式信息，构建用户兴趣模型。本文把电影作为要推荐的广告，从预料集合中获取用户对电影观看记录和评价，分析出用户兴趣。

(2) 广告特征建模，广告特征建模同样是推荐系统关键部分。广告与用户之间的作用是相互的，根据两者的相似度，才能计算出广告与用户之间的最相似的广告。

(3) 根据广告与用户的模型，通过一定的推荐算法向用户推送广告。目前有很多推荐算法，比如说协同过滤算法、关联规则等。本文使用的是协同过滤算法作为广告的推荐技术，并结合 Top-N 推荐。

### 2.3.2 推荐技术的常用方法

随着用户需求增大，很多推荐算法产生<sup>[39]</sup>，比较成功的推荐技术有三种，本文基于推荐算法形成机制进行划分<sup>[40]</sup>，下面对这几算法进行描述。

(1) 基于内容推荐技术<sup>[41]</sup> (CBR, Content-based Recommendations)。CBR 推荐技术一般通过深度学习和自然语言处理算法实现推荐内容。CBR 推荐技术，用户不参与对项目的评价。算法根据项目的内容，可能是一条新闻的内容，也可能是视频中的一张照片，通过深度的学习，构建出用户兴趣模型，然后将相似的项目内容推荐给用户。Google 著名的项目“猫脸识别”就是基于图片内容的识别技术。

CBR 推荐技术只需要根据用户自己的历史信息，无需参考对比其他用户信息。快速推荐将是 CBR 推荐技术的优势所在。但是 CBR 技术对项目的依赖性比较强，可移植性就比较弱，不能多项目使用，如自动提取项目特征，这种方法不适用于多媒体领域，不利于推荐系统的扩展性。二是基于内容推荐技术的内容仅仅对用户历史兴趣信息推荐，无法为用户提供新的兴趣推荐。三是基于内容推荐算法无法确保推荐质量。最后是新的项目无法给用户推荐。

(2) 基于人口统计学的推荐技术<sup>[42]</sup>。(DBR, Demographic-based Recommendation) DBR 技术仅仅利用了用户的基本信息作为推荐算法的数据源，比如说，用户年龄，性别，职业，国籍等基本信息。推荐系统以此信息计算用户之间的相似度，然后按相似性分类。给同集合类的用户推送相同的项目（广告）。DBR 技术不需要大量的信息，也不需要复杂的算法分析。DBR 算法核心思想就是同龄人（或则同性别等）具有相同的兴趣爱好。例如幼儿园的小孩子喜欢玩积木，北方的人喜欢吃面食，南方人喜欢吃米饭，编程人员喜欢浏览技术方面的信息。

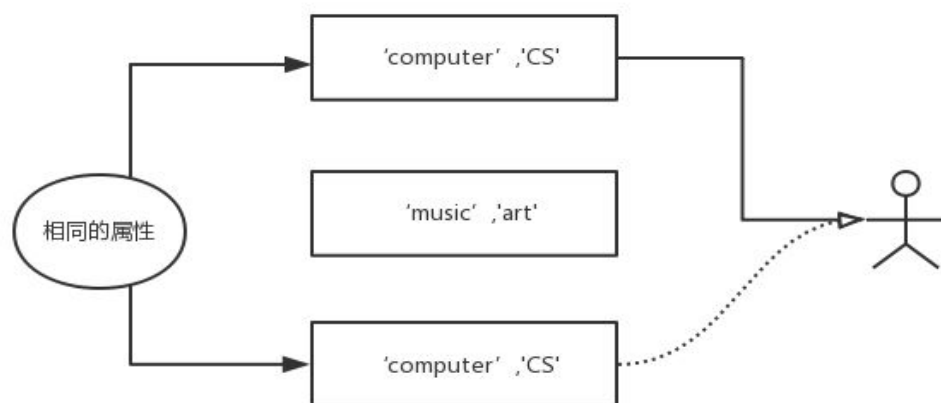


图 2-3 基于内容推荐示意图

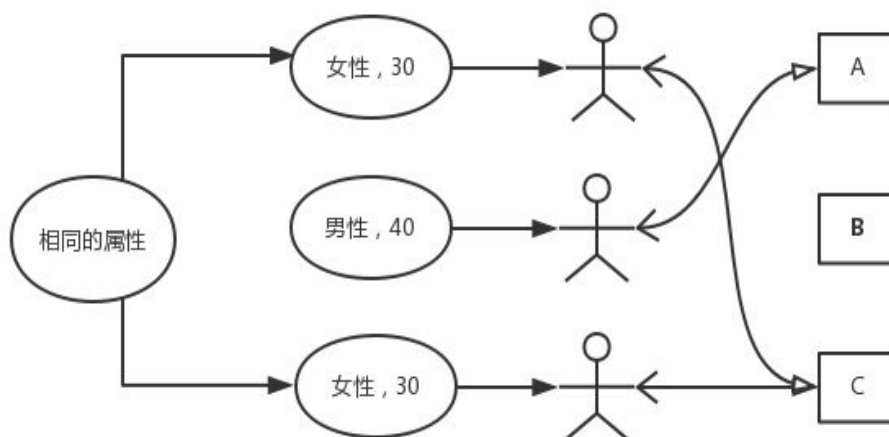


图 2-4 基于人口统计学的推荐示意图

DBR 算法虽然简单<sup>[43]</sup>，但是存在以下缺点：

1) 基于 DBR 系统需要的信息涉及了用户的隐私，很多人对此有抵触情节。一般的信息还可以，如果这些信息过于隐私，用户就会提交一些假的数据，这就会导致系统推荐质量的降低甚至推荐出错误的结果。

2) 这种方法对那些需要精确质量的个性化推荐而言太过简单，基于 DB 的分类方法是概括性的推荐方法，但是针对书籍、电影、音乐等差异性较大的项目时，该方法的弊端更加明显。

3) “局部冷落”问题，个别用户得不到相关推荐信息。

### (3) 基于关联规则推荐技术

基于关联规则推荐技术（RBR，Rule-based Recommendation），是数据挖掘领域中经典技术，大家最为熟悉的就是“啤酒-尿布”的案例<sup>[44]</sup>，就是通过挖掘用户购买商品记录来寻找商品之间的联系，根据这些商品组合为用户推荐。RBR 算法一目了然，就是基于大量的数据分析挖掘项目之间的关联性，关联性越强，基于 RB 的推荐准确度越高。但是规则的发现过程比较长，并且系统维护性高。

### (4) 协同过滤推荐技术

**协同过滤推荐**（CFR，Collaborative Filtering recommendation），CFR 算法的思想是通过计算相似用户集合，将相似用户喜欢的项目推荐给目标用户，建立在相近的用户有相同的爱好的假设之上（如图 2-5）。该技术在商业上最为成功<sup>[45]</sup>，给广告商带来了丰厚的回报。协同过滤算法有两种类型，一是基于用户相似性，二是基于项目之间的相似性。下面将分别介绍这两种算法。

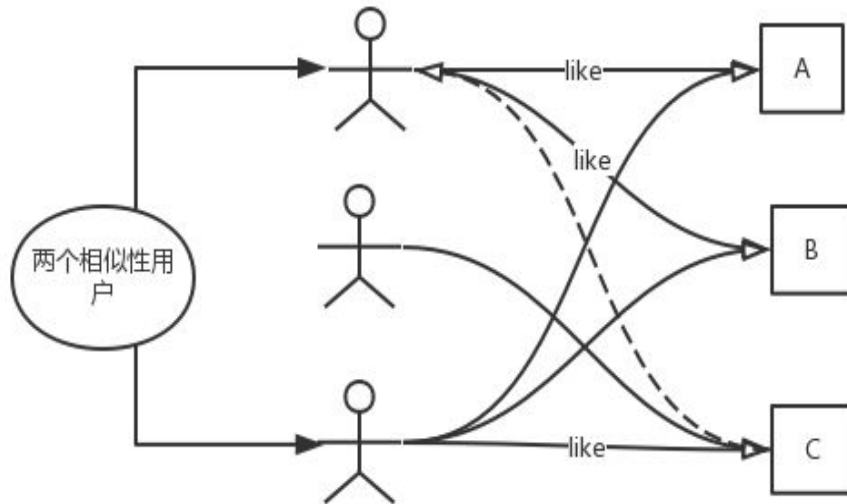


图 2-5 协同过滤推荐示意图

1) 基于用户的推荐，简称 UBCFR。该技术核心思想：在需要给用户 A 进行个性化推荐时，我们可以先找到和用户 A 有着相似兴趣爱好的用户 B。然后把用户 B 喜欢的，并且用户 A 没有发现的项目推荐给用户 A。我们把用户 B 变成一个集合，称之为 A 的近邻集合（NN，nearest-neighbor），就可将 NN 中的用户感兴趣的项目推荐给 A 用户。

当用户-项目集合中每个用户评价的项目都不同或者被评价的项目集合交集较少时，UBCFR 算法就会因为找不到近邻集合或则是近邻集合关联性小，使系统的推荐质量变得很差。当我们的用户集合数据量变得很大时，用户-项目矩阵就会变得很大，对系统硬件要求比较高。

2) 基于项目的推荐，简称 IBCFR。IBCFR 算法也是基于相似性，只不过该相似性是基于项目的，为相似的项目归为一个邻近集合。来通过用户对评价过的项目，找出该项目的近邻集合，并推荐给用户。

### 2.3.3 广告推荐技术

由于互网和电子商务的发展，互联网广告发展越来越快，本文结合前人的研究总结出了一下几种广告推荐技术。

#### (1) 随机投放式。

随机投放方式是指没有区别的推荐，对于每个用户，每个网页都是相同的推荐结果。随机投放方式，不需要复杂的算法，实现起来很简单，至今有不少广告商依然在使用此方式。此方法虽然简单，但是其收益预期只有 14% 左右。比如，当我们打开某

些广告的时候，很多网页都有游戏广告的推荐，这对于那些不喜欢游戏的人就是信息垃圾。严重影响了用户体验。

### (2) 基于网页信息投放方式。

基于网页信息投放的方式是基于某一个网页的信息，通过分析一个网页内容，将网页使用特征词标记，然后通过一定的匹配算法计算出网页与广告的相关度，通过网页的相关度推荐给用户。据统计网页上投放与该网页相关的广告，只有不到 30% 的网页适合根据网页内容展现，其余的如视频，小说等页面<sup>[46]</sup>。这种方式的广告投放虽然把广告与网页相结合。但是依然会出现用户不喜欢的广告。

### (3) 个性化广告推荐技术。

个性化广告就是针对不同的用户展现出不同的广告，由不同的平台展现不同的广告，从而实现广告的精准化推荐。按照推荐算法我们介绍几种常用的个性化广告推荐技术。

首先是基于内容的广告推荐技术。该技术只是仅仅利用用户主动提交的如 Age、Sex、Job 等个人信息，依据这些基本信息将相似的用户归为一类。基于 DB 推荐系统将为同一类用户推荐相同的广告。基于 DB 的推荐技术，无需深入挖掘 User 信息，只需要用户之间有相同的生理属性。如同一个职业的人会喜欢相同的书籍等。

其次是基于协同过滤算法的广告推荐技术<sup>[47]</sup>。根据上面章节介绍，用户之间或者是项目之间的相似性是基于项目评分矩阵计算而来，计算用户的近邻集合，将近邻集合中用户的兴趣项推荐给用户。基于协同过滤算法的推荐应用是目前最为成功的广告推荐技术。

通过以上广告推荐技术分析，本文提出了基于用户兴趣模型的个性化广告推荐技，将本文训练出的用户兴趣模型，结合协同过滤推荐算法，为用户提供广告推荐。

## 2.4 本章小结

第二章相关技术中，我们详细描述了用户兴趣模型，通过介绍用户兴趣模型的代表方法，技术分类，以及数据来源等相关技术。介绍了推荐技术原理，分别对基于人口统计和基于用户，基于项目的推荐算法做了详细介绍，重点讲解了协同过滤算法。并对当前的广告推荐技术如随机投放、基于网页内容、个性化投放等技术做了讲解。



### 第三章 基于显隐式信息结合的用户兴趣模型构建

基于用户主动提交的感兴趣关键字和个人基本信息创建的显式用户兴趣模型，由于数据的准确性，显式用户兴趣模型真实性比较高。但是随着时间的变化，用户兴趣也在变化，显式用户模型已经不能代表用户真正的兴趣方向。隐式建模技术的数据源是用户的实时的上网日志，隐式建模能够随用户兴趣而发生改变，具有实时动态性。所以本文提出了基于显隐式结合用户兴趣模型建模方法。

#### 3.1 基于显式信息的用户兴趣建模

由第二章我们可知，显式建模技术的信息来源主要是用户提交的个人信息和感兴趣的关键词。基于显式信息创建的用户兴趣模型真实性比较高，能代表用户初期的兴趣方向。

显式建模技术我们采用示例用户信息建模。用户提交自己感兴趣的兴趣词，每一种类型对应一种文档集，文档集总称为文档库。文档库分汽车、军事、科技等主题。显式获取用户兴趣处理过程如下图 3-1 所示。

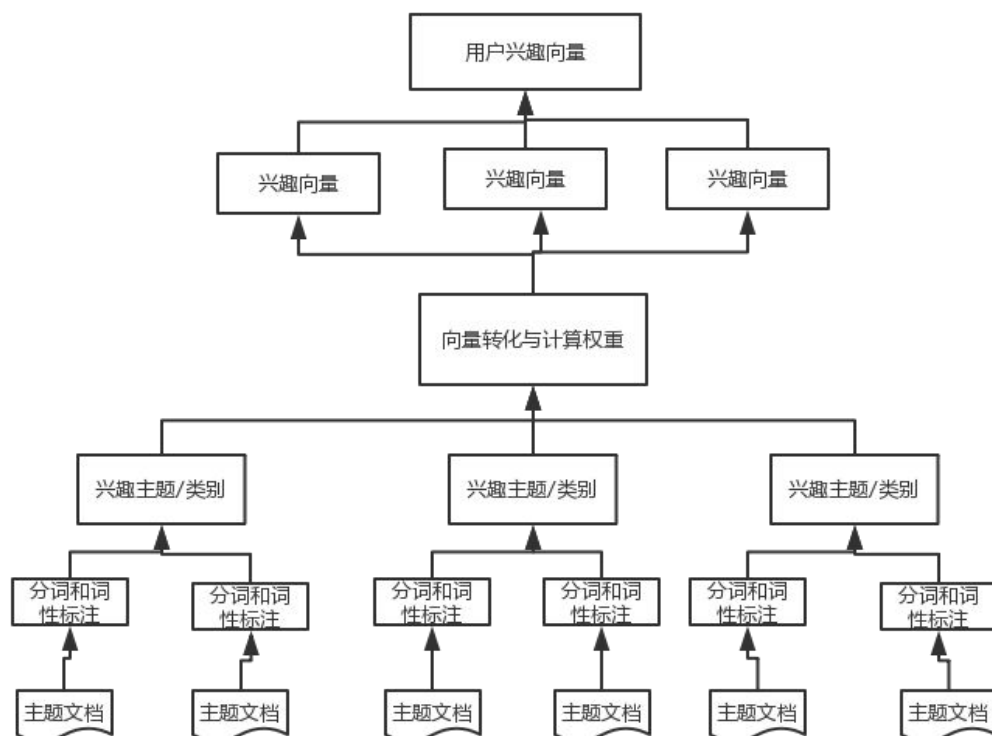


图 3-1 显式信息模型初始化示意图



显式信息用户兴趣建模算法如下：

**Step 1:**创建示例文档与兴趣词特征矩阵

首先，将用户提交的兴趣词以及示例文档均用向量空间模型表示。然后利用 **tf-idf** 公式求出每个兴趣词在每篇文档下的权重。从而生成一个反应兴趣词与示例文档关系的矩阵。这个矩阵带有示例文档的特点，用户提交的兴趣词用向量  $Q=(q_1, q_2, \dots, q_i, \dots, q_n)$  表示，用户提交的兴趣词对应的示例文档用  $D=(d_1, d_2, \dots, d_i, \dots, d_m)$  表示，这样由  $m$  个行向量和  $n$  个列向量可以构成一个  $m \times n$  的特征矩阵  $M_{DQ}$ ，每个兴趣词根据  $tf * idf$  算法得到一个值  $w$ ， $w$  就是这个兴趣词在特征矩阵的权重。公式（3-1）就是  $w$  值的计算公式：

$$w_{ik} = \frac{f_{ik} * \log(\frac{N}{n_j})}{\sqrt{\sum_{j=1}^M [f_{ik} * \log(\frac{N}{n_j})]^2}} \quad (3-1)$$

其中  $f_{ik}$  表示第  $i$  个兴趣词在第  $k$  篇文档中出现的次数； $N$  表示相关文档个数； $n_j$  表示相关文档。

最后形成表 3-1 的文档和兴趣词的特征矩阵  $M_{DQ}$ ，其中每行代表一篇文档，每列代表一个兴趣词，例如兴趣词为“丰田”，对应文档  $D_5$  中的特征权重为 0.039。

表 3-1 示例文档与兴趣词的特征矩阵  $M_{DQ}$  表

文档	平安基金	丰田	工商银行	篮球
D1	0	0	0	0
D2	0	0	0	0
D3	0	0	0.01	0
D4	0	0	0	0
D5	0	0.039	0	0
D6	0	0.038	0	0
D7	0.011	0	0.007	0

### Step 2:创建示例文档与主题类别特征矩阵

由于  $SVM$  即支持向量机文本分类算法对文本分类有较好的区分能力，因此本文使用  $SVM$  算法对示例文档进行分类。文档和主题类别特征矩阵  $M_{DC}$  就是通过  $SVM$  算法得出，如表 3-2 所示：

表 3-2 文档和主题类别特征矩阵  $M_{DQ}$  表

文档	财经	电子电气	军事	旅游	汽车	体育
D1	0.117	0.077	0.071	0.041	0.034	0.112
D2	0.112	0.158	0.053	0.151	0.065	0.139
D3	0.112	0.158	0.053	0.151	0.085	0.065
D4	0.134	0.092	0.056	0.150	0.087	0.073
D5	0.072	0.018	0.014	0.007	0.026	0.004
D6	0.073	0.018	0.014	0.003	0.003	0.033
D7	0.858	0.119	0.013	0.004	0.004	0.000

文档与特征类别矩阵  $M_{DC}$  也是为初始化用户兴趣模型提供中间数据。

### Step3:兴趣词和类别特征矩阵

通过以上两步可以得到文档和兴趣词矩阵  $M_{DQ}$  以及文档和特征类别矩阵  $M_{DC}$ ，从而我们可以发现两个矩阵之间通过文档具有相关性，进而得到用户兴趣词与特征类别矩阵  $M_{QC}$ 。本文使用文本分类中的简单自适应算法 Rocchio，公式 (3-2) 如下所示：

$$M_{QC}(i, j) = \frac{1}{N_i} \sum_{k=1}^m M_{DC(k, i)}^T M_{DQ(k, j)} \quad (3-2)$$

其中， $m$  表示矩阵  $M_{DQ}$  中文档的数量； $N_i$  与第  $i$  类文档相关的文档数量； $M_{QC}(i, j)$  表示第  $j$  个兴趣词与第  $i$  类相关文档的权重。

通过 Rocchio 算法我们就可以得到兴趣词和主题类别特征矩阵如表 3-3 所示，其中每一行代表一类，每一列代表一个用户提交的兴趣词，例如，查询词“丰田”在汽车类别的兴趣度为 0.9915。

表 3-3 兴趣词和主题类特征矩阵

类别	平安基金	丰田	工商银行	篮球
财经	0.74524	0.09337	0.99522	0.79759
电子电气	0.66679	0.02376	0.06388	0.51273
军事	0	0.01867	0.06724	0.21648
旅游	0	0.03395	0	0.04177
汽车	0	0.99150	0.00336	0.04177
体育	0	0.03056	0	0.07975

**Step4:初始化用户模型**

在兴趣词和类别特征矩阵中，我们计算每一行即对每一类的平均值，便可得到用户对每个主题类别的兴趣度值，进而得到用户的兴趣偏好，兴趣度数值如表 3-4 所示：

表 3-4 文档和主题类别特征矩阵

财经	电子电气	军事	旅游	汽车	体育
0.79	0.34	0.07	0.22	0.27	0.12

从表 3-4 中，我们可以看到用户在财经、电子电气、军事、旅游、体育等类别下的兴趣度，我们用用户向量形式来表示用户的兴趣模型  $U$  表示如下：

$U=[(\text{财经}, 0.79), (\text{电子电气}, 0.34), (\text{军事}, 0.07), (\text{旅游}, 0.22), (\text{汽车}, 0.27), (\text{体育}, 0.12)]$

**3.2 改进的隐式建模方法**

传统的用户兴趣建模<sup>[48]</sup>会将用户所有日志信息作为建模数据，但是用户兴趣不是一成不变的，而是实时变化的，这样构建出的用户模型会对个性化展示带来适得其反的效果。本文改进的隐式建模方法采取历史信息分类的方法，只将与本次查询相关的历史信息作为用户模型的文档集。改进的隐式建模方法能够反映用户实时兴趣信息。

### 3.2.1 传统隐式信息建模方法

隐式建模的数据来源是用户的检索词和浏览的网页内容，我们需要将这些信息保存到数据库中。通过对用户历史文档做充分的分析，采用数据挖掘和文本分析等相关技术，找到用户的兴趣偏好，采用向量空间模型表示用户兴趣模型。

首先，我们将用户查询历史文档转化成向量由  $d_i = (t_1, t_2, \dots, t_i, \dots, t_n)$  表示，其中  $t$  代表文档中的特征词。假设用户提交多个不同的查询，查询向量表示为  $M = (q_1, q_2, \dots, q_i, \dots, q_m)$ ，用户点击浏览的文档向量为  $N = (d_1, d_2, \dots, d_i, \dots, d_n)$ 。我们使用矩阵  $C_{m \times n}$  表示用户查询的文档点击分布，如下所示：

$$C_{mn} = \begin{pmatrix} c_{11} & \dots & c_{1n} \\ \dots & \dots & \dots \\ c_{m1} & \dots & c_{mn} \end{pmatrix}$$

上述矩阵中  $C_{m \times n}$  值代表用户在查询关键字  $m$  上点击了文档  $n$  的次数。由此得出检索词  $m$  的文档向量  $\vec{q}_m$  的计算公式<sup>[49]</sup>如下：

$$\vec{q}_m = \sum_{i=0}^n \frac{C_{m,i}}{\sum_{j=1}^N C_{m,j}} \vec{d}_i \quad (3-3)$$

下面详述公式 (3-3)：

(1)  $\frac{C_{m,i}}{\sum_{j=1}^N C_{m,j}}$  代表在使用  $m$  检索词时，查看网页  $d_i$  的概率，上述矩阵中  $m$  行的某一项除以  $m$  行所有项之和。

(2)  $\frac{C_{m,i}}{\sum_{j=1}^N C_{m,j}} \vec{d}_i$  表示在某个查询词下。将文档兴趣词的权重计算进去，以此来体现用户对每个点击浏览过的文档感兴趣程度是不一样的。

(3) 最后的  $\sum_{i=0}^n$  表示将使用检索词  $m$  点击查看过的所有文档  $d_1, d_2, \dots, d_i, \dots, d_n$ ，

在完成第二步之后进行文档向量合并工作，收集所有文档中的词汇构成一个针对检索

词  $m$  的向量，合并其中相同的兴趣词，作权重累加，保证向量中每个项的互异性。比如向量  $d_1$ ， $d_2$ ：

$$\begin{aligned} d_1 &= \{(t_1, w_{11}), (t_2, w_{12}), (w_3, w_{13})\} \\ d_2 &= \{(t_2, w_{22}), (t_3, w_{23}), (t_3, w_{24})\} \end{aligned} \quad (3-4)$$

合并向量  $d_1$ ， $d_2$  得出：

$$d_{12} = \{(t_1, w_{11}), (t_2, w_{12}+w_{22}), (t_3, w_{13}+w_{23}), (t_4, w_{24})\}$$

例如我们设计一个历史检索矩阵  $C_{3 \times 3}$ ，如下矩阵：

$$C_{3 \times 3} = \begin{pmatrix} 1 & 2 & 3 \\ 2 & 3 & 1 \\ 2 & 3 & 2 \end{pmatrix} \quad (3-5)$$

上述矩阵中的每一行分别用  $q_1$ ， $q_2$ ， $q_3$  表示，对应用户历史检索词；每一列分别用  $d_1$ ， $d_2$ ， $d_3$  表示，对应点击文档的次数。

我们得出检索词  $q_1$  查看网页  $d_1, d_2, d_3$  的概率为  $\frac{1}{1+2+3} = \frac{1}{6}$ ， $\frac{2}{1+2+3} = \frac{1}{3}$ ， $\frac{3}{1+2+3} = \frac{1}{2}$ ；

我们将文档转化为特征词的列向量表示，如下所示：

$$\vec{q} = \left\{ \begin{pmatrix} t_{11} \\ t_{12} \\ t_{1n} \end{pmatrix}, \begin{pmatrix} t_{21} \\ t_{22} \\ t_{2n} \end{pmatrix}, \begin{pmatrix} t_{31} \\ t_{32} \\ t_{3n} \end{pmatrix} \right\} \quad (3-6)$$

其中  $t$  代表文档中的特征词。利用公式 (3-1) 我们得出检索词  $q_1$  的文档向量计算如下：

$$\vec{q} = \sum_{i=1}^3 \frac{C_{1,i}}{\sum_{j=1}^3 C_{1,j}} \vec{d}_i = \frac{1}{6} \begin{pmatrix} t_{11} \\ t_{12} \\ \vdots \\ t_{1n} \end{pmatrix} \text{Sim}_1 + \frac{1}{3} \begin{pmatrix} t_{21} \\ t_{22} \\ \vdots \\ t_{2n} \end{pmatrix} \text{Sim}_2 + \frac{1}{2} \begin{pmatrix} t_{31} \\ t_{32} \\ \vdots \\ t_{3n} \end{pmatrix} \text{Sim}_3 \quad (3-7)$$

我们用  $Q_j$  代表一个查询词，用  $t_j$  表示查询词的次数，则用下面矩阵表示查询词被提交的次数。

$$Q_n = \{t_1, t_2, \cdots \cdots, t_j, t_n\}^T \quad (3-8)$$

通过对检索词的使用频率不同,分析挖掘用户对每个检索词下的网页兴趣度,  $w_m$  就是表示用户对不同的网页的兴趣,计算公式如下:

$$w_m = \frac{Q_m}{\sum_{i=1}^M Q_i} \quad (3-9)$$

公式(3-9)中,分子代表检索词被检索的频次,分母代表用户对所有的检索词的检索次数之和。

结合以上介绍,我们用  $U$  代表用户兴趣向量,公式如下:

$$U = \sum_{m=1}^m w_m \vec{q}_m \quad (3-10)$$

公式(3-10)中,  $\vec{q}_m$  是  $m$  主题词下的所有文档经过合并的向量  $w_m$  是主题词所对应文档的权重。

如此我们对上面的假设例子建模得出如下公式<sup>[50]</sup>:

$$U = w_1 \vec{q}_1 + w_2 \vec{q}_2 + w_3 \vec{q}_3 + \cdots + w_n \vec{q}_n \quad (3-11)$$

### 3.2.2 改进的隐式建模方法

本文在传统的基于用户查询历史的模型上做了更准确化的改进。传统的用户模型对用户上网日志不作区分,无论用户的历史兴趣是否与本次查询相关与否都参与兴趣建模。由于用户兴趣是动态变化的,如此计算会影响个性化效果,甚至适得其反。针对这个问题,本文提出改进方法,就是筛选出于本次查询相关的历史记录进行用户兴趣模型建模,提高用户兴趣建模的准确性。下面是对上一节原有的隐式建模方法做具体的改进工作。

为了在用户历史查询记录中找出与本次查询相关的历史记录,我们将本次查询与历史查询文档做相似度计算,其相似度计算公式如下:

$$\text{Sim}_m = \cos(d_j, q) = \frac{\langle \vec{d}_j, \vec{q} \rangle}{|\vec{d}_j| \times |\vec{q}|} = \frac{\sum_{i=1}^{|V|} w_{ij} \times w_{iq}}{\sqrt{\sum_{i=1}^{|V|} w_{ij}^2} \times \sqrt{\sum_{i=1}^{|V|} w_{iq}^2}} \quad (3-12)$$

得到上述相似度值  $\text{Sim}_m$  之后,接下来有两种操作方式,一是将  $\text{Sim}_m$  的值小于某个设定值的历史记录直接过滤掉,不再将这些文件进行用户兴趣建模;二是将  $\text{Sim}_m$  作为体现差异的值,参与到文档权重的计算。融入了  $\text{Sim}_m$  的  $\vec{q}_m$  计算公式为:

$$\vec{q}_m = \sum_{i=0}^n \left( \frac{C_{m,i}}{\sum_{j=1}^N C_{m,j}} \vec{d}_i, \text{Sim}_i \right) \quad (3-13)$$

之后本文得出最终的计算公式为：

$$U_{new} = \sum_{m=1}^M w_m \vec{q}_m = \sum_{m=1}^M \frac{Q_m}{\sum_{i=1}^n Q_i} \times \frac{\sum_{i=1}^n (C_{m,i} d_i \times \text{Sim}_i)}{\sum_{j=1}^N C_{m,j}} \quad (3-14)$$

由于第一种改进只是文档的过滤，所以本文只展示第二种改进方式。第二个操作将  $\text{Sim}_m$  值作为体现差异的值，参与用户文档权重的计算中。

首先我们计算本次查询与历史文档的相似度，假设本次的查询向量为  $\vec{q}$ ，则它与文档  $d_j$  的相似度计算公式如下：

$$\text{Sim}_j = \cos(d_j, q) = \frac{\langle \vec{d}_j, \vec{q} \rangle}{|\vec{d}_j| \times |\vec{q}|} \quad (3-15)$$

通过 (3-15) 公式， $\vec{q}$  与历史文档  $d_1, d_2, d_3$  的相似值分别是： $\text{Sim}_1, \text{Sim}_2, \text{Sim}_3$ 。  
(3-9) 公式得出的结果反映出了本次查询与历史文档的相关性，我们将其作为权重计算的因子。 $\text{Sim}_i$  值越大，表示当前的检索词与历史文件  $i$  的相关度比较大。我们利用公式 (3-8) 得到计算  $\vec{q}$  的新公式：

$$\vec{q} = \sum_{i=1}^3 \frac{C_{1,i}}{\sum_{j=1}^3 C_{1,j}} \vec{d}_i \text{Sim}_i = \frac{1}{6} \begin{Bmatrix} t_{11} \\ t_{12} \\ \vdots \\ t_{1n} \end{Bmatrix} \text{Sim}_1 + \frac{1}{3} \begin{Bmatrix} t_{21} \\ t_{22} \\ \vdots \\ t_{2n} \end{Bmatrix} \text{Sim}_2 + \frac{1}{2} \begin{Bmatrix} t_{31} \\ t_{32} \\ \vdots \\ t_{3n} \end{Bmatrix} \text{Sim}_3 \quad (3-16)$$

用户兴趣模型的计算公式更新为：

$$\begin{aligned}
 U_{\text{new}} &= \sum_{m=1}^n w_m \vec{q}_m = w_1 \vec{q}_1 + w_2 \vec{q}_2 + \cdots + w_n \vec{q}_n \\
 &= w_1 \sum_{i=1}^n \frac{C_{1j}}{\sum_{j=1}^n C_{1j}} \vec{d}_i \text{Sim}_j + w_2 \sum_{i=1}^n \frac{C_{2i}}{\sum_{j=1}^n C_{1j}} \vec{d}_i \text{Sim}_j + \cdots + \\
 &\quad w_3 \sum_{i=1}^n \frac{C_{li}}{\sum_{j=1}^n C_{1j}} \vec{d}_i \text{Sim}_i
 \end{aligned} \tag{3-17}$$

由公式(3-17)可以得出用户兴趣模型最后的表现形式为向量 $U$ ，是一个多维的空间向量，如下所示：

$$U = [(w_1, W_1), (w_2, W_2), \cdots, (w_n, W_n)] \tag{3-18}$$

其中 $w_1, w_2, w_n$ 代表的是兴趣词，而 $W_1, W_2, W_n$ 代表的是每个兴趣词的权重。

### 3.3 显隐式信息结合的用户兴趣建模方法

显式方法创建的用户模型，由于用户的主动参与，虽然更加贴近用户真实兴趣，效果也比较好，但是缺点也比较多。隐式信息建模能自动获取用户信息，对用户兴趣变化反应敏锐，具有极强的自动更新和学习能力，并且无需用户的参与，不足之处在于系统只能获取用户客观信息，对语义上的理解不够充分，因此难以准确反映用户真正兴趣所在。

最能表达用户兴趣的当然是用户自己，用户对信息的描述体现用户主观因素，而客观因素可以通过隐式建模获取用户信息得到补充。所以，本文采用显隐式结合的方法创建用户兴趣模型。

首先，在用户建模的初期要求用户显式描述自己对哪些内容感兴趣，需要用户自己选择。为了更加准确细致的收集用户兴趣爱好，用户可以添加兴趣词，也可以设置主题兴趣类别，这些信息将用户初始化用户兴趣模型。

其次，系统能在用户原有的历史兴趣中挖掘历史数据，自动分析用户信息，完善对用户兴趣的收集和描述。

本文采用显隐式信息相结合的用户建模方法，能够有效弥补两种建模方式的不足，发挥各自所长为用户提供了较好的解决途径。



### 3.3.1 显隐式信息结合的兴趣建模框架与流程

#### (1) 用户信息来源架构。

只有建立一个合适的用户模型才能精准的表示用户兴趣爱好。一个好的用户模型必须基于真实、全面的用户信息，运用有效的方法对用户日志信息进行挖掘分析，通过不断的学习中，更新和完善用户模型。图 3-2 为用户信息来源示意图。

从图 3-2 中我们知道用户建模有三个信息来源，一是用户兴趣类主题词集合，集合中的词是常见领域内具有代表性的词，词是按照主题类型进行分类，这些主题类型特征词将会为用户兴趣模型提供最为初始化的信息。还有一个就是用户主动提交或选择系统提供的兴趣项。最后就是用户的历史查询记录，这些信息随着用户兴趣偏好转移而变化。这三种信息共同构成了用户兴趣模型的数据源，在创建用户兴趣模型的各个时段，发挥着不同程度的作用。

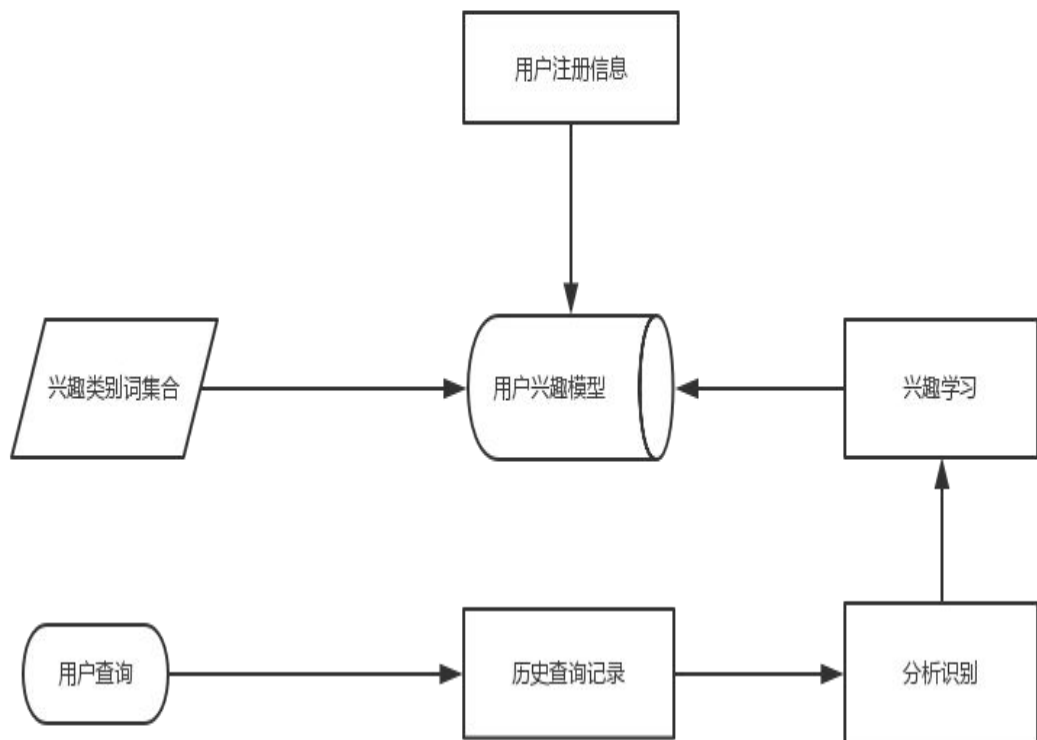


图 3-2 用户信息来源示意图

#### (2) 用户兴趣模型工作流程

通过图 3-3 用户兴趣模型工作流程图所示，系统需要对用户身份进行判断，如果是老用户，则提醒用户登录。若是新用户则需要跳转到注册页面，强制用户进行注册，

用户注册时需要根据系统提示选择自己感兴趣的主题词，以及用户个人信息。用户的基本信息作为系统中的标识信息，用来区分用户身份。用户提交的感兴趣的主题信息作为用户建模的初始化数据。在用户检索浏览网页信息的过程中，系统根据用户选择的兴趣信息初始化用户模型。

随着用户检索词和浏览的网页文档增多，系统针对用户的上网日志记录进行隐式建模。隐式建模能够通过分析挖掘用户的实时兴趣信息，并对初始化的用户模型进行更新，使用户模型更加能够代表用户兴趣。随着用户上网浏览信息，历史日志信息会过于庞大，大量日志信息需要分析挖掘，渐渐的影响用户兴趣模型的构建速度。因此我们需要对用户的兴趣主题词进行一定量的限制，并且对很早的日志文档以及那些与本次查询无关的文档进行过滤，使其不参加用户兴趣建模。

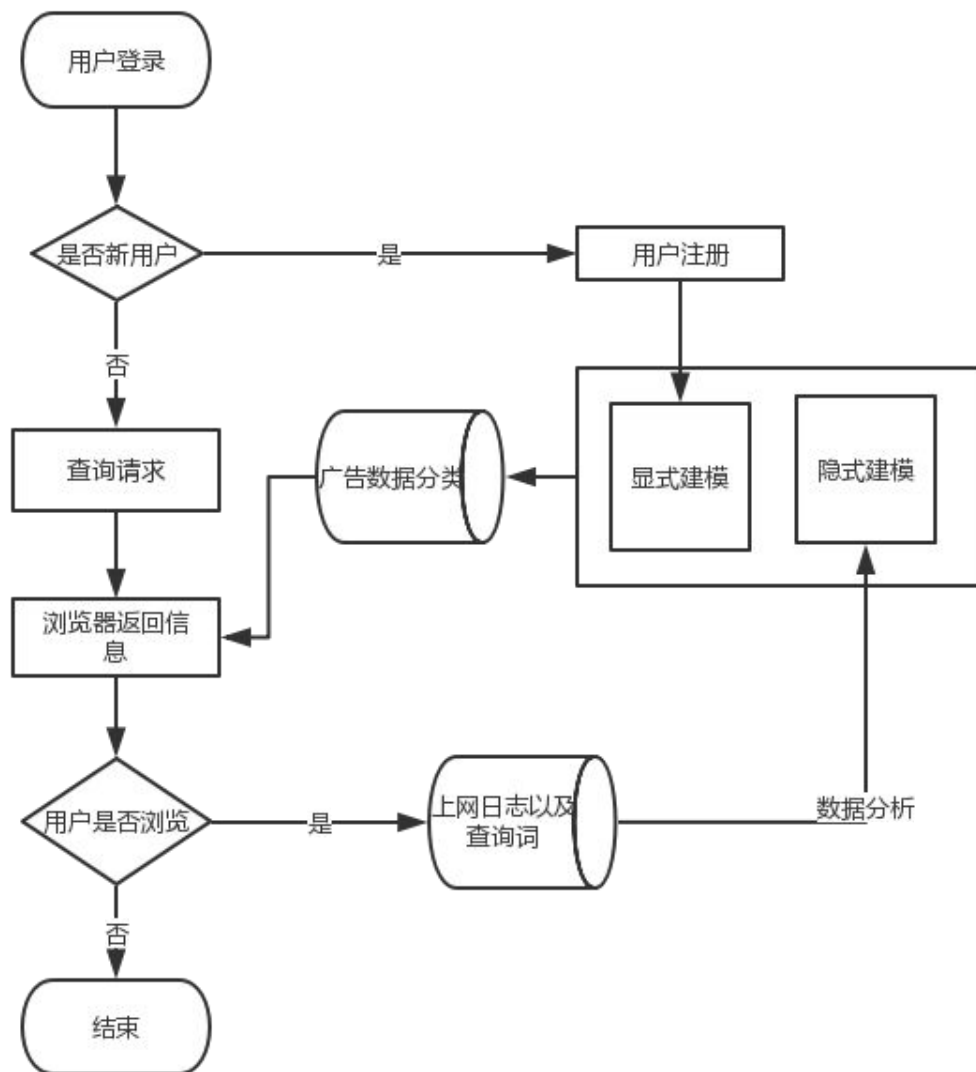


图 3-3 显隐式信息结合建模流程图

### (3) 显隐式结合的用户兴趣建模算法设计

根据用户兴趣模型的工作流程可知,用户兴趣模型的构建是整个个性化广告推荐的重要部分,只有构建出符合用户兴趣模型系统才能推荐出用户喜欢的广告,下面我们结合前面所描述的显式建模以及隐式建模方法,设计出显隐式信息结合的用户建模算法。算法如下:

**Step1:** 根据用户提交的兴趣词以及示例文档进行显式用户信息建模,构建初始化用户兴趣模型。这一步骤中主要是计算示例文档与兴趣的特征矩阵  $M_{DO}$  和示例文档与主题类别特征矩阵  $M_{DC}$  以及兴趣词与主题类别特征矩阵  $M_{OC}$ ,通过这三个矩阵生成显式用户兴趣模型。

**Step2:** 收集用户的上网日志信息以及查询信息,这一步骤主要是收集用户的隐式信息,比如查询词、点击浏览的文档等,将这些信息转化成用户的查询矩阵  $M_{mi}$ 。

**Step3:** 对用户的上网日志信息进行自然语言处理,对日志文档进行向量转化。这一步骤主要是对日志文档进行分词、关键词提取以及词频统计等。

**Step4:** 依据本文提出的改进式的隐式用户信息建模方法创建用户兴趣模型。

**Step5:** 结合 Step1 步骤生成的初始化用户模型,构建出符合用兴趣偏好的用户兴趣模型。

#### 3.3.2 显隐式结合的用户兴趣模型建立

本文的用户兴趣模型的实现主要有由 PreFrame、UserMatrix、Noun、Calculate 和 UserProfile 等主要部分。程序的数据源是主题分类文档库,用户检索历史文档和用户提交的关键词列表。程序输出的结果就是用户兴趣模型。本程序中用到了 JK Analyzer 分词工具,这里不再详细介绍此分词包工具。下图 3-4 即为本文系统结构图。

我们通过以下几步分析介绍本文设计的用户模型构建系统。

(1) 显式建模,示例建模方法中最为主要的部分为主题文档数据源,以及关键词列表。用户提交的兴趣爱好与示例建模中的主题文档进行匹配,从而能够全面的为示例建模方法提供数据源信息。

(2) PreFrame 模块,该模块的主要功能就是用户的身份识别,以及接收用户显式反馈的偏好,根据这些用户信息,通过显式信息建模方法初始化用户兴趣模型。

(3) Noun 模块,主要工作是通过用户选择的兴趣偏向,关联主题类文档,对主题类文档进行分词和词性标注,然后进行特征词的提取和权重计算。

(4) UserMatrix 模块,对用户查看的历史网页文档进行计算分析,构建用户查询矩阵,通过改进的查询词与历史文档相似度计算公式计算查询词与历史文档的相似度,过滤掉与本次查询无关的历史文档。

(5) Calculate 模块，主要对用户的历史网页文档进行自然语言处理，包括分词，关键词提取以及词频统计等。

(6) UserProfile 模块，主要工作就是对用户的访问历史记录文档做向量转化，然后与 Calculate 模块共同完成用户模型创建。通过计算本次检索词与历史网页文档信息的相似性实现本文提出的改进方法。

(7) 用户查询历史文档，作为日志信息，它存储着用户查询过的文档标题和摘要、查询词、点击时间和用户名等信息，保存在服务器端，以关系数据库的形式保存。

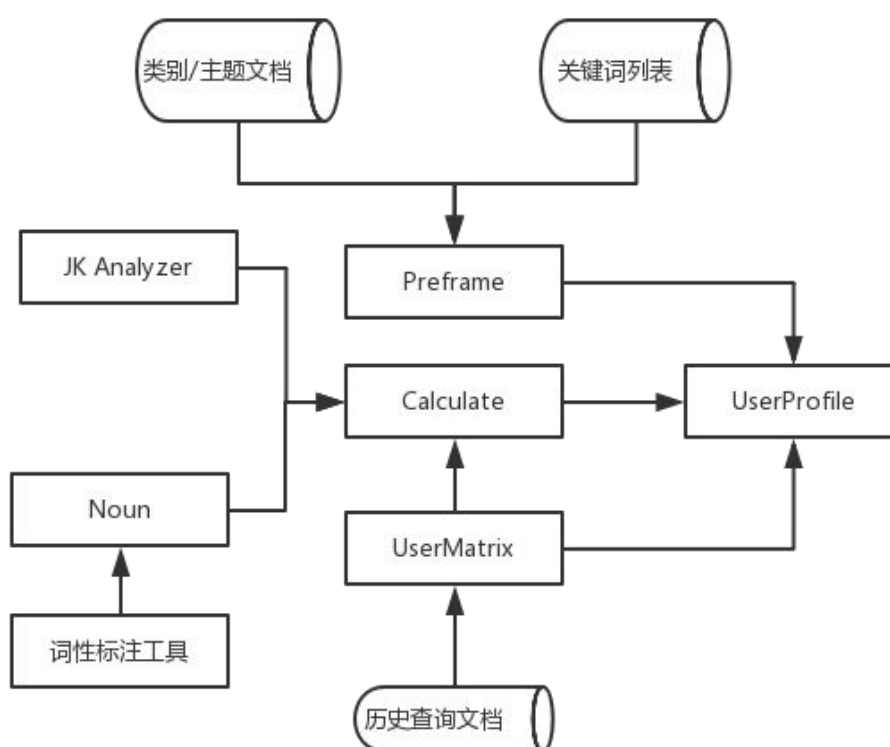


图 3-4 用户模型结构图

## 3.4 实验结果与分析

### 3.4.1 实验目的

本次实验的目的在于测试主题类别文档在初始化用户模型后能否在搜索中起到相应的作用。前面章节已经介绍了显式模型的初始化过程，本文提到了两个关键点，一是主题兴趣类别对用户兴趣特征词的涵盖能力，二是对相似文档匹配能力。

首先，我们测试随着某一领域下文档的增多，系统在此领域上用户建模后，最后的兴趣向量中该领域特征词的数量是否跟着增加，以此测试系统能否从多篇文档中充分收集相关特征词。

其次，建立用户兴趣模型以后，我们需要通过两个方面考察主题类别文档在用于建立用户模型后能否有效匹配出文档中的相似文档，第一个方面，在单一主题类别下，随着同一主题文档的增多，即用户模型中同一类特征词覆盖面的增大是否有助于提高用户模型匹配同类文档的概率；第二个方面，随着兴趣主题类别的增多，即用户兴趣如果在后来变的更加多样和广泛，那么模型对原先兴趣类别文档匹配数应该下降以满足其它兴趣文档的匹配出现。

本文在网上选择了 180 篇文档，分为 6 个主题类，有体育、政治、经济、文化、科学、汽车。将这 180 篇文档作为匹配文档库。另外选择 10 篇经济类以及若干篇其它领域的文档，将这些文档随机组合，模拟多个用户兴趣模型。具体测试实验如下：

第一环节，增加相同主题文档。从 6 个主题类中选择经济类作为用户兴趣描述文件。首先拿出一篇经济类文档进行模型创建，模型名称为 ex1，然后每次追加一篇经济类型文档构建用户模型（ex2-ex7）；

第二环节，增加不同主题类型文档。首先经济类 1 篇文档作为用户描述文件构建用户模型 ex1，然后每次增加 1 篇经济类文档，创建 4 个用户模型（ex1-ex4），然后每次增加一篇其它主题类型的文档，创建 3 个用户模型（ex5-ex7），完成以后，通过相似度计算函数对不同的模型与主题库中的 180 篇文档进行匹配打分。根据相似值对主题库中的文档进行排序，取出分别取出前 15 篇、20 篇、25 篇文档作为系统推荐给用户的匹配结果。

第一个环节中随着经济文档增加，统计了用户模型中经济类的特征词，其变化情况如图 3-5 所示：

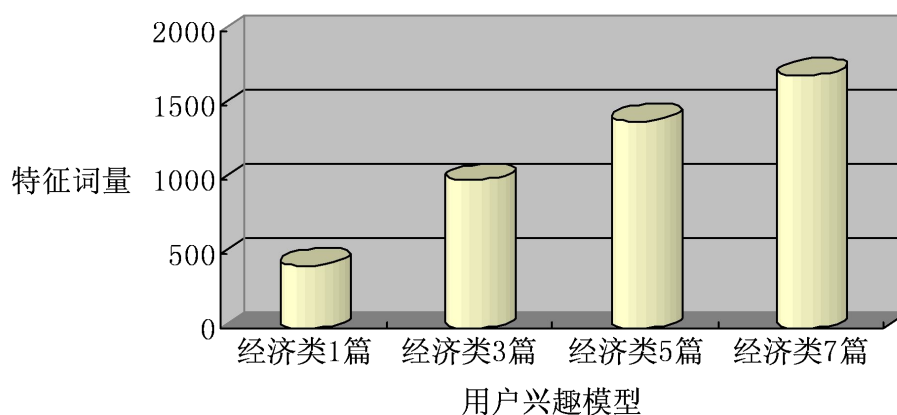


图 3-5 单一主题类型特征词数量变化效果图

图 3-5 分别对经济类描述文档在一篇、三篇、五篇、七篇时作了特征词的统计。其结果证明了增加单一主题用户兴趣描述文档，构建出的用户模型能捕获出新的用户兴趣特征词，提高该类型特征词的覆盖面，本文将提取到的特征词以向量形式保存，采用 HashMap 实现，如下所示：

{人民币=0.6667,美元=0.3338,通货膨胀=0.5558}

推荐系统的性能评价普遍采用查准率和查全率来衡量，它们能够很好地检测系统能否按照用户的要求，准确、全面的检索出相关文档。它们是反映检索效果的重要指标，本文用它们来测试系统中主题类库匹配和推荐相关文档的效果如何。查全率（Recall）和查准率（Precision）。其计算公式如下：

$$\text{查准率} = \frac{\text{被推荐文档中符合要求的文档}}{\text{被推荐的所有文档}} \quad (3-19)$$

$$\text{查全率} = \frac{\text{被推荐文档中数据类的文档数量}}{\text{资源库中属于经济类的文档总数}} \quad (3-20)$$

### 3.4.2 实验结果分析

根据以上查准率和查全率计算公式，在系统推荐的结果中，分别取推荐结果的前 15、20、25 文档的情况下得到实验数据，结果如下所示：

第一个环节测试数据：

表 3-5 单一主题类型下文档推荐匹配结果

	推荐数量	属于经济主题文档数量	总的经济主题文档数	Precision	Recall
1	15	6	35	0.40	0.17142
	20	12	35	0.60	0.34285
	25	14	35	0.56	0.4
2	15	8	35	0.53333	0.22857
	20	13	35	0.65	0.37142
	25	15	35	0.60	0.42857
3	15	10	35	0.66667	0.28571
	20	14	35	0.7	0.40
	25	16	35	0.64	0.45714
4	15	11	35	0.73333	0.34285
	20	16	35	0.80	0.45714

	25	18	35	0.72	0.51428
5	15	12	35	0.80	0.34285
	20	17	35	0.85	0.48571
	25	20	35	0.80	0.57142
6	15	13	35	0.86667	0.37142
	20	18	35	0.90	0.51428
	25	22	35	0.88	0.62857
7	15	14	35	0.93333	0.40
	20	19	35	0.95	0.54285
	25	24	35	0.96	0.68571

我们将这个表格转换为趋势图（3-7）如下：

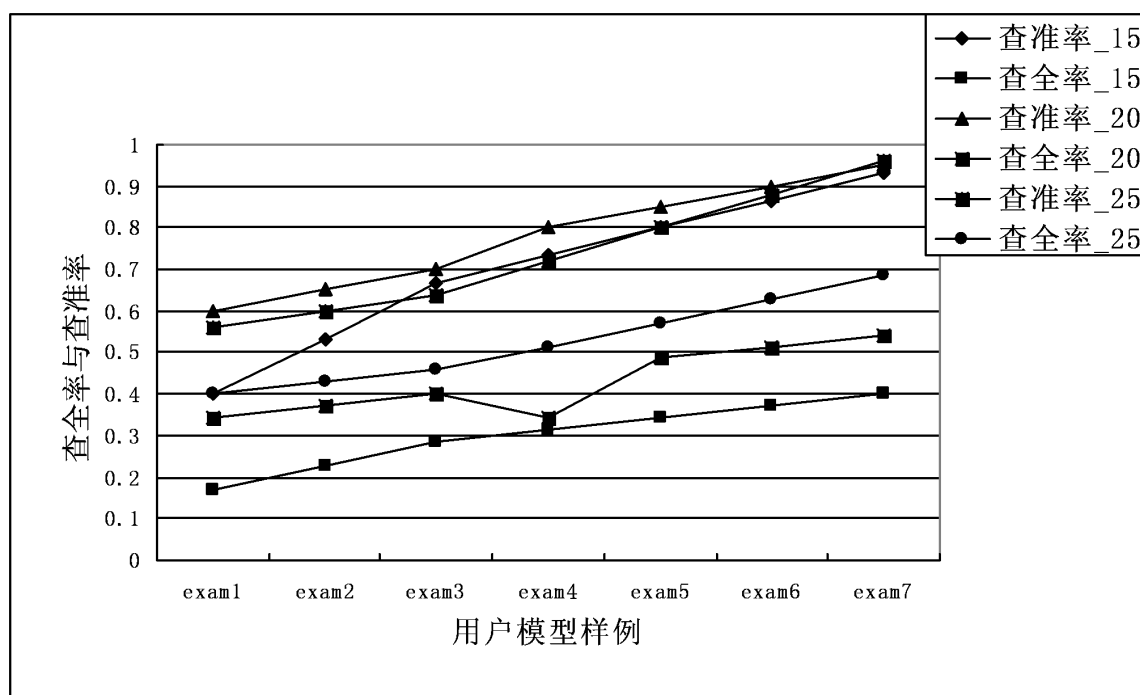


图 3-7 单一主题类型下查准率与查全率折线分布图

上图中每一个 exam（样例）代表一次文档相似得分计算，参与文档数量表示本次用于构建用户模型的用户描述文档数量。并通过查准率和查全率计算公式，计算出在不同的用户模型下文档推荐的 Precision 值和 Recall 值。

第二部分测试数据：

表 3-6 增加不同文档类型下推荐匹配结果

	推荐数量	主题数量	总主题文档数	Precision	Recall
1	15	6	35	0.40	0.17142
	20	12	35	0.6	0.34285
	25	14	35	0.56	0.4
2	15	8	35	0.53333	0.22857
	20	13	35	0.65	0.37142
	25	15	35	0.60	0.42857
3	15	10	35	0.66667	0.28571
	20	14	35	0.7	0.40
	25	16	35	0.64	0.45714
4	15	11	35	0.73333	0.34285
	20	16	35	0.80	0.45714
	25	18	35	0.72	0.51428
5	15	10	35	0.66667	0.28571
	20	15	35	0.75	0.42857
	25	17	35	0.68	0.48571
6	15	8	35	0.53333	0.22857
	20	13	35	0.65	0.37142
	25	15	35	0.60	0.42857
7	15	5	35	0.33333	0.14285
	20	10	35	0.5	0.28571
	25	12	35	0.48	0.34285

第二部分测试数据转换成趋势图如 3-8 所示：

图 3-7 中，随着经济类文档的累加，用户模型推荐的经济类文档比率以及数量均呈上升趋势，表明用户描述文档中所积累的该领域词汇在不断增多，用户模型能够动态捕获到该领域新的词汇，能够随着领域特征词的增多完善用户模型，有效提高匹配用户兴趣文档的概率。图 3-8 中，随着在 ex4 至 ex7 非经济类主题文档的增加，Precision 和 Recall 均是呈下降趋势。说明用户兴趣爱好发生了偏移。用户模型匹配出经济类主题文档数量下降，以使其它类别兴趣的文档被推荐出来。通过实验我们可以证明基于显隐式结合的用户兴趣建模方法是能够有效的表达出用户的兴趣模型的，同时也验证了本文提出的改进型的隐式用户建模方法。



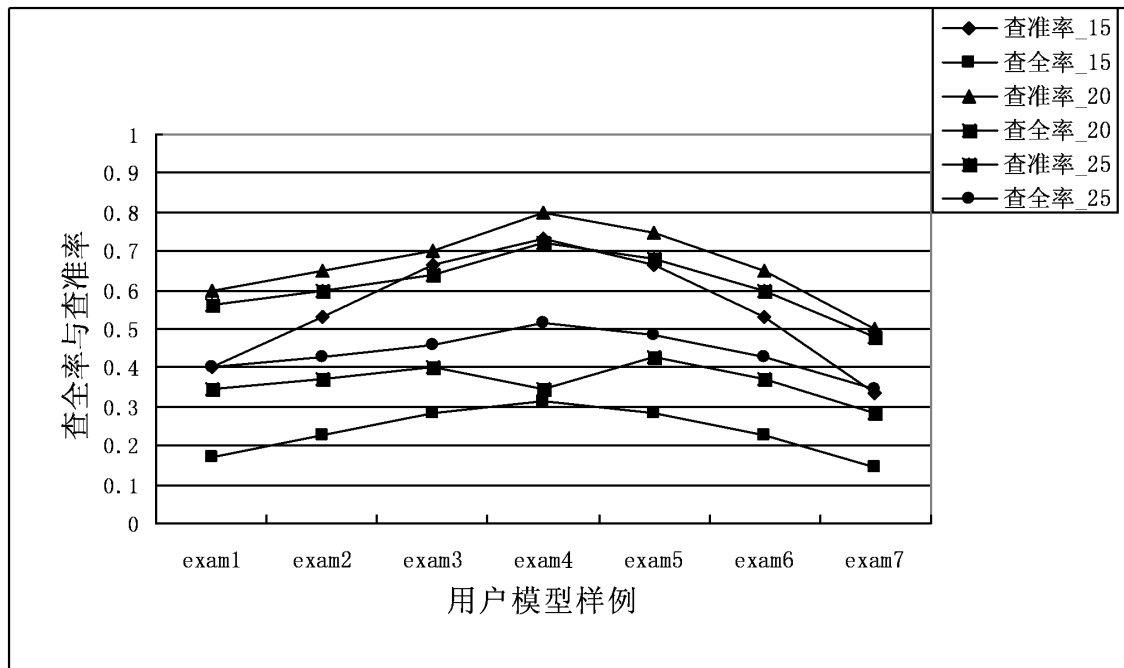


图 3-8 不同主题类型下查准率与查全率折线分布图

### 3.5 本章小结

本章通过分析显式和隐式用户建模的优缺点，提出了基于显隐式信息建模方法。完成了基于显隐式用户兴趣建模的流程步骤，基于原有的隐式用户建模方法，改进了隐式建模技术。实现了基于显隐式用户模型的建立。介绍了在建模过程中用到的分词技术。通过实验验证，证明了显隐式用户模型能更好的表达用户的兴趣信息。

## 第四章 基于用户模型的协同过滤个性化广告推荐技术

本章将基于上一章节创建的用户模型，使我们设计的广告推荐系统具有个性化推荐，本文将对基于用户模型的个性化广告推荐系统设计整体框架。从广告内容方面以及广告推荐具体方法做详细讲解。

### 4.1 协同过滤推荐技术

协同过滤技术（CF，Collaborative Filtering），协同过滤技术是在 1992 年由 DavidGoldberg 提出的，CF 技术是使用率最高的推荐技术。CF 最早是为了解决 PARC 机构的资讯超载问题。目前这一技术被广泛的应用到广告推荐系统中<sup>[46]</sup>。

在平常的工作和生活中，我们在做出决定之前，需要听取身边的朋友或同学的建议，融合多种观点以后，做出决定。CF 算法就是基于这样一种思想而实现的。算法的基础是相似用户具有相近的兴趣偏好，只要知道相似用户的兴趣喜好，就能预测目标用户的兴趣喜好。

协同过滤推荐过程分为三步，第一步是用户的兴趣表示，一般的表示都是用户-项目评分矩阵；第二步是近邻集合生成；第三步是项目推送。下面将以此介绍协同过滤推荐算法的三个步骤。

#### (1) 用户兴趣表示

我们把用户对项目（广告）的评价作为用户的兴趣偏好，CF 算法用一个  $m \times n$  维的用户-项目评分矩阵。其中， $m$  表示用户数目， $n$  表示广告类型数目，形式如表 4-1 所示， $R_{ij}$  代表用户  $User_i$  对项目  $Item_j$  评分。 $R_{ij}$  有多种表现形式，如 1 或 0（分别代表感兴趣或不感兴趣），具体分值等。

表 4-1 用户-项目评分矩阵

	$Item_1$	$Item_2$	...	$Item_n$
$User_1$	$R_{11}$	$R_{12}$	...	$R_{1n}$
$User_2$	$R_{21}$	$R_{22}$	...	$R_{2n}$
...	...	...	...	...
$User_m$	$R_{m1}$	$R_{m2}$	$R_{m3}$	$R_{mn}$

## (2) 近邻用户的选择

这一步的工作是寻找与目标用户有相同兴趣爱好的用户，将这些相似用户组合在一起称之为近邻用户集合，根据近邻用户对项目（广告）的评分，给目标用户提供个性化广告推荐。这一步的数据源就是第一步我们创建的用户-项目评分矩阵，通过计算用户之间的相似度，选择出  $K$  个相似度较高的用户作为近邻。用户之间的相似度计算方法很多种，下面本文列举几个常用的方法：皮尔森相关相似度（PCS）、余弦相似度(CS)和修正的余弦相似度(ACS)等。

### 1) 余弦相似度

余弦相似度是计算文档相似性常用的方法，常常被用于信息检索领域<sup>[52]</sup>。余弦相似度计算数据来源是用户-项目矩阵，用户之间的相似度通过两个向量之间夹角大小来体现。例如，设置用户  $u_i$  和用户  $u_j$  对项目的评分分别为向量  $I_i$  和向量  $I_j$ ，则用户  $u_i$  和用户  $u_j$  的余弦相似度为：

$$\text{sim}(u_i, u_j) = \cos(I_i, I_j) = \frac{I_i \cdot I_j}{|I_i| \cdot |I_j|} = \frac{\sum_{k=1}^n R_{ik} R_{jk}}{\sqrt{\sum_{k=1}^n R_{ik}^2} \sqrt{\sum_{k=1}^n R_{jk}^2}} \quad (4-1)$$

其中， $R_{ik}$ ， $R_{jk}$  分别是用户  $u_i, u_j$  对项目的评分。当用户-项目评分矩阵较为稀疏或者数据两极分化特别严重时，此时余弦相似度不能很好的衡量用户之间的相似性了。

### 2) 皮尔森相关相似度

PCS 方法是基于不同用户对相同两个项目之间的评分差来衡量用户之间相似度的，相关系数越高，则相似度越大，反之则越小。PCS 的取值范围时 $[-1,1]$ 。假设用户  $u_i, u_j$  都有评价的项目集合，皮尔森相似度计算公式为：

$$\text{sim}(u_i, u_j) = \frac{\sum_{c \in I_{ij}} (R_{i,c} - \bar{R}_i)(R_{j,c} - \bar{R}_j)}{\sqrt{\sum_{c \in I_{ij}} (R_{i,c} - \bar{R}_i)^2} \sqrt{\sum_{c \in I_{ij}} (R_{j,c} - \bar{R}_j)^2}} \quad (4-2)$$

$$\bar{R}_i = \frac{1}{n} \sum_{k=1}^n R_{ik}$$

公式中， $R_{i,c}$ 、 $R_{j,c}$  分别指两个用户对项目  $c$  的评分， $\bar{R}_i = \frac{1}{n} \sum_{k=1}^n R_{ik}$  表示用户对项目的评分均值。

皮尔森相关方法能够解决用户对广告评分严谨度不同的问题，这种方法要求变量值是连续的，具有线性相关等假设条件，当这些条件不满足时，相似度的准确性也会降低。

### 3) 修正的余弦值相似度

上面我们介绍了普通的 CS 方法计算用户之间的相似性，但是这种方法对用户评分严谨性要求比较高。因此，改进了 CS 计算方法，称为修正的余弦值相似性，计算公式为：

$$\text{sim}(u_i, u_j) = \frac{\sum_{k=1}^n (R_{ik} - \bar{R}_i)(R_{jk} - \bar{R}_j)}{\sqrt{\sum_{k=1}^n (R_{ik} - \bar{R}_i)^2} \sqrt{\sum_{k=1}^n (R_{jk} - \bar{R}_j)^2}} \quad (4-3)$$

其中， $\bar{R}_i = \frac{1}{n} \sum_{k=1}^n R_{ik}$  表示用户的项目评分均值。

### (3) 广告推荐

协同过滤推荐技术建立在相似用户对同一项目有相似评分这一假设之上，通过第二步得出的近邻集合中的用户对项目（广告）的评分，对目标用户进行项目推荐，一种推荐方式是用户对任意一个项目进行评分预测。另一种推荐方式是置顶推荐，即对用户没有评价过的项目进行预测评分，将评分值最高的项目推荐给目标用户。常用的预测方法有：

$$P_{ik} = \frac{\sum_{u_j \in NNU} R_{jk}}{n} \quad (4-4)$$

$$P_{ik} = \frac{\sum_{u_j \in NNU} \text{sim}(u_i, u_j) \times R_{jk}}{\sum_{u_j \in NNU} \text{sim}(u_i, u_j)} \quad (4-5)$$

$$P_{ik} = \bar{R}_i + \frac{\sum_{u_j \in NNU} \text{sim}(u_i, u_j) \times (R_{jk} - \bar{R}_j)}{\sum_{u_j \in NNU} \text{sim}(u_i, u_j)} \quad (4-6)$$

$P_{ik}$  代表用户  $u_i$  对项目  $k$  的预测评分， $\text{sim}(u_i, u_j)$  代表两个用户之间的相似度， $R_{jk}$  代表用户  $u_j$  对项目  $k$  的评分， $NNU$  表示用户  $u_i$  的近邻集合， $NNU$  中的用户数量为  $k$ ，用户  $u_i$  与用户  $u_j$  对项目的均值分别为  $\bar{R}_i$ 、 $\bar{R}_j$ 。

公式(4-4)是通过近邻用户对项目的评分均值，对用户没有进行评价过的项目进行评分。公式(4-5)体现出了近邻集合中用户之间的差异性，给近邻用户加以权重描述。

公式(4-6)考虑不同用户的评分严谨度不同，利用平均值修正该差异引发的预测差值，在评分的预测精度上优于前两个公式。

## 4.2 基于用户模型的协同过滤个性化广告推荐技术

### 4.2.1 个性化广告推荐系统框架设计

第三章中我们讨论了用户兴趣模型，以及用户模型的创建和模型的改进。用户兴趣模型中的主题词能够在一定程度上代表用户的兴趣，基于这些主题词，我们选择相应的广告推荐给用户。首先是发现用户兴趣，然后创建用户兴趣模型，最后实现基于用户兴趣模型实现个性化广告推荐，图 4-1 表示基于用户兴趣模型的广告推荐系统框架。

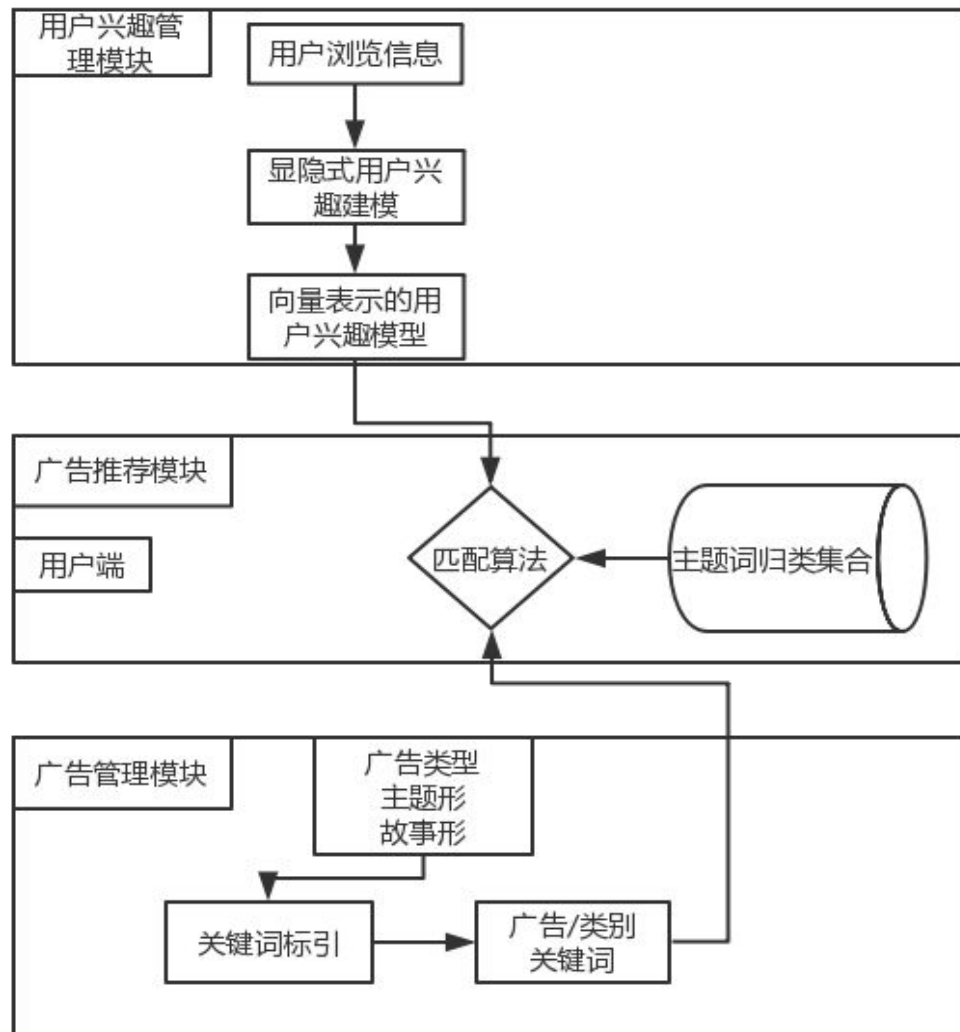


图 4-1 基于用户模型的个性化广告推荐系统框架图

通过上图可知，整个系统中心部分是用户模型、广告和推荐。

#### (1) 用户兴趣管理模块

用户兴趣模块主要是进行用户兴趣模型的创建以及实时更新，获取能够代表用户兴趣的最新的主题词，第三章提出用户兴趣模型将在这个模块实现。

系统将用户的上网日志记录、搜索关键字等信息处理成被模型识别的数据后，输入到基于显隐式用户兴趣模型中进行主题词的提炼。该模型能帮我们实现用户的上网日志信息转变为兴趣主题词。将我们生成的由主题词代表的用户兴趣模型直接转交给系统的推荐模块。

#### (2) 广告管理模块

该部分的主要功能是管理丰富多样的广告资源，本系统所涉及到的广告资源主要是标签性的广告。

为了实现广告与用户兴趣匹配的自动化，广告库中的所有广告必须有关键词，关键词的选择需要依据广告内容、样式、广告主题等方面综合考虑。将广告进行关键词或则是主题分类以后，就能够与用户兴趣模型完成匹配。

#### (3) 广告推荐模块

广告推荐模块是整个系统架构的核心，当用户兴趣管理模块和广告管理模块分别将“用户-兴趣主题词”与“广告-类别关键词”输出到推荐模块后，用户与广告的匹配就将在该模块中完成。首先，系统需要有一个主题词集合，该集合涵盖全部的广告类型，并且关键词集合要与广告建立关联。集合的作用是为了兴趣主题词与广告类型提供参考。在实际的应用当中，主题词库需要不断更新完善。

在广告主题词集合构建完成以后，推荐模块将基于此，选用性能更优的匹配算法将用户兴趣中对应的主题词与广告类型主题词进行匹配，输出“用户-广告”相对应的序列。

基于上述的三个模块构成的系统框架，我们实现了用户兴趣模型模块的工作，接下来，本文将从推荐模块实现本文的创新，实现在用户兴趣模型的基础之上，研发出个性化广告系统。

### 4.2.2 基于用户兴趣模型的个性化广告推荐算法

上面我们介绍了传统的协同过滤推荐算法，当系统中的广告类型急剧增多以后，导致用户-项目评分矩阵的极度稀疏性，文献<sup>[48]</sup>提出利用预测评分的方法将未评分项目进行估值来解决矩阵的稀疏性。这种方法实现起来比较复杂，同时预测值与用户真实态度具有一定的，不能从根本上解决矩阵稀疏问题。

本章提出在使用协同过滤算法的基础之上将用户兴趣模型融合到算法当中，利用第三章提出的用户兴趣模型来表示用户的兴趣爱好。个性化广告推荐算法如下：

**Step1:** 将用户注册的信息和上网历史信息利用第三章提出的方法创建用户模型。用户兴趣模型用向量形式表示。表 4-2 表示用户兴趣模型矩阵:

表 4-2 用户兴趣模型的矩阵表示

	$F_1$	$F_2$	$F_3$	.....	$F_n$
$User_1$	$W_{11}$	$W_{12}$	$W_{13}$	.....	$W_{1n}$
$User_2$	$W_{21}$	$W_{22}$	$W_{22}$	.....	$W_{2n}$
$User_3$	$W_{31}$	$W_{32}$	$W_{32}$	.....	$W_{3n}$
.....	.....	.....	.....	.....	.....
$User_m$	$W_{m1}$	$W_{m2}$	$W_{m3}$	.....	$W_{mn}$

其中,  $W_{ij}$  表示用户  $U_i$  对主题词  $F_j$  的兴趣度。

**Step2:** 使用余弦值相似性公式计算用户之间的相似度。对用户的兴趣度计算是客观属性, 无需要考虑用户之间对广告评分的差异化。

**Step3:** 计算用户对所有广告的兴趣度。利用用户兴趣模型向量建立线性回归方程, 计算用户对任何一个项目(广告)的兴趣度。如公式(4-7), 并使用 Top-N 推荐技术, 将项目推送给目标用户。

用户  $u_i$  对广告  $p_j$  兴趣度值的计算公式为:

$$I_{ij} = \sum_{t_k \in T} w_k \times p_k^j \quad (4-7)$$

其中,  $p_k^j = \begin{cases} 0 & t_k \notin \text{广告的特征词集合} \\ 1 & t_k \in \text{广告的特征集合} \end{cases}$

$t_k$  是主题类特征词,  $T$  为用户  $u_i$  的兴趣模型中兴趣项的集合;  $w_k$  为兴趣项  $t_k$  的权重。项目(广告)  $p_j$  的特征项集合是特征集合以及广告  $p_j$  的自身属性集合。

### 4.3 实验设计与分析

一个推荐系统是能否实现个性化, 最为重要是能够创建一个符合用户的兴趣模型, 并且使用精准的推荐算法将用户兴趣与广告主题相结合。在上一张中我们实现并验证了本文提出的基于显隐式信息创建用户兴趣模型的有效性, 并且能够随着用户兴趣发生偏移而改变。本章我们提出了基于用户模型的个性化推荐算法, 能够提升推荐系统的精度和准确度。本节实验将加以验证。

### 4.3.1 MovieLens 数据集介绍

实验中采用的数据是 MovieLens 数据集，该集合中包含用户对电影的评价信息以及电影的分类信息，是推荐算法常用的数据集。数据集合中的评分信息如表 4-3 所示。

表 4-3 用户项目评分数据集合

UserID	ItemID	Score	Time
1	12	2	1477295722
2	13	4	1477295724
3	58	1	1477295720

本实验使用的数据集是 100k 的数据集，数据集中有 980 个用户。集合中用户能够进行评价的电影数量为 1703。所有的用户对这些电影产生了大约有 1000000 条评分记录。评分值都是在[1,5]之间的数据。数据集合中的用户都对电影进行评分，并且评分条数不下于 30 条。数据集合中还保留了用户的基本信息。如表 4-4 所示：

表 4-4 用户信息表

UserId	age	Sex	Occupation
1	23	F	Teacher
2	30	W	IT
3	50	F	Worker

数据集收集了从 1997 年到 1998 年共 7 个月的时间。从表 4-3 中，我们发现每一条记录都包括用户 ID、项目 ID、评分值以及用户评分的时间。从表 4-4 用户的基本信息记录中，都会有一个用户 ID 作为用户在数据集合中的唯一标识。然后是一些用户的基本信息，主要包括用户的职业与年龄信息。

在实验数据集中，电影信息表中的信息主要包括电影的唯一 ID，还有就是电影的名称，以及电影属于哪个主题类等。例如表 4-5 中的两个例子：

表 4-5 广告信息表

ItemID	name	Time	Classes
1	Toy Story	01-Jan-1994	000111000000
2	American Beauty	07-May-2008	100100000001

在电影信息表中，一个电影可以属于多个类别。并且我们将电影类别信息转换成向量表示。



### 4.3.2 数据集的划分

本节实验中，由于需要用到训练集与测试集。因此需要将我们的数据集划分成两部分，一部分用于模型的训练称之为训练集，另一部分作为验证结果称之为测试集，验证实验的性能。本次实验中所用到的数据集是 MovieLens 数据集，利用切分数据集的脚本工具将数据集分成 5 分，切分后的数据集分成 5 个训练集与测试集对，每一个训练集与测试集对代表一次切分，按照 80%/20% 的原则把原来的数据集一分为二。一次切分只能保证得到测试集是原评分数据集的 20%，所以需要一共切分 5 次，并且保证各次切分所得到的测试集都是原评分数据集不相交的子集。

切分之后的集合，在做实验时为了能够在整个训练集合上得出均衡的结果，应该在切分之后的 5 个集合对上都运行一次，取 5 次结果的平均值。

### 4.3.3 评价方法

任何一个推荐系统完成以后，都需要对于其推荐质量进行评价，这是必不可少的环节。并且在不同的环境下，不同的推荐算法或则是不同的推荐内容，对系统推荐质量的评价标准和方法也是不相同的。下面介绍在推荐系统中常用的几种方法。

#### (1) 误差平均值法 (MAE)

推荐系统中对未进行评分的广告进行评分得出评分值，这个预估值与真实的值之间存在一定的差异。MAE 算法就是计算两者之间的差值，然后取绝对值。MAE 的绝对值小，说明预测的评分值与真实值比较接近，系统的推荐算法的准确性就比较高。MAE 的计算方法如下：

$$MAE = \frac{\sum_{i \in U, j \in I} |p_{ij} - r_{ij}|}{n} \quad (4-8)$$

其中， $p_{ij}$  代表用户  $i$  对 Item  $j$  通过推荐系统的预测评分值， $r_{ij}$  是用户  $i$  对广告  $j$  的真实评分。 $U$  是用户集合， $I$  是要推荐的广告集合。 $n$  为广告的评分总数。

#### (2) 查全率(recall)

查全率，也可称为召回率，表示被正确推荐的广告项目数量在全部测试集中所占的比值。我们假定用户  $i$  的测试集合为  $T_i$ ，正确推荐的项目集合为  $P_i$ ， $n$  为总的用户数量。则查全率计算公式为：

$$recall = \frac{1}{n} \sum_{i=1}^n \frac{|T_i \cap P_i|}{|T_i|} \quad (4-9)$$

#### (3) 查准率 (precision)

查准率，也称之为准确率，查准率主要针对 Top-N 推荐的时候，正确推荐的项目数量在 Top-N 中的比例。计算查准率的方式如下：

$$precision = \frac{1}{n} \sum_{i=1}^n \frac{|T_i \cap P_i|}{N} \quad (4-10)$$

公式中，N 表示系统推荐出的项目数量，n 代表总的用户数量。

#### 4.3.5 实验结果分析

本小节根据前面章节提出传统的 CF 算法和基于用户兴趣模型的 CF 算法简称为 UIM-CF，给出两个算法的性能指标。本文根据前面提到评测方法如准确率、查全率以及覆盖率三个方面评价实验结果。实验中我们使用 top-N 推荐，并且每次推荐给用户的广告个数为 10 个。下面给出传统的协同过滤算法和基于用户模型的协同过滤算法的性能指标，图 4-2 是准确率对比图，图 4-3 是覆盖率对比图。

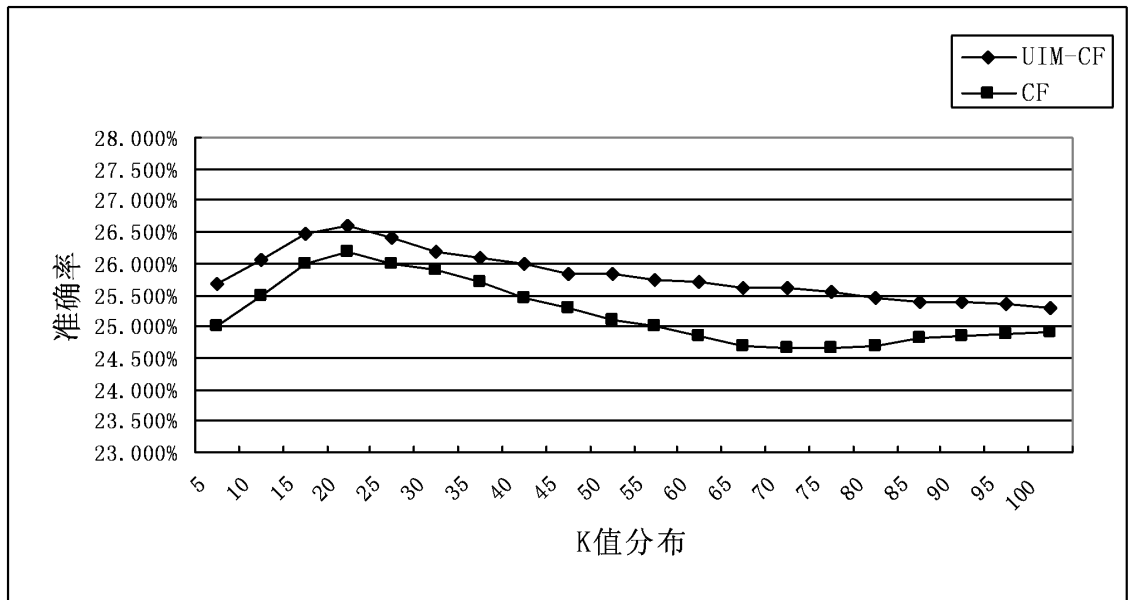


图 4-2 CF 与 UIM-CF 的准确率对比结果

从图 4-2 中可以看出，UIM-CF 相对于传统的 CF 在评价指标的两个参数 precision 和 recall 的值都有提升，并且两个参数的曲线分布相一致。

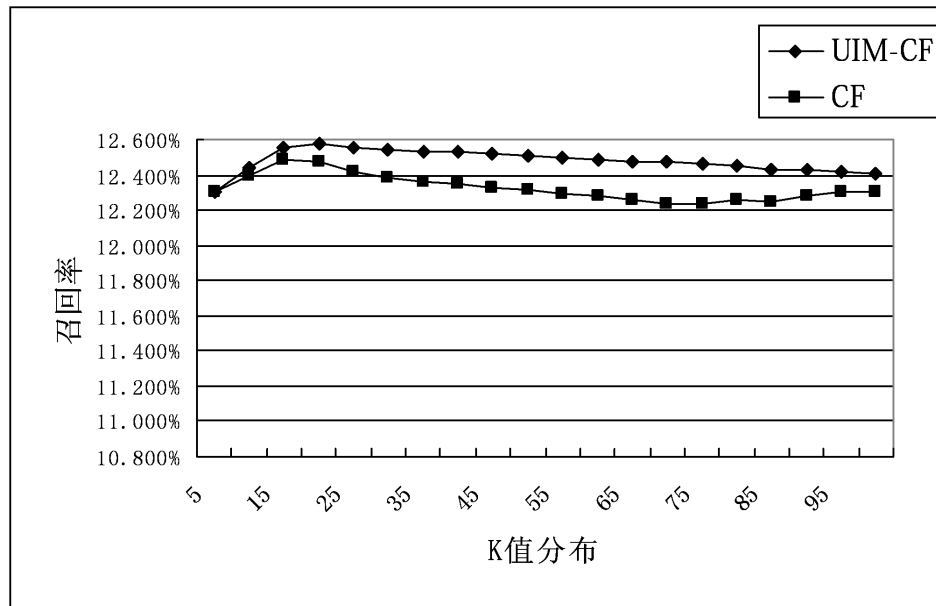


图 4-3 UIM-CF 与 CF 的召回率对比结果

从总体上得出，K 值不太大时，M-CF 算法的性能要比 CF 算法好很多。基于实验我们可以看出基于用户模型的 CF 算法比传统的基于用户-项目的 CF 算法性能要好。印证了本文提出的个性化推荐算法的正确性。

#### 4.4 本章小结

在第四章本文详细介绍了 CF 算法的原理和过程，将上一章实验并验证的用户兴趣模型融合到 CF 算法当中。提出一种新的个性化广告推荐方式 UIM-CF 推荐模型，基于 UIM-CF 算法利用用户模型计算用户之间相似性，使用模型的线性回归方程计算用户对广告的兴趣权重，通过 Top-N 推荐方法做出推送。提出该算法的目的是为了体现基于用户模型的个性化广告推荐比传统的广告推荐算法质量要好。实验证明，基于用户兴趣模型能够在一定程度上提高广告的推荐质量。

## 第五章 总结与展望

### 5.1 总结

本文从个性化广告的需求，产生，国内外的的发展以及个性化广告推荐系统的使用等方面做了全面介绍。本文将两个方面作为研究重点，一是用户的兴趣模型，另一个就是将用户兴趣与广告主题相结合的推荐算法。首先围绕用户兴趣模型进行分析研究，通过传统分析传统建模的优点与缺点，在传统的建模基础之上提出新的建模方法。其次，介绍了传统的协同推荐算法，分析其中的优缺点，将本文训练出的用户模型和协同推荐算法相结合，提出了基于用户模型的个性化广告推荐模式。

本文主要工作如下：

(1) 阐述了个性化广告的研究背景及意义。个性化广告基础是用户兴趣模型。本文研究分析了用户模型的建模信息来源、信息获取方法、用户模型的表示以及建模技术等。

(2) 提出了一种融合显式和隐式用户信息的建模技术。显式建模方法的信息源是用户的注册信息。这些信息能够比较准确的体现用户的兴趣，但是这种方法不利于用户的上网体验，有些用户不会主动提交过于隐私的信息。隐式建模方法不需要用户的主动参与，通过挖掘分析用户上网日志记录，得出用户的兴趣偏好，逐步完善用户初始化模型。由于对语义理解不够充分，很难反映出用户的兴趣所在。基于前两种建模方法优缺点，本文提出了显隐式信息结合的用户兴趣建模方法。

(3) 提出一种改进的隐式建模方法，并将改进后的隐式建模方法应用到上面提出的显隐式建模方法中。

(4) 提出了基于用户兴趣模型的协同过滤广告推荐技术。协同过滤技术是根据用户-项目评分矩阵，挖掘出用户相似集，通过相似集中的用户给目标用户推荐信息。本文将上面提出的用户兴趣模型应用到协同过滤算法中，提出了基于用户模型的协同过滤广告推荐技术。

### 5.2 展望

这篇文章通过对用户兴趣模型以及广告推荐的个性化研究中，提出一些新的方法和算法。经过大量实验论证了本文的观点，但是还有一些未考虑到的地方需要更加深入的研究。

(1) 用户模型数据来源问题。

通过什么方式，在不影响用户上网体验情况下收集用户感兴趣的而且比较真实的信息。这是一个急于待解决的问题。

(2) 用户兴趣模型的准确性。

文中对用户兴趣模型的表示采用使用量比较广泛的 VSM 表示方法，VSM 模型表示比较直接，没有考虑到太多的其它因素，比如中文的语义因素等。下一步可以从语义等方面找到一种全新的用户兴趣模型表达方式。

(3) 本文的实验数据采用的是 Movielens 数据集中，在初始化用户特征集合时仅仅使用了电影属性，并没有从其它类型的项目进行特征提取。下一步工作是的重点就是采用多种类型项目对系统进行验证，提高系统的可扩展性。

## 参考文献

- [1]Lucas J P, Luz N, Moreno M, et al. A hybrid recommendation approach for a tourism system[J].Expert Systems with Applications An International Journal, 2013, 40(9):3532-3550.
- [2]Raj J R,Sasipraba T.Quality Web Services Recommendation System Based on Enhanced Personalized Hybrid Collaborative Filtering Approach[J]. 2015, 10(3):249.
- [3]Lim C S,Shin J H,Lee S J.Efficient Book Recommendation System Based on a MapReduce Model[J]. Applied Mechanics & Materials, 2013, 284-287:3405-3408.
- [4]葛译泽.互联网广告精准投放平台设计与实现[D].成都:成都理工大学, 2015.
- [5]林霜梅,汪更生,陈弈秋.个性化推荐系统中的用户建模及特征选择[J]. 计算机工程, 2007, 33(17):196-198.
- [7]赵银春,付关友,朱征宇.基于 Web 浏览内容和行为相结合的用户兴趣挖掘[J]. 计算机工程, 2005, 31(12):93-94.
- [8]许海玲,吴潇,李晓东,等.互联网推荐系统比较研究[J].软件学报, 2009, 20(2):350-362.
- [9]Bollacker K D, Lawrence S, Giles C L. A system for automatic personalized tracking of scientific literature on the Web[C]// ACM Conference on Digital Libraries.ACM, 2013:105-113.
- [10]Agrawal R,Gupta A,Prabhu Y, et al.Multi-label learning with millions of labels:recommending advertiser bid phrases for web pages[J]. 2013, 28(5):13-24.
- [11]Foltz P W, Dumais S T.Personalized information delivery: an analysis of information filtering methods[J]. Communications of the Acm, 2013, 35(35):51-60.
- [12]Klein K,Scholl J H,Vermeer N S,et al.Traceability of Biologics in The Netherlands: An Analysis of Information-Recording Systems in Clinical Practice and Spontaneous ADR Reports[J]. Drug Safety, 2016, 39(2):1-8.
- [13]Bellegarda J R, Naik D,Silverman K E A. Method and apparatus for filtering email: US, US 7836135 B2[P]. 2010.
- [14]Zenil H, Soler-Toscano F, Dingle K, et al. Correlation of automorphism group size and topological properties with program-size complexity evaluations of graphs and complex networks[J]. Physica A Statistical Mechanics & Its Applications, 2014, 404(16):341-358.
- [15]Bartolucci F,Farcomeni A,Pennoni F.An overview of latent Markov models for longitudinal categorical data[J]. Longitudinal Research with Latent Variables, 2010:1-36.

- [16]Huang Y, Gao X, Gu S. Collaborative Filtering Recommendation Algorithm Based on User Acceptable Rating Radius[M]// LISS 2013.Springer Berlin Heidelberg, 2015:141-146.
- [17]宗成庆.统计自然语言处理[M].北京:清华大学出版社, 2013.
- [18]马晓迪.基于网络结构的个性化推荐系统的研究与开发[D].杭州:浙江工业大学, 2014.
- [19]Lemire D, Maclachlan A. Slope One Predictors for Online Rating-Based Collaborative Filtering[J]. Computer Science, 2008:21-23.
- [20]赵志荣.个性化搜索引擎的研究、设计与实现[D].成都:四川大学, 2002.
- [21]Kim H J,Zhu Y,Kim W,et al. Dynamic faceted navigation in decision making using Semantic Web technology[J]. Decision Support Systems, 2014, 61:59-68.
- [22]Daoud M,Tamine L,Boughanem M,et al.Learning Implicit User Interests Using Ontology and Search History for Personalization[C]// International Conference on Web Information Systems Engineering. Springer Berlin Heidelberg, 2007:325-336.
- [23]Fernandez F M H,Ponnusamy R.Categories of Web User Behaviour Models and Information Retrieval – A Survey[C]// 2014:31-35.
- [24]Codocedo V,Napoli A.Formal Concept Analysis and Information Retrieval – A Survey[J]. 2015, 9113:61-77.
- [25]Alsarem.A generic approach based on Linked Data to enhance Web information retrieval and increase user satisfaction[C]// CORIA. 2013.
- [26]Mamchich A A.Models and Algorithms of Information Retrieval in a Multilingual Environment on the Basis of Thematic and Dynamic Text Corpora[J]. Cybernetics & Information Technologies, 2016, 16(1):99-115.
- [27]Kurimo M.Multilingual Information Access Evaluation I - Text Retrieval Experiments, 10th Workshop of the Cross-Language Evaluation Forum, CLEF 2009, Corfu, Greece, September 30 - October 2, 2009, Revised Selected Papers, Part I.[M]. Springer Berlin Heidelberg, 2010.
- [28]Dewandaru A, Supriana S I, Akbar S. Evaluation on geospatial information extraction and retrieval: Mining thematic maps from web source[C]// International Conference on Information and Communication Technology. IEEE, 2015:283-288.
- [29]罗莉娟.基于网络生活方式的综合价值个性化推荐机制研究[D]. 北京:北京邮电大学, 2015.
- [30]Gretarsson B, Donovan J, Bostandjiev S, et al. TopicNets: Visual Analysis of Large Text Corpora with Topic Modeling[J].Acm Transactions on Intelligent Systems & Technology, 2012, 3(2):23.

- [31]Berendt B, Kralisch A. A user-centric approach to identifying best deployment strategies for language tools: the impact of content and access language on Web user behaviour and attitudes.[J]. Information Retrieval Journal, 2009, 12(3):380-399.
- [32]Steichen B, Ghorab M R, O'Connor A, et al. Towards Personalized Multilingual Information Access - Exploring the Browsing and Search Behavior of Multilingual Users[C]// International Conference on User Modeling, Adaptation, and Personalization. Springer International Publishing, 2014:435-446.
- [33]Li W C, Liu J G. Design and Implementation of the Personalized Search Engine Based on the Improved Behavior of User Browsing[J]. Research Journal of Applied Sciences Engineering & Technology, 2013, 5(4).
- [34]Zhang Z, Li Y. Research and implementation of the personalized meta search engine based on ontology[C]// Web Society, 2009. Sws '09. IEEE Symposium on. IEEE, 2009:180-183.
- [35]武成岗,焦文品,田启家,等.一种基于本体论和多主体的信息检索服务器[C]// 中国人工智能联合学术会议. 2001.
- [36]李勇.智能检索中基于本体的个性化用户建模技术及应用[D].长沙:国防科学技术大学, 2002.
- [37]Li Q S, Zou Y X, Sun Y C. Ontology based user personalization mechanism in meta search engine[C]// International Conference on Uncertainty Reasoning and Knowledge Engineering. IEEE, 2012:230-234.
- [38]Kumar N, Prasad S G R. Information Retrieval based on Content and Location Ontology for Search Engine (CLOSE)[J]. 2014.
- [39]宫玲玲, 乔鸿. 移动信息服务中用户兴趣建模研究[J]. 网络安全技术与应用, 2012(12):33-35.
- [40]Hu Y, Song L. Ontology based user model for personalized agriculture search[C]// Robotics and Applications. IEEE, 2012:475-479.
- [41]Balabanovic M. and Shoham Y Learning Information Retrieval Agents: Experiments with Automated Web Browsing[C], In: Proceedings of the AAAI Spring Symposium Series on Information Gathering from 1-Ieterogeneous, Distributed Environments, March, 1995: 13-18.
- [42]Pazzani M J, Muramatsu J, Billsus D. Syskill & Webert: identifying interesting Websites[C], In Proceedings of the 13<sup>th</sup> National Conference on Artificial Intelligence and 8<sup>th</sup> Innovative Applications of Artificial Intelligence Conference. Menlo Park, CA: AAAI Press, 1996: 54-61.



- [43] Billsus D. and Pazzani M. A Hybrid User Model for News Story Classification[C], In: Proceedings of the 7th International conference on user modeling (ITM'99), Baff, Canada: 1999: 98-108.
- [44] Papadimitriou C H. Computer and System Latent semantic indexing: A probabilistic analysis [J], Journal of Sciences, 2000, 61(2): 217-235.
- [45] 李振星, 陆大珏, 任继成, 等. 基于潜在语义索引的 Web 信息预测采集过滤方法[J]. 计算机辅助设计与图形学学报, 2004, 16(1): 142-147.
- [46] 吴志媛. 基于潜在语义索引的 Web 文本挖掘[D]. 无锡: 江南大学, 2013.
- [47] 李扬. 基于向量空间模型的信息检索技术的探讨[J]. 商情, 2013(18): 168-168.
- [48] 李晓笛. Web 文本挖掘技术研究及应用[D]. 北京: 北京交通大学, 2015.

## 攻读学位期间主要的研究成果

### 发表的学术论文和专著:

- [1] 朱艳辉, 田海龙, 张永平,等. 一种基于两因素相结合的自适应学习三支决策阈值的算法[J]. 小型微型计算机系统, 2016, 37(6):1303-1307.
- [2] 朱道杰, 朱艳辉. 基于用户模型的广告推荐系统[J]. 信息与电脑(理论版), 2016, (13):155-156.
- [3] 张永平, 朱艳辉, 朱道杰, 王天吉, 李飞. 基于本体特征的汽车领域命名实体识别[J]. 湖南工业大学学报, 2016, (06):39-43.

### 参与项目:

- [1] 网际星辰文化传媒公司的广告推荐系统开发
- [2] 快连无线路由器设备中 wifi 认证, 数据审计开发

### 获奖情况:

- [1] 校学业奖学金 三等奖

## 致 谢

三年硕士研究生生活即将结束，在这段时间内，我学习到了很多东西，也认识了对自己很有帮助的老师 and 同学们。回首研究生生活，自己每天在研究所、食堂和宿舍三点一线式的生活，看起来索然无味，但其中的乐趣只有我自己才能深刻的体会。每一个研究生的背后都有一个为你辛勤付出为你指导的默默无闻的导师，还有一群和你志同道合的师兄师弟，师姐师妹以及玩得好的同学，这些人都值得我感谢。

我最要感谢的就是我的导师朱艳辉教授，导师是一个很慈祥也很负责的人，为我的课题研究给出了很多启示与指导，使我很快进入到研究生的学习和生活中。导师给我的帮助融入到了生活和科研的点点滴滴中，在科研上，朱老师的指导使我进入到了数据挖掘以及算法研究这个全新的领域，自己在这个领域中收获颇丰。在生活中，只要有重大的节日朱老师都会组织聚会、聚餐，使得在外读书的自己，总能找到家的感觉，再次对朱老师说声，由衷的感谢。

感谢满君丰教授为我们提供了良好的学习和科研环境，感谢智能信息处理研究所所长文志强教授以及智能信息处理研究所的老师们在论文开题、论文中期检查等环节中给予的宝贵意见。

同时要感谢田海龙师兄和刘景师姐，在我遇到很多不懂的问题时给我的帮助和指导，这些经验使我在研究的道路上少走了很多弯路，同时感谢你们在很多事情上给予我的帮助，和你们在一起学习的日子使我终身难忘。

感谢每一个科研工作者，是你们的成就与成果，让我在这条丰硕的科研之树上继续开花结果。

感谢帮助和关心过我的同学们，和你们一起学习和生活的日子使我永远难忘。

我还要感谢我的父母、哥哥，是你们给了我大学本科毕业后继续深造的勇气和条件，是父母的辛勤劳动给了我经济上的支撑，每当遇到困难，是父母的安慰和鼓励使我重新继续奋斗。哥哥，我无话不谈的好兄长，在读研期间不仅给予我经济上的帮助，还指引我奋斗的目标。真心的谢谢你们！

最后，感谢百忙之中审阅本文而付出辛劳劳动的各位专家、教授！