浙 江 大 学

硕士学位论文文献综述

**(专业学位)**

文献综述题目： 团购商品个性化推荐系统的设计与实现

姓 名： 李朝阳

学 号： Z134325272

专 业： 软件工程

院 别： 软件学院

导 师： 朱小军

日期：二零一四年 十一月

目录

[一. 个性化推荐系统的演讲 3](#_Toc404330434)

[二．个性会推荐系统的成功案例 3](#_Toc404330435)

[电子商务-亚马逊 4](#_Toc404330436)

[电影和视频网站 4](#_Toc404330437)

[个性化音乐网络电台 5](#_Toc404330438)

[社交网络 5](#_Toc404330439)

[基于位置的服务 6](#_Toc404330440)

[个性化邮件 7](#_Toc404330441)

[个性化广告 7](#_Toc404330442)

[三． 推荐系统的推荐算法和系统架构 8](#_Toc404330443)

[推荐算法 8](#_Toc404330444)

[基于邻域的算法 8](#_Toc404330445)

[隐语义模型 9](#_Toc404330446)

[系统整体架构 10](#_Toc404330447)

[数据收集和存储 10](#_Toc404330448)

[基于特征的推荐系统架构 11](#_Toc404330449)

[参考文献 12](#_Toc404330450)

# 个性化推荐系统的演讲

什么才是好的推荐系统？这是推荐系统评测需要解决的首要问题。一个完整的推荐系统一般存在3个参与方：用户、物品提供者和提供推荐系统的网站。以图书推荐为例，首先，推荐系统需要满足用户的需求，给用户推荐那些令他们感兴趣的图书。其次，推荐系统要让各出版社的书都能够被推荐给对其感兴趣的用户，而不是只推荐几个大型出版社的书。最后，好的推荐系统设计，能够让推荐系统本身收集到高质量的用户反馈，不断完善推荐的质量，增加用户和网站的交互，提高网站的收入。因此在评测一个推荐算法时，需要同时考虑三方的利益，一个好的推荐系统是能够令三方共赢的系统。

在推荐系统的早期研究中，很多人将好的推荐系统定义为能够做出准确预测的推荐系统。比如，一个图书推荐系统预测一个用户将来会购买《C++ primer中文版》这本书，而用户后来确实购买了，那么这就被看作一次准确的预测。预测准确度是推荐系统领域的重要指标（没有之一）。这个指标的好处是，它可以比较容易地通过离线方式计算出来，从而方便研究人员快速评价和选择不同的推荐算法。但是，很多研究表明，准确的预测并不代表好的推荐。比如说，该用户早就准备买《C++ Primer中文版》了，无论是否给他推荐，他都准备购买，那么这个推荐结果显然是不好的，因为它并未使用户购买更多的书，而仅仅是方便用户购买一本他本来就准备买的书。那么，对于用户来说，他会觉得这个推荐结果很不新颖，不能令他惊喜。同时，对于《C++ Primer中文版》的出版社来说，这个推荐也没能增加这本书的潜在购买人数。所以，这是一个看上去很好，但其实却很失败的推荐。举一个更极端的例子，某推测系统预测明天太阳将从东方升起，虽然预测准确率是100%，却是一种没有意义的预测。

所以，好的推荐系统不仅仅能够准确预测用户的行为，而且能够扩展用户的视野，帮助用户发现那些他们可能会感兴趣，但却不那么容易发现的东西。同时，推荐系统还要能够帮助商家将那些被埋没在长尾中的好商品介绍给可能会对它们感兴趣的用户。这也正是《长尾理论》的作者在书中不遗余力介绍推荐系统的原因。

为了全面评测推荐系统对三方利益的影响，本章将从不同角度出发，提出不同的指标。这些指标包括准确度、覆盖度、新颖度、惊喜度、信任度、透明度等。这些指标中，有些可以离线计算，有些只有在线才能计算，有些只能通过用户问卷获得。下面各节将会依次介绍这些指标的出发点、含义，以及一些指标的计算方法。

# 二．个性会推荐系统的成功案例

和搜索引擎不同，个性化推荐系统需要依赖用户的行为数据，因此一般都是作为一个应用存在于不同网站之中。在互联网的各类网站中都可以看到推荐系统的应用，而个性化推荐系统在这些网站中的主要作用是通过分析大量用户行为日志，给不同用户提供不同的个性化页面展示，来提高网站的点击率和转化率。广泛利用推荐系统的领域包括电子商务、电影和视频、音乐、社交网络、阅读、基于位置的服务、个性化邮件和广告等。

## 电子商务-亚马逊

电子商务网站是个性化推荐系统的一大应用领域。著名的电子商务网站亚马逊是个性化推荐系统的积极应用者和推广者，被RWW（读写网）称为“推荐系统之王”。亚马逊的推荐系统深入到了其各类产品中，其中最主要的应用有个性化商品推荐列表和相关商品的推荐列表。

图1.亚马逊的个性化推荐列表



推荐结果的标题、缩略图以及其他内容属性 告诉用户给他们推荐的是什么。

 推荐结果的平均分 平均分反应了推荐结果的总体质量，也代表了大部分用户对这本书

的看法。

 推荐理由 亚马逊根据用户的历史行为给用户做推荐，因此如果它给你推荐了本金庸的小说，大都是因为你曾经在亚马逊上对武侠方面的书给过表示喜欢的反馈。此外，亚马逊对每个推荐结果都给出了一个按钮Fix this recommendation（修正这一推荐），单击后可以看到推荐理由。如图1-4所示，亚马逊的推荐结果中有一本关于机器学习的书（Introduction to Machine Learning），单击该推荐结果的Fix this recommendation按钮后，会弹出页面给用户提供了5种对这个推荐结果进行反馈的方式，包括Add to Cart（加入到购物车）、Add to Wish List（加入到心愿单）、Rate this item（给书打分 ）、 I own it（我已经有这本书了）和Not interested（对这本书没兴趣）。同时，在推荐结果的下面还展示了推荐原因，此处是因为我曾经给Probabilistic Graphical Models：Principles and Techniques和Data Mining: Practical Machine Learning Tools and Techniques，Second Edition这两本书打过5分。亚马逊允许用户禁用推荐理由，这主要是出于隐私的考虑。有些用户可能不喜欢他对某些物品的行为被系统用来生成推荐结果，这个时候就可以禁用这些行为。

## 电影和视频网站

在电影和视频网站中，个性化推荐系统也是一种重要的应用。它能够帮助用户在浩瀚的视频库中找到令他们感兴趣的视频。在该领域成功使用推荐系统的一家公司就是Netflix，它和亚马逊是推荐系统领域最具代表性的两家公司。Netflix原先是一家DVD租赁网站，最近这几年也开始涉足在线视频业务。Netflix非常重视个性化推荐技术，并且在2006年起开始举办著名的Netflix Prize推荐系统比赛。该比赛悬赏100万美元，希望研究人员能够将Netflix的推荐算法的预测准确度提升10%。该比赛举办3年后，由AT&T的研究人员获得了最终的大奖。该比赛对推荐系统的发展起到了重要的推动作用：一方面该比赛给学术界提供了一个实际系统中的大规模用户行为数据集（40万用户对2万部电影的上亿条评分记录）；另一方面，3年的比赛中，参赛者提出了很多推荐算法，大大降低了推荐系统的预测误差。此外，比赛吸引了很多优秀的科研人员加入到推荐系统的研究中来，大大提高了推荐系统在业界和学术界的影响力。

## 个性化音乐网络电台

个性化推荐的成功应用需要两个条件。第一是存在信息过载，因为如果用户可以很容易地从所有物品中找到喜欢的物品，就不需要个性化推荐了。第二是用户大部分时候没有特别明确的需求，因为用户如果有明确的需求，可以直接通过搜索引擎找到感兴趣的物品。在这两个条件下，个性化网络电台无疑是最合适的个性化推荐产品。首先，音乐很多，用户不可能听完所有的音乐再决定自己喜欢听什么，而且每年新的歌曲在以很快的速度增加，因此用户无疑面临着信息过载的问题。其次，人们听音乐时，一般都是把音乐作为一种背景乐来听，很少有人必须听某首特定的歌。对于普通用户来说，听什么歌都可以，只要能够符合他们当时的心情就可以了。因此，个性化音乐网络电台是非常符合个性化推荐技术的产品。



## 社交网络

最近5年，互联网最激动人心的产品莫过于以Facebook和Twitter为代表的社交网络应用。在

社交网络中，好友们可以互相分享、传播信息。社交网络中的个性化推荐技术主要应用在3个方面：

 利用用户的社交网络信息对用户进行个性化的物品推荐；

 信息流的会话推荐；

 给用户推荐好友。

Facebook最宝贵的数据有两个，一个是用户之间的社交网络关系，另一个是用户的偏好信息。因此，Facebook推出了一个推荐API，称为Instant Personalization。该工具根据用户好友喜欢的信息，给用户推荐他们的好友最喜欢的物品。很多网站都使用了Facebook的API来实现网站的个性化。表1-1中是使用了Facebook的Instant Personalization的具有代表性的网站。图1-14是著名的电视剧推荐网站Clicker使用Instant Personalization给用户进行个性化视频推荐的界面。

除了利用用户在社交网站的社交网络信息给用户推荐本站的各种物品，社交网站本身也会利用社交网络给用户推荐其他用户在社交网站的会话。如图1-15所示，每个用户在Facebook的个人首页都能看到好友的各种分享，并且能对这些分享进行评论。每个分享和它的所有评论被称为一个会话，如何给这些会话排序是社交网站研究中的一个重要话题。为此，Facebook开发了Edge-Rank算法对这些会话排序，使用户能够尽量看到熟悉的好友的最新会话。



## 基于位置的服务

随着移动设备的飞速发展，用户的位置信息已经非常容易获取，而位置是一种很重要的上下文信息，基于位置给用户推荐离他近的且他感兴趣的服务，用户就更有可能去消费。

## 个性化邮件

我们每天都会收到大量的邮件，这些邮件有些对我们很重要（如领导交代任务的邮件），有些比较次要（比如别人邀约周末打球的邮件），还有些是垃圾邮件。垃圾邮件可以通过垃圾邮件过滤器去除，这是一个专门的研究领域，这先浏览，无疑会大大提高用户的工作效率。其实，目前在文献中能够查到的第一个推荐系统Tapestry就是一个个性化邮件推荐系统，它通过分析用户阅读邮件的历史行为和习惯对新邮件进行重新排序，从而提高用户的工作效率。Google的研究人员在这个问题上也进行了深入研究，于2010年推出了优先级收件箱功能。Google的研究表明，该产品可以帮助用户节约6%的时间。在如今这个时间就是金钱的年代，6%的节约无疑是一大进步。

## 个性化广告

广告是互联网公司生存的根本。很多互联网公司的盈利模式都是基于广告的，而广告的CPC、CPM直接决定了很多互联网公司的收入。目前，很多广告都是随机投放的，即每次用户来了，随机选择一个广告投放给他。这种投放的效率显然很低，比如给男性投放化妆品广告或者给女性投放西装广告多半都是一种浪费。因此，很多公司都致力于广告定向投放（Ad Targeting）的研究，即如何将广告投放给它的潜在客户群。个性化广告投放目前已经成为了一门独立的学科——计算广告学——但该学科和推荐系统在很多基础理论和方法上是相通的，比如它们的目的都是联系用户和物品，只是在个性化广告中，物品就是广告。

个性化广告投放和狭义个性化推荐的区别是，个性化推荐着重于帮助用户找到可能令他们感兴趣的物品，而广告推荐着重于帮助广告找到可能对它们感兴趣的用户，即一个是以用户为核心，而另一个以广告为核心。目前的个性化广告投放技术主要分为3种。

 上下文广告 通过分析用户正在浏览的网页内容，投放和网页内容相关的广告。代表系统是Google的Adsense。

 搜索广告 通过分析用户在当前会话中的搜索记录，判断用户的搜索目的，投放和用户目的相关的广告。

 个性化展示广告 我们经常在很多网站看到大量展示广告（就是那些大的横幅图片），它们是根据用户的兴趣，对不同用户投放不同的展示广告。雅虎是这方面研究的代表。

广告的个性化定向投放是很多互联网公司的核心技术，很多公司都秘而不宣。不过，雅虎公司是个例外，它发表了大量个性化广告方面的论文。在个性化广告方面最容易获得成功的无疑是Facebook，因为它拥有大量的用户个人资料，可以很容易地获取用户的兴趣，让广告商选择自己希望对其投放广告的用户Facebook的广告系统界面，该界面允许广告商选择自己希望的用户群，然后Facebook会根据广告商的选择告诉他们这些限制条件下广告将会覆盖的用户数量。

# 三． 推荐系统的推荐算法和系统架构

## 推荐算法

### 基于邻域的算法

#### 基于用户的协同过滤算法

基于用户的协同过滤算法是推荐系统中最古老的算法。可以不夸张地说，这个算法的诞生标志了推荐系统的诞生。该算法在1992年被提出，并应用于邮件过滤系统，1994年被GroupLens用于新闻过滤。在此之后直到2000年，该算法都是推荐系统领域最著名的算法。本节将对该算法进行详细介绍，首先介绍最基础的算法，然后在此基础上提出不同的改进方法，并通过真实的数据集进行评测。

**基础算法**

每年新学期开始，刚进实验室的师弟总会问师兄相似的问题，比如“我应该买什么专业书啊”、“我应该看什么论文啊”等。这个时候，师兄一般会给他们做出一些推荐。这就是现实中个性化推荐的一种例子。在这个例子中，师弟可能会请教很多师兄，然后做出最终的判断。师弟之所以请教师兄，一方面是因为他们有社会关系，互相认识且信任对方，但更主要的原因是师兄和师弟有共同的研究领域和兴趣。那么，在一个在线个性化推荐系统中，当一个用户A需要个性化推荐时，可以先找到和他有相似兴趣的其他用户，然后把那些用户喜欢的、而用户A没有听说过的物品推荐给A。这种方法称为基于用户的协同过滤算法。

从上面的描述中可以看到，基于用户的协同过滤算法主要包括两个步骤。

(1) 找到和目标用户兴趣相似的用户集合。

(2) 找到这个集合中的用户喜欢的，且目标用户没有听说过的物品推荐给目标用户。

#### 基于物品的协同过滤算法

基于物品的协同过滤（item-based collaborative filtering）算法是目前业界应用最多的算法。无论是亚马逊网，还是Netflix、Hulu、YouTube，其推荐算法的基础都是该算法。本节将从基础的算法开始介绍，然后提出算法的改进方法，并通过实际数据集评测该算法。

**基础算法**

基于用户的协同过滤算法在一些网站（如Digg）中得到了应用，但该算法有一些缺点。首先，随着网站的用户数目越来越大，计算用户兴趣相似度矩阵将越来越困难，其运算时间复杂度和空间复杂度的增长和用户数的增长近似于平方关系。其次，基于用户的协同过滤很难对推荐结果做出解释。因此，著名的电子商务公司亚马逊提出了另一个算法——基于物品的协同过滤算法。基于物品的协同过滤算法（简称Item-CF）给用户推荐那些和他们之前喜欢的物品相似的物品。比如，该算法会因为你购买过《数据挖掘导论》而给你推荐《机器学习》。不过，Item-CF算法并不利用物品的内容属性计算物品之间的相似度，它主要通过分析用户的行为记录计算物品之间的相似度。该算法认为，物品A和物品B具有很大的相似度是因为喜欢物品A的用户大都也喜欢物品B。图2-9展示了亚马逊在iPhone商品界面上提供的与iPhone相关的商品，而相关商品都是购买iPhone的用户也经常购买的其他商品。

基于物品的协同过滤算法可以利用用户的历史行为给推荐结果提供推荐解释，比如给用户推荐《天龙八部》的解释可以是因为用户之前喜欢《射雕英雄传》。Hulu在个性化视频推荐利用Item-CF给每个推荐结果提供了一个推荐解释，而用于解释的视频都是用户之前观看或者收藏过的视频。

基于物品的协同过滤算法主要分为两步。

(1) 计算物品之间的相似度。

(2) 根据物品的相似度和用户的历史行为给用户生成推荐列表。

### 隐语义模型

自从Netflix Prize比赛举办以来，LFM（latent factor model）隐语义模型逐渐成为推荐系统领域耳熟能详的名词。其实该算法最早在文本挖掘领域被提出，用于找到文本的隐含语义。相关的名词有LSI、PLSA、LDA和Topic Model。本节将对隐含语义模型在Top-N推荐中的应用进行详细介绍，并通过实际的数据评测该模型。

隐语义模型是最近几年推荐系统领域最为热门的研究话题，它的核心思想是通过隐含特征(latent factor)联系用户兴趣和物品。

首先通过一个例子来理解一下这个模型。图2-16展示了两个用户在豆瓣的读书列表。

从他们的阅读列表可以看出，用户A的兴趣涉及侦探小说、科普图书以及一些计算机技术书，而用户B的兴趣比较集中在数学和机器学习方面。

那么如何给A和B推荐图书呢？

 对于User-CF，首先需要找到和他们看了同样书的其他用户（兴趣相似的用户），然后给他们推荐那些用户喜欢的其他书。

 对于Item-CF，需要给他们推荐和他们已经看的书相似的书，比如作者B看了很多关于数据挖掘的书，可以给他推荐机器学习或者模式识别方面的书。

还有一种方法，可以对书和物品的兴趣进行分类。对于某个用户，首先得到他的兴趣分类，然后从分类中挑选他可能喜欢的物品。

总结一下，这个基于兴趣分类的方法大概需要解决3个问题。

 如何给物品进行分类？

 如何确定用户对哪些类的物品感兴趣，以及感兴趣的程度？

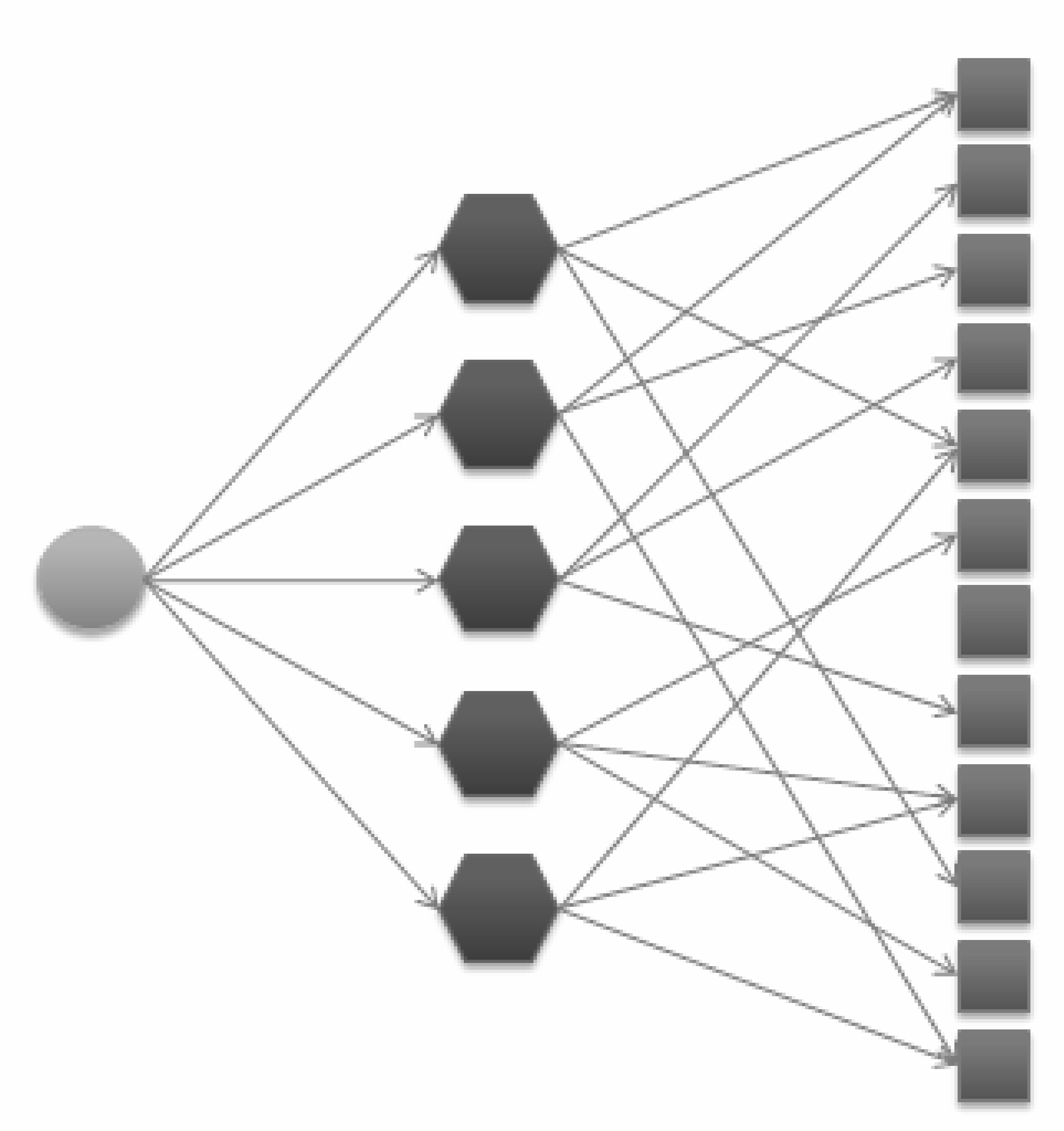
 对于一个给定的类，选择哪些属于这个类的物品推荐给用户，以及如何确定这些物品在一个类中的权重？

## 系统整体架构

### 数据收集和存储

个性化推荐算法依赖于用户行为数据，而在任何一个网站中都存在着各种各样的用户行为数据。那么如何存取这些数据就是推荐系统需要解决的首要问题。从规模上看，浏览网页、搜索记录的规模都很大，因为这种行为所有用户都能产生，而且平均每个用户都会产生很多这些行为。购买、收藏行为规模中等，因为只有注册用户才能产生这种行为，但购买行为又是电商网站的主要行为，所以它们相对于评论来说规模更大，但相对于网页浏览行为来说规模要小得多，最后剩下的行为是注册用户里的小部分人才有的，所以规模不会很大。从实时存取的角度上看，购买、收藏、评论、评分、分享等行为都是需要实时存取的，因为只要用户有了这些行为，界面上就需要体现出来，比如用户购买了商品后，用户的个人购买列表中就应立即显示用户购买的商品。而有些行为，比如浏览网页的行为和搜索行为并不需要实时存取。数据能否实时存取在推荐系统中非常重要，因为推荐系统的实时性主要依赖于能否实时拿到用户的新行为。只有快速拿到大量用户的新行为，推荐系统才能够实时地适应用户当前的需求，给用户进行实时推荐。

### 基于特征的推荐系统架构



用户的特征种类非常多，主要包括如下几类。

 人口统计学特征 包括用户的年龄、性别、国籍和民族等用户在注册时提供的信息。

 用户的行为特征 包括用户浏览过什么物品、收藏过什么物品、给什么物品打过什么样

的分数等用户行为相关的特征。同时，用户行为从时间上也可以分为用户近期的行为和长期的行为。

 用户的话题特征 可以根据用户的历史行为利用话题模型（topic model）将电视剧和电影聚合成不同的话题，并且计算出每个用户对什么话题感兴趣。比如用户如果看了《叶问》、《新龙门客栈》和《醉拳》，那么可以认为用户对“香港武侠电影”这个话题感兴趣。

 将最新加入的物品推荐给用户。

 将商业上需要宣传的物品推荐给用户。

 给用户推荐不同种类的物品，比如亚马逊会推荐图书、音像、电子产品和服装等

 给用户混合推荐，有时需要将图书和音像制品放到一个推荐列表中展示给用户

 对于不同的产品推荐不同新颖度的物品。比如在首页给用户展示比较热门的推荐结果，

在推荐系统页面给用户展示长尾中的物品

 考虑到用户访问推荐系统的上下文

# 参考文献

[1] 项亮. 推荐系统实践[M]. 北京:人民邮电出版社, 2012.

[2] 刘超峰. 团购商品个性化推荐系统的研究与实现[D]. 杭州：浙江工业大学, 2012. 12-14

[3] 易明. 基于web挖掘的电子商务个性化推荐机理与方法研究[D]. 武汉：华中科技大学, 2006. 3-5

[4] 应中运. 基于用户情境的论坛个性化推荐模型研究[D]. 重庆：西南大学, 2014.