基于Knn的相亲网站推荐算法研究

学生姓名：龙瑜

指导老师：冶红

摘要

随着国家的经济飞速发展，生活节奏日益加快，网络社交开始融入到我们每个人的生活中，也开始慢慢的融入到我们传统的恋爱生活中。“网恋”这个词语开始出现在人们的视野，逐渐开始改变人们生活中的恋爱观，爱情观。婚恋网站，相亲网站的出现，通过网络把人们连结起来，打破了传统的相亲模式，让广大的单身同胞能够有更多的机会或者更广阔的视野去选择自己的配偶，许多成功的案列也让人们愈来愈相信网络婚恋的可行性。网络婚恋的是否可行，在于其能否对用户精准和高效的进行推荐，而不是只是对用户进行扩大交友范围。

在广大的用户集当中选出推荐集给予用户，这是网络婚恋的关键部分，也就是网络婚恋推荐。传统的相亲中的推荐往往都有很大的主观成分，人们在推荐的时候多多少少都是结合了一些推荐人的喜好，而通过计算机，在广阔的用户群和大数据的支持下，通过算法能够客观的堆被推荐人进行推荐。在许多的婚恋网站中，推荐算法百家争鸣，有使用传统的基于内容的推荐算法（CB算法），也有有使用kNN（k-近邻算法）算法，也有使用基于kNN的item-based kNN算法，在一些比较前沿的公司中有使用深度神经网络推荐算法。

本论文主要是基于kNN算法的婚恋推荐算法研究，分析kNN在推荐算法的优点与缺点。通过婚恋推荐了解 kNN算法并将其应用到婚恋推荐当中，利用该算法，通过网站世纪佳缘的用户模型集对用户进行推荐，得出准确率较高的推荐理由

关键词：网络婚恋社交；推荐算法；kNN算法

Abstract

With the rapid development of the country's economy, the pace of life is accelerating, and social networking has begun to integrate into the lives of each of us, and has begun to slowly integrate into our traditional love life. The word "online love" began to appear in people's vision, and gradually began to change the concept of love and love in people's lives. Marriage websites, the emergence of dating websites, connecting people through the Internet, breaking the traditional blind date model, allowing the majority of single compatriots to have more opportunities or a broader perspective to choose their spouse, many successful cases also Let people more and more believe in the feasibility of online marriage. The feasibility of online marriage and marriage is whether it can recommend users accurately and efficiently, instead of just expanding the range of friends.

Selecting a recommendation set from the majority of the user set to give the user, this is the key part of the network marriage and love, which is the network marriage love recommendation. The recommendation in the traditional blind date often has a large subjective component. When people recommend it, it is more or less combined with the preferences of some recommenders, and through the computer, with the support of a broad user base and big data, through The algorithm can objectively be recommended by the recommender. In many dating websites, there are hundreds of recommended algorithms, including traditional content-based recommendation algorithms (CB algorithms), kNP (k-nearest neighbor algorithm), and kNN-based item-based kNN algorithms. Deep neural network recommendation algorithms are used in leading companies.

This thesis is mainly based on the kNN algorithm of marriage and love recommendation algorithm, and analyzes the advantages and disadvantages of kNN in the recommendation algorithm. Through the recommendation of marriage and love to understand the kNN algorithm and apply it to the marriage and love recommendation, this algorithm is used to recommend the user through the user model set of the website Century Jiayuan, and the recommendation reason with higher accuracy is obtained.

Keywords: network marriage and social interaction; recommendation algorithm; kNN algorithm

1. 绪论

1.1课题背景

中国的婚恋网络交友需求基数非常巨大，随着中国人口数量的增长，国内单身人口数量日益增多。根据数据显示，2013年-2018年我国的单身人口数目从1.3亿人增长到了2.4亿，占全国人口的17.3%，同时，我国的离婚率高涨也给中国的婚恋行业创造需求，具体的数据显示，中国的结婚率已经从2010年的9.3‰下降到2018年的7.3‰，而离婚率则是从2010年的2.0‰上升到2018年的2.7‰[1]（2018年中国网络婚恋行业发展现状与市场前景分析—前瞻经济学人）。伴随着年轻一代的崛起，80后，90后开始成为互联网婚恋大军，相比于70后，60后的传统婚恋选择方式，年轻一代更倾向于使用互联网平台进行婚恋交友，助力不断释放的婚恋产业市场的潜力。根据数据显示，随着我国移动网民的不断提高，我国的网络婚恋渗透率到2018年已经达到惊人的49%。在这样的巨大市场潜力下，婚恋市场规模愈来愈大，国内也涌现出众多的婚恋相亲网站，比如三巨头珍爱网，百合网，世纪佳缘网。婚恋市场由于拥有庞大的适婚人群的刚性需求所创造的良好的变现潜质，近几年中移动端的快速增长和崛起也给婚恋市场带来了全新的机遇，在核心企业的频繁运转和高度关注下，行业投资讯速增长。资本的涌入，带来的是网络婚恋移动端的新进入者的激增，PC+移动的模式成为行业的主要增长模式[2艾瑞咨询—关于中国婚恋行业发展的前景展望]。

移动端作为新进入者的加入，行业的竞争进一步加剧，但也为行业带来了更多的发展方式，不在是之前婚恋网站三巨头垄断。但由于婚恋本身就是一个比较严肃的问题，因为这是一个人的人生大事，来不得半点马虎。对于婚恋用户而言，寻找到自己满意的对象是根本，婚恋推荐在其中便起到举足轻重的重要性，准确的针对用户进行推荐。如何推荐，这便是婚恋网站的一个主要的问题，网站拥有大量的用户群，在大数据的支持下，通过推荐，能够让用户在众多的用户群当中寻找相应的适合人群，用户没有快速找到自己心仪的人群是网站推荐算法是否成功的关键，能够有效的减少因为冗余信息的增多导致时间的浪费，让用户对网站有更高的粘性，让更多的用户加入到用户集当中，从而最大化的提高网站推荐算法的精确度，一个好的推荐，不仅能够帮用户更快的找到自己心仪的对象，更能提高用户的满意度，才能够提高自身的品牌认知度和品牌信任度。能够让用户更加安心的去付费，去使用网站更多的功能，从而提高企业的利润，让企业发展更加稳定和繁荣。

1.2 研究的目的及意义

在网络婚恋发展中，婚恋推荐作为主项，现在的推荐算法虽然百家争鸣，却没有一个统一的算法趋势，然而在众多的婚恋推荐算法中，kNN算法作为经久不衰的算法一直长存至今，无论是网络婚恋刚刚起步，又或者到现在的百家争鸣，kNN算法都一直的伴随着网络婚恋发展的历史长河中

kNN算法作为最基础的机器学习算法，随着计算机技术的飞速发展，kNN算法不断的完善，kNN已经能够很好的适应到各种的推荐算法当中，研究推荐算法，能够更加的了解和深入到kNN算法当中。本论文基于kNN进行婚恋网站的推荐算法研究，承前启后，通过在巨人的肩膀上眺望，扩大kNN算法在婚恋推荐算法当中的应用，提高婚恋推荐的准确度，推进行业的发展。让国内广大的单身人士终有一天能够通过婚恋网站找到属于自己的另一半。

1.3 推荐算法国内外发展现状

推荐算法实在社交网络中非常热门的研究点，除了网络婚恋社交的推荐之外，许多的电子商务网站（淘宝，京东，亚马逊），社交网站（微博，百度贴吧），视频网站（优酷，爱奇艺，bilibili）都有自己相应的推荐算法，网络中的信息量非常庞大，如何让用户在如此庞大的信息当中寻找到自己感兴趣的东西，这便是推荐系统的由来。

推荐算法的兴起与互联网的发展息息相关。最早的自动化协同过滤系统可以追溯到 1994 年，明尼苏达大学双城分校计算机系的 GroupLens 研究组设计了名为 GroupLens 的新闻推荐系统。该工作不仅首次提出了协同过滤的思想，并且为推荐问题建立了一个形式化的模型，为随后几十年推荐系统的发展带来了巨大影响。该研究组后来创建了MovieLens推荐网站，一个推荐引擎的学术研究平台，其包含的数据集是迄今为止推荐领域引用量最大的数据集。

伴随着推荐算法的发展，人们不仅仅满足于分析用户的历史行为对用户进行建模，转而研究混合推荐模型，致力于通过不同的推荐方法来解决冷启动、数据极度稀疏等问题，国内知名新闻客户端今日头条采用了内容分析、用户标签、评估分析等方法打造了拥有上亿用户的推荐引擎。移动互联网的崛起为推荐系统提供了更多的数据，如移动电商数据、移动社交数据、地理数据等，成为了社交推荐的新的尝试。

近年来，机器学习和深度学习等领域的发展，为推荐系统提供了方法指导。RecSys会议自2016年起开始举办定期的推荐系统深度学习研讨会，旨在促进研究和鼓励基于深度学习的推荐系统的应用。

历史上将推荐算法的研究推向高潮的事件非 Netflix 百万美金大赛莫属。2006 年，Netflix 宣布提供一百万美元奖金给第一个能将现有推荐算法的准确度提升 10%以上的参赛者。重赏之下必有勇夫，这项百万美金大赛一经宣布，立即吸引了 186 个国家的 4 万多支参赛队伍鹿逐中原。随后几年，大量经典的推荐算法涌现出来。其中最有影响力的当属 SVD++模型，这篇论文也在 2018 年的 SIGKDD大会上被评选上了Test of Time Award，实至名归。Netflix 比赛对推荐系统的影响，可以类比到 ImageNet挑战赛对图像识别技术发展的影响。当人们有共同的目标和共同的数据集、公平、激烈地竞争时，对推动该领域的技术发展是最快、最完善的[The Netflix Prize,2009]。

推荐算法发展至今，其背后的技术大致的可以划分为三大类：基于内容的模型，基于协同过滤的模型，以及混合模型。推荐算法具有两大特性：1）主动化：用户不需要提出明确的需求，而是根据分析用户的特征，对用户进行建模，从而主动地为用户提供信息。2）个性化：推荐系统能够更好的发现和挖掘用户的信息，把每个用户的长尾信息发掘出来，从而能够更好的堆用户进行个性化的推荐。

目前，推荐算法已经广泛的应用到诸多领域，其中最典型的电子商务领域。同时，伴随着神经网络，机器学习的不断兴起，推荐算法也越来越融入到各行各业当中，工业界和学术界对推荐算法的研究也越来越深入，形成了一门独立的算法分支。

1.4 论文内容的安排

本论文的主要章节内容安排如下：

第一章为绪论部分，主要为阐述课题的研究背景以及研究的目的和意义，对国内外的推荐算法发展以及现状进行分析。

第二章为理论部分，主要为推荐算法和数据挖掘分类的研究

第三章为数据部分，主要为数据挖掘与数据处理

第四章为算法部分，主要为kNN婚恋推荐算法的设计与实现

2.相关理论技术研究

2.1 推荐算法研究

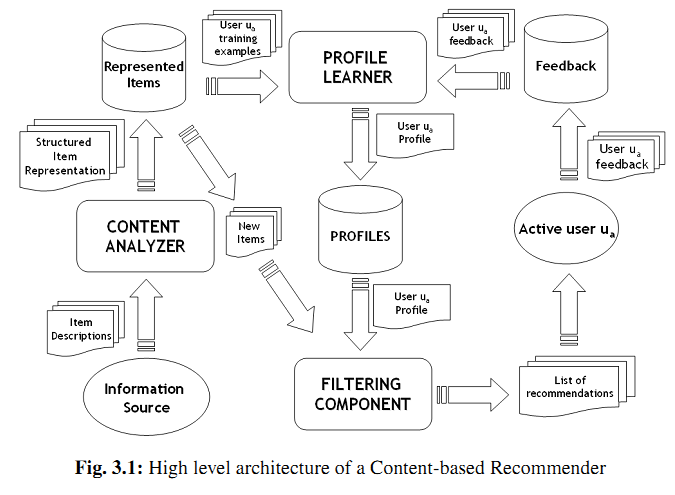
推荐算法由来已久，在第一章绪论的1.3节推荐算法国内外发展现状也提到，推荐算法发展到现在，其背后的技术也基本形成了3个形式：1)基于内容模型推荐，2）基于协同过滤模型推荐，3)混合模型推荐

2.1.1 基于内容的模型推荐

基于内容模型推荐（Content-based Recommendations）是最早使用的推荐算法，它是根据用户过去喜欢的物品，为用户推荐他过去喜欢的产品的类似的产品。例如，一个推荐饭店的推荐算法可以根据用户以前喜欢去火锅店而为用户进行推荐火锅店。基于内容的模型他推荐算法过程一般包括以下三步：

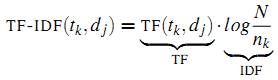
1. Item Representation：为每一个用户的item抽取出用户相应的用户特征
2. Profile Learning：利用用户item的特征数据放入算法学习用户喜欢的
3. Recommendation Generation：根据Item Representation与Profile Learning得出的用户相似喜欢物品对用户进行大量推荐

对于上面的流程有一个很详细的流程图[3]( Pasquale Lops, Marco de Gemmis and Giovanni Semeraro, Chapter 3 in Recommender Systems Handbook, 2011)



1. Item Representation

在推荐算法中往往都要将用户的属性提取出来，也就是特征提取。一般来说，特征属性的提取分为两种：一种是结构化的属性，一种是非结构化的属性，结构化的属性是属性的意义比较明确，结构化属性的值域是连续性的。而非结构化的属性的意义有意义比较明确的，也有意义不比较明确的，但是非结构化的属性的是离散型的数值，值与值之间不存在连续的关系。就好比本论文的婚恋网站数据中，每个用户的属性便是一个非结构化的属性，用户的年龄身高学历收入等等属性，虽然其表达的意义都是比较 明确的，但是都是一些离散型的数据。结构化连续性的属性往往我们都是可以直接量化后来进行使用，但是非结构化的属性就需要将这些属性进行结构化后才可以加以使用。比较经典的将非结构化数据转化为结构化的数据的算法是词频-逆文档频率（TF·IDF算法）



这个算法是将分词后的文章通过该算法将一篇文章用向量表示出来，这就将文章作为一个非结构化的数据用向量表示成结构化的数据。

1. Profile Learning

假设用户已经有过了自己喜欢过的项目和不喜欢的项目，有了这些我们就需要为用户建立和用户相关的判断模型，我们在通过这个模型，利用算法来推断用户喜欢什么样的项目，这样就是一个对用户的有监督的学习分类过程，也有很多的算法支持用户模型的判断。基于内容的推荐的算法使用一般使用下面的方法

1.k-近邻算法：也就是本论文所使用的算法，计算项目两者之间的距离的到相似度，对于结构化的数据都是计算项目之间的欧几里得距离，而对于从非结构化转化而来的结构化向量数据一般是计算余弦相似度，在这不做详细介绍。

2.Rocchio算法：Rocchio算法是20世纪70年代在Salton的SMART系统中引入并广泛流传的一种相关的反馈算法，在应用上，也就是会根据用户的每一个行为会进行向量分析，用户在进行每个行为过后都会修改向量得分，就好比用户在百度上百度了微软，用户在结果中点击了微软相关的Xbox，这样网站就会根据用户的这次点击结果所得到的反馈重新计算网站评分，可能下一页的结果为微软，Xbox，索尼，PS4等等结果。这就是通过Rocchio算法来对用户反馈来修改用户向量。

通过类比，在基于内容推荐的推荐算法中，我们也可以用过Rocchio来对用户反馈来做相似计算：

http://images.cnblogs.com/cnblogs_com/breezedeus/201204/201204101405391270.png

其中Image(13)表示项目item的属性Image(14)与Image(15)分别为用户喜欢的项目和不喜欢的项目集合，Image(16)与Image(17)为给出的权重值，计算出Image(11)后我们可以与Image(13)计算相似度来表示用户对Image(13)的喜好程度。

除了上面常用的两个算法，作为典型有监督的机器学习分类问题，很多的机器学习算法都可以用上，比如线性回归算法，决策树和朴素贝叶斯。

1. Recommendation Generation

第三步便是推荐，通过上面两步的算法后，得出的推荐结果根据相似度排序后推荐给用户即可，但是推荐的时候会遇到一个问题，人是是感性的，兴趣总是会改变的，很少有人能够对一件事或者一个物品一直感兴趣下去，兴趣迁移的问题需要引入一个兴趣衰减的机制，加入一个系数，模拟用户的兴趣衰减，当衰减到一定程度后便清除。

基于内容推荐算法算上是第一个世代的个性化推荐算法，本身拥有优点但是也不乏缺点，CB的优点：

用户之间的独立性强：每个人的特征都是其个人行为创造的，不受他人的行为影响

好的可解释性：因为这是根据相似度推荐，可以很好解释用户喜欢这个所以才推荐这个

CB的缺点：

用户的特征提取难：在实际应用中，我们所提取的特征值一般都只能表示项目的一部分特征，并不是很全面，这就导致有可能两个ietm的特征有可能是一样的的，这就会造成很大的判断失误

无法挖掘出用户的潜在兴趣：基于内容的推荐是基于用户过去的行为所作的推荐，产生的推荐结果也是与用户过去的行为相似，并不会推荐用户的潜在兴趣

无法为新用户产生推荐：基于内容推荐既然是通过用户的历史行为进行推荐，新用户并没有，自然无法推荐

现在工业上的推荐算法也已经很少使用基于内容推荐，很多时候都是使用CB来弥补算法上的一些不确定性，更多的是使用接下来介绍的基于协同过滤算法。

2.1.2 基于协同过滤的推荐算法

协同过滤算法（Collaborative Filtering）作为最经典的推荐算法，在很多工业级应用上都有使用，协同过滤算法包括在线协同和离线过滤两个部分。在线协同是通过在线数据找到用户喜欢的item，离线过滤就是过滤掉一些不值得为用户推荐的一些item，比如在评分上比较低的一些item

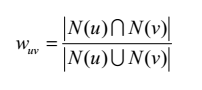
协同过滤的模型一般为有x个item，有y个item的用户评分数据，并且y<x，，需要使用y个的稀疏数据来预测那些空白item和数据之间的评分关系，并将最高的评分推荐给用户。一般来说，协同过滤分为三种类型，第一种是基于用户的协同过滤（user based），一种是基于项目的协同过滤（item-based），第三种是基于模型的协同过滤（model-based）

1. 基于用户的协同过滤（user based CF）

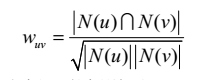
通俗地来说，要给用户推荐一件物品，首先先计算这个用户与其他用户相比的相似度，

比如说，向A用户推荐item，就首先将A用户与其他用户计算相似度，发现B,C用户与A用户的相似度最高，这样就可以从B和C用户中喜欢的东西并且A用户没有浏览或者标记过的东西进行推荐，实现对A用户的个性化推荐。

计算用户的相似度可以通过Jaccard公式：



或者余弦相似公式：



计算得出，其中N(u)为用户u的兴趣集合，N(v)为用户v的兴趣集合。

通过算法得出相似用户的高相似兴趣后，再通过算法来计算用户u对item的感兴趣程度

https://img-blog.csdn.net/20180303114727333?watermark/2/text/aHR0cDovL2Jsb2cuY3Nkbi5uZXQvc2luYXRfMzU4NjY0NjM=/font/5a6L5L2T/fontsize/400/fill/I0JBQkFCMA==/dissolve/70

其中，S（u，K）包含了K个与用户u相似的用户集合，N(i)为对item I 有过行为的用户集合，Wuv是用户u与用户v的兴趣相似度,Rvi表示用户v对I item的感兴趣程度

基于用户的协同过滤需要在线地去找用户与用户之间的相似程度，可以为用户推荐一些用户意想不到的东西，但是随着数据的增多，用户的数量不断的增大就会带来计算复杂度提高的问题，计算的时间也会大大的延长。而且另外基于用户的算法也有一个弊端，就是大众化问题，大家喜欢的东西如果有个用户不喜欢，这样也会将item推荐给用户

1. 基于项目的协同过滤（item-based）

Item-based的协同过滤大体上也是和user-based是差不多的，user-based是根据用户

的兴趣相似的用户来进行推荐，别的用户与你兴趣相投，自然你也会喜欢他喜欢的东西，而item-based则是用户曾经看过或者标记喜欢过这项物品，所以根据推荐与这个物品相似的物品相比你也是会喜欢的。这样就会有一个问题，这样不就和上面所说的基于内容的推荐是一样的吗？在这里就需要说说基于内容的推荐算法CB与item-based CF算法的区别。其实举一个很简单的例子就可以：

假设我们已经知道：

用户A喜欢电影a,b,c

用户B喜欢电影a,c,e,f

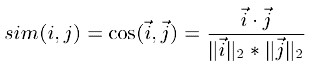
用户C喜欢电影b,d

现在我们的推荐算法需要决定是否将电影f推荐给用户A。

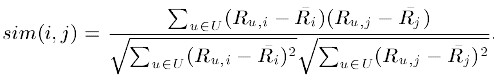
首先，我们先使用基于内容的推荐算法，根据上面所说的步骤，先分析用户A的item即电影a,b,c的特征属性，假设a—喜剧片，b—喜剧片，c—爱情片。如果分析的出f电影的特征为喜剧片，那么该推荐算法很大的几率会将电影f推荐给用户A，反之则相反。这就是基于内容的推荐，需要对item的特征进行建模，

再到item-based CF算法，该算法可以完全不用考虑电影的特征，可以完全忽略，该算法以来的是user-item之间的向量关系，我们不需要知道这些电影是什么片子，我们只需要知道a,b,c这三部电影与A之间的向量关系，电影f与用户B之间的向量关系再通过算法公式计算出着两个向量关系的相似程度，相似程度高，推荐算法就会把电影f推荐给用户A，反之亦然。

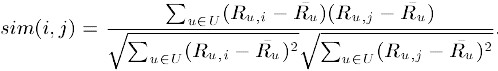
上面的比较中，我们可以知道，基于项目的协同过滤的过程非常简单，就是计算user-item之间的关系的相似程度，相似度的计算算法也可以是计算余弦的相似程度：



其中分子为两个向量的内积。也可以使用关联相似度的计算，计算两个向量之间的Pearson-r的关联度，公式如下：



其中Ru,i表示用户u对物品i的评价，Ri表示i物品的打分平均值。也可以使用调整余弦相似度计算：



其中Ru为用户u的打分平均值，这个公式算法可以避免有的用户对物品偏向于的打高分，而有的用户则是偏向于给物品打低分，可以消除这些用户对相似度计算的影响

从上面这些算法不难看出，基于物品的协同过略对比基于用户的协同过滤是要好的item-based的算法预测通常会要比user-based的算法预测质量要高一些，并且在大数据的情况下降低了数据的计算量。

对于上面的两种协同过滤算法，基于用户的协同过滤需要在线找用户和用户之间的相似度关系，计算复杂度肯定会比基于基于项目的协同过滤高。但是可以帮助用户找到新类别的有惊喜的物品。而基于项目的协同过滤，由于考虑的物品的相似性一段时间不会改变，因此可以很容易的离线计算，准确度一般也可以接受，但是推荐的多样性来说，就很难带给用户惊喜了。

1. 基于模型的混合协同过滤

基于模型的混合过滤是自从机器学习兴起后，逐渐成为主流的协同过滤算法，其相关的

算法也有许多，主流的算法大致可以分为：关联算法，聚类算法，分类算法，回归算法，矩阵分解，神经网络，图模型以及隐语义模型加以解决。

其中，基于矩阵分解做协同过滤是目前使用得最广泛的一种方法，由于传统的奇异值分解SVD要求矩阵元素不能够稀疏，必须是稠密的，然而协同过滤的推荐的user-item评分一般来说都是一种稀疏矩阵，现在基本使用的都是FunkSVD，BiasSVD和SVD++算法。

假设下面有m=3个用户和n=3个物品，并有用户对物品的评分

|  |  |  |  |
| --- | --- | --- | --- |
| 用户\物品 | 物品A | 物品B | 物品C |
| 用户1 | 1 |  |  |
| 用户2 |  | 3 |  |
| 用户3 |  |  | 5 |

对于每个用户，我们都希望能够预测出用户对空白物品的评分。

1. FunkSVD模型推荐算法

传统的SVD矩阵分解在数据量比较大的时候分解计算效率也会比较，而且存在稀疏矩阵问题，我们在使用SVD进行矩阵分解时候，既然将一个矩阵分解为多个解决，怎么解决稀疏矩阵的问题，于是就会有尝试能不能把一个矩阵只分解成2个矩阵，就如下公式：

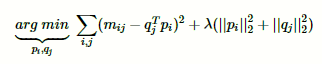
C:\Users\99605\AppData\Local\Temp\1558510976(1).jpg

对于某个用户的评分，也就是矩阵M中的一个矩阵元素，使用FunkSVD矩阵按照上述公式分解后对应，采用均方差作为损失函数：

我们期望这个损失函数能够尽量的小，如果考虑到矩阵M的所有情况的话，则公式就变为：

只要能够最小化上述的公式，并且能够求出公式中极值所对应的,的值的话，就可以得到最终的矩阵P和矩阵Q，对于矩阵M中的任何空白位置都可以通过来进行填补了

在实际应用中，为了防止过拟合，会加入一个L2的正则化项，因此最小化公式为：



其中为正则化的参数，对公式进行求导可以得到：

使用梯度下降法求极值，迭代公式为：

通过迭代求出,的极值，便可以求出P，Q矩阵了。

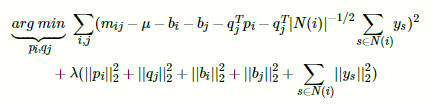
1. BiasSVD模型推荐算法

BiasSVD算是FunkSVD算法的一个改进算法，BiasSVD在用户的评分基础上添加了偏置项，一个是用户偏置项，理解为故意刷差评的用户，物品的评分就会与这个用户无关，另外一个是物品偏置项，理解为一个物品为劣质产品，物品本身就就不可能评分高的与用户无关的。加入偏置项后的公式：

其中为物品评分系统的平均分，为第i个用户的用户偏置项，为第j个物品的物品偏置项。BiasSVD的迭代公式也与FunkSVD差不多，的初始值可以设定为0。

1. SVD++算法

SVD++算法是在BiasSVD的基础上再加以改进的方法，加上了用户的隐式反馈，对于某个用户i，系统提供了隐式反馈的物品集合为N(i)，这个用户的对隐式反馈的修正的评分值为，则该用户的评分修正值为，加入隐式反馈后的目标函数为：



当需要考虑用户的隐式反馈的时候，使用SVD++是一个不错的选择。

<https://www.cnblogs.com/pinard/p/6349233.html>

2.1.3 混合推荐

混合推荐就是通过两种或者多种推荐算法融合，使其能够尽量的避免单推荐算法的缺点[基于婚恋网站数据的推荐系统研究]，本论文所使用的算法便是混合推荐算法，采用两种算法进行推荐，采用kNN算法和逻辑回归混合算法进行推荐。

2.1.3.1 kNN算法:

kNN算法或者说是k-最邻近算法(kNN，k-NearestNeighbor)是数据挖掘分类技术和机器学习技术最简单的方法之一[百度百科]，是由Cover T和Hart P在1967年提出的一种基本分类和回归方法[Cover, T., Hart, P., "Nearest neighbor pattern classification," IEEE Trans Information Theory 13. 21-27.1967.]

kNN的算法思想非常简单，在一个样本的特征空间中，一个样本i的k个最近邻样本大多数属于一个类别A的话，那么这个样本i也是属于类别A，并且拥有这个类别上的特性。。其实质是每个样本都可用它最接近的K邻居来代表, 这倒也应了那句古话, “物以类聚、人以群分”, 其实在数学上的许多方法就是运用了类似思想[[基于KNN与逻辑回归对海难乘客生还率的预测探讨——以泰坦尼克为例](http://kns.cnki.net/kns/detail/detail.aspx?QueryID=1&CurRec=2&recid=&FileName=XDSM201916039&DbName=CJFDPREP&DbCode=CJFQ&yx=&pr=&URLID=&bsm=QS0503;)，陈宇皓，现代商贸工业，2019-5-20，期刊]。

假设存在一个样本数据的集合，也就是样本训练集，每个样本训练集上面都有其所属类别的分类标签，输入没有标签的测试数据后，将新输入的新数据的每一个特征都与训练样本集中的对应特征进行相对应的比较，然后选取样本训练集当中前K个最相似的数据来判断测试数据的类别，这就是kNN算法的工作原理。

举一个简单的例子，我们使用k-近邻算法来判别一个游戏是男性向游戏还是女性向游戏：

|  |  |  |  |
| --- | --- | --- | --- |
| 游戏名称 | 男生游玩人数 | 女生游玩人数 | 游戏类别 |
| 游戏1 | 100 | 200 | 女性向 |
| 游戏2 | 650 | 350 | 男性向 |
| 游戏3 | 300 | 800 | 女性向 |
| 游戏4 | 1080 | 480 | 男性向 |

表2.1.3.1.1：每个游戏男女生的游玩人数及游戏类别

上述的表2.1.3.1.1是我们已知的样本训练集，这个数据集有两个特征[男生的游玩人数，女生的游玩人数]，同时也有每个游戏的所属标签[游戏类别]。现在有一个游戏，如果将这个游戏给我来判断的话，我会根据这个游戏的信息来判断，这和游戏是属于男性向游戏还是女性向游戏，k-近邻算法也可以向我们人一样做出判断，但和我们不同的是，我们的经验更多，更广，k-近邻算法则是只能够依靠数据来判断这个游戏是什么，比如现在这个游戏的数据是男生游玩人数是5，女生游玩人数是100，那么k-近邻算法算出这个是一个女性向游戏，我也会很自然的给你说这是一个女性向的游戏。但是假如一个游戏的数据是男生游玩人数50，女生游玩人数51，这个数据在我看来这个游戏既不是男性向游戏也不是女性向游戏，只能说是综合性游戏，但是k-近邻算法不会告诉我们这是一个综合性游戏，因为k-近邻算法按照表2.1.3.1.1来判断的话，它只能判断出女性向游戏或者男性向游戏，所以此时的分类标准就不会特别标准，就需要在样本训练集当中加入新的特征或者新的数据来更新了。

k-近邻算法是通过特征进行比较的算法，我们继续以表2.1.3.1.1为例来：

从上面的散点图不难看出，绿色点类别的游戏大概率是女性向游戏。k-近邻通过计算两点间的间距来进行判断的，也就是两点间的距离公式：

不过一般常用欧几里得距离公式：

不难看出，绿色点与两个黄色点的距离是最小的，k-近邻算法就会判断这个为女性向游戏，此时k=2。所以k-近邻算法的算法步骤一般为：

①特征提取：特征提取出样本训练集。

②算相似：利用距离公式算出距离排序

③分类：返回前k个点的出现频率最高的类别当作当前的预测点

kNN算法自身操作简略，有效，易于理解，易于完成，无需预计参数，也无需训练。[窦小凡.KNN算法综述[J].通讯世界,2018(10):273-274.]，是一种懒惰的半监督的机器学习算法，对异常值不敏感，但是由于其是通过对所有的样本进行比较，计算复杂度高，空间复杂度也会大大增加，当样本不平衡的时候，其预测的样本就会显得意义不大。

2.1.3.2逻辑回归

逻辑回归也是一个分类算法，通常，逻辑回归用于二分类问题。利用sigmoid函数阈值在[0,1]的特性，主要思想是，根据现有的数据进行分类边界线建立回归公式来进行分类。逻辑回归的sigmoid函数如下：

=

整合后得：

Sigmoid函数为：



其中z矩阵中θ是需要求解得参数，x矩阵是样本列向量，表示θ矩阵的转置，实现能够将矩阵x分布到[0,1]的映射，表示输出为1的概率，当有合适的列向量θ的时候，就可以通过这个训练集所提供的θ，对样本进行计算分类，如果这个样本的输出概率大于0.5，那么就可以说明这个样本为正样本，否则就为负样本。所以接下来就是如何推导出最佳的系数θ。

根据sigmoid函数的特性，我们可以推断出一个假设：

这是一个条件概率，假设已知样本向量x和最佳系数θ，样本x属于正样本的概率和负样本的概率。将公式合并得：

合并成一个代价函数，当y=1时，代价函数得值为，当y=0时，代价函数值为。随后求对数：

根据这个代价函数，给定一个样本，我们就可以求出其相对应得概率，我们希望这个概率越大越好，也就是分类完全正确。转而就变成了求这个代价函数得最大值。自然我们就会想到最大似然估计法，则公式变为：

求取这个函数得最大值，有很多种方法，在这里采用梯度上升算法求解，对θ求偏导得：

其中：

由：

可得：

化简得：

综上得：

根据上面得公式计算得到最佳拟合系数后便可对样本数据进行推测。逻辑回归是二元分类很常见的算法模型，训练速度很快。虽然跟支持向量机(SVM)还是有点差距，但是解决普通的分类问题，逻辑回归算法跟支持向量机相比还是有很大的优势。

2.2开发工具及语言介绍：

开发语言使用python开发，python拥有numpy矩阵运算模块，对于算法开发来说能够提升很大，开发工具使用pyCharm，由jetBrain公司研发的pyCharm能够很好的兼容python语言。

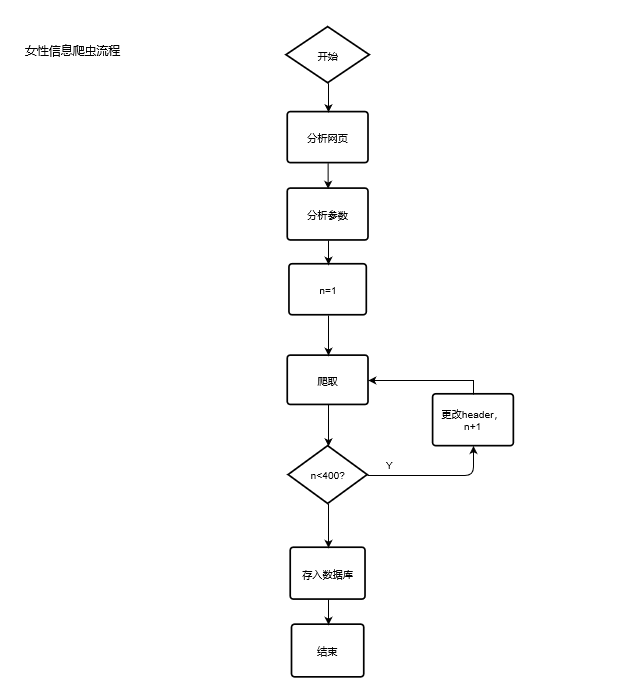
2.3本章小结

本章主要对推荐算法进行研究，从基于内容的推荐到混合推荐，对比比较不同的推荐算法的优劣，从中选取最佳的推荐算法进行推荐。之后阐述了本论文涉及的算法进行介绍和研究，最后介绍本论文的开发工具与开发语言。

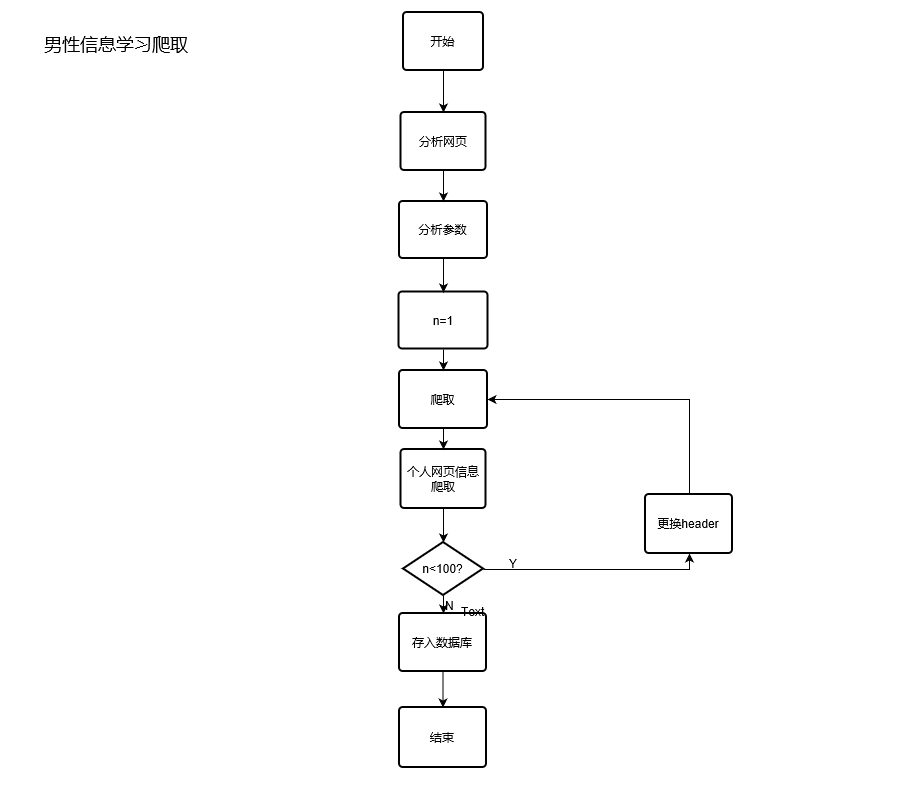
3.数据爬取与数据处理

3.1数据爬取

本次论文所需要的数据均通过爬虫爬取，爬取世纪佳缘网站男生的数据和女生数据。爬取的流程图如下所示：



女性的信息爬取网站url为：<http://search.jiayuan.com/v2/search_v2.php>，该url所返回的信息便是含有信息的json格式文件，每页有25名会员信息，爬取400页共10万条信息。但是由于世纪佳缘由于某种原因，导致100页之后的重复率就会开始提升，所以需要在爬取的过程中去重，去除掉重复的数据，防止其影响算法的准确度。



男性的信息爬取与女性信息的爬取大同小异，但是多了一个步骤，男性的某些信息需要点进其个人的主页上爬取，故需要在爬取信息的同时，在分析每个人的信息之后都会根据其相关信息进入其主页爬取所需要的信息。同样的，世纪佳缘网站上的男性信息都会有重复，在爬取数据的过程需要对数据进行去重处理。

使用爬虫能够很容易的爬取到相关网站的信息，选取世纪佳缘网站爬取也是有原因，实世纪佳缘网站也可以说是国内最老牌的相亲婚恋网站了，其推荐系统成熟程度也是发展得比较成熟。这样通过爬虫爬取得信息特征处理起来也是比较轻松，也更简单。

3.2数据处理

3.2.1数据预处理及特征分析

通过爬虫爬取完信息后就需要对信息进行处理，处理成我们算法能够接受得数据格式，由于数据都是一维数据，故不需降维，首先我们先通过数据得预处理来处理我们爬取到得数据。

我们先来看男性得信息数据表：

表3.2.1 男性信息表

|  |  |  |
| --- | --- | --- |
| 属性 | 类型 | 长度 |
| age | Int | 11 |
| edu | varchar | 50 |
| height | Double | 50 |
| Lng | double | 10 |
| Lat | double | 10 |
| marriage | Varchar | 50 |
| income | Double | 10 |
| matchCondition | Varchar | 255 |

本论文得代码是以男生为视角进行推荐的一个推荐算法，所以男性的信息表不需要太多的信息，原本男性的信息表中有：

(uid,nickname,sex,age,edu,height,location,marriage,nationality,matchCondition,image,shortnote)共12个属性，男性信息作为第一步kNN算法的样本训练集，在第二章有关kNN算法的介绍中，我们也已经知道，kNN算法的样本训练数据集是由特征数据和样本标签组成，灾男性的样本训练集当中，择偶标准就自然而然的成为了样本数据集中的标签数据，现在就需要分析上述的信息表中那些数据是能够影响到标签数据的，那些是不影响的，需要在数据的处理上就要必须的将一些对算法无用的数据剔除。

第一个属性为uid，uid是世纪佳缘网对每一名用户的编号，作为样本数据特征来进行估算的意义不大，因为每个人的uid都是不一样的，所以uid的数据可以作为无意义数据去除。

第二个属性为nickname，nickname是世纪佳缘用户的每个用户的自己昵称，这种自我个人因素大的数据不能作为训练特征，所以nickname数据也不能作为数据特征。

第三个属性为sex，这是每个用户的性别特征，由于这是男性的训练样本，故性别在此的意义也不大，也可去除。

第四个属性为age，age是世纪佳缘每个用户的年龄，年龄是一个很重要的特征，不同年龄段的人的择偶标准也大能不不同，能够影响到训练的结果，age可以保留。

第五个属性为edu，这是每个用户的教育程度，这个数据特征一共有6种数据类型，高中及高中以下，大专，本科，双学士，硕士，博士。不同教育程度的人所决定的择偶标准也会不一样，本科的可能会喜欢大专，硕士，博士，什么样的都有，这个也是一个关键性的因素。

第六个属性为height,这是用户的身高数据，身高自然不用多说，在如今的相亲中，身高也已经成为了一项标准，这个是必须纳入的。

第七个属性为location，这是用户的地域信息，这也是一个非常重要的属性，南方和北方人的择偶标准，东部和西部的择偶标准差距也是非常大的，也是在相亲当中起决定性的作用的一个属性。

第八个属性为marrige，这是用户的婚姻信息，这个数据特征一共有3种数据特征，未婚，离婚，丧偶。不同的婚姻状况也会有不同的择偶标准。这个也是一个必要的属性。

第九个数据为民族，这个民族其实也是一个比较关键性的数据，但是我们在第二章介绍

kNN算法的时候也介绍了其优缺点，其中kNN算法有一个缺点就是当样本的数据特征不平衡的时候，kNN算法所预测的数据也不会有很大的意义。实际上的世纪佳缘的用户中，用户有80%上都为汉族，在如此悬殊的样本数据中，民族的意义就不会显得大了。所以民族的数据特征可以去除。

第十个属性为matchCondition，这是用户的择偶标准，这是作为样本训练集的标签，也是需要保留的。

第十一个属性为image，这使用户的照片数据，保存格式为url网站链接，image照片数据作为预测也是没有太大的意义。

第十二个属性为shortnote，这是用户的一个简短的自我介绍，这个属性也是可以用来做用户预测的特征数据，因为有些用户可能会在自我介绍中介绍自己的择偶观，自己的性格，喜欢什么样的性格，性格在择偶标准的选择上也是起了决定性的作用，自我介绍可以通过中文分词处理技术分词出来一些性格数据或者一些其他的择偶数据。但是也有很多人的自我介绍并没有介绍，导致有可能在中文分词的处理上会有缺失，导致数据缺失，性格特征数据的缺失很难补上，因为性格带有很强的个人因素，而且也考虑到世纪佳缘上一些用户上报虚假信息的原因，性格这个方面的特征缺失填补的工作会有很大的阻碍。所以综合考虑下，通过分词处理得出自我介绍中的信息特征不作为数据特征，故只能去除。

上述的数据分析完后便是表3.2.1的数据表样式，在这里要特意说明以下income属性，这是一个收入属性，在3.1数据提取中有提到选用世纪佳缘的原因，因为其推荐算法也是比较成熟了，爬取下来的数据也已经是特征化处理较高的json数据，也就是上面说分解的各种数据，都是处于json数据当中，但是这些json数据中并没有收入这一数据特征，收入在择偶标准中也是占比较重要的一环，不同的收入个人婚恋观，个人价值观都会有所不同，所以对收入数据就需要进入到每个用户的用户信息界面进行爬取爬到，所以男性用户的信息爬取会比女性用户的爬取多一部爬取收入。

在第二章介绍推荐算法的时候我们也提到，对于这些非结构化的离散型的数据我们要处理成结构化的数据，表3.2.1中的edu教育程度和marriage婚姻状况都是一些非结构化的数据，需要进行量化处理，其中edu教育程度的有6种属性，高中及高中以下，大专，本科，双学士，硕士，博士，这6种属性进行打标签量化，采用最简单的线性量化处理，高中及高中以下为1，之后一次递增。同理婚姻状况也是3种情况一次递增。地域的特征量化就不能通过简单的线性函数来量化处理，但是地域特征有一个其本身具有特色的量化特征——经纬度，经纬度作为量化特征就无需使用函数来量化。择偶标准作为标签向量可以用字符串保存下来。

男性信息通过数据预处理将不必要的信息去除和量化之后可以处理成算发能够接受的数据格式了，现在再来看女性信息表：

表3.2.1 女性信息表

|  |  |  |
| --- | --- | --- |
| 属性 | 类型 | 长度 |
| uid | Int | 11 |
| nickname | Varchar | 50 |
| sex | char | 10 |
| age | int | 11 |
| edu | varchar | 50 |
| height | Double | 50 |
| location | varchar | 50 |
| marriage | Varchar | 50 |
| nationality | varchar | 50 |
| haveimage | int | 10 |
| image | varchar | 255 |
| shortnote | varchar | 255 |
| matchCondition | Varchar | 255 |

本论文的测试方向是向男生推荐，所以女生的信息表就会齐全很多，女性信息表的预处理不需要太多，主要处理的是择偶标准，应用到逻辑回归算法上。所以主要的特征处理为择偶标准的处理。择偶标准为一个字符串，在男性信息表中，择偶标准作为kNN算法的标签向量存在，而女性信息表中的择偶标准就需要当作逻辑回归的训练模型训练出最佳的回归系数。由于世纪佳缘的特征处理已经很成熟了，择偶标准中也将也特征化成一个个特征，其择偶标准的格式为：

择偶标准 = ‘年龄’+‘身高’+‘民族’+‘学历’+‘照片’+‘婚姻状况’+‘地域’+’诚信’

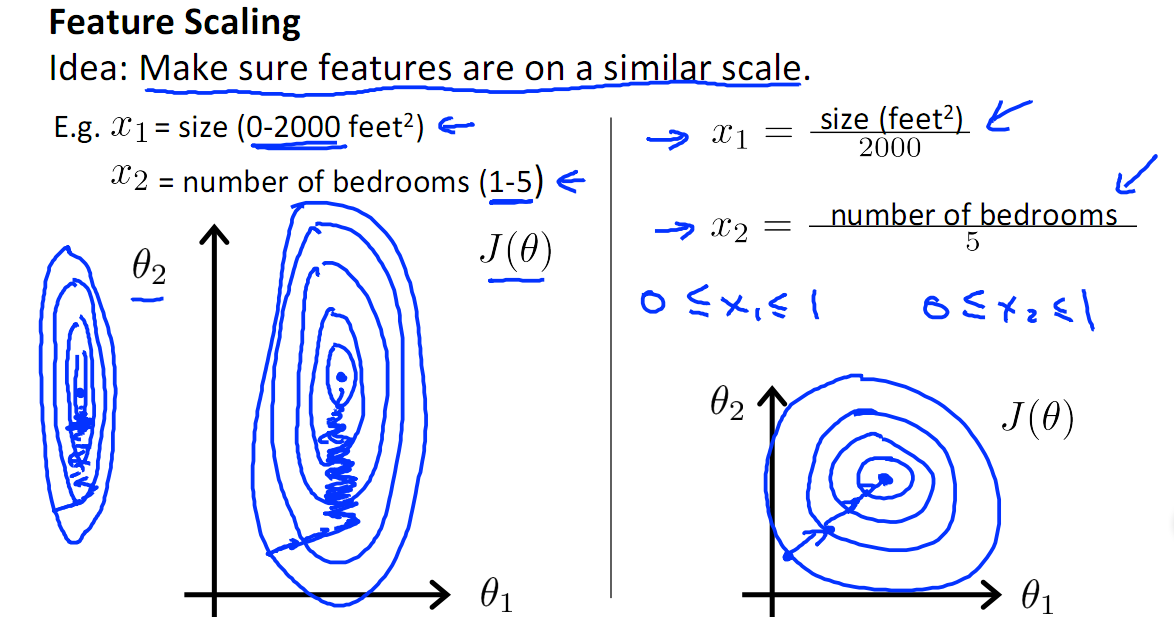
择偶标准由上述的特征组成，但是当用户的择偶标准中是不限的时候，就比如当年龄不限的时候就会将这个省略后放入到json数据中，然而在接收数据的表现就为数据缺失，就需要将这个的数据缺失加以填补上去，数据填补的方式在这先不说，在设计与实现章节将会有详细的介绍。在择偶标准中也是有非结构化的数据，民族，照片和诚信这些都是一些无用的数据，民族因为样本不平衡，照片个人因素过大，诚信是与世纪佳缘网站有关的一些属性，主要表现是付费用户，在此不予以考虑。结构化数据与男性信息大同小异，都可以通过线性函数量化。

3.2.2数据归一化

经过上面的数据预处理，我们得到了量化的数据，可以对数据进行算法处理。但是，这些数据用算法处理起来往往误差都会显得比较大，因为每这些数据都是离散型的数据，数据与数据之间联系不大，就比如地域的经纬度与教育程度，中国国内的经度一般都处于90-120左右的较大数据，而教育程度都是线性函数量化的1，2，3等小数据，数据大小相差过大，导致各个特征的权重就不同，导致严重影响到计算结果。

在处理范围不同的数据特征值的时候，通常使用的方法就是将数据进行归一化处理，将数据特征的取值处理到某个范围之间。去除数据的单位，将数据变成无量纲的纯数值，其中工业上和学术上最常用的规划范围是[0,1]或者[-1,1]，数据归一化后有两个好处：

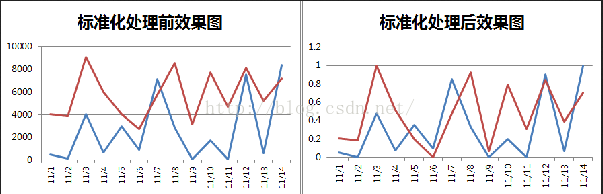
1. 提升模型的收敛速度



上图是未进行归一化和归一化后的数据的梯度下降算法迭代速度图，左图是为未归一化的数据迭代，其迭代曲线是Z型，而右图是归一化后的迭代曲线。从上图就可以很明显的看出有图的迭代速度会更快，收敛的速度也会加快，这样就可以减少迭代的次数，进而提升程序的性能。

1. 提高模型的精度

对数据进行归一化的另一个好处就是提高模型的精度，特别是一些需要通过计算距离来实现的算法，就比如本论文所使用的kNN算法中所使用的欧式距离公式。归一化数据后可以减少精度的损失，让各个特征对结果做出的贡献都是相同的，让不同维度的数据在数值上能够有比较性。



常用数据归一化方法有min-max标准化/0-1标准化，log函数转换，atan函数转换，z-score标准化，decimal scaling小数定标标准化，logistic/softmax变换和模糊量化公式。

这里主要介绍常用的前四种，

1. min-max标准化/0-1标准化

这个归一化方法也叫极差归一化，对原始数据值进行线性函数变换，使得结果可以落到[0,1]的区间，归一化公式如下：

其中min是样本每个列向量的最小值，max-min是每个列向量的极差。归一化的区间也可以变成[-1,1]，此时公式就变成:

其中将分子中min改成样本每列的平均值，min-max标准化/0-1标准化最大的缺陷就是每当有新数据加入的时候都要重新计算样本列向量的最小值或者平均值或者标准差。

1. log函数变换

log函数变换公式如下：

其中max是每列函数的最大值，这个归一化也有一个缺陷就是每个样本特征的特征值必须大于1才能够进行归一化，归一化的区间是[0,1]

1. atan函数转换

atan函数是反正切函数，公式为：

反正切函数的归一化函数区间有所不一样，当x>0时，函数的归一化区间落在[0,1]区间，当x<0时，函数的归一化区间落在[-1,0]这个区间上。

1. z-score归一化

z-score归一化是最常见的归一化方法了，在SPSS上也是一个最常用的归一化方法，z-score归一化是将特征值的均值和标准差进行标准化，经过归一化的数据符合正态分布，转化公式为：

其中为所有样本数据的平均值，为所有样本的标准差，z-score比较适合的数据特征格式是数据的最大值和最小值不知道的情况，或者超出范围的离群数据。

数据归一化是一个很重要的过程，本论文采用的归一化算法是min-max标准化/0-1标准化，这个归一化算法对于距离计算来说优化相对来说是比其他的优化算法要好，而且实现简单。数据归一化完成后数据处理阶段就算是基本完成。

3.3本章小结

本章节主要为数据的处理，通过爬虫将世纪佳缘网站的数据信息爬取下来，随后对数据进行数据预处理，由于世纪佳缘网站的数据特征分析程度高，特征分析的工作也会比较轻松。对数据进行预处理完成后便对数据进行归一化处理，主要介绍了归一化数据的一些方法和本论文使用的归一化方法。

4.kNN婚恋推荐算法的设计与实现

3.1kNN婚恋推荐算法的设计

首先我们先来看以下本论文算法的算法流程图：