基于Knn的相亲网站推荐算法研究

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。

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

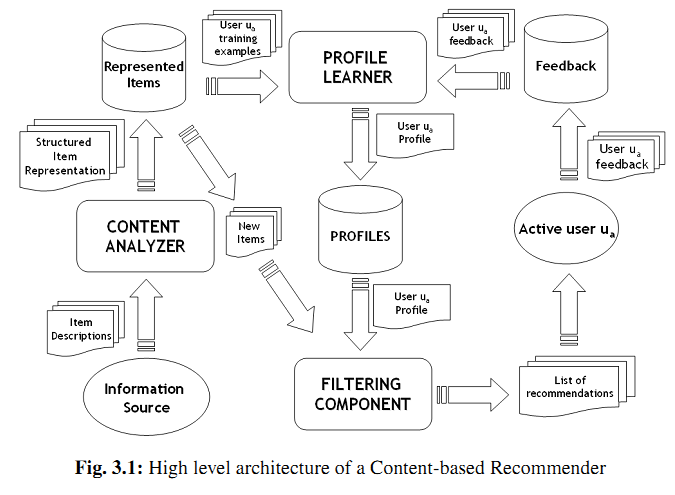
推荐算法由来已久，在第一章绪论的1.3节推荐算法国内外发展现状也提到，推荐算法发展到现在，其背后的技术也基本形成了3个形式：1)基于内容模型推荐，2）基于协同过滤模型推荐，3)混合模型推荐

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：

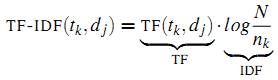
1. Item Representation：为每一个用户的item抽取出用户相应的用户特征
2. Profile Learning：利用用户item的特征数据放入算法学习用户喜欢的
3. Recommendation Generation：根据Item Representation与Profile Learning得出的用户相似喜欢物品对用户进行大量推荐

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

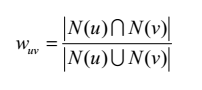
协同过滤的模型一般为有x个item，有y个item的用户评分数据，并且y<x，，需要使用y个的稀疏数据来预测那些空白item和数据之间的评分关系，并将最高的评分推荐给用户。一般来说，协同过滤分为三种类型，第一种是基于用户的协同过滤（user based），一种是基于项目的协同过滤（item-based），第三种是基于模型的协同过滤（model-based）

1. 基于用户的协同过滤（user based CF）

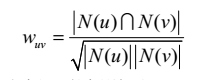
通俗地来说，要给用户推荐一件物品，首先先计算这个用户与其他用户相比的相似度，

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

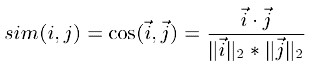
用户C喜欢电影b,d

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。

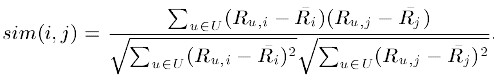
首先，我们先使用基于内容的推荐算法，根据上面所说的步骤，先分析用户A的item即电影a,b,c的特征属性，假设a—喜剧片，b—喜剧片，c—爱情片。如果分析的出f电影的特征为喜剧片，那么该推荐算法很大的几率会将电影f推荐给用户A，反之则相反。这就是基于内容的推荐，需要对item的特征进行建模，

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。

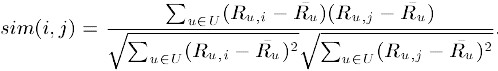
上面的比较中，我们可以知道，基于项目的协同过滤的过程非常简单，就是计算user-item之间的关系的相似程度，相似度的计算算法也可以是计算余弦的相似程度：



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