学　　号 2020302181189

武汉大学本科课程论文

《社会计算》结课论文:
基于社区网络的新闻精准推荐系统研究

院 (系) 名 称:　国家网络安全学院

专 业 名 称:　信息安全

学 生 姓 名:　李勃衡

指 导 教 师:　石小川　副教授

二〇二一年十二月

COURSE ESSAY
OF WUHAN UNIVERSITY

Research on News Recommendation System Based on Community Network

School (Department): School of Cyber Science and Engineering

Major: Information Security

Candidate: LI Boheng

Supervisor: A.P. Shi Xiaochuan

Wuhan University

December, 2021

郑 重 声 明

本人呈交的课程论文, 是在导师的指导下, 独立进行研究工作所取得的成果, 所有数据、图片资料真实可靠. 尽我所知, 除文中已经注明引用的内容外, 本课程论文的研究成果不包含他人享有著作权的内容. 对本论文所涉及的研究工作做出贡献的其他个人和集体, 均已在文中以明确的方式标明. 本论文的知识产权归属于培养单位.

本人签名: _____ 日期: _____

摘　　要

本文主要讨论了一种基于社区网络发现算法的新闻精准推荐算法模型. 首先对现有的推荐系统算法和社区发现算法进行概述，并对现有的推荐系统和社区发现算法进行了介绍，总结了前人的工作。随后提出本文的基于 Fast-Unfolding 算法的新闻推荐系统模型的构建方法，并对算法做了实验验证和评价，最后指出了算法存在的考虑不周到之处和改进方向。

关键词: 新闻推荐系统; 社区发现算法; Fast-Unfolding 算法

ABSTRACT

This article mainly discusses a news accurate recommendation algorithm model based on the community network discovery algorithm. First, the existing recommendation system algorithm and community discovery algorithm are summarized, and the existing recommendation system and community discovery algorithm are introduced and summarized. The work of predecessors. Subsequently, this article proposed the construction method of the news recommendation system model based on the community discovery algorithm, and did experimental verification and evaluation of the algorithm, and finally pointed out the imperfect considerations and improvement directions of the algorithm.

Key words: News recommendation system; community discovery algorithm; Fast-Unfolding algorithm

目　　录

1 引言

1.1 推荐系统概述

随着互联网时代的飞速发展，网络空间的信息呈爆炸式增长，人们的生活已然进入大数据时代。以新闻为例，在互联网时代之前，人们获取新闻信息的主要渠道只有报纸、杂志和人与人之间的口口相传等等，获取信息的数量和质量都极其有限，效率也低。互联网时代后，人们通过互联网新闻网站能迅速获取到大量新闻信息，但信息爆炸的后果在于众多繁杂纷多的信息也给人们带来一种难以从大量信息中获取到自己所需要信息的信息超载问题。

为解决信息超载问题，相关学者和专家提出了包括搜索引擎、目录分类等方案在内的解决方案。搜索引擎方法允许用户通过输入自己想要搜索的关键词来获取包含有这些关键词的网页；目录分类方案将信息依据关键词分为很多板块，比如网易新闻[1] 将他们的新闻分为如图1.1所示的国内、国际、数读、军事等等板块，方便需要查找相应信息的用户快速找到相应板块。

图 1.1 网易新闻的目录分类方案

搜索引擎的好处在于可以让用户从大量信息中筛选出自己想要的信息，但不足之处也显而易见-—当用户无法准确描述自己的需求，抑或是只想看一些可能感兴趣的新信息时，搜索引擎将无法发挥作用。网站目录分类也难以解决类似的问题。因此，推荐算法应运而生。

推荐算法[2, 3] 是指根据用户的历史操作信息提取用户特征，主动的为用户推荐有用信息和建议的软件工具，以此来帮助用户对海量信息进行挑选。推荐系统的研究在上世纪 90 年代中期开始成为了一个独立的研究方向，近年来人们对这个

领域的研究兴趣与日俱增进而出现了个性化推荐系统。

个性化推荐系统[4, 5] 根据每个用户的历史操作记录构建符合每个用户的兴趣模型，然后将待推荐的信息和用户独有的兴趣模型进行匹配，当达到一定阈值时推荐给用户，这样就实现了因人而异的个性化需求，使得推荐的信息对于用户来说更加精准、贴切。个性化推荐系统应用广泛，以今日头条、百度新闻、Google 为代表的个性化新闻推荐系统都在各自领域中发挥着不可替代的作用。

1.2 社区发现概述

瑞格尔德 (Rheingold) 是最早对社区概念进行描述的学者，他将社区定义为“一群通过计算机网络进行沟通交流的人们，他们彼此有某种程度的认识、分享某种程度的知识和信息、在很大程度上像对待朋友般关心彼此，从而形成的团体”[7]。在真实世界里，用户可以看作抽象的节点，用户之间的关系看作抽象的边，大量的节点与节点之间边的关系抽象成一个由点集与边集构成的复杂网络。复杂网络中的社区发现问题最早由 Girvan 和 Newman 在 2002 年提出，他们基于实证数据发现，复杂网络中普遍存在社区结构，即复杂网络的结构普遍具有“同一社区内的节点相互紧密连接，而不同社区间的节点相互稀疏连接”的特点 (Girvan，Newman，2002)。传统的社区概念通常认为网络中的个体只能属于一个社区，即非重叠社区；而在许多复杂网络中，节点可能同时属于多个社区，如一个科学家可能会对好几个研究领域感兴趣。Palla 等 (Palla,Derenyi et al.,2005) 对传统社区概念进行了扩展，允许一个节点同时属于多个社区，社区之间可以有交叠重合的部分，并基于此提出了重叠社区发现算法。

社区的发现算法主要有基于模块度优化的算法、基于概率模型的算法以及基于信息论的算法等。

基于模块度的划分算法最早由 Girvan 和 Newman 提出 [8]，通过移除社区间边介数最大的边来进行社区划分，后又引入模块度的计算方法来评价社区划化的优劣。模块度的提出推动了社区发现技术的研究，不断有学者针对基于模块度的方法进行优化改进，社区发现的问题转为优化问题，通过优化目标函数来发现社区结构，如: 基于模块度优化的算法，极值优化算法、模拟退火算法等。

基于概率模型的算法主要是基于混合模型的算法，Newman 等人基于社团结

构期望最大化建立混合模型 [9]，运用概率计算方法研究复杂网络中的社团，从而避开社团定义的问题。基于信息论的算法是由 Martin Rosvall 等人基于信息论提出 Infomap 算法 [10]，使用随机游走得到信息流并找寻最优压缩信息流的方式来进行社区发现。

1.3 本文组织结构

本文主要研究基于社区网络的新闻精准推荐系统，同时结合社区发现算法和传统的协同过滤推荐算法构建用户关系网络进行推荐，并尽力提高推荐算法的准确性和多样性实现精准推荐，提高用户留存率和召回率。文章组织结构如下:

1. 现有技术简介：简要介绍国内外关于新闻推荐系统的发展现状、介绍常见新闻推荐算法和社区发现算法的具体内容，并分析当前的推荐算法的不足之处。
2. 基于社区发现算法的推荐算法：介绍本文提出的结合社区发现算法和传统协同过滤推荐算法的新闻推荐算法模型。该模型通过用户信息和社区发现算法构建相似度关系网络，并在用户相似度关系网络中运用社区发现算法寻找并划分新闻兴趣相似的社区群体，然后对不同的社区群体中运用协同过滤算法进行推荐。
3. 算法验证和评价：本文采用 CCF 协会提供的新闻数据集对本文算法进行验证和评价，对实验结果进行分析，并和其他传统算法进行比较。
4. 总结与展望：首先总结了本文的研究工作，然后对论文中的不足之处与缺陷进行分析，展望改进方法和未来的研究方向。

2　现有技术介绍

2.1　推荐系统简介

随着推荐系统的不断发展，在各大网站都能看见推荐算法的身影——如淘宝、京东等购物网站的“猜你喜欢”页面、百度知道、知乎等平台等推荐回答、QQ 音乐、网易云音乐等音乐软件的私人电台等等功能都离不开推荐系统的支持。推荐系统通过获取用户兴趣和待推荐物品的信息构建抽象模型，并通过算法匹配用户可能感兴趣的物品进行推荐。推荐算法的工作流程如图2.1 所示。

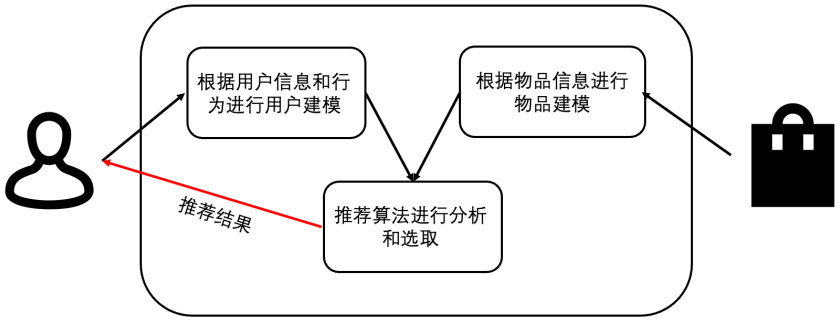

图 2.1　推荐算法的工作流程

2.2　常见推荐算法概述

目前常见的推荐算法包含基于内容的推荐算法、协同过滤算法、基于关联规则的推荐算法和混合推荐算法等等，下文将对他们做简要介绍，并分析其优点及不足之处。

2.2.1 基于内容的推荐算法

基于内容的推荐算法是一种建立在项目内容和用户兴趣信息模型上作出推荐的推荐算法。它将用户兴趣和信息内容抽象化成特征向量矩阵，并计算两者之间的相似度，然后选取和用户兴趣相似度较高的信息推荐给用户。算法执行流程如流程图所示。

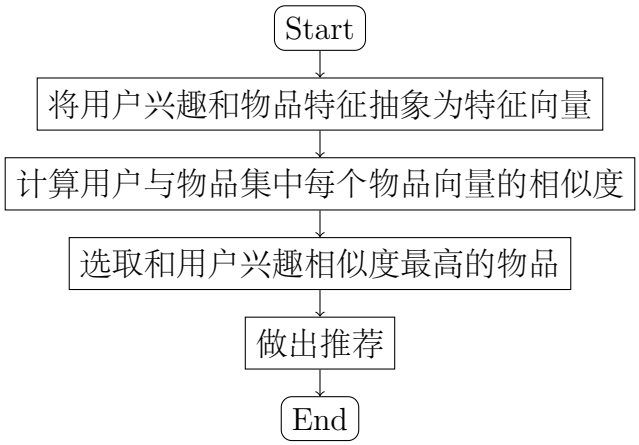

图 2.2 基于内容的算法流程图

基于内容的推荐算法优点在于算法原理简单且成熟，采用的决策方法具有较高的可解释性。由于只需要当前用户和物品的特征向量，该算法不需要其他用户的信息就能完成推荐，并且能根据用户的喜好推荐相似度高的信息，能满足需求特殊、爱好小众的用户。缺点在于需要对每个用户和信息内容都抽象成有意义的数学特征模型，这要求内容有良好的结构性或者明显的标签，而现实世界的用户兴趣和信息内容往往都是复杂的、不断改变的，想要设计出优秀的特征抽取算法是比较困难的。除此之外，基于内容的推荐算法完全依靠用户当前的兴趣来做出推荐，无法挖掘出用户可能感兴趣的潜在信息，容易形成信息茧房。

2.2.2 协同过滤算法

协同过滤算法基于这样一个事实——品味相似的人会对相似的事物感兴趣。算法通过对用户历史行为数据的挖掘发现用户的偏好，基于不同的偏好对用户进行群组划分并推荐品味相似的商品。协同过滤算法通常采取最近邻技术，通过用户历史偏好信息计算不同用户间的距离，再运用与目标用户距离最近的用户对某

项目的加权评价值以推测目标用户对相应项目的偏好程度，最后系统将根据偏好程度进行推荐。其基本思想是：利用用户间的相似性进行推荐，挖掘用户新的兴趣点，最终做出推荐。协同过滤算法的工作原理如图2.3所示。

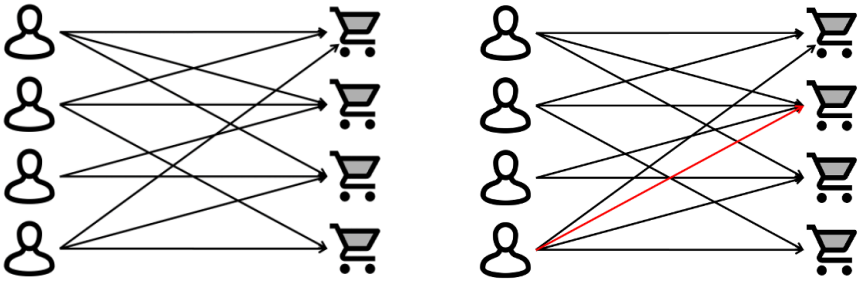

图 2.3　协同过滤推荐算法的工作流程

协同过滤算法需要首先用数学模型表示用户的兴趣，计算用户间的距离后形成近邻用户集，然后利用目标用户的最近邻居用户对商品评价的加权评价值来预测目标用户对特定商品的喜好程度，系统从而根据这一喜好程度来对目标用户进行推荐。

协同过滤算法的最大优点在于不需要对物品进行特征划分，而是从用户角度出发寻找相似的用户来进行推荐，能处理难以直接用语言描述、难以直接结构化的兴趣（比如艺术品爱好，音乐爱好等），也有推荐新信息的能力。由于协同过滤算法是通过相似用户的角度出发进行推荐，有较大概率能挖掘出用户的潜在爱好。且当使用时间足够长、用户数据集足够大时，协同推荐算法的表现会越来越好。

协同过滤算法的缺点在于由于完全依靠用户进行推荐，推荐质量完全依赖用户历史行为和数据集的数量和质量。因此当初期用户量较少、信息也不明朗时，推荐算法的表现会不尽人意。但当用户数量较多、行为信息也逐渐增加时，用户之间的相似度计算和集合划分的性能又会成为瓶颈。

2.2.3　基于关联规则的推荐算法

基于关联规则的推荐算法通过在大量数据中分析并挖掘出物品与其他物品之间潜在的关联性，然后再通过关联集合做出推荐。最典型的案例莫过于著名的“啤酒与纸尿布”问题，商场通过啤酒和纸尿布在同一购物篮中反复出现这一现象挖

掘出了两个看似不相似的物品之间存在的潜在联系——都是婴儿父亲易购的商品，从而通过捆绑销售和推荐提高了物品的销量。

如图2.4所示，推荐算法通过挖掘历史数据中的信息发现了苹果和香蕉的潜在关联规则，并以此作为依据向已经购买香蕉的用户推荐苹果。

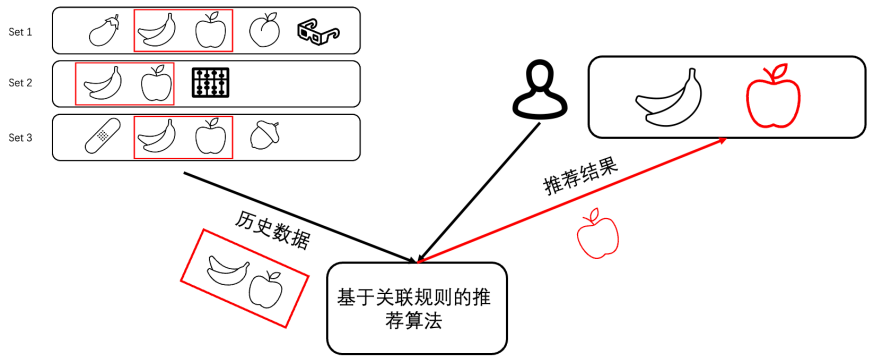

图 2.4　基于关联规则的推荐算法工作流程

关联规则的概念最先由 Agrawal 等人在 1993 年提出，其中最具代表性的算法 Apriori 算法作为关联规则研究领域的重要方法被大量讨论与研究。基于关联规则的推荐主要思想是通过对大量历史数据的分析与挖掘，发现物品和物品之间存在的关联关系，进而根据用户的历史行为记录，推荐用户曾经选购的某件物品相关联的物品。基于关联规则的推荐算法简单，易于理解，但是提取关联规则的过程会花费大量的时间，随着项目集数量的不断增加，整个挖掘过程会极大的消耗系统性能；随着系统中规则数量的上升也会给系统的管理带来隐患。

2.2.4　混合推荐算法

混合推荐算法结合前文介绍的多种算法，在不同需求、同一需求的不同时期采用不同的推荐方法和模型，或是同时采用多种推荐模型，然后通过决策算法得到最终的推荐结果，从而避免单一推荐算法的缺陷与不足，达到最好的推荐效果。在实际应用过程中一般都采用以协同过滤为主的混合推荐算法。

2.3 社区发现算法概述

社区发现方面的研究是为了揭示网络的聚集行为。复杂网络的社区发现实际上是一种网络聚类的方法，它通过对网络图的分析与重建构造社区网络，使得网络社区内部的结点连接性较社区间结点连接性要强。将网络划分为社区的好处在于，可以发现更为相似、连接性更强的群体，利于研究其局部特性和关联关系。

常见的社区发现算法（包含重叠社区和非重叠社区的发现算法）包括层次聚类算法、图分割算法、基于模块度优化的算法、谱分析算法和基于动力学的算法、基于标签传播的算法、InfoMap 算法、派系过滤算法等。本文讲简要介绍层次聚类算法、基于模块度优化的算法和派系过滤算法。

2.3.1 层次聚类算法

层次聚类方法的基本思想是: 通过某种相似性测度计算节点之间的相似性，并按相似度由高到低排序，逐步重新连接各节点。该方法的优点是可随时停止划分，主要步骤如下:

Step 1. 移除网络中的所有边，得到有 n 个孤立节点的初始状态;

Step 2. 计算网络中每对节点的相似度;

Step 3. 根据相似度从强到弱连接相应节点对，形成树状图;

Step 4. 根据实际需求横切树状图，获得社区结构。

层次聚类方法不需要指定网络的社区个数和社区规模，但不能确定网络的员忧划分。此外，层次聚类方法依赖于节点相似度的衡量标准，可能会将某些节点划分成单独的社区，或不能正确划分网络的外围节点。

2.3.2 基于模块度优化的算法

2004 年，Newman 和 Girvan 提出了一个用于刻画网络社区结构优劣的量化标准——模块度 (Modularity) 函数 Q，其基本思想是将划分后的社区与相应的零模型进行比较以确定划分的质量。并以此衍生出很多依靠模块度进行优化的算法。下文将对它们做简要的介绍。

2.3.2.1 社区模块度

社区发现算法的主要目的是将图划分为多个社区集合，使得划分后社区内部的链接较为紧密，而社区与社区之间的链接较为稀疏。模块度 Q 是 Newman 和 Girvan 于 2004 年提出的一个用于刻画网络社区结构优劣的量化标准，其基本思想是将划分后的社区与相应的零模型进行比较以确定划分的质量。模块度越大，代表划分后集合在社区内部的节点相似度越高，而在社区外部节点的相似度越低，划分算法的质量越高。社区模块度的计算方法如公式2.1所示：

$$Q = \frac{1}{2m} \sum_{l=1}^{k} \sum_{i \in C_l, j \in C_l} (A_{i,j} - \frac{d_i d_j}{2m}) \tag{2.1}$$

公式中，m 是网络 C 中的总边数，社区发现算法将网络 C 划分为 k 个社区 $C_1, C_2, \ldots, C_l, \ldots, C_k$，$A_{i,j}$ 表示结点 i 和结点 j 之间是否存在连边，若存在连边为 1、不存在连边为 0。d_i, d_j 分别表示结点 i 和结点 j 的度，系数 $\frac{1}{2m}$ 将模块度的值标准化为 $(-1, 1)$ 之间的值。模块度 Q 越大，代表社区发现算法的划分质量越高。一般认为，Q 大于 0.3 时，网络中存在明显的社区结构。

2.3.2.2 G-N 算法

G-N 算法是典型的分裂型非重叠式社区发现算法，其基本思想是不断地删除网络中具有相对于所有源节点的最大边介数的边，然后再重新计算网络中剩余的边相对于所有源节点的边介数，重复这个过程，直到网络中所有边都被删除。在分裂过程中，可在任意时刻终止，并输出当前结果作为发现的社区结构。

GN 算法由于每次剔除最大边介数的边后需要重新计算边介数值，导致计算时间复杂度较高，并不适用于大型网络中。

2.3.2.3 FN 算法

针对 G-N 算法在大型网络中计算时间复杂度较高的缺点，Newman 对其进行了改进，提出了基于模块度优化的快速社区发现算法 FN 算法。该算法的基本思路为：初始化时，候选解中每个社区仅包含一个节点，在每次迭代时，选择并合并两个现有的社区，根据优化函数的方向，FN 算法选择使得 Q 函数值增加最大 (或

减小最少) 的社区对合并；当候选解只对应一个社区时，算法结束。通过这种凝聚式的层次聚类过程，FN 算法输出一棵层次聚类树，并将 Q 函数值最大的社区划分作为最终聚类结果。

2.3.2.4 Fast-Unfolding 算法

Fast Unfolding 算法是基于模块度优化的一种社区发现算法，该算法的主要目标是通过迭代的方法不断划分社区，使划分后的模块度不断增大。该算法主要包括两个阶段: 第一个阶段将每个节点作为一个初始独立社区，然后计算节点合并到邻接节点的社区后模块度的变化量 ΔQ ，如果 $\Delta Q > 0$ 则将节点与邻接节点合并，如果 $\Delta Q < 0$; 则保持节点不变，第二个阶段根据第一阶段形成的新的社区结构重复上面的节点的合并过程，直到变化量 ΔQ 不再增加为止。

2.3.3 派系过滤算法

CPM 派系过滤算法是最早的重叠社区发现算法，算法的主要目标是寻找网络中的 k-派系，k-派系表示网络中含有 k 个节点的完全子图。社区内边的连接密度较高，使得社区内部容易形成团，这些团一般表现为全连通子图；而社区之间的边形成团的可能性较小。派系过滤算法通过合并全连通子图来构建重叠社区。首先从网络中找出所有大小为 k 的团；然后将每个 k 团作为节点构建一个新的图，当两个 k 团共享 $k-1$ 个节点时，新图中两个对应的节点之间才有边。则新图中每个连通子图所对应的 k 团集合即构成了一个社区。

2.4 本章小结

本章主要介绍了现有的常见推荐系统算法，并分析了它们的原理、特性和不足之处；接着对社区发现方面的研究做了简要论述，阐述了社区发现的重大意义和现有进度，并介绍了目前常见的社区发现算法思路及其优缺点。

3 基于社区发现算法的新闻推荐系统模型

传统的协同过滤推荐算法的一大缺点在于，当用户数量较多、行为信息也较多时，用户之间相似度计算和集合划分将会变得缓慢且困难。规模不断增长的用户群不仅会带来计算开销的增加，也会导致用户项目评分矩阵中用户维度的增加，带来数据稀疏性的问题。而研究 [4] 发现，网络上兴趣爱好相似的用户群体往往符合网络社区的特性——兴趣爱好相似的用户群体往往聚集成小团体和社区，社区内部的用户有着共同或相似的兴趣；而不在同一社区内的用户则具有不同的兴趣。因此，使用社区发现算法发现网络中潜在的社区群体并对其进行划分，能有效缩小用户的推荐计算范围到一个社区内，对于减少计算开销、提高推荐系统的准确率有着重要的作用。因此，本文提出了一种结合传统协同过滤算法和社区发现算法的算法模型，并期望能提高传统协同过滤推荐算法的准确性和效率。

3.1 推荐模型介绍

传统的社区发现算法在计算时只考虑节点之间连边的关系，忽略了节点自身的相似性，在社区划分中可能两个相似性较高但是没有连边的结点被划分在了不同社区中。因此，本算法主要思路是计算用户浏览新闻的内容相似度，然后利用该相似度建立用户网络，在建立的用户关系网络中采用社区发现算法进行社区划分，再通过划分后的社区内部用户浏览过的新闻对新未浏览用户进行推荐。算法模型的流程图如图3.1 所示。

3.2 新闻文本模型构建

要通过用户浏览新闻的内容相似度建立用户关系网络，首先需要计算用户浏览的新闻的相似度，而新闻文本本身是非结构化的文本信息，因此需要先构建新闻文本向量空间模型，本模型运用 TF-IDF 算法来计算文本中的特征词权重，并

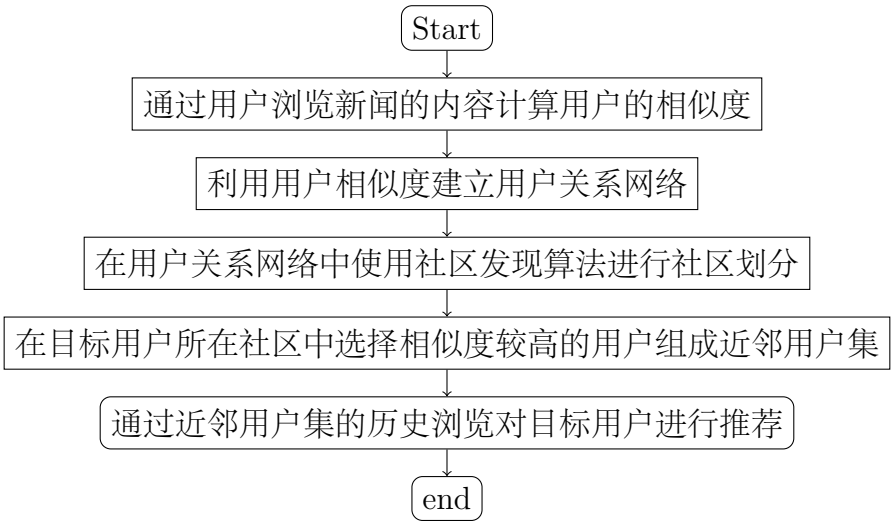

图 3.1　基于社区发现的新闻推荐算法流程图

以此构建用户的新闻文本特征词向量矩阵作为用户的新闻兴趣模型。

3.2.1　新闻文本预处理

3.2.1.1　中文文本分词

中文新闻的文本由于没有空格分割，无法像英文一样直接筛选关键词，而是需要先进行中文分词，将大段新闻文本信息切割成一个一个词语，方便进行词性标注、关键词筛选和情感分析、语义分析等操作。本文使用 python 开源库 jieba 来进行中文分词。jieba 是 python 的一个先进的中文分词库，特别适合中小型文本的分词、词性标注和关键词筛选，同时还支持基于 TF-IDF 算法的关键词抽取。

3.2.1.2　无效词语过滤

由于新闻是完整的文章段落，经过分词处理之后会出现大量的零散词语。但这些零散词语并不都是有效的——比如“的”、“了”、“因为”、“但是”等词语不具有实际意义，而是作为功能性词汇组织文章段落，这些词语对于具体新闻的关键阐述对象没有直接的联系，也不能有效体现用户兴趣。因此，在文本预处理过程中本文将过滤掉这类无效词语。

经过对文本的简要分析，名词、动词、形容词和副词较能反映出新闻的主体、

动作和程度，而介词、连词、代词等词性的词语具备的实际意义有限。因此对新闻预处理过程中，首先将文本进行分词处理和词性标注，之后只保留名词、动词、形容词和副词，而其他词性的词将不予保留。

新闻文本预处理的过程如图3.2 所示。

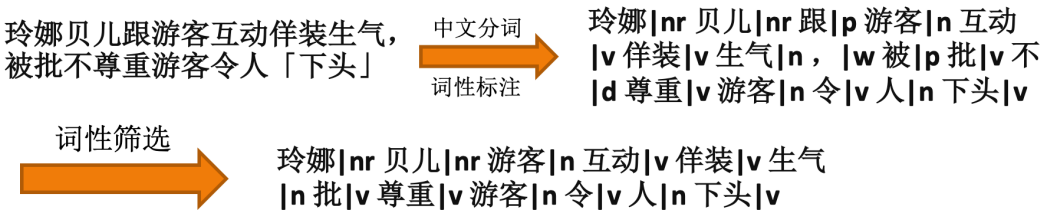

图 3.2　新闻文本预处理过程

3.2.2　新闻文本模型构建

目前常见的文本模型构建方法有向量空间模型、概率检索模型、布尔模型、语言模型等。向量空间模型使用简单且效果理想，因此本文在构建用户新闻文本模型阶段采用向量空间模型来进行新闻文本模型构建。

向量空间模型的基本思想是将文本内容表示为空间中的向量，每一段文本被处理为一个以特征词和特征词权重为分量的 N 维向量。这样，文本表示为 N 维空间中的一个向量，比较文本之间的相似度就可以通过比较向量之间的相似度来完成。

要将文本处理为以特征词和特征词权重为分量的 N 维向量，首先需要对文本进行特征词权重的计算。常见的特征词权重计算方法包括 TextRank、词频法和 TF-IDF 算法。本文采用 TF-IDF 算法进行文本特征词处理。TF-IDF（Term Frequency - Inverse Document Frequency）是一种思想简单但高效可靠的特征词权重计算方法，它的基本思想是，在一篇文章中重复出现、但在所有文本中出现频率低的词汇即是该文章的关键词。TF（Text Frequency）表示一段文本中一个词语的出现频率，一个显而易见的结论是，在同一段文本中反复出现的词是这篇文章中较为重要的词。TF 的数学表达式如公式3.1 所示。

$$TF(k) = \frac{n_k}{\sum_i^n n_i} \tag{3.1}$$

其中，$TF(k)$ 表示第 k 个特征词在该篇文本中的 TF 值，n_i 表示第 i 个特征词在本文本的出现次数。结合公式容易看出，一段文本中，一个特征词的 TF 值实际就是该特征词在此文本的出现次数比上该文本中所有特征词的出现次数。

IDF 表示逆文档频率，TF-IDF 认为，在所有文本中出现频率都很高的特征词，即使其在一段文本中出现频率很高，也不一定说明其就是这段文本的关键特征词。因此，需要将特征词在所有文本中出现的频率作为修正因子来平衡关键词的筛选权重。IDF 的数学表达式如公式3.2 所示。

$$IDF(k) = log\frac{\mid D \mid}{\mid \{j : t_k \in d_j\} \mid} \tag{3.2}$$

公式中，$IDF(k)$ 表示第 k 个特征词的 IDF 值，D 表示文档集合中文档的总数量，分母 $\mid \{j : t_i \in d_j\} \mid$ 表示所有文本中，出现该特征词的文本个数。

最终，TF-IDF 的计算公式如公式3.3 所示。

$$TF - IDF(k) = TF(k) * IDF(k) \tag{3.3}$$

对于给定的新闻数据集 $N = \{n_1, n_2, \ldots, n_k\}$ 针对每篇文本中的每个特征词，均使用 TF-IDF 算法计算其权重，最终形成如3.4 所示的向量空间模型：

$$NM = \begin{bmatrix} w_{11} & w_{12} & w_{13} & \ldots & w_{1j} \\ w_{21} & w_{22} & w_{23} & \ldots & w_{2j} \\ w_{31} & w_{32} & w_{33} & \ldots & w_{3j} \\ w_{41} & w_{42} & w_{43} & \ldots & w_{4j} \\ \ldots & \ldots & \ldots & \ldots & \ldots \\ w_{i1} & w_{i2} & w_{i3} & \ldots & w_{ij} \end{bmatrix} \tag{3.4}$$

其中 $[w_{i1}, w_{i2}, w_{i3}, \ldots, w_{ij}]$ 表示第 i 段文本的特征向量，w_{ij} 表示第 i 段文本中第 j 个关键词的权重。

3.3 社区模型构建

构建好新闻文本模型后，即可通过新闻模型构建用户的新闻内容特征模型，并通过模型矩阵计算用户之间的相似度。之后通过用户相似度对用户结点进行连边操作构造社区网络，然后使用社区发现算法在社区网络中进行社区划分，构造好社区模型。

3.3.1 用户兴趣相似度

本文通过用户浏览过的新闻相似度作为用户的兴趣相似度的评判标准，根据用户浏览的历史阅读数据集，使用上文提到的向量空间模型，使用 TF-IDF 算法计算特征词权重，并以此构建用户新闻内容特征矩阵模型，然后通过计算用户特征矩阵模型的相似度来代表用户兴趣相似度。

假设用户 u、v 的特征矩阵分别为 NM_u、NM_v，采用余弦相似度作为计算方法，则 u、v 用户的兴趣相似度 $sim(u,v)$ 的计算方法如公式3.5 所示。

$$sim(u,v) = \frac{NM_u \cdot NM_v}{\mid NM_u \mid\mid NM_v \mid} \tag{3.5}$$

3.3.2 建立用户结点网络

在前面的步骤中我们已经计算出了用户兴趣相似度，接下来通过兴趣相似度来构建用户关系结点网络，具体步骤是：通过实验确定用户关系网络的用户兴趣相似度阈值 α，如果两用户 u、v 之间的相似度 $sim(u,v) > \alpha$，则认为两用户结点之间相似度较高、可能在同一近邻集群中，则对两结点连边；否则认为两用户之间相似度较低，不做处理。

3.4 基于社区网络发现算法的新闻精准推荐系统

构建好社区网络后，我们通过在社区网络上应用社区发现算法发现网络中的社区，并以此为依据作为新闻精准推荐。推荐系统的具体构建步骤包括通过社区发现算法划分用户兴趣关系网络社区，和社区内推荐两步。

3.4.1 社区发现算法

考虑到推荐系统中，用户数量和推荐项目的数量较大，且出于优化传统协同过滤算法性能的考虑，本文将采用时间复杂度较低的社区发现算法进行社区划分。传统的 G-N 等算法划分耗费时间较长，本文选择前文介绍过的 Fast-Unfolding 算法作为社区划分算法。

Fast-Unfolding 算法的具体执行步骤如下：

Step 1. 移除网络中的所有边，得到有 n 个孤立节点的初始状态；将每个孤立结点 i 作为独立的初始社区；

Step 2. 对于每个节点 i，将该节点从其社区中移除并将其放入邻近节点 j 构成的社区中，计算改变前后模块度变化量 ΔQ，找出模块度变化量最大的邻近节点，如果该模块度的变化量 $\Delta Q > 0$，则将节点 i 分配到邻近节点 j 所在的社区中，否则节点 i 留在原来的社区中不变。

Step 3. 对于网络中的每个节点都如此操作一次；

Step 4. 将 Step 3. 中形成的社区压缩为一个新节点，原社区之间的边权重变为新节点间的边权重，原社区内部节点间的权重之和为新节点环的权重；

Step 5. 重复 Step 4.，直到整个网络图的模块度不再变化。

3.4.2 社区内推荐

传统的协同过滤算法中，近邻用户的选择需要在所有用户中寻找，而引入社区划分算法后，方法中近邻用户的选择不再是所有用户中寻找，而是依靠前文中社区发现算法所划分的社区，先在目标用户所在社区中寻找相似度较高的用户作为目标用户的近邻集，在近邻集用户阅读过的新闻中选取目标用户未阅读过的新闻作为待推荐新闻集，在新闻集中按兴趣度大小排序后推荐给目标用户。

目标用户 u 对新新闻 d 的兴趣程度 $P(u,d)$ 的计算方法如公式3.6 所示。

$$P(u,d) = \sum_{v \in U} sim(u,v) * P(v,d) * \delta(v,d) \tag{3.6}$$

其中，U 代表用户 u 的近邻用户集，v 是近邻用户集中的某位用户，$sim(u,v)$ 代表前文提到的用户相似度，$P(u,d)$ 是用户 v 对新闻 d 的感兴趣程度，$\delta(v,d)$ 表

示用户 v 是否浏览过本新闻，如果没浏览过为 0，浏览过则为 1，通过此参数有效筛选掉没有浏览过此新闻的近邻用户。

社区内推荐的算法流程如下：

Step 1. 对于用户 u 找到目标用户所在社区，寻找目标用户的近邻用户集，在目标用户近邻集中根据相似度大小进行排序，选择前 K 个用户作为近邻集用户；

Step 2. 将目标用户未浏览过的新闻组成待推荐新闻集，按照公式3.6求得用户对于每篇待推荐新闻的兴趣度，取前 n 篇组成推荐新闻列表；

Step 3. 将新闻推荐列表推荐给用户。

3.5 本章小结

本章介绍了基于社区发现算法的新闻推荐系统，从新闻文本预处理、计算用户相似度矩阵、建立用户相似度网络、实施社区发现算法进行社区划分、组成近邻用户集、社区内推荐几步做了详尽的阐述。

最终，基于社区网络发现算法的新闻精准推荐系统算法流程图如图3.3 所示：

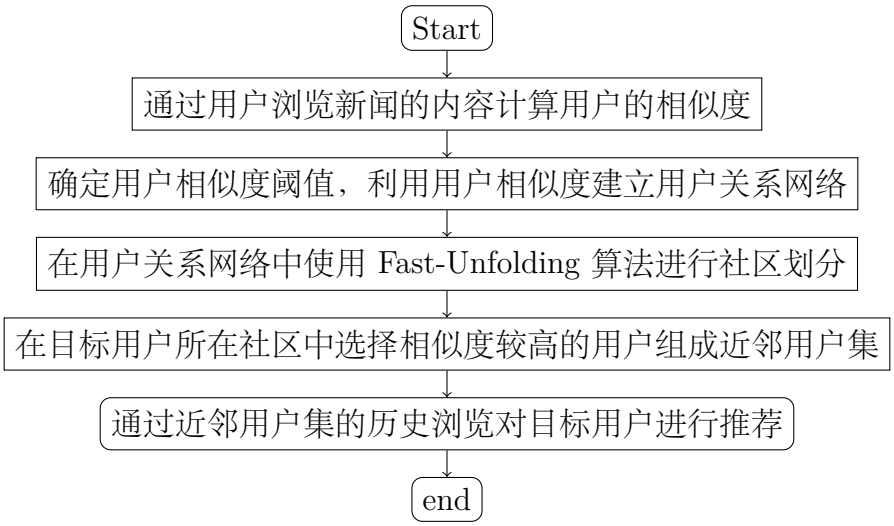

图 3.3　基于社区发现的新闻推荐算法流程图

4 算法评价与实验验证

本文采用 CCF 协会大数据竞赛提供的“用户浏览新闻的模式分析及个性化新闻推荐”数据集进行实验验证。数据集中包含从某财经网站中收集到的 10,000 名用户在一个月内阅读新闻的历史记录。每条记录包括用户编号、新闻编号、浏览时间、新闻标题以及详细内容等。同时，本文将采用预测准确率和召回率两项指标，并与传统的协同过滤算法进行对比来对算法做综合评估。

准确率和召回率的定义如公式4.1 和公式4.2 所示：

$$precision = \frac{\sum_{u_i \in U} hit(u_i)}{\sum_{u_i \in U} L(u_i)} \tag{4.1}$$

$$recall = \frac{\sum_{u_i \in U} hit(u_i)}{\sum_{u_i \in U} T(u_i)} \tag{4.2}$$

其中，U 为数据集中 10,000 个用户所组成的集合，$hit(u_i)$ 表示算法推荐给用户 u_i 的新闻中，确实在测试集中被用户浏览的个数。由于在此数据集中每个用户在测试集中只有一条测试记录，因此 $hit(u_i)$ 只有 0 和 1 两种取值可能。$L(u_i)$ 表示算法给用户 u_i 提供的新闻推荐列表的长度，$T(u_i)$ 表示测试集中用户 u_i 真正浏览的新闻的数目。在此数据集中，$T(u_i)$ 为 1，$\sum_{u_i \in U} T(u_i) = 10,000$.

最终将两个评价指标组合为综合指标 F，使用 F 值作为算法评价的标准。F 值的定义如公式4.3 所示：

$$F = \frac{2}{\frac{1}{precision} + \frac{1}{recall}} \tag{4.3}$$

F 值是一个 $(0, 1)$ 之间的数，F 越大，代表算法的推荐效果越好。

本次实验从数据集中选取阅读新闻数大于 30 篇新闻的用户作为实验集，随机选取 200 名用户，将每个用户的后 15 篇新闻作为测试集，其余作为训练集，并将通过推荐方法得到的推荐结果与这些用户的最后 15 篇新闻浏览纪录进行对比，计算相应的准确率、召回率，并以此计算得到最终的 F 值。

4.1 确定相似度阈值 α

前文提到，对用户网络连边时，需要根据实验确定的相似度阈值 α 完成。当两用户的兴趣相似度 $sim(u,v) > \alpha$ 时，认为两用户的相似度较高，在网络中对两结点连边；否则认为两用户的相似度不够高，不做处理。

相似度阈值 α 不能随意确定，如果 α 太大，则用户网络中连边少，大多数用户都会被当作独立的社区，社区划分对网络分割的作用有限；若 α 太小，则较多用户之间都有连边，用户之间兴趣区分不明显，也会影响到推荐系统的结果。因此，需要首先将 α 作为自变量进行实验，确定合适的相似度阈值 α。

首先通过二分法确定合适的阈值大约在 $0.1 \sim 0.2$ 之间，然后对此区间内的每个步长为 0.2 的阈值进行实验，绘制出如图4.1 所示的 F 值折线图：

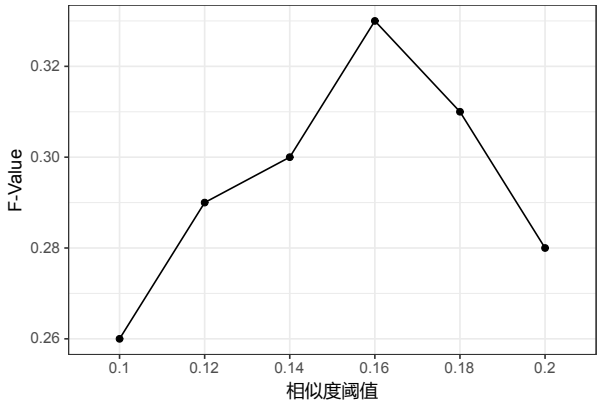

图 4.1　不同阈值下的 F 值结果

通过实验发现，$\alpha = 0.16$ 时 F 值达到最大，是较为合适的相似度阈值。因此后续实验将使用 $\alpha = 0.16$ 作为相似度阈值。

4.2 不同算法比较实验与对比分析

根据前文确定的算法超参数，我们将在数据集上对比本文算法和传统的基于用户行为的协同过滤算法，并对本文算法的效果进行分析与评价。当对用户推荐的新闻列表长度为 5，10，15，20 时，两算法推荐的准确率、召回率和 F 值如图4.2、4.3和 4.4 所示：

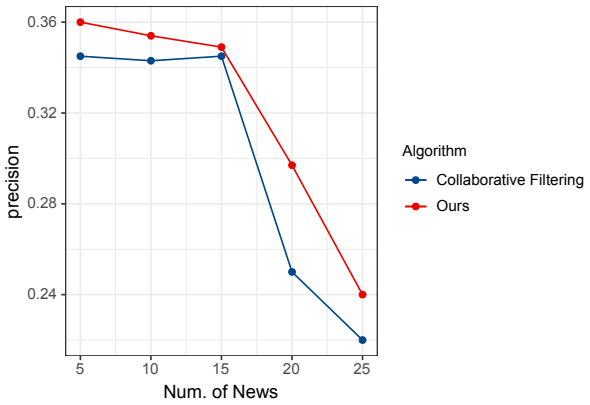

图 4.2　算法预测准确率对比

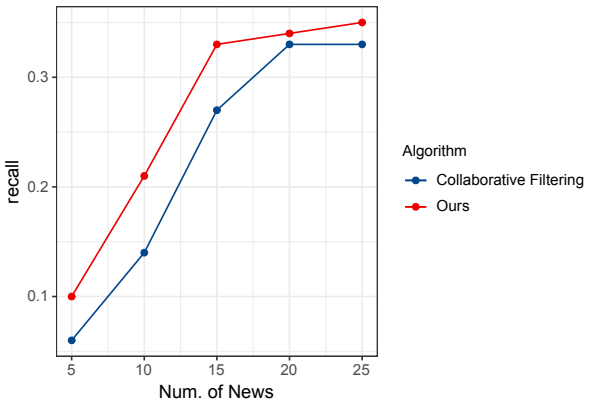

图 4.3　算法预测召回率对比

从实验数据可以看出，本文提出的基于社区发现的推荐算法在预测准确率、召回率和 F 值上均较传统的协同过滤算法有不同程度的提升，表明将新闻内容的相似度也作为用户兴趣相似度的评判标准，而非仅像传统协同过滤算法一般只考虑是否浏览相同新闻，能更有效地筛选出兴趣相关的近邻用户，提高传统协同过滤算法的准确度。

同时，引入社区发现算法的一大好处在于能快速筛选出近邻群体，而无需像传统协同过滤算法一样在所有样本中进行筛选。当数据集逐渐调大时，本文算法的推荐速度快于传统的协同过滤算法，如表4.1 所示：

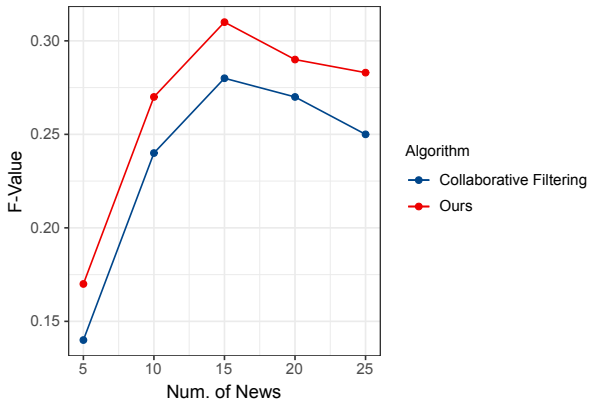

图 4.4　算法预测 F 值对比

表 4.1　在不同大小数据集上算法的运行速度对比

Algorithm	500 news	5,000 news	50,000 news
Collaborative Filtering	0.03s	0.37s	3.62s
Ours	0.03s	0.28s	2.08s

4.3　算法评价与不足之处分析

经过前文实验，本文算法在准确率、召回率和运行速度上皆优于传统的基于用户行为的协同过滤算法。在准确率和召回率上的领先是因为本文算法在计算用户相似度时，将新闻内容的相似度也作为评价指标纳入用户向量矩阵，而非仅仅像传统的基于用户行为的协同过滤算法一般只考虑用户是否浏览量相同的新闻。在运行速度上的领先是由于引入了社区发现算法，将用户近邻集的筛选范围从整个用户集缩小到社区范围内，由此提高了推荐速度。

但由于缩小了用户近邻集的筛选范围，也损失了部分的推荐精度——由于社区划分时用户的相似度是人为确定的，很难保证没有兴趣和目标用户相似、但未被划分到相同社区的漏网之鱼。因此社区划分算法的准确率和相似度算法、相似度阈值的确定是提高推荐准确率和召回率的重要因素。可以遇见的是，如果将传统过滤算法的相似度评价指标改为和本文算法相同的内容相似度，推荐精度将比本文算法好。

同时，用户相似度的计算也有可改进之处——除了用户浏览的新闻内容，用户的行为相似度也是值得参考的指标数据，例如，对新闻文章进行点赞、转发、评论等操作的用户，明显会比只浏览新闻的用户对新闻文本更加感兴趣，在实际操作中，结合用户行为和用户浏览新闻的内容相似度计算用户的综合相似度，可能可以提高新闻预测的成功率。

除此之外，文献 [11] 指出，热度和时效性是协同过滤算法中，使用余弦相似度计算用户相似度的一个不足之处。由于新闻的热点性和时效性，不同兴趣的用户也可能会浏览内容的热点时效新闻，而余弦相似度可能会因此判定用户对新闻文本感兴趣而非对热点感兴趣，从而造成用户兴趣的误判。因此在实际应用中，应当对热点新闻的文本矩阵计算过程乘以某一惩罚因子，来以此修正热点内容对用户兴趣的影响。

4.4 实验过程中的缺陷分析

4.4.1 在测试集上验证

可以发现，在前面的实验过程中，本文没有将训练集按比例划分为训练集和验证集，而是在训练集上训练、在测试集上验证。这可能会出现过拟合的问题：由于超参数都是基于测试集调整的，自然得到的是表现好的结果。然而真正在未知数据集上的表现未知。这就好比一个学生没有学习知识，而是将所有练习题和测试题都背了下来，由于平时模拟题都是测试题，该学生可能会有较好的表现，但当真正到高考考场面对没有原题的试卷时，该学生的表现将极为糟糕。

但值得提及的是，测试集的结果仅仅用于调试超参数，而并未加入测试集中。因此对验证的影响不算很大。

这启示我们以后一定要将训练集分为训练集和验证集两部分，训练集用于学习训练，验证集用来验证结果和调整超参数，测试集仅用于测试最终结果，不能作为调试超参数的依据，更不能放入训练集中，否则未来仍会出现过拟合的问题。

4.4.2 没有交叉验证

交叉验证是指，假设在划分测试集时将测试集分为等量的 5 份，4 份用于训练 1 份用于验证，那么交叉验证需要循环取 5 份数据中的 4 份，最后取五次训练的平均值作为最终的训练结果，如下图所示。

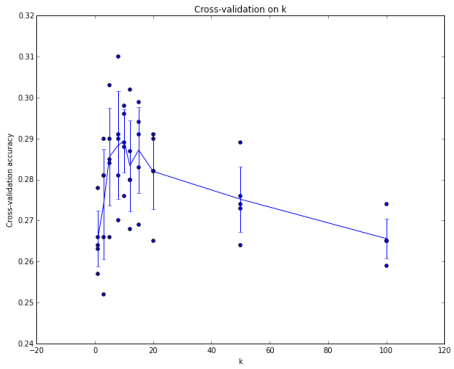

图 4.5　交叉验证的例子

交叉验证能极大降低数据偶然性和噪音的影响，从而选取出最为合理的超参数。但交叉验证需要消耗更大的算力和更多的时间，而本实验中为节省时间没有使用。

4.5 本章小结

本章介绍了本文提出的基于社区发现算法的新闻精准推荐模型的实际实验表现，分析了实验结果并对算法优劣之处做出评价，同时指出了算法可能存在的部分考虑不周之处，并提出了可能的改进措施。

5 总结与展望

5.1 本文工作总结

本文首先对现有的推荐系统算法和社区发现算法进行概述，并对现有的推荐系统和社区发现算法进行了介绍，总结了前人的工作。随后提出本文的基于社区发现算法的新闻推荐系统模型的构建方法，并对算法做了实验验证和评价，最后指出了算法存在的考虑不周到之处和改进方向。主要研究内容有：

1. 改进了传统协同过滤算法通过用户行为计算用户相似度的方法，引入 TF-IDF 算法对新闻文本进行预处理和向量模型构建，以此内容相似度作为计算用户相似度的方法，提高了用户相似度评判的准确性；
2. 引入社区划分算法到寻找近邻集用户的过程当中，提高了用户近邻集的寻找速度，进而提高了算法的运行效率；

5.2 未来展望

本文提出的算法较最为基本的基于用户行为的协同过滤算法有一定程度上的改进，但推荐的准确率、召回率和速度仍然处于较低水平，和数据竞赛中排名靠前的用户乃至于工业级推荐算法有较大的差距。在未来的研究工作中，仍需要改进用户相似度的计算方法和社区划分的精度与速度，同时解决前文提到的不足之处，进一步提高推荐的效率、准确率和召回率，提高实验的准确性和说服力。

致　　谢

感谢本课程的任课老师石小川副教授，是您在我发邮件询问能否加入您的社会计算班级后快速回复我并同意我的加入，否则选课掉了社会计算的我将无课可上；您课程教授过程中严谨和负责的态度也让我受益良多，几次问您问题都得到了详尽的解答；《社会计算》作为一门自然语言处理和机器学习入门相关的开拓视野的课程，我在本课程中收获了大量的知识，开拓了视野，也借此机会结合实验室的学习内容学习了大量的机器学习相关知识；

感谢 nis&p 实验室的王骞老师、博士生龚雪鸾学姐和网安大三的孔维翰学长，是你们在我确定论文主题后对我的实验过程进行细致的指导、提供算法讲解资源，让我在练习过程中有效提高了代码能力；

最后感谢有毅力坚持下去的自己。在大多数人眼里的“水课”中倾注 3 个周末的时间查阅文献、写代码、写论文，是一件不容易的事情，但获得的收获也完全是值得的。希望在寒假和更远的未来，我能继续坚持下去，学习感兴趣的知识。

感谢我的好朋友余希。最开始的实验结果很不尽人意，是她鼓励我更换 baseline、和我一起思考模型的不足之处，最后得到了一个不错的结果。希望她也能在这门课里取得好成绩。

参考文献

[1] 网易新闻，https://news.163.com/

[2] 朱扬勇, 孙婧. 推荐系统研究进展 [J]. 计算机科学与探索, 2015, 9(5):513-525.

[3] Lü, Linyuan, Medo M , Yeung C H , et al. Recommender Systems[J]. Physics Reports, 2012, 519(1):1-49.

[4] 孙鲁平, 张丽君, 汪平. 网上个性化推荐研究述评与展望 [J]. 外国经济与管理, 2016, 38(6):82-99.

[5] 刘凯, 王伟军, 黄英辉, et al. 个性化推荐系统理论探索: 从系统向用户为中心的演进 [J]. 情报理论与实践, 2016, 39(3).

[6] 万雪飞. 基于社会网络的协同过滤推荐技术研究 [D]. 电子科技大学, 2010.

[7] Rosvall M, Bergstrom C T. Maps of Information Flow Reveal Community Structure In Complex Networks[J]. Proceedings of the National Academy of Sciences Usa, 2007:1118–1123.

[8] 冷亚军，陆青，梁昌勇. 协同过滤推荐技术综述四. 模式识别与人工智能,2014,27(08):720-734.

[9] 李建华, 汪晓锋，吴鹏. 基于局部优化的社区发现方法研究现状四. 中国科学院院刊，2015(2).

[10] Horvitz E . The Lumiere project : Bayesian user modeling for inferring the goals and needs of software users[J]. Proc. of Fourteenth Conf. in Artificial Intelligence, 1998

[11] 冯文杰, 熊翱. 基于新闻时效性的协同过滤推荐算法 [J]. 计算机系统应用,2018，