文本复制检测报告单(全文标明引文)

№:ADBD2018R_2018053015312720180530154817440173843211　　检测时间:2018-05-30 15:48:17

检测文献:　53140120_朱海洋_计算机科学与技术_基于社交网络的好友关系可视化及社交推荐

作者:　朱海洋

检测范围:　中国学术期刊网络出版总库

中国博士学位论文全文数据库/中国优秀硕士学位论文全文数据库

中国重要会议论文全文数据库

中国重要报纸全文数据库

中国专利全文数据库

图书资源

优先出版文献库

大学生论文联合比对库

互联网资源(包含贴吧等论坛资源)

英文数据库(涵盖期刊、博硕、会议的英文数据以及德国Springer、英国Taylor&Francis 期刊数据库等)

港澳台学术文献库

互联网文档资源

CNKI大成编客-原创作品库

个人比对库

时间范围:　1900-01-01至2018-05-30

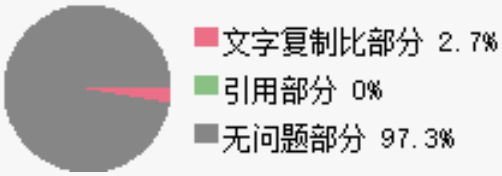
检测结果

总文字复制比：2.7%　　跨语言检测结果：0%

去除引用文献复制比：2.7%　　去除本人已发表文献复制比：2.7%

单篇最大文字复制比：0.7%

重复字数：	[725]	总段落数：	[6]
总字数：	[26487]	疑似段落数：	[2]
单篇最大重复字数：	[177]	前部重合字数：	[0]
疑似段落最大重合字数：	[485]	后部重合字数：	[725]
疑似段落最小重合字数：	[240]		



指　标：☐疑似剽窃观点　☑疑似剽窃文字表述　☐疑似自我剽窃　☐疑似整体剽窃　☐过度引用

表格：0　　公式：0　　疑似文字的图片：0　　脚注与尾注：0

0%（0）	中英文摘要等（总2810字）
5.3%（240）	第一章绪论（总4534字）
6.2%（485）	第二章相关概念与技术（总7833字）
0%（0）	第三章好友推荐算法的实现（总5968字）
0%（0）	第四章好友关系可视化的实现（总4169字）
0%（0）	第五章总结与展望（总1173字）

（注释：无问题部分　文字复制比部分　引用部分）

1. 中英文摘要等　　总字数：2810

相似文献列表　文字复制比：0%(0)　疑似剽窃观点：(0)

原文内容　红色文字表示存在文字复制现象的内容; 绿色文字表示其中标明了引用的内容

摘要

基于社交网络的好友关系可视化及社交推荐

随着互联网时代近些年来的快速发展，社交媒体平台的数据量与日俱增，用户借助于社交媒体进行社交，而在面对大量数据时，用户如何在“信息过载”的情况下更好的找到自己的潜在好友成为了问题，基于社交网络的好友推荐能够帮助用户分析

用户之间的潜在联系，挖掘潜在好友，社交关系的可视化同样是研究重点之一，通过好友关系可视化的方法，能够更为清晰的对用户的社交圈进行更好的展示以及帮助用户更好的了解自身的社交状况。

首先，本文通过新浪微博采集的数据进行分析，提出了一种好友推荐的方法，首先选取了二度好友作为好友推荐候选集，接着通过好友参考因素的选取，以及对相关因素的处理，对数据进行加权消除好友数量上的差距。之后通过对好友相似度进行计算，以及数据的标准化处理，对用户进行好友推荐。

其次，对用户与其他用户之间的好友关系，进行好友关系网络可视化的构建，在构建的同时，加入用户社交影响力的因素，用户社交影响力的因素主要参考用户的活跃度以及用户所拥有的好友数量，以可视化的方式对用户现有社交状态进行展示。

最后，为使用户更好的了解自身社交圈地理位置分布情况，对用户地理位置信息可视化的构建，首先通过用户的地理位置信息与地图shape数据文件做模糊匹配，选取最优匹配，对相应的数据进行数据标准化的处理，与色阶卡进行比对，最终对地图进行着色处理。

关键词：好友推荐,可视化,社交网络,相似度,影响力

ABSTRACT

Social network-based friend relationship visualization and social recommendations

With the further development of the Internet in recent years. Information has increased greatly on social media platforms every day, users use social media to socialize. When users face a large amount of data, how to better find potential friends in the “information overload” situation has become a problem. Friends recommendations based on social networks can help users analyze potential connections between users and tap potential friends. Visualization of social relationships is also one of the research focuses. Through the visualization of friends, we can better display the user's social circle and help users better understand their social situation.

First of all, this paper analyzes the data collected by Sina Weibo and proposes a method for recommending friends. First, the second-degree friend is selected as a friend recommendation candidate set, and then the selection of friend reference factors and the processing of relevant factors are performed. The data is weighted to eliminate the gap in the number of friends. Afterwards, through the calculation of the similarity of friends and the standardization of data processing, the user is recommended to friends.

Secondly, the user relationship between the user and other users is visualized with the construction of a friend relationship network. At the same time as constructing, the social influence factor of the user is added. The social influence of the user is mainly related to the activity of the user and the number of friends the user owns. We could display the user's existing social status in a visual way.

Finally, to help users better understand the geographical distribution of their social circles, we would visualize the user's geographic information. First, we make a fuzzy match between the user's geographic location information and the map shape data file, select the optimal match, and perform data standardization on the corresponding data, and compare with the color gamut card, finally, the map is shaded.

Key words: Friend recommendation, Visualization, Social network, Similarity, Shadow strength

目录

2. 第一章绪论

总字数：4534

相似文献列表　文字复制比：5.3%(240)　疑似剽窃观点：(0)

1	基于地理位置的POI推荐算法设计与实现 田元昊 - 《大学生论文联合比对库》- 2017-04-27	3.9%（177） 是否引证：否
2	戴志鹏_2013081105_基于个性化推荐新闻资讯系统设计与实现 戴志鹏 - 《大学生论文联合比对库》- 2017-06-05	3.3%（148） 是否引证：否
3	基于空间数据挖掘的热门景点及线路推荐研究 刘勇(导师：张文元) - 《华中师范大学博士论文》- 2017-05-01	2.2%（101） 是否引证：否
4	基于嵌入式Web服务器的无线远程测温系统研究 朱丽叶(导师：王淑蓉;谢俊屏) - 《西安工业大学博士论文》- 2012-04-30	0.8%（38） 是否引证：否
5	南开社区学院教务管理信息系统的设计与实现 任鹏(导师：许春香;孙克泉) - 《电子科技大学博士论文》- 2013-03-25	0.7%（33） 是否引证：否

原文内容　红色文字表示存在文字复制现象的内容; 绿色文字表示其中标明了引用的内容

第一章绪论

1.1研究背景与研究意义

1.1.1研究背景

在互联网的进一步的发展下，人们的社交圈逐渐扩大，网络社交在人们生活中有着相当重要的影响。国内更加多样的社交媒体，例如：新浪微博、脉脉、米聊等等，都以自己的方式影响着社交圈。社交网络平台在不同年龄阶层之间的渗透程度有着很大程度的增加，社交网络平台更多的得到了用户的认可，大量的用户与大量的信息构成了如今的社交网络，用户之间相互传递分享着自己的兴趣爱好，线上交流的方式变得更加普遍。

随着互联网的发展，社交成为如今信息化时期的很重要的一个方面，它不像以前那样传递信息，它将进行真实的沟通，人与人之间联系更加紧密，对信息的理解更加深刻，同时对信息加以利用和预测。在如今这个互联网的时代，我们拥有了很多技术，社交状态逐渐的呈现出便携性，可操作性强的状态，社交平台同时成为了很多方面的切入点，例如广告和电商平台，他们利用社交平台的优势扩大自己的影响力。目前，电子商务，公众号的运营，直播节目，教育、等等一些领域也都逐步的加入社交性的元素，通过这种方法来扩大用户的数量，加强用户的粘性，在我国的发展前景逐步向好的方向发展。

如下表格所示：

表1-1 2017凯度中国社交媒体影响报告

2016年（%）2017年（%）增长率增长百分比

15 – 19 66.9 69.4 2.5 3.7 %

20 – 29 75.8 77.3 1.5 1.9 %

30 – 39 57.7 61.4 3.7 6.3 %

40 – 49 56.9 63.8 6.9 12.1 %

50 – 59 26.7 34.2 7.5 28.3 %

60 + 9.7 13.4 3.7 38.2 %

对于我国来说，社交网络的发展趋势十分可观，从90年代迈出第一步，到如今已经发展将近三十年，2017年全球社交网络调查显示：社交媒体使用人数达到了惊人的30亿，并且我国的社交用户也即将迈入7亿大关，近几年来，社交网络以更加迅猛的态势增长，连续三年以来市场范畴都在进一步的增长，不论用户增长方面或者社交热度方面都在前行进步，其发展势头依旧十分强劲。

随着信息技术的进一步发展，微博成为我们日常生活中相当重要的网络交流平台之一。2018年微博数据新鲜出炉，2018年前三个月微博每天的在线人数平均将近2亿，对于单个月来说，在线人数超过4亿，与此同时，微博广告季度收入将近20亿。用户的增长带来了相应的社交信息，用户对相应的兴趣爱好的关注产生了需求，与此同时用户的社交圈在逐步的增长，对潜在用户信息的使用成为了社交媒体平台关注的热点之一。以及用户社交关系如何以一种更加清晰的方式进行展示同样成为了问题之一，不同用户的社交影响力不同，用户之间的纷繁复杂的关系也不尽相同，如何将用户之间的关系进行可视化展示也逐渐成为了需求。

当前社交网络已经存在相应的好友推荐方式，但仅仅将用户注册信息相似度高的用户进行推荐，而忽略用户之间的内在联系是十分不明智的，其间过多内容不符合用户的兴趣爱好，使得使用者对其的推荐结果不是很满意。找出用户的更深层次的联系，不仅仅包括单纯的使用者的注册信息，找出用户与用户之间的关联同样重要。用户可以以主动的方式关注自己的好友、感兴趣的人，同样，社交媒体提供了相应的扩大用户社交圈的办法，通过设计内部的算法或者特定的功能使得用户可以结识更多的好友。从开始的“摇一摇“，”附近的人“，”感兴趣的人“，都在以某种方式扩大用户的社交圈。

好友推荐是一项较为基础的社交媒体提供的服务功能，通过现有用户以及用户之间的相互联系，以此为基础拓展社交圈，使得社交媒体尽可能的提高使用者与平台的粘度，更好的维持社交他们的活跃程度。而将好友关系以更为直观的方式展现使得用户能够清楚的了解自己的社交关系现状也是我们思考的问题。

数据可视化是指以图形图像的方式将信息表达出来，这能够帮助用户能够更快更好的地对现有状况进行一定程度的了解。好友关系的可视化同样作为好友了解自己社交状况的一种方式，以更清楚明显的方式展示出用户当前的社交状况。人脑对视觉信息的敏感程度比书面信息高很多，使用可视化的方式展示好友关系使得用户能够以一种更为清晰的方式甄别并处理信息。

从现有的数据中尽可能的提取出用户的潜在信息，在相关的用户中找出最为适合的用户作为好友，从而扩大用户的社交范围，而从中筛选出的信息将为我们进行的好友推荐提供更多的参考依据。通过对好友关系可视化可以进一步了解用户的社交现状，以及用户之间的社交关系，提高好友推荐的准确率，具有一定的创新性与实用性。

1.1.2 研究意义

社交网络的快速进入一个兴盛状态的同时，社交媒体也慢慢显示了其所占的重要作用。社交媒体变得多种多样让人眼花缭乱。

如图所示：

图1-1 2017年中国社交媒体示意图

属于现在较为时兴的社交媒体，微博拥有大量的用户群体。微博是提供使用者之间的联系，交友，共享兴趣爱好信息的一个交际舞台。同时它提供了客户端，用户可以借助它形成自己的关系圈，它提供了更多人能够并且愿意使用的网络平台。近些年来，社交网络不断发展，微博逐渐成为人们交友、聊天的工具，实现即时分享。微博使得用户之间可以相互关注，成为好友关系，其作为一种大众之间可以共同交流与分享，展示自我的舞台，其更注重时间效应，使得用户更能表达出当前的想法和状态。

随着越来越多使用者的加入，如何在海量用户中找到自己的潜在好友成为了问题，而好友推荐可以帮助他们以便捷的方式找到自己的关注好友，节省了大量的筛选时间[1]。从用户的个人信息以及与好友的相互联系中挖掘出更多更具有价值的信息以帮助好友进行筛选出感兴趣的用户集，增强用户之间的联系紧密度，将好友关系可视化使得用户更加清楚的了解自己的好友关系同样显得重要。本文重点将考虑如何通过多个因素对好友进行推荐，并且通过数据可视化的方法将其清晰的展示出来，提高使用者查找潜在好友的效率，了解当前的交际关系现状。

1.2 研究现状及发展趋势

由于平台使用者数量的逐渐增长，好友推荐的功能逐渐的发展起来，从最开始的六度分隔理论到三元闭包理论再到协同过滤算法的提出，都给好友推荐提供了很大的理论支持。国外有名的教授斯坦利米尔格兰姆描述一个人与人之间的关系网图，他曾经做过相关的研究。在此之后他发现了“六度分隔”现象,我们可以将其较为简单的说明：通过6个人的关系，你就可以结识任何一个你并不认识的人”[2]。它说明了在当前社会中存在着弱联系的现象，但是却存在着十分重要的影响力，许多人都在受这种弱联系的影响，通过弱联系使得人与人之间的联系变得更加的紧密。三元闭包理论就是描述在一个社交圈子之中，如果某两个人之间存在同样的朋友，那么这两个人有一定的几率成为朋友[3]。

协同过滤推荐算法是出现相对较早，而且它的流行度相对较高。它的主要的功能是做出相应的预测和对应的推荐，该算

法首先通过对用户之前所产生的信息进行更深度的分析，通过这样来发现用户的兴趣所在，然后通过分析得到的结果对用户进行群体的分割归类，最终对不同的群体推荐得到不同的兴趣商品。

协同过滤推荐算法分为两个不同的类别，分别是基于用户的协同过滤算法(user-based collaborative filtering)，和基于物品的协同过滤算法(item-based collaborative filtering)[4]。通过对用户进行相似度的分析，对相似较高的用户进行相关的推荐，但该方法存在一定的缺陷，在于数据的稀疏性，以及对于一些较为复杂的属性不方便进行处理，推荐结果好坏取决于之前产生的信息，在刚刚开始的时候，它的推荐效果不好。

由于社交网络的复杂性，单一的社交网络推荐方法对于特定的场合或许效果比较良好，而对于各种不同场景来说，很难找到一个对所有场景都适用的方法，考虑多种因素，挖掘更加隐蔽的信息关系能起到更好的效果。

近几年来，社交推荐更加成熟，推荐准确率有了很大的提高以及推荐方式也有了很大改变。好友推荐也将发展为信息挖掘更为深入，更基于海量数据进行甄别，同样通过语义分析，情感分析，文本挖掘的方式来提供好友推荐的依据[5]。

尽管上述方式在特定的场景下同样能取得较好的实验效果，但是这些方法同样通过单一的元素或者属性来进行推荐算法的研究，而实际情况中对于每个用户他们的社交关系以及对不同属性的重要程度不同，此类方法受到较大的局限性。有些时候并不能满足较多用户的一些较为比较特别的好友推荐需要。我们考虑将更多因素考虑进来，挖掘用户之间的潜在联系，通过这种方式来更加准确的进行好友推荐，同时通过可视化的方法展示好友之间的社交关系[6]。

1.3 主要工作

在社交平台上，推荐系统为其拥有的用户提供了很好的数据过滤的功能，而好友推荐是其中十分重要的一部分。对于使用者而言，在面对大量的数据时如何找到自己的潜在好友是尚待解决的一项问题，本文对于该问题提供了一种将多个参考因素用于对用户进行潜在好友推荐的方法，并且对当前好友的社交关系进行可视化的展示。本文进行的工作具体如下：

1．实验数据集的收集与处理。

本文实验数据均来自于本人在新浪微博官方网站上收集的真实数据，然后对其进行一些处理，最后得到了我们所使用的数据集。

2．用户好友推荐方法所涉及的参考因素。

通过分析与考量，对用户数据提取出对应的参考因素（具体参考因素以及数据集的具体字段设置将在后文给出），通过这些参考因素对用户好友推荐方法提供了相关的基础。

3. 用户好友推荐方法的实现。

根据对用户数据的分析，构建出用户好友推荐的具体方案，并给出方案的具体实现方法。

4. 好友关系可视化的实现。

通过对用户社交关系分析，进一步的对好友的社交关系进行可视化的实现，更加直观的了解用户当前的社交关系，分为好友关系网络可视化，以及用户地理信息数据的可视化。

1.4 论文的组织结构

本篇论文具体组织结构如下所示：

第一章，绪论。绪论介绍了好友推荐功能的研究背景，对当前社交媒体的发展以及国内外研究现状以及发展趋势做了进一步概述，同时，对本篇论文的组织结构进行了阐述。

第二章，相关概念与技术。本章介绍了社交网络推荐的兴起缘由，以及相关的推荐算法的介绍，以及一些当前流传度较高的社交平台所用的推荐算法的阐述，为之后的社交网络推荐方法奠定理论基础。

第三章，好友推荐算法的实现。好友推荐算法的实现介绍了具体如何使用用户数据进行好友推荐的依据与方法，包括好友推荐候选集的选取以及相关依据，好友相似度参考因素，以及相关的计算，和对相关数据的处理方式。

第四章，好友关系可视化的实现。本章介绍了好友关系可视化的具体方案，分为两个部分，好友关系网络图的构建，以及好友地理信息数据可视化的构建。包括对用户数据信息的处理方式，可视化构建采用的方法以及呈现的效果。

第五章，总结与展望。本章将本文的研究内容做出了归纳，并且指出了本文进行的工作的尚需改进的方面，同样对进一步的研究工作做出了展望。

指　标
疑似剽窃文字表述

1. 发现用户的兴趣所在，然后通过分析得到的结果对用户进行群体的分割归类，最终对不同的群体推荐得到不同的兴趣商品。
2. 1.4 论文的组织结构
   本篇论文具体组织结构如下所示：
   第一章，绪论。绪论介绍了好友推荐功能的研究背景，对当前社交媒体的

3. 第二章相关概念与技术　　总字数：7833

第二章相关概念与技术

2.1 社交网络好友推荐算法概述

在当前这个信息化的时代，随着web2.0逐渐的兴起，信息数据呈现出巨大的增长，越来越多的人开始在社交平台上发表自己的看法与理解，然而随着数据信息在经过大量的增长之后，社交平台上的用户已经无法以人工的方式开始寻找自身感兴趣的方向，因为如此巨量的数据已经超过了他们的预期，同时也超出了他们的承受能力。

使用人数的急剧增加再加上用户所产生的数据信息的增长，当使用者在面对如此巨量的信息时，传统的搜索引擎只能提供一定量的信息过滤方法，他们将同样的信息推荐给不同的用户，这样的过滤方法非常缺少个性化以及人性化的体验，基于社交网络的推荐算法应便是在这种情况下出现的。

所谓推荐系统，顾名思义，其目的就是为了帮助用户找到想要的东西，可是在面对大量数据时，帮助用户找到想要的东西是相当困难的一件事情，同样，在经济学中，存在一个相对来说比较有名的逻辑：长尾理论（The Long Tail）[7]。

我们将用互联网领域中讨论长尾效应，就是在数据众多的时候，小部分的资源将会得到大部分的关注，剩下的大部分资源却少有关注，与此同时，这造成了资源上的大规模浪费，而且对于那些兴趣偏向小众爱好的群体来说无法找到自己感兴趣的内容。

如下所示：

图片2-1 长尾效应示意图

互联网信息过于繁杂且规律性不强，若是所有的信息放入用户首页，那么不是任何一个用户能够接受的，信息利用率极其低下，因此我们需要推荐算法来帮助用户筛选出相应的信息。筛选信息的方式多种多样，对无意义数据的过滤，对重复数据的删除，对相关数据的合并等等，最主要的工作则是通过算法的方式进行个性化的筛选以期望满足我们更个性化的需求。在社交网络中，信息错综复杂，使用推荐算法的方式进行筛选是非常行之有效的方式之一。

年龄职业性别浏览收藏评论推荐算法推荐列表Item1Item2Item3UserItem年龄职业性别浏览收藏评论推荐算法推荐列表Item1Item2Item3UserItem 图2-2 推荐系统示意图

上图对推荐系统有一个大概的说明，首先，平台通过用户的基本信息以及用户行为进行分析，之后设计推荐算法，将推荐算法运用到用户的数据信息当中，形成最终的推荐内容。

基于社交网络的好友推荐对于社交网络中的推荐系统而言是非常重要的一部分。社交网络的好友推荐是通过用户的一些基本数据以及发现出的用户所拥有的潜在关系，并根据这些信息进行相对应的计算，寻找出部分符合要求的用户作为推荐好友用户。通过社交网络的好友推荐功能大规模的减少了用户因为数据量过大而在寻找好友时所面临的窘迫情况。

好友推荐的功能已经不仅仅应用在各大即时通讯软件（新浪微博、脉脉、米聊等）上面，该功能在各种电商平台（亚马逊、淘宝等）同样扮演着重要的角色，各平台通过扩大用户在自己平台的社交范围来提高用户对平台的依赖度或者说粘度。

目前有很多好友推荐算法，但大致都能分为几种类型。首先，基于流行度的算法，它能够根据一个方向的热度生成热度

列表，然后将热度较高的问题向用户进行推荐，例如我们在新浪微博的，将最流行的，较高的热点的事情推荐给所有用户，在较热的关注里面，将一些大V用户推荐给普通用户，但缺点是无法提供个性化的推荐。协同过滤算法，是经常使用的一个算法，尤其是在电商平台应用比较广，通常包括基于用户的协同过滤算法和基于物品的协同过滤算法。还有基于内容的算法，比如TF-IDF算法，通过该算法，将能够计算出部分词语的权重，在计算相似度的时候引入权重的影响，通过利用某些工具将这些词语进行聚类分析，最后终根据话题将这些东西进行向量化的处理[8]。如能够将黄蜂队，洛杉矶快船队，鹈鹕队归类到“篮球”的话题，将灌汤包，饺子，馅饼归类到“食物”的话题，之后根据话题将这些信息与用户作相似度的计算。还有混合算法，而在现实的条件下我们比较少的采用混合算法的使用，对于不同的情况有不同的算法，单纯面向各种场景的推荐算法相对来说效果没有专用场景的算法效果出色。

2.2 二度人脉好友推荐算法

在好友推荐算法的开始阶段，更多的考虑用户之间的关联，开始研究好友之间的关系，在现实生活中好友通过相互联系，并且相互介绍好友来增加在自己的好友，在社交平台上，目的是希望模拟这一现象进行好友推荐，通过二度好友的方式来进行好友的推荐[9]。但是部分用户的好友数目很多，那么相应的二度好友同样数据量很大，我们不能单纯的将所有的二度好友推荐给某个用户，因为其数据量过大，我们更希望的是通过一定的方式，再进行一次筛选以减少数据量，于是基于一些规则的二度好友推荐方式出现在了人们的视野之中，例如基于投票规则的二度好友推荐。

2.2.1 传统二度好友推荐

社交网络中为了使得用户能够认识更多的朋友推出了“间接好友推荐”等推荐功能，其中用户之间的相互关注可以抽象如下图：

BCHEDFAGBCHEDFAG图2-3 社交网络用户关注抽象图

通过上述抽象图，介绍将如何利用用户关注关系来实现好友推荐的功能。节点A、B、C到H为社交平台中的用户，节点与结点之间的边表示用户之间的关注关系，节点与边构成了上述的社交网络抽象图。其间我们假设，如果A与B是相互关注的关系，那么我们认为他们为好友关系，或者至少有着一定的交际关系。用户A与用户B为相互关注关系，同时用户A与用户C也为相互关注关系，D用户与用户B和用户C同样具有这种关系，那么认为用户D为用户A的二度好友。而对于用户E，C、D、G、F都与用户E存在相关联系，那么A、B、H都为用户E的二度好友，于是我们将用户的二度好友推荐给他。

2.2.2 基于投票规则的二度好友推荐算法

对于上面的分析 ，我们使用了用户的二度好友作为其推荐好友，这种方式的推荐更像是一种图的宽度优先搜索，与此对应的还有N度好友，即宽度优先搜索通过逐步的对关系的扩展来查询相应的数据信息，我们以用户A为起点，逐层扩散，在以相对应的关系为基础构建的网络上，向外部进行扩散，即不进行往复行为地向前N次行走到达的地方，就是A所拥有的N度好友。改进的二度好友推荐采用了投票机制，选择出其中的较多选票的用户作为推荐好友，二度人脉只能单纯的看出在社交网络中的人际关系链接，而基于投票规则的好友推荐才能取得较优的推荐效果[10]。

User1User3User2User4User5User1User3User2User4User5图2-4 投票机制示意图

以研究user1的二度好友为例，图中可以看出，user1与user2，user3为一度好友关系，对于user1的二度好友关系存在：user1与user4，user1与user5，其中user4分别与user2，user3为好友关系，那么user2，user3都会给user4投票，而user5仅仅得到user3的投票，即{user4：2，user5：1}，显然user4更适合作为向user1推荐的二度好友。

2.2.3 MapReduce实现二度好友推荐

在数据量大规模增加的现在，任务的分析往往要比传统的数据分析任务复杂很多，主要原因在于巨大的数据量，例如大型的电商平台，其间包含大量的销售数据，搜索引擎的日志等等。对于我们的单机笔记本来说，它可以完成很多不同的事情，听音乐，撰写文档，观看视频等等，这些事情可以在同一台笔记本电脑上来完成，在数据量加大的情况下，使用单机来对这些大量的数据进行处理，往往显得力不从心。在大数据的时代，对于大规模的数据处理不再是由单机来完成许多不同类型的任务，而是许多电脑来完成相同类型的任务。假设一个程序，如果用单个笔记本来完成较大数据量的处理，那么单个主机将花费很长的时间来完成这项工作，于是有人考虑到用多台电脑来完成同一件事情，由此而引出了并行计的概念。MapReduce通过相应的分布式技术对问题进行处理，与此同时，将简单的工作分别分开来进行处理来达到快速批量处理的效果。

图2-5 MapReduce处理经典流程

许多电脑同时来完成同一件事情，涉及问题有很多，例如将任务怎么分开进行处理，分开的处理如何进行，如何将分开的任务重新组合，以及在单个主机出现问题应该如何处理。在社交网络中往往会遇到大规模的数据需要分析，而传统的方法对如此巨量的数据进行分析所消耗的时间也是很大的，如何减少分析时间，进一步的将分析提取信息时间大量减少，MapReduce被适时的提出，面对社交网络中海量的社交数据，MapReduce能大量的节省时间[11]。

LucyTomJackLucyTomJack使用MapReduce的方法进行二度好友推荐，我们将对MapReduce实现二度好友进行简要的说明，如下好友关系图所示：

图2-6 MapReduce方法说明

如上图所示：Jack分别与Tom，Lucy为好友关系，我们将Jack-Lucy这样的格式叫做key-value格式，以Jack为例，可以形成Jack-Lucy，Jack-Tom两个key-value对，那么我们能够得到Jack{Lucy，Tom}这样的集合，对该集合进行笛卡尔计算。笛卡尔乘积通俗来进行解释就是将2个任意的群集进行乘积计算得到的结果，将任意两个元素结合在一起得到的[12]。

通过笛卡尔计算，我们得到{ Lucy-Lucy，Tom-Tom，Lucy-Tom，Tom-Lucy},同时我们对得到的key-value值进行以字母

顺序进行逆操作，得到{Lucy-Lucy：1，Tom-Tom：1，Lucy-Tom：2}，在将重复元素进行去重，删除类似Lucy-Lucy这样的关系，就得到{Lucy-Tom：2}的集合，将Lucy-Tom重新定义为key值，其后的2代表二度好友，这样，借助Jack我们就能得到Lucy-Tom的二度好友关系。

2.3 协同过滤推荐算法

协同过滤推荐算法是在早期形成，而且流传程度比较高的推荐算法。它的主要的功能是做出相应的预测和对应的推荐。协同过滤算法分为两种：基于用户的协同过滤、基于物品的协同过滤。基于用户的协同过滤是通过用户对过往历史的行为数据，其中包括内容或者物品的评分，同时根据用户之间浏览或收藏的喜好程度或者评价进行相似度的分析。评论，浏览次数，收藏转发等等一些行为都可以作为评级的内在属性提供参考。以上这些行为都可以在某些方面展现了了用户对物品的喜好程度。

2.3.1 基于用户的协同过滤算法

对于基于用户的协同过滤算法我们可以直观的用下图表示：

图2-7 基于用户的协同过滤算法示意图

如上图所示，user为平台用户，product为平台商品，实线箭头表示用户与商品的购买关系，虚线箭头表示对用户的推荐购买。

解释如下：user1与user2都购买了product1与product3，相对于user2，user3和user2只共同购买了product2，显而易见，user1和user3相似度更高，我们可以通过user1给user3推荐产品，推荐对象为user1曾经购买过的而user3未曾购买过的商品，例如product1和product4。

有很多用户相似性的计算方法，如下所示：

表2-1 部分相似度计算的方法

类别方法

Correlation based Coisne, Pearson Correlation, Adjusted Cosine, OLS coefficient

Distance based Euclidean distance, Manhattan distance, Minikowski distance

Hash based Mini Hash，Sim Hash

Topic based PLSA, LDA

Graph based Shortest Path， Random walk, Item rank

对于基于用户的社交推荐，较为常用的相似度计算方法有欧几里德度量（Euclidean Metric）和皮尔逊相关度（Pearson correlation coefficient）[13]。

欧几里德（Euclidean Metric）距离评价：

欧几里德度量，我们将其作为一种在距离上度量的方法。其中它提到了在多维空间下两个点之间的真正的间隔长度。d=x1-x22+y1-y22+…+yn1-yn22#2-1

从意义上来说，欧几里德值比较小时，两个用户的相似度相对越大，同样的情况，欧几里德值比较大时，两个用户的相似度相对越小。

皮尔逊相关度（Pearson correlation coefficient）评价：

皮尔逊相关系数是衡量向量相似度的一种方法。其值大小范畴为-1到+1，0表示二者之间没有相近关系，值为负时代表负的关系，值为正时表示正的关系。

皮尔逊相关系数公式如下：ρX,Y=EX-μxY-μyσxσy=EX-μxY-μyi=1nXi-μx2i=1nYi-μy2#2-2

其中E表示数学期望，μ表示均值，σ表示标准差，Σ表示求和。

皮尔逊相关系数适用范围：

当两个变量的标准差都不为零时，相关系数才有定义，皮尔逊相关系数应用合适的情况有：

1. 两个变量的表现是线性的，并且都是接连的信息。

2. 两个变量表现为正态的分布，或者接近正态的单峰分布。

3． 两个变量两两出现，两两之间不相互影响。

我们通过皮尔逊相关系数来计算用户之间的类似程度，然后按照近似程度大小进行筛选。假设用户A、 B、C近似程度较高，那么我们认为这些用户属于同一个群体，即他们拥有着相似的兴趣，因此如果我们将向用户A做推荐，那么可以将用户B与用户C的喜好推荐给A用户。我们对不同属性进行加权排序，按照排序的顺序向用户进行推荐，这样该用户就被推荐得到了与他偏好相似的用户的商品，以上就是基于用户的协同过滤推荐算法基本概念。

然而该算法依靠用户历史的行为数据作为参考，而对于那么数据较少的平台或者说新平台，基于用户的协同过滤推荐算法的推荐结果将大打折扣，因为它十分受制于一定的信息量的大小，在较多的信息量的情况下基于用户的协同过滤算法具有相对不错的推荐结果。同时基于用户的社交推荐算法还存在其他的缺点，例如：信息有时候比较稀薄，一个较为常见的推荐系统通常都会有比较繁复的物品，顾客有时候购买的只是其中很少的一部分物品，不同的顾客相同的物品的概率不高，导致算法无法找到一个用户的邻居，即喜好较为近似的顾客。另外还有算法扩展性。最近邻居算法依赖于数据量的大小，如果数据量较大，那么相应的计算时间也会大幅度延长，它不太符合信息量较大的方面的处理工作。

2.3.2 基于物品的协同过滤算法

基于物品的协同过滤算法与基于用户的协同过滤算法具有很大的相似性，我们将商品与顾客的位置对调，通过同一顾客

对不同商品的评分，我们推断这些物品具有一定的相似性，对不同物品进行相似度的比较，我们将相似度高的物品作为一个群体，当用户对某物品产生一定兴趣时，我们将向其推荐同一个群体之中的其他物品。其主要思想就是首先通过所有用户的过往信息来进行计算物品之间的类似程度，接着将与用户较为喜爱的物品类似程度较好水平的物品向其他用户推选。基于物品的协同过滤算法计算物品相似度有以下几种方式。

基于余弦（Cosine-based）的相似度计算，在向量空间之中，将两个向量夹角的余弦值当作比较两个个体间差别大小的凭据[14]。余弦值越接近1，就表示他们的夹角越接近重合，也就是两个向量越类似程度越好。我们通过计算两个向量之间的夹角余弦值来计算物品之间的类似程度，

图2-8 余旋相似度示意图

公式如下：假设向量a、b的坐标分别为(x1,y1)、(x2,y2)cosθ=x1x2+y1y2x12+y12x22+y22#2-3

余弦值越接近1，就表明二者比较接近重合，也就是两个向量类似程度越高，为平角，即两个向量完全相等。

此外，余弦距离使用二者之间的角度的余弦值作为比较两个个体之间差别程度，和欧氏距离比较，余弦距离比较偏向于两个向量在方向上的差别，通过三维图像来展示欧式距离与余弦距离的区别，如下图：

图2-9 三维下的欧式距离与余弦距离

从上图可以看出，欧氏距离比较的是各个节点之间的真实距离，而这个值的大小取决于节点在当前坐标系下的位置。然而余弦距离比较偏向的是向量之间的角度的大小，更加偏向于展示二者之间方向上的不同，而不是实际的距离值的大小。

假如将A点不变化当前状态，而B点在原来的方向的基础上增加自身的长度，在这种情况下余弦距离的大小和之前没有任何区别（因为角度没有区别），而A、B两点的真实长度在逐渐产生了不同，它们两者之间的区别就在这里。他们二者之间都有着不一样的处理方法与不同的特点，所以他们分别用于不同情况下对于数据的处理。

从欧式距离来说，它展现了不同的维度之间数据的真实差别之处，如果我们需要对一个单独的行为进行分析，而且它属于不同的维度，那么将能够体现出二者之间的真实差别。而对于另一种近似程度的考量，余弦距离对于真正的数字大小反应较为稳定，主要是用于方向上的不同导致的差别之处，若用户之间信息量的不同导致的差别，那么它可以产生很好的应用效果。

2.3.3 两种推荐方式的区别

从实现的能力方面来说，如果某个平台信息量过小，那么基于用户的算法相对来说更加的合适，因为人数越少，需要比对的时间就会越少，那么相应的消耗就会越少。而信息量比较大的时候，基于用户就不太合适，因为人数量太多，那么比对信息需要的时间越长，我们可以采用基于物品的方案进行分析。

两种算法所面对的情况也不尽相同，基于用户来说，其对人数的变化较为敏感，人数越少，越容易比对，而当用户出现了不一样的行为，也能很好的捕获，及时的进行信息的处理。相对来说，基于物品不适用于这种情况，当物品信息量比较大时，它就派上了用场，能够准确的分析出物品差距来进行对比分析，得到较为良好的效果。

从冷启动来看，基于用户不能很好的利用现有行为信息进行分辨处理，每次用户之间的差异性以及类似程度需要等待一段时间进行处理[15]。基于物品能够应付这一情况，一旦有新的购买或者其他信息出现可以实时比对，并且迅速处理出新的结果。

从推荐理由看，对于基于用户的协同过滤相对较难用相关的数据的给出理由，而对于基于物品的协同过滤分析利用过往信息可以对其给出出相应的理由，以使得用户比较相信。

2.4 本章小结

本章主要介绍了社交网络推荐算法的概述，之后举例了相应的算法，并对算法进行了相应的解释，具体介绍了二度好友推荐算法的概念以及基于投票规则的二度好友推荐方式，同时介绍了MapReduce方法实现二度好友推荐的概念。详细介绍了基于用户的协同过滤算法以及基于物品的好友推荐算法的基本概念，以及二者之间的联系与区别，并介绍了用户相似度计算的几种方法，其中包括欧几里德度量（Euclidean Metric）、皮尔逊相关系数（Pearson correlation coefficient）、基于余弦（Cosine-based）的相似度。

第三章好友推荐算法的实现

3.1 实验环境

实验环境如表所示：

表3-1 实验运行环境

CPU Intel(R) Core(TM) i5-6300HQ CPU @ 2.30GHz (4 CPUs), ~2.3GHz

内存（RAM）8192MB RAM

操作系统 Windows 10 家庭中文版 64-bit

运行平台 JetBrains PyCharm 2017.1.3 x64

开发语言 Python3.6

3.2 好友推荐候选集的选取

选取好友推荐候选集的目的是避免不必要的全局搜索所带来的时间上大规模的浪费，这样可以在很大程度上提高我们工作的效率。在用户数据集合中大部分的用户并不是我们希望选取推荐作为好友的用户，假设某人有100个朋友关系，那最终会有将近20000个二度好友，数据表明，相当大的一部分的好友关系都建立在二维的空间之中。

基于三元闭包理论进行好友推荐算法，对用户进行对应的推荐，也在以另一种方式来采用二度好友实现推荐效果。在对好友候选集选取的时候，考虑到了在尽量不损失有效好友的情况下在尽可能的缩小候选集，以减少计算量，所以我们将二度好友作为好友推荐的候选集，继而为提高社交网络好友推荐的效果。

通过好友之好友的方式来进行二度好友的计算，对用户信息数据进行处理，来得到用户好友推荐候选集，而且该方法与日常生活中我们结识好友的方式基本相同，都是通过现有的好友来扩大自身的社交圈。在此之外，对好友候选集的计算当中，加入了投票规则。算法如下：

Algorithm：calculate the second friends Input： the all of users id, all followers of every user, the user id of who want to access his or her second friends（user1）. Output： the list of second friends of the user you have inputted, the list of users who is not in user dictionary (notInDict)Function：FOR every user IN user1' friend list: IF user2 in user dictionary: FOR user3 IN the dictionary of user2: IF user3 is not belong to user dictionary or user3 has been in user1' s friends list: continue IF user3 has been in the list of user1's second friend: value of user3 = value of user3 + 1 ELSE: IF user3 is not in the list of users not in user dictionary: continue ELSE: Append user3 to the dictionary of users who is not in user dictionaryAlgorithm：calculate the second friends Input： the all of users id, all followers of every user, the user id of who want to access his or her second friends（user1）. Output： the list of second friends of the user you have inputted, the list of users who is not in user dictionary (notInDict)Function：FOR every user IN user1' friend list: IF user2 in user dictionary: FOR user3 IN the dictionary of user2: IF user3 is not belong to user dictionary or user3 has been in user1' s friends list: continue IF user3 has been in the list of user1's second friend: value of user3 = value of user3 + 1 ELSE: IF user3 is not in the list of users not in user dictionary: continue ELSE: Append user3 to the dictionary of users who is not in user dictionary

图3-1 算法示意图

我们用以下示意图对该算法进行说明：

U1U2U3U4U5图3-2 用户关系示意图U1U2U3U4U5图3-2 用户关系示意图

如图所示，每个椭圆表示一个用户，椭圆内部的字符表示用户标识，箭头表示好友关系。我们首先假定将为用户u1来进行好友推荐集的选取。由图我们可知，用户u1与用户u2，u3为好友关系，而用户u3与用户u2都为用户u5的好友，用户u4仅仅与用户u3为好友关系，那么我们将（u4，u5）作为用户u1的好友推荐候选集，而且，因为用户u2，u3都将u5作为推荐用户，则用户u5在该推荐集合中得到2次投票，对比用户u4来说，用户u4获得用户u3的一次投票，即我们得到的推荐集合为{u4:1，u5:2}。

3.3 用户相似度的计算

在进行好友推荐的过程中，我们考虑如何将现有用户之间进行对比，同时我们将用户对比的范围约束在我们前文提到的好友推荐候选集之中，如何进行用户之间的匹配对比面临着如何进行好友相似度的计算。

3.3.1 用户相似度参考因素的选取

在用户数据中包含了很多不同类型的元素，我们将从中挑选出部分元素作为我们对好友推荐的研究参考因素。

（1）用户粉丝集合。用户存在关注与被关注的关系，例如单项关注或者双向关注，显然双向关注相比单项关注来说，好友之间的联系紧密度更高，这些都是无法从肉眼直接能看出的内在联系。单纯的从用户方面看，社交网络上的用户节点在很少的情况下，用户与用户之间的联内在联系在数量上来说却是远远大于用户数量的。用户之间同样以此种方式存在着联系，我们借此希望从中能够找到用户之间的潜在联系，并通过这些联系找出用户之间的更为密切的关系，我们将通过这些数据构建用户集的社交网络。

（2）用户累计微博数量。不同用户的发送微博的数量不同，那么一定程度上说明了用户基于社交媒体平台活跃度的不同，当然，由于微博账号创建时间的长短，我们采用月均微博数对其进行比较更能说明用户活跃程度的大小。

（3）用户地理位置。显而易见，相对较近的地理位置有着得天独厚的便利使得用户之间相互交流，所谓"近水楼台先得

月”，更容易以此来扩大自己的交际圈，例如，微信现如今存在的“附近的人”功能，其初衷就是想要通过地理位置这一优势来扩大用户自身的社交网络。

（4）用户好友数以及共同好友数。由于社交能力或者社交兴趣度以及活跃度的不同，不同的用户拥有着不同数量的好友。首先，我们对每一个相同好友进行同样的看待，同样，在拥有相同好友数目的情况下，当某个用户的好友数较少时，那么，这个共同好友相对来说更加重要。所以用户的共同好友数当作好友推荐的部分参考因素。

3.3.2 对于数据的处理

（1）计算共同好友。在好友推荐候选集中将用户数据集进行对比，找出不同用户对于待推荐好友的用户逐一比较，然后选出其中的共同好友。sameFriendsuser1,user2=friendsListuser1∩friendsListuser2#3-1

其中user代表用户，sameFriends代表共同好友，friedsList(user)代表user的好友列表。

（2）通过杰卡德系数对共同好友数进行加权处理，借用杰卡德系数[16]（Jaccard index）来进行用户之间类似程度的计算。我们将通过杰卡德系数来对用户双方好友数进行加权操作，以此来消除双方好友数处于不同水平而引起的差异。杰卡德系数定义如下：Coefficient= |friendsList(user1)∩friendsList(user2)||friendsList(user1)∪friendsList(user2)|#3-2

和以往传统相似性计算方式对比来说，杰卡德系数比余弦相似度的方式更加适合对于某些情况的处理，比如信息量较少的情况。其中，coefficient表示杰卡德系数，friedsList(user)代表user的好友列表。分子位置|friendsList(user1)∩friendsList(user2)|表示user1与user2的好友交集，分母位置|friendsList(user1)∪friendsList(user2)|表示user1与user2的好友并集。同样的，杰卡德相似系数较高时，二者的类似程度就越大高。

（3）对共同好友进行加权。在此之前，我们对所有用户好友平等的看待，但是实际情况并非如此，对于同一个共同好友，他相对于其他不同的用户来说有着不同程度的重要性。在之上的式子中，我们通过将共同好友数除以二者好友交集的数量，以此来平衡二者好友数之间的差距。同样我们可以采用相同的方式对共同好友进行加权操作。公式如下：Value= userfriendlistuser1∩friendlistuser21friendlistuser#3-3

上式有些时候仍然不能够弥补用户之间的差距，例如好友数目相差过大的情况，我们还可以采用开方，对数等方式进行处理。如下公式所示：value= userfriendlistuser1∩friendlistuser21friendlistuser#3-4

或者如下对数的方式进行处理：value= userfriendlistuser1∩friendlistuser21log2friendlistuser#3-5

对于我们通过下述图表进行必要的说明：

可以明显的看出，在开方或者对数的情况下，我们可以非常有效的对因为数量差距过大引起的影响，从而对共同好友数进行加权处理。

图3-3 函数示意图

从上图可以明显看出开方，取对数的方法对于平衡用户好友数差距过大的情况取得了有效的效果。

3.4 实验结果分析

对实验结果使用TOPK的评价主要参考准确率与召回率[17]，如下：准确率=TPT#3-6召回率=TRT#3-7

在上述式子中T表示推荐成功即已经是用户粉丝的的人数，PT表示推荐的人数，RT表示用户的粉丝人数。我们采用的参考因素有二度好友得分，用户平均每月发微博数量，共同好友比例，地理位置信息，我们将其分别赋予不同的权重，之后对准确率，召回率做一个比较，结果如下所示：

表3-2 实验结果总结表

权重/TOPK TOP40 TOP80 TOP120

[100，0，0，0] (0.175,0.875) (0.875,0.875) (0.0666,1.0)

[0，100，0，0] (0.0,0.0) (0.0,0.0) (0.0, 0.0)

[0，0，100，0] (0.075,0.375) (0.05,0.5) (0.05, 0.75)

[0，0，0，100] (0.025,0.125) (0.0125,0.125) (0.0167,0.25)

[25，25，25，25] (0.1,0.5) (0.0875, 0.875) (0.0583,0.875)

colSpan="4" >续表表3-2实验结果总结表

[40，30，20，10] (0.175,0.875) (0.0875, 0.875) (0.0583,0.875)

[30, 40, 20, 10] (0.175,0.875) (0.0875, 0.875) (0.0583,0.875)

[40, 30, 10, 20] (0.175,0.875) (0.0875, 0.875) (0.0583, 0.875)

[40, 30, 20, 10] (0.175,0.875) (0.0875, 0.875) (0.0583, 0.875)

[40, 20, 30, 10] (0.175,0.875) (0.0875, 0.875) (0.0583, 0.875)

[40, 10, 30, 20] (0.175,0.875) (0.0875, 0.875) (0.0583, 0.875)

[40, 10, 20, 30] (0.175,0.875) (0.0875, 0.875) (0.0583, 0.875)

[40, 20, 10, 30] (0.1,0.5) (0.0875, 0.875) (0.0583, 0.875)

我们将四个参考因素进行赋权重，四个参考因素分别是：用户二度好友出现的次数，用户活跃度，好友共同好友比例，地理位置，由上表可以看出用户二度好友出现次数效果较好，在其不变时，地理因素和共同好友比例作为参考因素效果也相对较好。

我们选择权重为[50，10，25，15]进行数值分配，之后进行两个数据的计算，得出下图：

图3-4 TOPK推荐结果示意图

可以看出当推荐用户增加时，准确率在最开始的增加之后进入下降状态，而召回率在随着推荐人数的增加，逐步上升到较为平稳的状态，这也符合二者之间的正确变化[18]。

使用我的id进行测试时，得到的结果如下所示：

图3-5 推荐结果示意图

数据分别对应用户的uid和昵称，上述图片中的19个人中有12个人是我的同学，也能够进一步说明算法的正确性。

3.5 本章小结

本章主要介绍了好友推荐算法的实现以及实验结果的展示，通过推荐集的选取以及对用户信息进行相关的处理，之后通过参考因素的选取，计算好友的相似度，并进行推荐，最后对实验的结果用准确率和召回率进行分析。

5. 第四章好友关系可视化的实现

总字数：4169

相似文献列表　文字复制比：0%(0)　疑似剽窃观点：(0)

原文内容　红色文字表示存在文字复制现象的内容; 绿色文字表示其中标明了引用的内容

第四章好友关系可视化的实现

4.1 实验环境

实验环境如表所示：

表4-1 实验运行环境

CPU Intel(R) Core(TM) i5-6300HQ CPU @ 2.30GHz (4 CPUs), ~2.3GHz

内存（RAM）8192MB RAM

操作系统 Windows 10 家庭中文版 64-bit

运行平台 JetBrains PyCharm 2017.1.3 x64

开发语言 Python3.6

在实验的实现过程中，我们使用了基于python的第三方模块，如下表所示：

表4-2 第三方模块及功能

Networkx 用于构建网络图像，对网络图进行处理。

Matplotlib 用于可视化的工具，能够很好的进行图像的绘制。

Basemap 用于地图绘制的可视化工具。

DataFrame 用于二维表格的处理。

Pandas 基于numpy的数据分析工具。

Levenshtein 计算文本内容的相似度。

Numpy 用于对多维数组的处理。

Polygon 制作多边形的工具。

4.2 好友关系网络图的构建

好友关系网络采用python语言实现，其中用到的主要模块有networkx与matplotlib,通过读取用户的数据信息以及用户之间的关系信息，将相关的用户以及用户关系进行图的绘制。如下图所示：

图4-1 好友关系可视化示意图

如图所示，每个节点代表一个用户，每条边代表用户之间具有好友关系。为了使得图像更加清晰化的展示，对每一个节点进行了颜色处理，使用networkx的spring_Layout布局方式以展示的好友关系网络图，可见好友关系网络图以近似聚类的方式汇聚在一起，我们将其部分展示出来，如下图所示：

图4-2 好友关系网络局部图

图像中节点的标签为用户所使用的昵称，通过图像展示的方式，可以清晰地看出好友之间的关联关系以及联系的紧密程度。更多的，我们能够观察出好友所处于不同的群体，以及所属群体的大小。

在此基础上，我们希望通过好友关系网络能够展示好友在当前社交群体之中的社交影响力的大小，选取社交影响力的参考因素有用户累计发送微博的数量以及用户所拥有的好友数，首先，好友数目的不同代表着用户在社交群体中拥有的基本影响力，用户的月均微博数目的大小，代表了用户所在群体之中的活跃度，之后由于用户创建微博的时间不同，所以累计发送微博的数量可能有明显的差距，我们采用创建微博后，月均发微博数量作为参考依据。对二者分别进行标准化处理之后，采用评分的机制，来评判用户社交影响力的大小[19]，节点的大小代表着用户在社交群体中的影响力大小。加入社交影响力后，更能清晰的观察到用户与用户之间的不同，从一定程度上体现了用户的社交影响力的大小。加入社交影响力之后的好友关系网络图如下所示：

图4-3 加入社交影响力的好友关系网络示意图

加入社交影响力的好友关系关系网络示意图显示了好友在社交圈中所占的比重的不同，同时也一定程度上表现出其社交

圈的大小。

4.3 好友地理位置可视化的构建

获取地理文件信息形成mapData开始获取用户数据地理信息形成friendData对信息进行标准化处理合并mapData与friendData对两个数据进行模糊匹配对数据进行最优匹配并进行去重结束开始设置经纬度获取中国地图区域轮廓读取中国行政区shape文件并导入绘制基本地图导入经过处理的dataFrame数据用色阶卡对相应地区进行填色处理结束获取地理文件信息形成mapData开始获取用户数据地理信息形成friendData对信息进行标准化处理合并mapData与friendData对两个数据进行模糊匹配对数据进行最优匹配并进行去重结束开始设置经纬度获取中国地图区域轮廓读取中国行政区shape文件并导入绘制基本地图导入经过处理的dataFrame数据用色阶卡对相应地区进行填色处理结束

图4-4 好友地理信息可视化流程图

为了能够使得用户更好的了解自身好友的地理位置情况，我们将对用户的好友地理位置信息做可视化的处理。好友地理位置信息可视化的构建通过两部分进行处理，第一步是对数据的提取以及处理，对用户的地理位置信息和地图shape文件进行数据的提取，之后对用户相应的数据进行标准化处理，将获得的二者信息进行合并，并对二者数据进行模糊匹配，去重，选取最优匹配，最终得到我们的所需要的数据，第二步进行图像的绘制，首先对经纬度进行定位，选取合适的经纬度，将大致的地图图像绘制出来，接着读入相关地图文件的信息，绘制具体形状，导入之前经过处理的数据，并根据数据采用色阶卡对相应地区进行填色处理。

4.3.1 好友地理位置信息可视化方案

通过以上的流程图对好友地理位置信息可视化方案作进一步的阐述。主要包括两个方面，用户地理信息数据方面，地图绘制方面，用户信息的处理主要是将提取出的用户信息与地图shape文件相匹配，再进一步对数据进行标准化处理，地图绘制方面，主要是提取地图shape文件的信息，进行基本的地图绘制，之后将经过处理的用户数据与色阶图进行对比，完成对地图的填色处理。

4.3.2 对相关数据的处理

首先统计各地区的人数，并将对其进行标准化处理，公式如下：数据标准化= x-minmax-min#4-1

X为各地区人数，min为各地区人数最小值，max为各地区人数最大值，这样处理后，使得其数值范围都在[0，1]之间，而且数据只是在数值上出现了改变，但是它并不能够对整体来说，占比例是没有变化的[20]，我们通过该数值对颜色深度进行设置。

为了使地理位置和shape文件中的地理位置相匹配，我们通过dataFrame对数据进行相关设置。首先，先对统计信息并将统计信息进行数据标准化的处理，将统计数据和经过处理的数据利用dataFrame进行存储,得到dataFrame的初试化示意表格，我们将其定义为friendData，如下所示：

表4-3 dataFrame初始化示意表

人数地区透明度

地区

云南 26 云南 0.168919

内蒙古 30 内蒙古 0.195946

广东 69 广东 0.459459

其他 149 其他 1.000000

山东 44 山东 0.290541

辽宁 38 辽宁 0.250000

福建 55 福建 0.364865

湖南 29 湖南 0.189189

香港 13 香港 0.081081

黑龙江 48 黑龙江 0.317568

重庆 30 重庆 0.195946

图片中的透明度就是上文我们通过数据标准化处理得到的数值，之后，我们对地理数据集进行读取得到相应的dataFrame，我们将其定义为mapData，如下：

表4-4 地图数据集dataFrame初始化示意表

上海|上海上海|上海

內蒙古自治區|内蒙古自治区內蒙古自治區|内蒙古自治区

北京|北京北京|北京

吉林吉林

四川|四川

天津|天津天津|天津

安徽|安徽安徽|安徽

寧夏回族自治區|宁夏回族自治区寧夏回族自治區|宁夏回族自治区

山東|山东山東|山东

山西山西

之后，为了将shape文件中的地理位置与好友信息的dataFrame进行匹配我们将dataFrame合并，通过索引可以将两个dataFrame表格进行合并，但二者索引并不完全相同，我们对其进行模糊匹配，计算二者之间的莱恩斯坦(Levenshtein)比例，即二者索引之间的相似程度，将筛选出的数据导入列表之中，如下所示：

图4-5 模糊匹配结果示意图

由上图可知，在一些列里面，同一省份重复出现，例如云南重复出现，原因是“模糊匹配”将匹配结果不为0的情况全部输出，但是通过我们可以得到，正确匹配的都是匹配度最高的那个，上图中将其红框标识了出来。之后我们在模糊匹配的基础上作最优匹配，通过对上述dataFrame过滤掉无效值，并进行排序降序的处理，得到如下结果：

图4-6 排序后的mapData示意图

由上图可以看出排序后的dataFrame的状况，我们用rn代表排序后的顺序，suitRatio代表匹配度，我们可以选出rn等于1.0的所在行，我们同样对其用红框进行了标注，并将其他的重复行删除，这样我们就得到了将用户信息与shape文件中的地理位置信息做了匹配，在接下来对地理位置可视化的处理奠定了基础。

图4-7 合并后的dataFrame示意图

之后，我们将friendData与mapData进行合并，得到匹配完成的dataFrame，如图4-7所示。

图4-8 匹配完成后的dataFrame示意图

可以看到，mapData与friendData合并的结果，我们只保留了匹配度最高的省份，之后只需要对不需要的数据进行删除，得到最终的结果， 得到上图所示的数据列表，我们就可以将该dataFrame加入绘制图像。

4.3.3 地理位置信息可视化

好友地理位置可视化的构建同样采用python实现，首先，我们先对地图进行基本的绘制，调用Basemap模块，设置相应的经度和纬度，对地图的大致轮廓即对基本的海岸线进行绘制。

图4-9 基本地图的绘制示意图

由于Basemap中不包含中国省份的经纬度信息，我们采用GADM网站提供的中国省份经纬度数据集，得到中国行政区划shape文件，读取shape文件中的具体的经纬度先进行中国地图的绘制，如下图所示：

图4-10 地图的基础绘制示意图

基本的地图绘制完成之后，在基础地图的基础上进行着色处理，在着色的过程中采用上述数据处理得到的各地区的透明度值，即通过数据归一化处理的好友地理信息数据，在着色过程中加入自制的色阶卡，色阶卡具体的指数如下所示：

图4-11 色阶卡示意图

色阶卡模拟热度色阶图，从左到右表示热度逐渐增大。

图4-12 好友地理位置信息可视化

4.4 本章小结

本章主要完成了两部分的工作，对好友关系网络可视化的构建以及用户地理位置信息的构建，对两部分的数据处理方式分别进行了相关的阐述，并举例展示了实验的效果图像。

第五章总结与展望

5.1 论文工作总结

随着互联网时代的进一步发展，社交媒体平台数据量与日俱增，越来越多的人将社交媒体平台作为联系好友，寻找有着共同兴趣朋友的平台。在社交平台上扩大自己的社交圈，而随着数据量大量增长，如何在海量用户中找到自己的潜在好友成为了问题之一，为用户提供个性化的推荐成为了趋势，社交好友推荐能够帮助用户在更短时间内发展自己的社交圈，增加用户对社交平台的粘度。为了能够给用户更好的了解自己的社交状况，以及了解自己的好友地理分布情况，我们将用户好友关系进行可视化的处理，包括好友关系网络的可视化以及好友地理位置信息的可视化，对这两部分进行好友关系可视化的处理都能够使得用户能够更好的了解自身的社交现状，我们将对以上的主要研究内容所做的具体工作解释如下。

本文所做的主要研究如下：

1. 基于社交网络的好友推荐。通过对用户现有的数据的分析，选择二度好友作为好友推荐候选集，计算二度好友的排行，通过好友推荐参考因素的选择，进行好友相似度的计算，以及对用户相似度的判断，进行TOPK推荐。

2. 好友关系网络图的构建。首先通过对用户之间关系数据的分析，初步构建好友关系网络图，在此基础上，加入用户社交影响力的因素作为参考，用户社交影响力通过平均每月发布的微博的数量，好友数因素来进行比较，通过好友关系网络图的构建，清楚的展现了用户的社交圈，以及用户社交影响力的大小，实现了好友关系网络的可视化。

3. 用户地理位置信息的可视化构建。对用户地理位置信息的初步处理之后对其进行数据标准化的处理，将该数据与地图shape文件的中的信息做模糊匹配，之后最优匹配。将匹配最终的数据比对色阶卡加入到地图绘制的填色处理，形成最终的用户地理位置信息的可视化构建。

5.2 未来研究展望

虽然本文就基于社交网络的好友推荐算法以及好友关系可视化方面做了一部分的工作，但是仍存在很多不足之处，有需要进行更深入的研究与探讨。

1. 由于新浪微博API的限制，数据源的获取变得更加的困难，因此获取的用户数据量以及用户的基本信息并不十分完整，仍旧需要进一步的工作来研究如何获取较大的数据量作为分析数据，同时还有较为完全的用户个人数据信息。

2. 在对用户信息进行相似度的计算时，只考虑了用户基本信息的比对以及用户之间关系的考虑，并未将用户的文本信息加入计算当中，如用户的微博文本，以及用户之间的互动信息，进一步将加入这些信息作为参考因素，以对用户相似度的计算进行优化。

3. 在好友关系可视化方面，用户社交影响力的考量因素过少，用户社交影响力的大小不应当仅仅包括微博数量以及好友数量的多少，其评论数，转发数，点赞数都应当加入考虑之中，由于数据的限制，该工作有待进一步的研究，同时好友地理位置信息可视化方面应当加入用户位置签到信息的考虑。

参考文献

[1]孙晓晨,徐雅斌.位置社交网络的潜在好友推荐模型研究[J].电信科学,2014,30(10):71-77.

[2] Paul, Jaccard. THE DISTRIBUTION OF THE FLORA IN THE ALPINE ZONE.

[3]高杨,张燕平,钱付兰,赵姝.基于三元闭包的节点相似性链路预测算法[J].计算机科学与探索,2017,11(05):822-832.

[4]彭敏,席俊杰,代心媛，何炎祥.基于情感分析和LDA主题模型的协同过滤推荐算法[J].中文信息学报,2017,31(02):194-203.

[5]李金海，何有世，熊强.基于大数据技术的网络舆情文本挖掘研究[J].情报杂志,2014,33(10):1-6+13.

[6]张琳. 电子商务网站个性化推荐的多样性对推荐效果的影响研究[D].北京邮电大学,2017.

[7]陈力丹,霍仟.互联网传播中的长尾理论与小众传播[J].西南民族大学学报(人文社会科学版),2013,34(04):148-152+246.

[8]徐冬冬,吴韶波.一种基于类别描述的TF-IDF特征选择方法的改进[J].现代图书情报技术,2015(03):39-48.

[9]陈思憬,骆冰清,孙知信.基于混合好友路径信任度的社交好友推荐算法[J].计算机技术与发展,2018,28(02):74-77.

[10]邹本友,李翠平,谭力文,陈红,王绍卿.基于用户信任和张量分解的社会网络推荐[J].软件学报,2014,25(12):2852-2864.

[11]向林泓,陈芋文,张昱琳.基于Hadoop平台的高阶矩阵相乘MapReduce算法研究[J].计算机科学,2013,40(S1):96-98.

[12]黄宏图,毕笃彦,查宇飞,高山,覃兵. 基于笛卡尔乘积字典的稀疏编码跟踪算法[J]. 电子与信息学报,2015,37(03):516-521.[2017-08-26].

[13]熊回香,叶佳鑫.一种双层的微博用户相似度算法[J/OL].情报杂志:1-7[2018-05-20].http://kns.cnki.net/kcms/detail/61.1167.G3.20180507.0917.002.html.

[14]王行甫,付欢欢,王琳.基于余弦相似度和实例加权改进的贝叶斯算法[J].计算机系统应用,2016,25(08):166-170.

[15]于洪,李俊华.一种解决新项目冷启动问题的推荐算法[J].软件学报,2015,26(06):1395-1408.

[16]张晓琳,付英姿,褚培肖等.杰卡德相似系数在推荐系统中的应用[J].计算机技术与发展,2015,(4):158-161,165.

[17]陈冲. 基于新浪微博的好友推荐系统设计与实现[D].西南交通大学,2017.

[18]马力. 基于聚类分析的网络用户兴趣挖掘方法研究[D].西安电子科技大学,2012.

[19]房旋,陈升波,宫婧,孙知信.基于社交影响力的推荐算法[J].计算机技术与发展,2016,26(06):31-36.

[20] Chen Bao, Lixun Cai, Kaikai Shi, Chen Dan, Yao Yao. Improved normalization method for ductile fracture toughness determination based on dimensionless load separation principle[J]. Acta Mechanica Solida Sinica,2015,28(02):168-181.