分类号	TP391	密级	公开
UDC	004.8	学位论文编号	

重庆邮电大学硕士学位论文

中文题目	融合用户文本语义和情感分析的好友推荐方法研究
英文题目	**Research on Friend Recommendation Method Based on fusion of Users' Text Semantic And Sentiment Analysis**
学号	**S130231041**
姓名	孙红涛
学位类别	工程硕士
学科专业	计算机技术
指导教师	刘群教授
完成日期	**2016 年 4 月 10 日**

独创性声明

本人声明所呈交的学位论文是本人在导师指导下进行的研究工作及取得的研究成果。尽我所知，除了文中特别加以标注和致谢的地方外，论文中不包含他人已经发表或撰写过的研究成果，也不包含为获得重庆邮电大学或其他单位的学位或证书而使用过的材料。与我一同工作的人员对本文研究做出的贡献均已在论文中作了明确的说明并致以谢意。

作者签名：　　日期：年月日

学位论文版权使用授权书

本人完全了解重庆邮电大学有权保留、使用学位论文纸质版和电子版的规定，即学校有权向国家有关部门或机构送交论文，允许论文被查阅和借阅等。本人授权重庆邮电大学可以公布本学位论文的全部或部分内容，可编入有关数据库或信息系统进行检索、分析或评价，可以采用影印、缩印、扫描或拷贝等复制手段保存、汇编本学位论文。

（注：保密的学位论文在解密后适用本授权书。）

作者签名：　　导师签名：

日期：年月日　　日期：年月日

摘要

随着 Web 2.0 技术的出现和互联网技术的迅速发展，社交网络服务(SNS)在现实生活中得到了广泛的应用。好友推荐作为在线社交网络（OSN）中一个重要的功能，增强了用户之间的黏性。目前 SNS 网站中的好友推荐一般采用用户的静态特征，但是现实生活中用户的兴趣和情绪波动很大，采用静态特征不能很好的表征用户，推荐效果不好。用户的微博中包含了丰富的用户信息，鉴于此，本文分别从用户的微博文本语义分析和情感分析两个角度对用户的兴趣特征进行分析和研究。

首先对微博文本中的语义和程度副词进行分析，提出一种采用二阶段的好友推荐方法。本方法的关键是通过语义分析技术对特征词进行替换来计算特征词的相似度。由于用户的特征是随时间变化的，所以特征词相似度计算中同时引入了时间因素。在获得语义特征相似的用户之后，又进一步考虑用户的情感特征，微博文本中往往含有表述用户情感的词汇，通过这些词汇对用户的情感特点进行分析，进而对上一步产生的结果做优化筛选，得出最终的结果。通过在真实的微博数据上进行验证分析，实验结果表明基于文本语义和情感程度的推荐方法比传统的推荐方法要好。

其次，基于上述的研究，进一步对文本和情感的分析进行更深层次的分析，提出一种融合文本语义分析和情感分析的好友推荐方法。本方法考虑了用户获取信息的时间先后顺序，使用了交叉相似度计算方法计算用户之间的文本相似度。由于文本语义研究中，程度副词可以表现用户的情感强烈程度，而否定词则可以改变用户文本的情感倾向。所以在对用户发布的微博进行情感分析时，通过程度副词和否定词的分析考虑了用户的情感倾向。通过在真实的数据集上与传统的好友推荐算法进行对比，得出的推荐结果在各项指标上都有提高。

最后，由于时间因素的影响，SNS 网站中在不同时段对不同类型的用户的推荐结果往往不同。所以本文对不同时间段的微博内容进行分析，设计了融合时间因素的文本语义和情感分析的好友推荐系统。

关键词：社交网络服务，文本语义，情感分析，时间衰减， 好友推荐

Abstract

With the emergence of Web 2.0 technology and development of the Internet, the SNS has been widely used in our real life. As the friend recommendation can enhance the stickiness of users, it becomes an important technology in the online social networks. Almost all SNSs have the function of friend recommendation, and in general use the users' static characteristics. However, the interests and emotions of users are often varied in their real life so that they can't describe the users' characteristics very well and get the bad results. On the contrary, the text of micro-blog contains many rich information of users'. In view of this, we start analysis and research on the users' characteristics from the semantic analysis and sentiment analysis of the text of micro-blog respectively in this thesis.

Firstly, the text semantic of micro-blog and degree adverbs are considered and then a two stages method of friend recommendation is proposed. In the model, the key point is to use the text semantic technology to change the key words to compute the similarity of friends. Due to the characteristics of the users are changed by time, a time factor is also introduced into the computation of users' similarity. Then we take further consideration on the user's emotional characteristics to compute the users' similarity through analyzing the emotional words in micro-blog text and get the final results. The fusion method of the text semantic analysis and degree adverbs analysis are better than traditional methods.

Secondly, we have further deeper analysis on the text and sentiment analysis based on the above research and then a friend recommendation method based on fusion of users' text semantic and sentiment analysis is proposed. Considering the different time that the user accesses to the information, the cross similarity calculation method are used to calculate the text similarity between users. In the research of text sentiment analysis, the degree adverbs can show emotional intensity and the negative words can change the tendency of emotion. So the users' emotional tendency with the effect of degree adverbs and negative words with the sentiment analysis of micro-blog of users are considered. We have comparison with some traditional methods on real datasets and find that the recommendation results are increased on various indicators.

Finally, due to the influence of time decay, the recommendation results of SNSs are different from different types of users in different periods. We analyze the text of micro-

blog of different period in this thesis and then design a friend recommendation system merged text semantic with sentiment analysis and time factor.

Keywords: SNS, text semantic, sentiment analysis, time decay, friend recommendation

目录

注释表

SNS	Social Networking Service，社交网络服务
OSN	Online Social Network，在线社交网络
ICML	International Conference on Machine Learning,国际机器学习大会
CNNIC	China Internet Network Information Center，中国互联网络信息中心
FOF	Friend Of Friend，朋友的朋友
TF-IDF	Term Frequency-Inverse Document Frequency，词频，逆向文件频率
RSS	Really Simple Syndication，简易信息聚合
LDA	Latent Dirichlet Allocation，文档主题生成模型
LSA	Latent Semantic Analysis ，隐含语义空间
VSM	Vector Space Model，向量空间模型
AHP	(Analytic Hierarchy Process，层次分析法
FOF+tag	Friend Of Friend plus Tag，朋友的朋友结合标签
FOF+BP	Friend Of Friend plus Birthplace，朋友的朋友结合出生地
UMFR	Unified Microblog Friend Recommendation，结合微博的好友推荐
RMSE	RootMean Square Error，均方根误差
MAE	Mean Absolute Difference，平均绝对误差
API	Application Programming Interface，应用程序编程接口
GPS	Global Positioning System，全球定位系统

表目录

图目录

第1章 引言

1.1 研究背景和意义

随着移动互联网的出现，人与人之间可以随时随地的进行交流沟通，交流变得更加方便快捷。社交网络(SNS)正在以一种新兴的姿态进入人们的生活，各种社交网站层出不穷，它们给人们的生活带来便利的同时也逐渐的改变着人们的生活方式和社交方式。SNS 为互联网用户提供了多方面的信息和服务，这些信息和服务同时不仅使人们更加紧密的联系在一起，促使现实社会不断的向虚拟社会过渡，同时也极大的促进了移动互联网的发展。国外的 Twitter，YouTube，Flickr，LinkedIn 和国内的微博、豆瓣等社交网站迅速占领了互联网社交市场，渗透到人们的生活中。人们对互联网和社交网络的态度在逐渐发生转变，每天都有数亿人通过社交网站分享内容、图片或者视频，而且越来越多的网民通过各种社交网站来获取信息、沟通交流和商贸交易，社交网络给人们的生活带来了很多便利。另外，社交网络实现了线上交流和线下交流、虚拟社会和现实社会的融合，是人们日常生活和工作中的必备品。

社交网络的理论基础是六度分割理论[1]，互联网的高速发展使用户的交友方式发生了巨大的改变，由以往的线下交友，逐渐向线上交友过渡。用户可以根据自己的需求查询和添加好友来建立和扩大自己的社交圈，实现好友之间的信息交流和互动，通过在线社交网络建立起全新的沟通交流方式[2]。庞大的用户规模使用户很难通过查找来寻找志同道合的朋友，用户已经不满足于和已经认识的好友聊天，而是希望通过网络认识更多的人，扩大自己的社交圈[3]。好友之间是相互影响的关系，网络市场调查公司 Sociable Labs 曾在 Facebook 上发起过一项实验：通过对 1088 名用户进行调查，发现在经常上网的用户中，接近 50%的用户会倾向于购买朋友推荐的产品，同时 75%以上的人会点击朋友在社交网站中分享的产品链接①。由此可见，在社交网站中进行精准的好友推荐非常有意义，而且用户的活跃度对社交网站的成功与否也是至关重要的，因此如何提高用户和社交网站的黏性成为当前社交网络中亟待解决的问题。

①http://www.199it.com/archives/33877.html

在大数据时代，网络上的信息繁杂，存在过多冗余信息。在一些社交软件中，用户的使用量、活跃度和信息量的高速增长，给整个网络环境带来了巨大压力，同时给社交网络平台带来了严峻的考验，如何从海量用户中给用户进行准确的好友推荐尤为关键。以 Twitter、Facebook、微博为主的社交网络平台吸引了大量的用户。在国内，新浪微博的用户量是很成功的社交网络平台之一，新浪微博官方统计的数据显示，2012 年 3 月，新浪微博用户规模为 3.24 亿；2013 年 3 月，用户规模增长到 5.365 亿，同比增长 65.5%左右。2012 年 3 月，新浪微博活跃用户规模为 3016 万；2013 年 9 月，该数值增长到 6020 万，一年半时间里接近翻番①。新浪微博里庞大的用户群体和平台上海量的用户微博信息，为用户对信息的获取和用户之间的交流带来方便，人们可以以更加快捷、低成本的方式获取所需的信息。物极必反，信息超载使用户在寻找感兴趣的话题和感兴趣的好友时变得越来越困难。为了更好地为用户提供服务，社交网站开始采用信息过滤和个性化推荐技术。推荐系统和传统的检索系统不同，不需要用户主动提供信息，而是利用数据挖掘技术从用户的历史信息中提取出用户的兴趣，并进行推荐，属于主动性的推荐。用户的兴趣可以是人，也可以是物或者是地点，挖掘出用户和它们之间的关系，给用户提供个性化推荐满足用户需求是推荐系统的主要任务。

社交网络和互联网信息传播的主导因素是人与人的关系，提升社交网站中用户的满意度，能够增强用户和社交网站之间的黏性。在社交网站平台中，用户只能看到自己的关注列表和自己选取的感兴趣话题，这就是信息过滤的一种体现，帮助用户合理的管理自己的社交圈。用户的这种需求推动了个性化推荐研究领域的发展，目前的好友推荐算法主要是通过分析用户的一些历史信息或轨迹，来进行个性化推荐，帮助用户找到志同道合的人。现有的社交网站都实现了个性化推荐服务，能够帮助用户主动的发现并推荐感兴趣的人，提升用户对网站的满意度，并增强相互之间的黏性，个性化推荐服务对社交网站和用户来说都很有意义。

①http://www.donews.com/net/201402/2711464.shtm

1.2 国内外研究现状

随着数据挖掘和机器学习研究领域的推动，个性化推荐技术已经成为了热门的研究方向，在学术界得到很多学者的关注，推荐系统进而成了独立的研究领域。在工业界，各大互联网公司如腾讯、豆瓣、阿里巴巴、百度和各大研究院如微软亚洲研究院、IBM 研究院、网易盛大研究院都有独立研究推荐算法的团队，努力给用户提供满意的推荐结果，提升用户的使用体验。在学术界，自 2009 年起，每年 ACM 推荐系统年会(ACM Conference on Recommender Systems,简称 RecSys)都会举办专门的推荐系统研讨会，同时在 KDD、ICML、ICDM 和 ACM TOIS、ACM TKDD、IEEE Intelligent System 等著名的数据挖掘与机器学习领域的顶尖会议或期刊上，出版很多推荐技术方向的 paper。移动互联网技术的发展，推动了个性化推荐技术的应用和研究，各种个性化推荐技术也得到了广泛的发展，相关研究和应用如雨后春笋般层出不穷。同时随着微博在中国的推广和使用人数的增长，微博中的推荐技术也得到了发展，在自然语言处理和中文计算会议上提出了很多自然语言处理方面的评估任务，其中就包含了中文微博文本的情感分析项[4]。

Jon Kleinberg[5]认为好友推荐就是“链路预测”问题，因此研究和对比了各种链路预测方法。在传统的好友推荐研究领域，一般通过注册时填写的信息对用户的关系进行预测，然后进行好友推荐，这种方法产生的推荐结果往往是相互认识的朋友关系。Gou L 等人[6]利用标记生成树和关系图的社交网络来发现好友，文中考虑了用户的动态信息，利用听音乐时的音乐标签相似性来进行好友推荐。Chin A 等人[7]主要关注 Facebook 和 LinkedIn 上的好友推荐，他们根据用户工作和参加相同会议的次数等行为特征来表明目标用户潜在的好友信息。YANG T 等人[8]利用社会标签系统的广泛使用，把好友推荐转化为链路预测问题，对用户的兴趣网络和社会网络信息同时进行处理，随后进行个性化好友推荐。Chu C H 等人[9]利用用户在某个地点停留时间的长短构建泰森多边形，并分析用户之间的位置相似度和用户的各种社交网站账户的兴趣列表，通过正则匹配找到最长公共子序列，最后达到高质量用户的推荐效果。Silva,N.B[10]利用遗传算法对复杂网络理论中的拓扑特征、拓扑信息和指标进行优化来解决好友推荐的问题，这种方法基于 FOF 机制进行推荐。好友的推荐在社交媒体中是一个重要的推荐功能，Xiao Yu 等人[11]提出了两阶段的更准确的好友

推荐方法，在不同的社交网络中，根据用户对网络相册的标签和朋友圈信息来生成可能成为朋友的待推荐列表，然后通过标签和图像信息的共簇来对列表进行过滤，达到推荐的效果。Zhibo Wang 等人[12]利用智能手机中的感应设备来获取以用户为中心的用户数据，发现用户的生活方式具有更高相似性的人更可能成为朋友。通过分析用户日常生活的生活文档，利用 LDA 模型提取用户信息来进行用户匹配。

在上述的好友推荐算法中，都会存在算法只适合相应的系统使用，应用型较低的情况，同时国内外研究中也存在以下几个方面的问题：

1. 社交网络中，用户的信息不能很好的利用，一般的社交网站用户的信息往往分为两类，静态信息和动态信息，在好友推荐领域，都是只关注了用户的不易变因素。在现实生活中，用户容易受到外界的影响，各种需求较为广泛，人类的情感认知比较丰富，而常见的好友推荐模型都不能及时地满足用户的需求。在线社交网络中，微博或博客等自媒体上发表、转发或评论的信息，不仅包含了用户的兴趣所在，而且隐含用户的情感特征，这些都可以作为好友推荐的一个非常重要的参考数据。

2. 数据的稀疏性和系统的冷启动问题，数据的稀疏性是推荐领域和数据挖掘领域一直存在的问题，相关的研究论文也提出了很多方法来解决稀疏性问题，但只是在一定程度上减少稀疏性对推荐的影响。如可以使用线性代数矩阵分解中的奇异值分解技术来减少数据稀疏问题[13,14]。冷启动问题主要是每天存在很多的新用户去注册和使用推荐系统，在推荐系统中不存在用户的历史数据和兴趣标签，很难对用户进行推荐，随着互联网的发展，各个网站都开放了对应的接口，冷启动问题带来的影响越来越小，冷启动问题应该被忘记[15]。

3. 用户信息的隐私和安全，数据挖掘的出现，使用户的个人信息变得不再个人。用户每天都在互联网上进行信息的贡献，很多互联网公司通过各种技术手段将用户的各种小数据进行结合，通过数据挖掘技术构造出用户清晰的行为图谱，对用户的偏好进行准备定位和预测，这种技术会对用户造成直接伤害。在推荐系统中如何解决用户的隐私性问题的解决方案还没有[16]，在未来隐私保护会变得越来越昂贵。

4. 数据集和评价指标，推荐系统的好坏不仅仅依赖于推荐算法和理论方面的优化，评价指标的选定也很重要。推荐系统的评价可以分为离线和在线两种方式，在科研工作中，最常用的方法是离线评价。但是现有的离线评价方法和指标存在考虑不全面的缺陷，评价指标较为简单，现有的科研论文中，大部分采用的是以准确

率为主，但是准确的预测不一定代表好的推荐[17]。推荐系统应该帮助用户扩大现有的交际圈，帮助用户主动的去探索新的领域的好友，帮助用户推荐不活跃的用户。要达到这些目的，必须要增加其他的评价方法，如推荐系统的覆盖率、实时性、多样性、新颖性等[15,18-20]。

1.3 研究目标和内容

在社交网络中，用户的需求变化波动和情绪变化较大，通常用户会通过文本表达个性化的需求和特征。针对这种情况，本文的研究目标是对用户的微博文本内容进行提取，利用语义分析等技术，提出和构建用户的个性化标签模型，并同时考虑用户的情感。用户的兴趣随着时间的变化具有衰减性，根据其特点，融合的时间因素，利用层次分析法来计算权重并验证权值的合理性，挖掘出用户列表，从而使得推荐的好友更准确，更具有个性化。

推荐技术是数据挖掘和链路预测下的研究领域，能够在海量信息中给用户推荐感兴趣的信息，好友推荐能够增加社交网站与用户之间的黏性而成为研究热点。现有的好友推荐系统主要是通过获取用户特征标签来进行推荐。在科研论文中，结合用户的文本语义和情感来进行好友推荐的较少，而在文本中又富含表征用户的信息，所以本文的重点是用户的文本和情感。本文主要针对用户文本语义和情感进行研究，主要工作及贡献包括：

1. 介绍推荐系统的主要工作原理，并对目前常用的推荐算法进行介绍，并简要分析了现有算法的不足之处，为下一步的研究奠定基础。

2. 对用户在网络中的行为进行了分析，用户作为在社交网站中的主体，通常会在社交网站中不定期的发表富含用户情绪的微博文本内容。这些内容均属于用户的动态特征，能够很好的表征用户，本文对文本语义和情感进行了详细的分析和讨论。

3. 通过对用户的文本内容提取，进行语义分析，得出用户的文本标签，采用交叉的相似度计算方法对相似度进行计算。在对用户的文本进行分析时，重点考虑了用户文本的情感倾向、程度副词和否定词。在考虑程度副词时，按照表达情感的轻重分配权重，否定词能够改变用户的情感倾向，在中文中，根据负负得正的原理，本文通过计算文本中出现的否定词的奇偶个数来决定用户的情感表示是否改变。

4. 在推荐系统中，时间上下文因素对推荐效果起着非常关键的作用，用户的兴趣随着时间的变化处于波动状态。为了考虑这一重要的上下文因素，以运筹学中层次分析法为决策方法，在计算用户之间的相似度时，根据距离当前时间的前后顺序进行权重决策分析，提高推荐质量和准确度。实验证明加入时间因素和权重分析能够得到更好、更符合用户兴趣的推荐结果。

5. 本文在对数据采集和算法评价时，与传统的推荐算法不同。由于数据集为微博数据，现有公开的数据集较少而且官方 API 限制较多，采用了自行编写的爬虫程序进行爬取，并使用长尾分布对数据集进行了验证。同时从不同角度对多种评测指标进行实验对比，可以比较全面和客观的对实验结果进行分析，得出规律性的结论。

1.4 论文结构

本文总共分为 6 个章节，主要的内容如下：

第一章，介绍了本文的研究背景和意义，总结国内外科研机构在好友推荐方向的研究现状，引出本文的主要研究内容。

第二章，主要介绍了推荐系统的主要原理并分析了相关的技术研究，以及现有的一些个性化推荐技术，分别对这些方法进行总结归纳，为后续的研究提供强有力的理论基础。

第三章，基于用户文本语义和情感程度的好友推荐，本章介绍基于用户微博的文本内容的用户信息发掘，中文文本中往往包含有大量的隐含信息，而微博文本内容虽然比较短小，但信息量却较大，基本上是比较精简和重要的词汇。在文本相似性计算的方法中首先从文本中提取关键词，把提取的关键词作为用户的个性化标签，然后计算用户的相似性以便推荐。而中文中经常会含有同义词等容易混淆的词汇，所以需要对文本进行语义分析进而作为用户的个性化标签，同时程度副词的作用可以增强用户的表达情绪的强烈程度，以计算文本语义相似度为第一阶段，情感程度相似度为第二阶段，当用户的文本语义具有相似的情况下，再对用户的情感进行考虑。

第四章，基于用户文本语义和情感分析的推荐方法，针对微博文本的推荐中，考虑用户最近几天的微博文本，采用交叉的相似度计算方法来计算文本相似度。同时用户获取信息的先后顺序也不尽相同，用户的情感的影响因素有程度词、情感词

和否定词。基于以上因素，提出了融合两者的好友推荐算法，并通过实验证明，这种融合后的算法能够得到较好的效果，用户满意度高。

第五章，融合文本语义和情感分析的好友推荐系统，本章将从用户的需求考虑，从微博文本中提取用户的文本和情感作为用户的标签，对用户进行深层次分析，设计了融合文本语义和情感分析的好友推荐系统。

第六章，总结和展望，对本文工作进行总结，并进一步讨论未来工作的设想和研究内容。

第2章 好友推荐相关理论基础

2.1 推荐系统和算法

2.1.1 推荐系统模块

推荐系统是解决信息超载问题一个非常实用的方法，从用户的个人信息、历史轨迹、社交关系等因素中，提取出用户的信息需求、兴趣等，进行个性化计算。好的推荐系统不仅能够为用户提供好的服务，还能增加用户与推荐系统的黏性，让用户产生依赖。虽然现在很多社交网站都提供推荐服务，但是推荐系统的核心思想都是一样，推荐系统通用模型如图 2.1 所示。

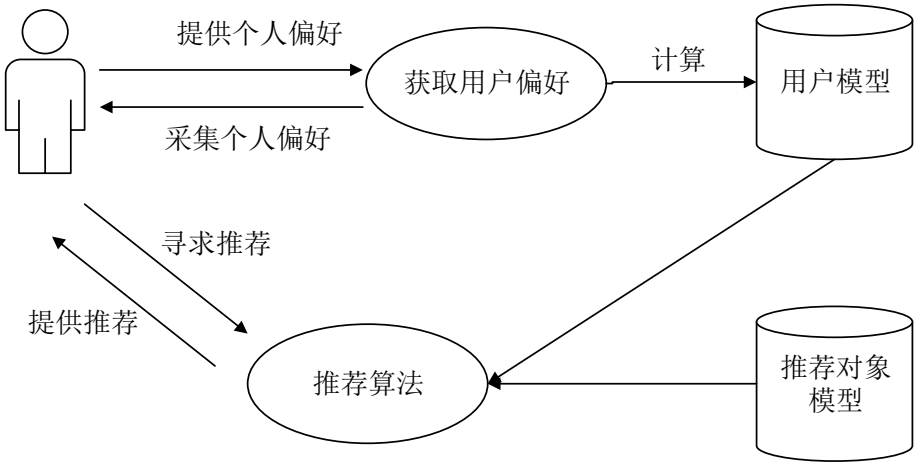


图 2.1 推荐系统通用模型

推荐系统有用户建模、推荐对象和推荐算法三个重要模块。给用户提供个性化的、高效的、准确的推荐的前提因素是能够获取用户的兴趣偏好。在用户建模模块中，能够获取用户的个性化兴趣爱好，并对用户进行分类，使推荐系统能够更好的识别用户的兴趣、需求和爱好。用户的信息一般包括用户属性、手工输入的信息、历史轨迹信息等。、推荐对象模块就是对推荐对象的特征进行提取，在对用户进行推荐时，需要对参考的因素进行准确定位。不同的对象所表现的特征也不相同，主要有基于内容的方法和基于分类的方法。在三个模块中，最为核心的部分就是推荐

算法模块，推荐算法决定推荐系统性能的优劣，根据不同的推荐系统、不同的推荐对象和不同的需求采用不同的推荐算法。

2.1.2 协同过滤推荐

协同过滤推荐算法在推荐领域是最著名的也是最常用的一种推荐算法，协同过滤算法可以分为基于用户和基于物品。GroupLens 在论文中首次提出了基于用户的协同过滤算法[21]，后来由亚马逊提出了基于物品的协同过滤方法[22]。该算法可以根据用户的行为偏好预测用户的兴趣点，如果两个用户有相同的兴趣（比如共同关注了某人），则他们在以后的行为中也会出现同类现象。基于这种思想，该算法适用范围较广。由于限制较少，可以应用在多种类型的推荐系统中。协同过滤算法在使用时，主要分为三个步骤[23]。

1. 构建评分矩阵

在很多推荐系统中，都存在打分机制，利用评分的相似度来计算用户之间的相似度进而产生推荐结果。对于 a 个用户和 b 个项目，可以构建 $a \times b$ 的 user-item 评分矩阵，其中 a 行代表用户，b 列表示项目。第 i 行 j 列 S_{ij} 表示用户 a 对项目 b 的打分，如表 2.1 所示。

表 2.1 评分矩阵

item / user	1	…	j	…	b
1	S_{11}	…	S_{1j}	…	S_{1b}
...	…	…	…	…	…
i	S_{i1}	…	S_{ij}	…	S_{ib}
...	…	…	…	…	…
a	S_{a1}	…	S_{aj}	…	S_{ab}

2. 相似度计算

准确计算用户相似度在协同过滤中属于重点，根据评分矩阵，计算用户之间的相似度，常用的相似度计算方法主要有以下三种：

(1) 余弦相似度：用户 i 和用户 j 在评分矩阵中的评分可以看作向量 $\vec{i}$ 和 $\vec{j}$，则用户之间的相似度可以通过两者之间的夹角得出，如公式 2.1：

$$sim(i,j)=cos(i,j)=\frac{\vec{i}\cdot\vec{j}}{\|i\|\cdot\|j\|} \tag{2.1}$$

(2) 修正的余弦相似度：为了消除不同用户评分在评分标准的不同，在余弦相似度的基础上，将用户的评分减去用户的平均分，如公式 2.2 所示。

$$sim(i,j)=\frac{\sum_{c\in I_{ij}}(R_{i,c}-\overline{R_i})(R_{j,c}-\overline{R_j})}{\sum_{c\in I_i}(R_{i,c}-\overline{R_i})^2\cdot\sum_{c\in I_j}(R_{j,c}-\overline{R_j})^2} \tag{2.2}$$

其中，I_{ij} 表示用户 i 和用户 j 共同评分的集合，I_i 和 I_j 分别表示用户 i 和用户 j 各自的评分集合，$R_{i,c}$ 表示用户 i 对项目 c 的评分。$\overline{R_i}$ 和 $\overline{R_j}$ 表示用户 i 和用户 j 各自评分项目的平均值。

3. 推荐结果

推荐结果通常采用 TOP-N 和相关阈值（Correlation Threshold）两种方法描述，TOP-N 是指通过计算用户与用户之间的相似度，给每个用户推荐 N 个最可能符合用户意愿的用户或物品。相关阈值是指根据相似度计算公式得出的结果，当超过一定的阈值，则加入到推荐列表。其中 N 和阈值都是指定参数或者通过多次实验得出。

2.1.3 基于内容的好友推荐

基于内容的推荐方法最早利用在信息检索和信息过滤领域[24,25]，是一种根据用户的文本内容进行推荐的方法，通过对用户的历史数据和标签信息进行分析，挖掘出相似的特征，推荐出有类似特征的用户。基于内容的相似度计算方法和协同过滤算法中的相似度计算方法类似，该方法特别适合有文本的社交网络平台，如微博，RSS 等。当两个用户发表的内容相似，我们可以认为两个用户具有相同的兴趣点和关注点，这样两个用户成为好友的概率就高。基于内容的推荐过程需要三个步骤：

1. 内容分析器：从原先的用户信息（例如注册信息、微博等）中提取有用的信息，并用一种适当的方式表示。例如从微博文本中提取关键词，以向量的表示形式作为其后两个步骤的数据输入。

2. 文件学习器：该步骤通过获取并处理表征用户的数据，处理用户的特征信息。一般情况下，是以用户以前的历史数据，利用机器学习的方法选出一个用户喜好模型。

3. 过滤部件：通过学习用户的属性信息，查找相似的好友信息，并推荐相似的用户。最后得出用户比较感兴趣的潜在好友列表。这种计算方法是通过计算原型向量，或者通过余弦相似度计算。

2.2 好友推荐相关技术

2.2.1 文本语义分析和情感分析

文本分析的主要目标之一就是通过对文本综合分析，从文本材料中提取出影响用户的因素。分析一篇文档中所描述的主题和通过阅读文档能得到什么总结具有紧密的联系，文档的认知模型与原文档可以进行比较。基于这种考虑，用户必须通过阅读内容找到与主题相匹配的信息，摘要的信息可能是从文章中获取，但是要获取这些信息并不容易，这需要扫描原始文本的来确定位置信息。同时主题的词汇和文本中的一些词汇并不一定相同，不可能进行精确的匹配。所以必须通过语义分析技术来进行匹配和总结。

在微博的热点发现领域，通常使用词-文档来建立特征词矩阵，一般使用 TF-IDF 来选择特征词，构建 LDA、LSA 或者结合 VSM 模型等。但是微博的词汇一般是 140 个字组成，较为短小，所以当选择上述两个模型时，信息缺失严重，不能很好的选择特征词，构建的 VMS 模型[26]的矩阵维度很大，特征提取较为困难，算法的复杂度高。LSA 模型将高维的文档矩阵投射到低维的潜在语义空间，可以减少维度，便于处理大规模的数据[27]。

情感分析技术一直是心理学和行为科学领域的研究热点，因为它们是人性的一个重要因素。同时在计算机科学领域也得到了关注，特别是人机交互方面，如面部表情[28]和通过各种传感器来识别情感[29]。在计算机语言学中，自动检测文本情感从应用的观点来说越来越重要，在意见挖掘、市场分析和在线学习环境（教育和娱乐）的自然语言接口中应用较广。另外，在以下领域情感分析可以提供帮助：

1. 计算机辅助创造力，在自动个性化广告和有说服力沟通方向，对某些地方有极性的偏见或意见可以自动生成评论语句。情感分析是不可或缺的因素。

2. 人机交互上的语言表达能力，未来的人机交互强调的是自然和有效性。因此会集成很多人类认知功能的模块，包括情感分析和情感生成。例如有学习能力的宠

物的情感表达被当作它们可信度高低的关键因素。情感词的正确选择和理解在实现恰当的和富有表现力对话中至关重要。

3. 情感分析，根据情感相关性进行文本分类和舆论挖掘来做市场分析等都是这些技术应用的体现。当具有积极和消极情感的舆论同时出现在一个较为活跃的社区时，我们认为一个细粒度的情感分析可以增加一些应用的有效性。

2.2.2 基于关键词的空间向量模型

用户偏好文档和推荐项目文档都采用关键词表示表征,进而采用 TF-IDF 方法为每个特征分配权重[30]，采用 k 维向量 $\overline{\mathbf{d}_j}=\left(w_{1j},w_{2j},...,w_{kj},\right)$ 和 $\overline{\mathbf{d}_c}=\left(w_{1c},w_{2c},...,w_{kc}\right)$ 分别表示项目文档和用户 c 的偏好文档,k 是关键词的个数。关键词 k_i 在文档 d_j 中的词频 TF_i 定义为：

$$TF_{ij}=\frac{f_{ij}}{\sum_k n_{kj}} \tag{2.3}$$

关键词 k_i 在文档集中出现的逆向文件频率 IDF_i 定义为：

$$IDF_i=\log\frac{M}{n^i} \tag{2.4}$$

最终的权值为：

$$w_{ij}=TF_{ij}\cdot IDF_i=\frac{f_{ij}}{\sum_k n_{kj}}\cdot\log\frac{M}{n^i} \tag{2.5}$$

其中 M 表示文档集包含的文档数目，关键特征词 k_i 在文档集中的文档数目为 n_i，在文档 d_j 中关键字 k_i 出现的次数为 f_{ij}。

2.2.3 层次分析法

层次分析法可以解决复杂多目标下定性与定量相结合的问题，它可以作为多目标（多指标）和多方案优化决策的系统方法，对权重系数进行有效地确定[31]。主要分为两大步骤：构建判断矩阵和一致性检验计算。

1. 构建判断矩阵

在判断矩阵的构建过程中，指标的确定是以五分位标度为基础，如表 2.2 所示。

表 2.2 相对重要性的比例标尺表

A 指标相对 B 指标	极重要	很重要	重要	略重要	同等重要
A 指标取值	9	7	5	3	1

在构建矩阵**M**的过程中，A 指标对 B 指标重要性与 B 指标对 A 指标的重要性是成反比关系。即构建的矩阵是正交矩阵，其中对角线的位置为 1，其余对称的位置元素均互为倒数，如下图 2.2 所示。

$$\mathbf{M}=\begin{bmatrix} 1 & \dots & a_{1j} \\ \dots & \dots & \dots \\ a_{i1} & \dots & 1 \end{bmatrix} \quad \Rightarrow \quad \mathbf{M}=\begin{bmatrix} 1 & \dots & a_{1j} \\ \dots & \dots & \dots \\ 1/a_{1j} & \dots & 1 \end{bmatrix}$$

图 2.2 判断矩阵

2. 一致性检验

层次分析法是尽可能对主观判断进行客观性转化的一种形式化表示，构建判断矩阵是考虑其中的客观成分是否足够合理化，由于客观成分相对复杂并且人们的认识比较主观，为了判断权值是否合理，需要进行一致性检验。检验指标包括一致性指标(CI)和一致性比率(CR)，其公式如 2.6 和 2.7 所示。

$$CI=\frac{\lambda_{max}-n}{n-1} \tag{2.6}$$

$$CR=\frac{CI}{RI} \tag{2.7}$$

其中，λ_{max}是判断矩阵的最大特征根，n是比λ_{max}小的最大整数。RI是随机一致性指标，其值参照表如表 2.3 所示。

表 2.3 随机一致性指标 RI 值

n	1	2	3	4	5	6	7	8	9	10	11
RI	0	0	0.58	0.90	1.12	1.24	1.32	1.41	1.45	1.49	1.51

当一致性比率$CR<0.1$时，认为构建的判断矩阵满足条件，可以作为权重参与计算。

在用户的微博文本和情感相似性公式中，权重 wf_k 的计算由 AHP 进行计算得出。为了保证推荐的实时性，每次计算都是以用户的 5 条微博为基准，构建的判断矩阵如表 2.4 所示。

表 2.4 微博文本的判断矩阵

微博文本	第一条	第二条	第三条	第四条	第五条
第一条	1	3	5	7	9
第二条	1/3	1	3	5	7
第三条	1/5	1/3	1	3	5
第四条	1/7	1/5	1/3	1	3
第五条	1/9	1/7	1/5	1/3	1

为了计算权重，我们使用规范列平均法进行求解，首先对判断矩阵 **M** 每一列归一化得到矩阵 **N**，再对矩阵 **N** 的每一元素进行相加求和并求平均值。最后得出一个一列 5 行的向量 **C**，该向量 **C** 即所求权重向量，代表计算文本相似性和情感相似性时对每一项分配的权值。通过公式(8)可以计算得出 λ_{max}=5.2375，$CI=\frac{\lambda_{max}-n}{n-1}$=0.0593，$RI=1.12$，则 $CR=\frac{CI}{RI}=\frac{0.0593}{1.12}$=0.053<0.1，符合一致性条件。计算的权重结果如表 2.5 所示。

表 2.5 实验权重的设置

wf_1	wf_2	wf_3	wf_4	wf_5
0.543	0.226	0.109	0.065	0.057

2.3 推荐系统评价指标

推荐系统中，评价推荐方法的好坏是至关重要的，评价指标可以用来衡量推荐算法的综合性能，但有效的评测推荐算法的好坏是无法定论的。在一些让用户打分的提供推荐服务的网站中，可以通过打分来获取用户的兴趣模型，预测用户对其它物品的兴趣程度，这种行为统称为评分预测。在评分预测系统中一般通过均方根误差(RMSE)和平均绝对误差(MAE)计算[15,18]。

对于数据集T中的用户u和物品i，用户u对物品i的真实评分为r_{ui}，而$\hat{r}_{ui}$是通过推荐算法计算出的预测评分，那么 RMSE 的定义如公式 2.8：

$$RMSE=\sqrt{\frac{\sum_{u,i\in T}\left(r_{ui}-\hat{r}_{ui}\right)^2}{|T|}} \tag{2.8}$$

MAE 是通过绝对值的计算来预测误差，定义如公式 2.9：

$$MAE=\frac{\sum_{u,i\in T}\left(r_{ui}-\hat{r}_{ui}\right)}{|T|} \tag{2.9}$$

在好友推荐领域常用的指标一般是采用 TOP-K 推荐中的准确率、召回率和 F-measure 值，同时为了考虑推荐结果中的用户顺序，加入了新的评价指标 P@N。在文献[32-35]中，首先把准确率和召回率的评价方法加入了推荐系统里面。当我们对用户进行好友推荐时，会出现四种情况，即推荐给用户且用户满意度高，推荐给用户但用户不满意，没有推荐用户感兴趣的用户，没有推荐用户不感兴趣的用户。这 4 种可能性情况如表 2.6 所示。

表 2.6 待推荐好友的 4 种可能情况

用户满意度	推荐	不推荐
满意	True-Hit(N_{th})	False-Hit(N_{fh})
不满意	True-Not-Hit(N_{tnh})	False-Not-Hit(N_{fnh})
总数量	$N_{th}+N_{tnh}$	$N_{fh}+N_{fnh}$

$N_{fh}+N_{fnh}$

在实验中，获取用户的现有好友列表 L，当推荐的好友存在于现有好友列表 L，则表明是用户满意的好友，反之亦然。根据表 2.6 可以得出用户现在的好友个数为$N_{th}+N_{fh}$，准确率可以表示为在不考虑物理系统的限制条件下，能够正确推荐出的好友个数与推荐好友总数的比值，计算方法如公式(2.10)：

$$Precision(u)=\frac{N_{th}}{N_{th}+N_{tnh}} \tag{2.10}$$

其中N_{th}表明推荐好友是用户真实好友的个数，$N_{th}+N_{tnh}$是总共推荐的好友数量。

召回率即查全率，可表示为在不考虑物理系统的限制条件下，能够正确推荐出的好友个数与用户好友总数的比值[31]，计算方法如公式 2.11：

$$Recall(u) = \frac{N_{th}}{N_{th} + N_{fh}} \tag{2.11}$$

其中 N_{th} 含义同上，$N_{th} + N_{fh}$ 是用户的好友总数量。

很多推荐系统里面存在推荐评分机制，由于评分稀疏性和推荐列表的长度等因素，很多研究人员不建议使用准确率和召回率单独的对推荐系统进行评价，特别是单一考虑一种指标。所以需要一种平衡准确率和召回率的一种评测方法，使用包含准确率和召回率的二维向量来反应系统的表现。文献[36]使用了 F-measure 指标，计算如公式 2.12：

$$F - measure(u) = \frac{2 \times Precision(u) \times Recall(u)}{Precision(u) + Recall(u)} \tag{2.12}$$

P@N 本身是 Precision at Position 的简称，经常用在搜索引擎算法的评测中，指的是在推荐列表中，推荐的前后顺序在推荐系统里的影响，排在最靠前的用户决定了用户的满意度，计算公式如 2.13：

$$P@N = \frac{T_N \bigcap R_N}{N} \tag{2.13}$$

其中，T 是根据抓取的前后顺序生成的用户列表。T_N 表示前 N 个好友。R_N 表示给用户推荐的前 N 个好友。

在推荐系统中，新颖性和惊喜度有时也作为一项推荐指标。给用户进行推荐的目的，就是为了提高用户的体验，增加用户的黏性。当结果呈献给用户之后，在准确率高的情况下，还必须要考虑用户的体验以及推荐的新颖性。当一个用户对电影感兴趣的时候，只推荐单个明星，用户的准确率可能很高，但是这种结果不新颖，因此还需要考虑多样性的结果。新颖性就是推荐用户没听过的用户，或者关注点相悖的用户。给用户推荐很准确的用户或物品，在一定程度上，用户可能已经获取到了该用户的信息或物品信息。这样的信息对用户而言就没有了价值，这样用户满意度就会很低。在考虑新颖性的同时，可以通过用户和物品的流行度或知名度来进行判断。如果用户的知名度很高，可能这些用户的推荐价值就变得很低，属于大众知名人物。为了减少准确度的不足，可以采用新颖性和多样性来综合平衡。

2.4 本章小结

本章首先介绍了通用的推荐系统的流程和架构，对文本的语义分析和情感分析做了阐述，并对常见的好友推荐算法的思想进行了总结和优缺点对比，然后对推荐系统的评价指标进行描述和说明。以上介绍为后续开展研究做理论基础。

第3章 基于用户文本语义和情感程度的好友推荐

3.1 微博内容研究

微博是最近几年发展起来的一种在线社交网络平台，发展态势相比传统社会媒体态势强劲。社交网络的良好发展离不开广大用户，而增加用户与社交网站之间的黏性则需要好友推荐功能来扩大用户的交际圈。常用的推荐算法都是采用好友的关系进行好友推荐，没有过多考虑用户的文本信息。在微博中，虽然用户只能发布短文本信息，内容比较短小，但信息量却很大，基本上都是比较精简和重要的词汇。这些短文本消息是用户描述近况、发表评论或抒发情感的一种手段。而且随着移动设备的流行,用户发表观点和心情的方式更加快捷,人与人之间的社交距离在缩短。

通过对大众微博内容的分析研究发现，微博短文本的数据规模很大，同时包含高价值的隐含信息。其中最重要的一点就是微博文本长度短小，信息含量较小，但数据量大，给短文本的表示造成了严重的稀疏现象和特征空间高维型的特点[37]。同时微博文本具有实时更新性和动态变化的特点，用户在网络的交互过程中会产生大量的短文本信息，使预处理过程变得复杂。微博短文本具有不规范性，噪音数据偏多。但是不可否认，微博短文本中蕴含的丰富社交内容具有较高的研究价值。

3.2 用户微博文本语义分析和情感分析

在线社交网络中的用户一般对注册信息的一些资料很少有改动，所以单纯地利用用户的注册信息只能反映用户的过时情况，这样的推荐就会造成偏差。在现实生活中，用户的兴趣和爱好都会随着时间的变化而产生波动，因此关注用户最近的动态越能反映用户的关注点。以 Facebook，Tweeter 和微博为代表的在线社交网络，用户可以在社交网站中发表动态、转发和评论信息，这些微博文本信息能实时反映用户兴趣所在和情感性格特征。

用户的文本语义分析的前提条件是对用户的微博文本进行处理。从网站上爬取下来的文本杂乱无章，包含各种脏信息和特殊符号等。根据抓取的文本特点，必须

对文本进行分词。本文使用的分词系统是 ICTCLAS（中科院分词系统）。分词之后，还需要对数据进行进一步处理。预处理过程一般需要以下三个步骤：

1. 去除文本中的特殊词和特殊符号，比如“的，得，#，@”等词和符号。

2. 去除超链接，爬取的内容中这些 URL 可以对微博内容的说明起辅助作用，由于本文研究是文本内容，所以不考虑超链接带来的影响。

3. 删除无用词，中文词语分为动词、形容词和名词等词性，而在对文本的处理中，只保留名词、动词、时间词、方位词和地点词。这些词性具有明显的特征，可以简要的对叙述的事情进行概括。

对文本预处理的流程图如图 3.1 所示。

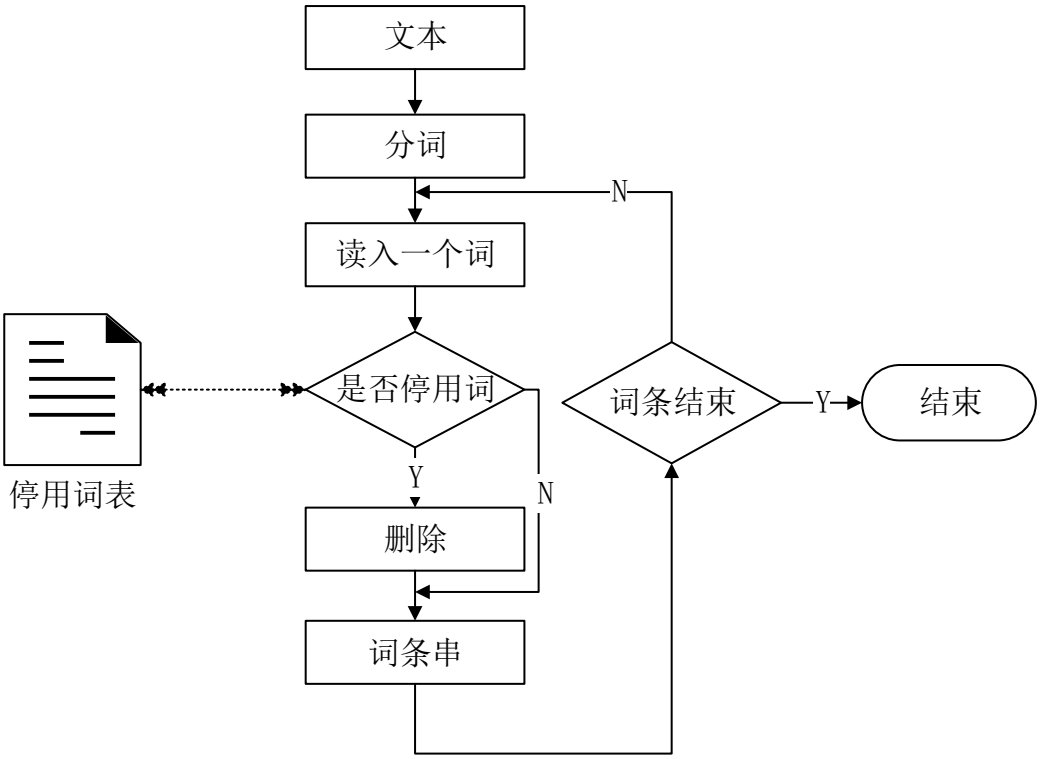


图 3.1 预处理流程图

对微博文本进行预处理之后，剩下的词汇基本可以作为用户的特征标签，但是在中文词语里面往往含有很多同义词、反义词以及词语相近等容易混淆的词汇，特别是存在“一词多义”和“一义多词”的现象。所以本文基于现有的《同义词词林》语义知识词典对用户的文本进行语义分析。在同义词词林中，每个有同义词的词汇都会被简化为各种词汇编号，对于相同意义的词汇具有同样的编号。对提取的关键

词进行相似性比较之前，先对词汇做同义词的转换处理。表 3.1 列举了同义词词林中的词语表示方法。

表 3.1 同义词词林词语表示方法

词语	编号	词语	编号
字迹	Bg09B07	喊声	Bg07A39
笔记	Bg09B07	吼声	Bg07A39
油渍	Bg09B06	医师	Ae15A01

在《心理学大辞典》中，情感被描述为一种是否满足自己的需要而产生的态度体验。文献[38]证明在在线社交网络中不同的用户在通过微博进行交互时，能相互影响情绪。所以，在好友推荐时，是否满足用户的需求在根本上就是用户的情感分析问题。

情感是人类互动的核心和人类社交中的一个重要因素，一直是心理学和行为科学研究的热点，深入理解用户的情感活动可以帮助我们更好的了解用户的需求。在社交媒体中，用户的性格特征决定了他会在不同的事情和问题中具有不同的情感反应，从而对其行为也会产生一定的影响，所以对用户的情感进行深入分析才能为其找到志趣相投的朋友。用户的微博信息不仅表达了他对不同问题的看法，同时在文本中也包含了他们的情感信息[39]，这些都有助于对用户情感进行分析。具有相似情感的人具有一定的黏合力，对用户的情感分析已经被用在产品推荐里面，而且具有显著的效果[40]。情感分析的首要任务是提取用户在文本中的情感词汇作为用户的情感特征。

3.3 好友推荐算法描述

3.3.1 基于用户文本的语义分析

在对用户的文本进行语义分析时，需要对用户的微博内容进行预处理。用户发表的微博中通常包含一些干扰词汇和符号，在计算用户的文本相似度时，首先需要对用户U_i和U_j的微博内容进行向量化描述：

$$\mathbf{mb}_i = \{wb_{i1} = (T_{11}, T_{12}, ..., T_{1n}), ..., wb_{ik} = (T_{k1}, T_{k2}, ..., T_{kn})\}$$

$$\mathbf{mb}_j = \{wb_{j1} = (T_{11}, T_{12}, ..., T_{1n}), ..., wb_{jk} = (T_{k1}, T_{k2}, ..., T_{kn})\}$$

其中 wb_{ik} 表示用户 i 的第 k 天的微博特征集合，T_{kn} 表示用户第 k 天的第 n 个关键词。

用户 U_i 和 U_j 的文本相似度的交集记作 T_{com}，$T_{com}=T_{wb_{ik}}\bigcap T_{wb_{jk}}$，选取文本时，必须选取相同时间的文本进行分析比较，这样在进行相似度计算时，才具有可比性。当 T_{com} 为空时，表示两个用户的文本不存在相似性，即文本相似度为 0，则用户相似性计算如公式 3.1 所示。

$$sim\left(U_i,U_j\right)=\begin{cases}\sum_{n=1}^{k} sim\left(mb_{in},mb_{jn}\right), T_{com}\neq\varnothing \\ 0 \qquad\qquad , T_{com}=\varnothing\end{cases} \tag{3.1}$$

其中 mb_{in} 和 mb_{jn} 分别表示用户 U_i 和 U_j 的第 n 天的文本内容。在文本相似性计算中采用 Jaccard 距离[5]，如公式 3.2 所示，

$$sim(mb_i,mb_j)=\frac{\left|N(wb_{in})\bigcap N(wb_{jn})\right|}{\left|N(wb_{in})\bigcup N(wb_{jn})\right|} \tag{3.2}$$

其中 $N\left(mb_{in}\right)$ 和 $N\left(mb_{jn}\right)$ 表示用户 U_i 和 U_j 的第 n 天微博中的特征词的集合。

3.3.2 基于用户的情感程度分析

情感相似度计算的主要任务是提取出文本中用户所产生的情感词汇。其中情感词的提取主要是基于语料库和词典的两个方法[41]。由于本文针对的对象都是中文词汇，所以采取的是基于词典的抽取方法。通常用户会通过文本表达对某种事件的意见和态度，根据微博的这一特点，课题组自定义了一个程度副词词典，对常见的程度情感词进行统计和标记，其中包含常见的程度副词 19 个，并按照程度的轻重顺序进行排列。在词典里每一个程度情感词都被单独设定下标，在比对两个程度副词相似性时，可以根据下标的距离来判定相似性。若微博中含有多个情感词，在提取时我们选择下标最大的关键词，这样能比较好地反映用户情绪。与文本相似度计算类似，情感词也考虑了时间的因素，所以同样根据时间的先后，分配权重，计算方法类似于文本相似度的权重计算。其中情感词词典由以下结构组成：$Emotion_{dict}=\{em_1,em_2,...,em_i\}$，则两个用户的情感相似性度计算方法如下：

$$Sim(U_i,U_j)=\sum_{k=1}^{m}\left(wf_k\times\frac{1}{1+\alpha\left|E_i-E_j\right|}\right) \tag{3.3}$$

其中α是衰减参数，当两个情感程度词的下标差值越大，则表示两个越不相似，反之亦然。若差值为 0，则相似度为 1，同时引入参数α对取值进行分析。在得出最后的好友推荐列表时，需要对列表进行过滤，以保证推荐的质量。

3.3.3 融合时间因素的好友推荐模型

社交网络中，时间因素可以被当作一种重要的上下文信息，用户的兴趣点和关注方向会随着时间的变化而产生一定的变动，所以必须考虑时间带来的影响。用户的兴趣是变化的，可以分为自身内部原因和外部原因，年龄、居住地和外部条件的变化都会对用户的兴趣产生影响。比如，用户在一年前对某一个方向a有兴趣，但在最近几个月又对另一个方向b感兴趣，而上个星期又对c产生了兴趣，那么，我们可以认为当前用户更希望和同样对c感兴趣的用户成为朋友。当我们对用户的文本内容进行分析时，距离当前时间越近的文本内容，对用户的表征就越强。所以计算用户之间的文本相似性时，需要对用户近期的文本内容分配更高的权重因子，给用户推荐近期具有相似行为的用户。

由于用户的文本内容存在时间先后的特性，不同时间的文本内容对用户的表征能力不同，距离当前时间越近，表征能力越强。为减少时间变化带来的偏差，本文在对用户的文本内容计算相似性时添加了相应的权重值以反映时间变化的影响。在经济学中的层次分析法(AHP)可以用来解决多因素下的决策问题，所以本文利用层次分析法对每个因素（本方法中分别是微博文本分析中产生的不同词汇和微博文本产生的时间）进行权重的计算。

首先将用户U_i和用户U_j的微博文本表示为$wb_i=(T_1,T_2,...,T_i)$和$wb_j=(T_1,T_2,...,T_j)$，假设要给用户U_i进行好友推荐，首先对所有用户最近的微博文本进行去除停留词以及提取关键词等处理。同时考虑时间特性，选取相同时间段的用户微博，这样能够保证在进行文本相似性计算时，对应比较的用户微博文本所产生的日期尽可能相同或接近的。通常微博的发布距离当前时间越近，越能反映用户的情感特征，因此在文本相似性计算中被分配的权重越大，采用公式如公式 3.4：

$$sim(U_i,U_j)=sim(mb_i,mb_j)=\sum_{k=1}^{m}\left(wf_k\times Sim\left(T_{ik},T_{jk}\right)\right) \tag{3.4}$$

文本语义和情感程度的推荐（SEM）模型框架如下图 3.2 所示，

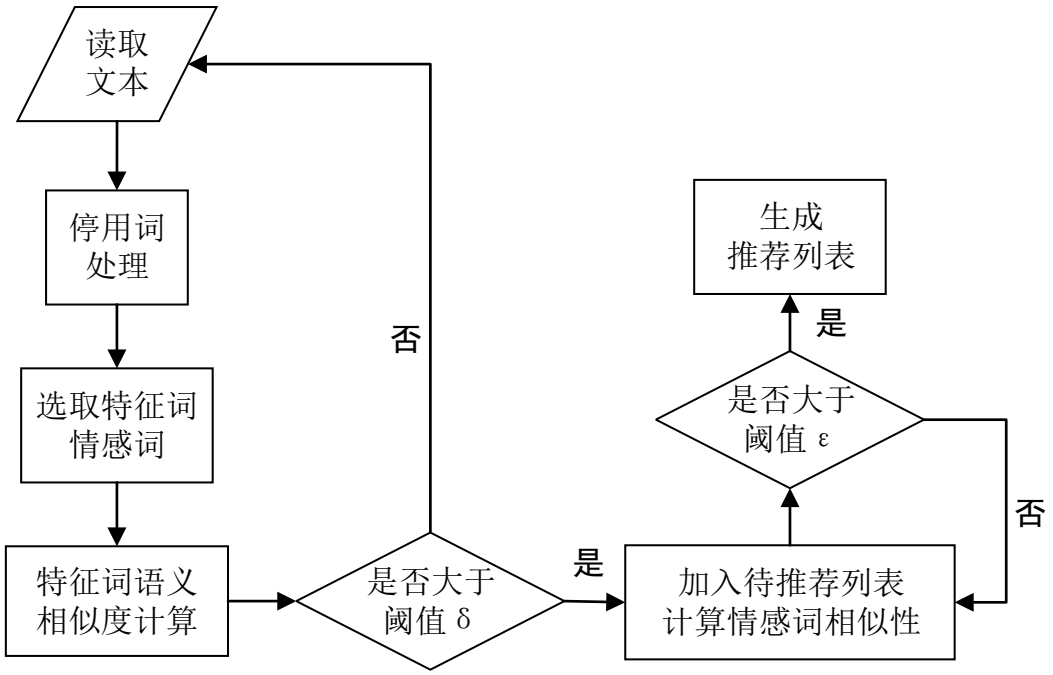


图 3.2 文本语义和情感程度的推荐(SEM)模型框架

其中 wf_k 是表示对用户的第 k 条文本信息分配的权重， m 代表选取文本的数量。

$\sum_{k=1}^{m} wf_k = 1$，且随着 k 的增大，wf_k 的值越大，表示距离当前时间越近，对用户的表征能力越强。通过公式 3.2 计算出的两个用户的文本相似性，可以得出第一个待推荐用户好友的列表。由于不同的用户在表述事情存在不同情感的差异，所以需要继续对用户的情感进行分析，在微博文本中经常包含了一些情感词汇，我们根据自定义的情感词典，提取出用户的情感词，然后计算情感相似性，从而过滤第一个推荐列表。

融合文本语义和情感程度的好友推荐算法步骤如下所示：

算法 1. 基于用户文本语义和情感程度的好友推荐算法

输入：用户 U 的 ID

输出：推荐列表 Rm

步骤：

1. 初始化 $Rmlist=\varnothing$， $Rm=\varnothing$，用户集合 $U_{list}=\{U_1,U_2,...,U_n\}$；

2. 提取用户 U 的文本特征词和情感特征词 $\{(T_1,E_1),(T_2,E_2),...,(T_5,E_5)\}$；

3. 分别提取用户集合中用户的文本特征词和情感征词：$\{(T_{i1},E_{i1}),(T_{i2},E_{i2}),...,(T_{i5},E_{i5})\}$；

4. FOR $U_i \in U_{list}$　　//遍历用户列表

5. $Sim(U,U_i)=\sum_{k=1}^{m}\left(wf_k \times Sim(T_k,T_{jk})\right)$

6. IF $Sim(U,U_i)>\delta$

7. 将U_i添加到$Rmlist$中

8. END IF

9. END FOR

10. FOR $U_i \in RmList$　　//遍历待推荐用户列表

11. $Sim(U,U_i)=\sum_{k=1}^{m}\left(wf_k \times \frac{1}{1+\alpha|E-E_i|}\right)$

12. IF $Sim(U,U_i)>\varepsilon$

13. 把U_i添加到Rm中

14. END IF

15. END FOR

16. 返回 Rm

3.4 微博数据采集

在推荐系统中，需要分析用户的兴趣和爱好，更好的得出精确的推荐结果。分析的前提是数据。而数据采集是推荐系统中的第一步也是最重要的一步，只有采集到数据才能进行后续的分析和推荐。数据采集的过程又可以称为网络爬虫，根据设定的规则对微博数据进行爬虫。本文选择新浪微博数据作为数据源，由于新浪微博

的公共 API 对数据抓取做了很大的限制，所以如何使用合理的决策并采集对我们有用的数据是至关重要的。

新浪微博的数据量较为庞大，采用合适的采集策略能够帮助我们快速高效的获取数据。数据采集模块的主要工作是从获取数据并保存到本地，整个采集过程的要点就是：首先，确定采集策略，在爬取部分数据的前提下，能够保证数据的合理性；其次，根据研究的内容，选取合适时间段的数据；最后为了以后的推荐，需要对数据进行相应的保存策略，选择并建立合适的数据库。

3.4.1 数据获取方式

目前，获取新浪微博数据的方式有两种，爬取分析网页和调用官方公开的微博 API。其中第一种方法爬取分析网页，为了爬取更多的数据信息，需要编写模拟登陆模块，否则只能爬取用户首页的数据。而调用官方的微博 API 接口可以方便获取用户信息，新浪公开了大部门的接口方便用户进行研究，但同样为了保护用户的隐私和信息安全，针对开放的接口也做了很多限制。

1. 调用官方微博 API

新浪微博逐渐成为国内最为活跃的在线社交网站，平台使用 Oauth 的授权方法公开了大部分的接口。用户通过编写程序调用公开的 API 就可以直接获取新浪服务器上的数据，但是在公开的 API 中也进行了等级划分来保证数据的安全性并对获取数据的数量进行了限制。接口的开放状态根据用户的需求随时进行变更，常见的接口类别有位置服务接口、支付接口、粉丝服务接口和地理信息接口。下图 3.3 展示了新浪微博常用的一些接口类别：

图 3.3 新浪微博常用接口类型

每个接口类型又有详细的划分，例如微博接口类型又具体分为获取用户发布的微博、获取用户发布的微博的 ID 等，如图 3.4 展示了微博 API 的接口信息。

用户		
读取接口	users/show	获取用户信息
	users/domain_show	通过个性域名获取用户信息
	users/counts	批量获取用户的粉丝数、关注数、微博数

关系		
关注读取接口	friendships/friends	获取用户的关注列表
	friendships/friends/in_common	获取共同关注人列表
	friendships/friends/bilateral	获取双向关注列表
	friendships/friends/bilateral/ids	获取双向关注UID列表
	friendships/friends/ids	获取用户关注对象UID列表
粉丝读取接口	friendships/followers	获取用户粉丝列表
	friendships/followers/ids	获取用户粉丝UID列表
	friendships/followers/active	获取用户优质粉丝列表
关系链读取接口	friendships/friends_chain/followers	获取我的关注人中关注了指定用户的人
关系读取接口	friendships/show	获取两个用户关系的详细情况
写入接口	friendships/create	关注某用户
	friendships/destroy	取消关注某用户

图 3.4 微博 API 详细接口

使用微博公开的API获取的数据是经过数据处理之后的JSON文件，方便使用。同时，新浪微博使用登录授权和频次限制的机制来规范用户对微博数据的不正当获取，限制每段时间只能请求一定的次数，限制的维度有单授权用户和单 IP。对抓取的要求越来越高，获取的数据也变得越来越有限。

2. 爬取分析网页

通过爬虫获取信息相对于直接调用 API 的方式复杂很多，但限制很少，用户可以根据自己的需求进行定制爬虫。在设计爬虫时，最主要的步骤一般包括以下几部分：

(1) 模拟登陆网站，保存输入的用户名和密码到 cookie 中；

(2) 自定义要爬取的地址放入待抓取队列；

(3) 从待抓取队列中取出待抓取的链接地址，进行域名解析，并将链接地址所对应的网页下载下来；

(4) 通过设定规则，分析链接地址下载的网页内容，提取信息；

(5) 把新获取的 URL 放入待抓取队列，并重复步骤(3)；

通用的网络爬虫的框架如图 3.5 所示：

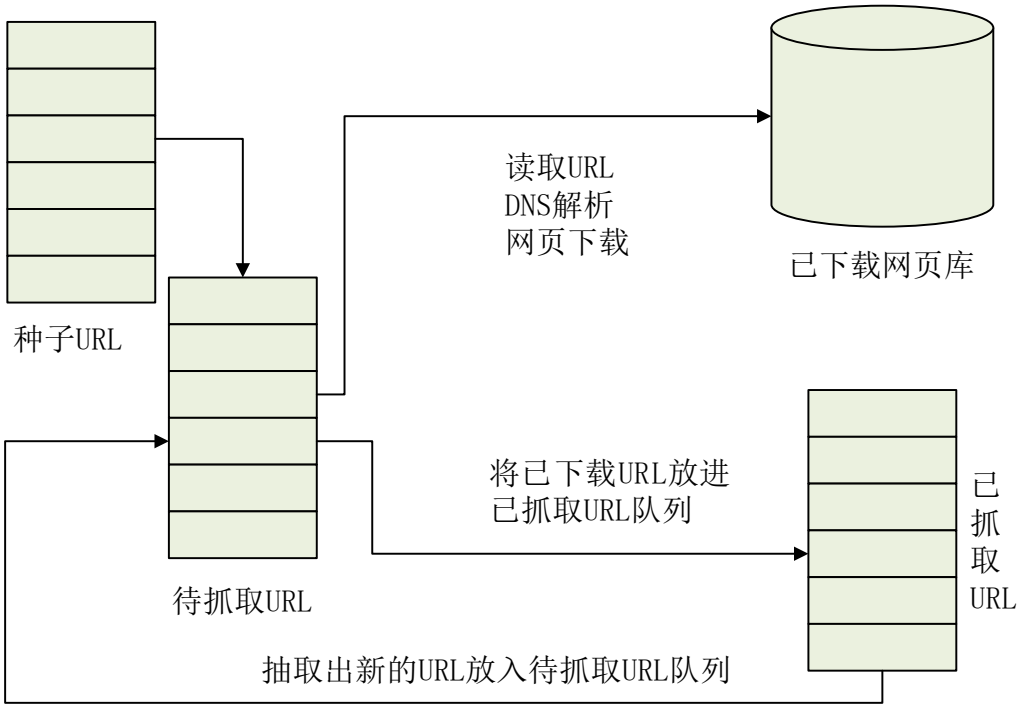


图 3.5 通用的网络爬虫框架

3.4.2 数据抓取

新浪公开的 API 的优点是接口丰富，同时可以根据自己特定的需求选择对应的接口，数据是以 JSON 的格式保存，后续不需要处理，较为方便。但是缺点是需要授权，爬取数据频次也有限制。考虑到数据的完整性，经过衡量，本文采用爬取网页的方式抓取数据，其关键是模拟登陆新浪微博，然后下载 WAP 端的 HTML 源代码，通过正则表达式来分析和提取自己需要的数据，保存到 MySQL 数据库中。

由于目前 PC 服务器配置较高，可以采用多线程的方式进行数据爬虫。在爬取好友关系和微博数据时分为两步，首先好友关系是通过“滚雪球”式的爬取方法，首

先选定目标用户 A，根据爬虫规则获取用户 A 的关注列表 B，再依次爬取列表 B 中的关注列表，依次进行。为了保证数据的有效性，只对 3 度的好友关系进行爬取，得到关系数据之后，再依次爬取用户最近的微博数据。用户的好友列表和用户的微博数据如图 3.6 和图 3.7 所示。

suid	tuid
1043325954	1035997554
1043325954	1039916297
1043325954	1047288737
1043325954	1084765950
1043325954	1093311041
1043325954	1182389073
1043325954	1182391231
1043325954	1187986757
1043325954	1188552450
1043325954	1191808911
1043325954	1193491727

图 3.6 用户好友列表

userid	date	content
1043325954	201601031802	人物男北京加关注
1043325954	201601031118	认证：《人物》杂志官方微博
1043325954	201601031100	《人物》创刊于1980年，人民出版社主办，致力于提供中文世界最...
1043325954	201601021719	Bravo年度盘点回归现实感的出版业今年的纽约伦敦法兰克福三大书
1043325954	201601021044	#YellowedPortrait#加入橄榄球队的少年卷福。
1043325954	201601021003	问地球35人张震：真正兴奋的角色是坏坏的好人2015年是演员张震的
1043325954	201601011706	Bravo年度盘点艺术是一种耻辱2015年中国话剧舞台上的外国戏最精
1043325954	201601010806	问地球35人谢念祖《康熙来了》落幕，象征一个时代的过去2015年，
1043325954	20151231195914	#YellowedPortrait#摄影师镜头下上个世纪五六十年代的巴黎，属于
1043325954	20151231151230	Bravo年度盘点古典音乐界发生了哪些事？《人物》新年的第一篇长文
1043325954	20151231113950	#YellowedPortrait#有些事只适合收藏它们不能变成语言，它们无法
1043325954	20151230212237	年度科学家颜宁：在迷雾旷野中寻找真理之路颜宁带领着她平均年龄
1043325954	20151230192422	年度作家刘慈欣：投向广阔宇宙的最后目光刘慈欣成为第一位获得雨果
1043325954	20151230141106	水墨中国奥地利摄影师JosefHoflehner偏好将照片进行黑白处理或者
1043325954	20151230112758	#YellowedPortrait#杰奎琳卡罗琳和花。

图 3.7 微博文本数据

3.4.3 微博用户分析

在微博网络中，由于存在很多僵尸粉和组织机构账户，这样的用户虽然拥有很多的粉丝，但是数据集中若含有大量的该类用户会对实验结果产生一定的影响，为了保证数据集的合理性和可用性，并对数据的真实可用性进行长尾分布验证，分别从用户的入度和出度两个角度对微博用户的群体行为进行分析验证。

1. 微博用户的入度分布

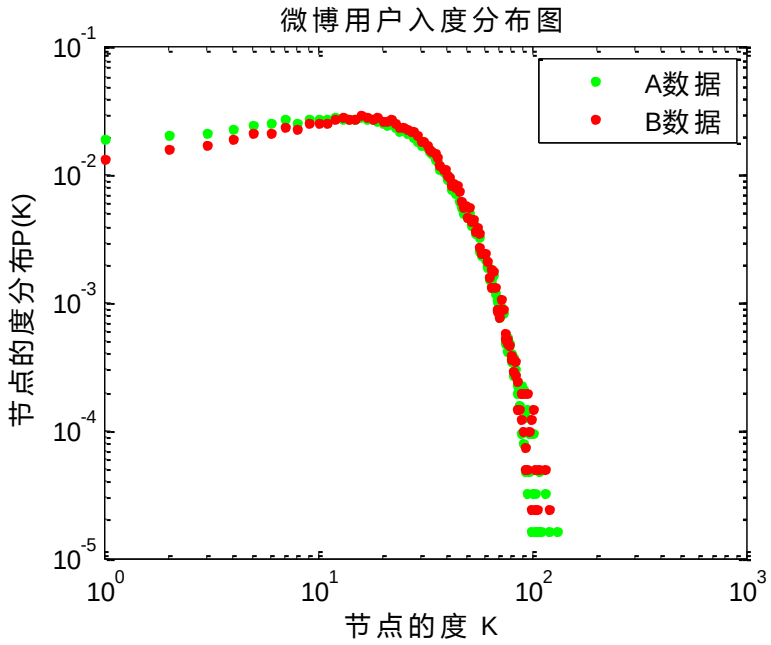


图 3.8 微博用户的入度分布图

微博用户的入度分布情况如图 3.8 所示，其中横坐标节点的度 K 表示微博用户的粉丝数量，纵轴 P(K)表示网络中度为k的节点在整个网络中所占的比例。从图 3.8 可知用户的入度分布符合复杂网络中节点度的幂律分布特点[42]，说明该数据集的网络是无标度的。符合复杂网络的特点，呈现出长尾现象。说明在微博网络中受到关注较多的微博用户只占少部分，而受到较少关注的用户占大部分。

2. 微博用户的出度分布

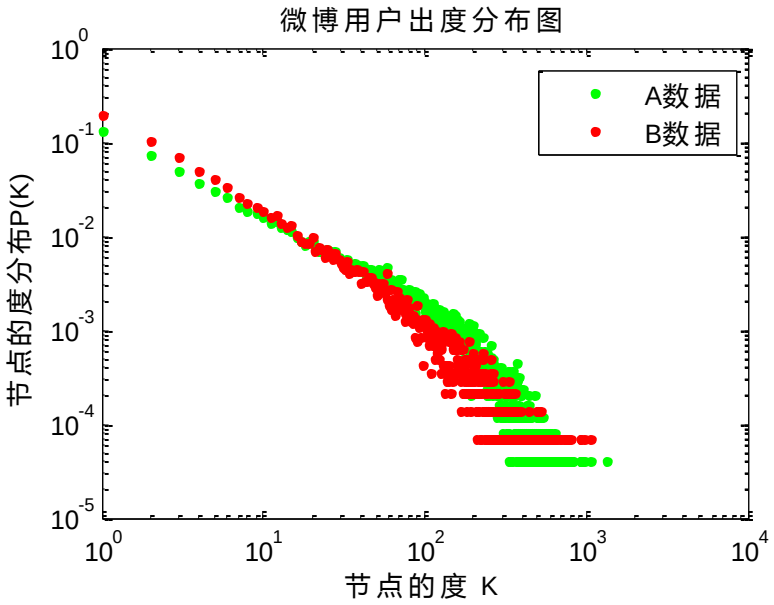


图 3.9 微博用户出度分布图

在新浪微博中，官方规定每个用户的最大关注量是3000个，根据图3.9所示，微博用户的出度均没有超过3000；此微博用户的出度分布符合实际的，并且符合幂律分布的特点，存在明显的长尾现象，表明较多的微博用户关注其他微博用户的数量较少，较少的微博用户关注其他微博用户的数量较大。

3.5 实验及性能分析

3.5.1 实验数据

实验数据的获取是通过自行编写的爬虫程序进行抓取，在爬取数据时我们采取“滚雪球”式的爬取策略，随机选择单个用户，爬取此用户的好友信息，根据好友的 ID 再依次爬取这些好友的好友微博等信息，然后重复此过程。总共分别爬取两组新浪微博数据，A 组数据是从 2014 年 5 月 3 日至 2014 年 5 月 17 日的数据，B 组数据是 2013 年 11 月 13 日至 2013 年 11 月 22 日的数据，数据中包含了用户的注册信息，微博数据信息和好友关系。数据统计信息如表 3.2 所示。

表 3.2 爬取数据集的统计信息

数据特征	A 组数量	B 组数量
用户数	63641	6038
微博数	84168	10569
用户关系	1391718	80692

3.5.2 评测指标

在第二章已经提到，一个推荐算法的好坏需要全方面的进行评价，用户和社交网站的利益都要考虑，对于社交网站来说，用户的体验尤为重要。现有的推荐算法可能在某些情况下的准确率较高，但是存在推荐不新颖，推荐的次序不一，考虑的因素不够全面，用户的满意度偏低。为了全面的评价算法的推荐质量，本章节以及下一个章节都会对各种算法进行全面的评价。在预测准确度时，本文采用 Precision、Recall 来进行评价，再使用 F-measure 对结果综合评估，弥补了 TOP-N 推荐的问题。为了考虑推荐结果中用户先后顺序给用户带来的影响，最后加入了 P@N 评价方法，可以准确的预测用户的满意度。

3.5.3 实验分析

实验环境中采用的硬件环境为 Inter(R) Xeon(R) CPU，4.00GHz 主频，16G 内存，操作系统为 windows Server 2003，Centos 7.0，软件环境为 Python、Matlab、Eclipse、MySql、Shell 等。

各个好友之间的推荐算法对比，采用的数据集分为两组，A 组数据和 B 组数据，对比的推荐算法描述如下：

1. FOF+Tag，此方法是根据用户的共同好友进行标签相似性匹配[8]。
2. FOF+BP，这种方法对共同的好友结合注册信息中的出生地进行推荐[44]。
3. UMFR(Unified Microblog Friend Recommendation)，此方法中结合了标签，位置、签到、和热点话题的信息进行推荐[45]。

在对用户的情感进行分析时，利用公式 3.3 来进行计算。在实验之前需要对α值进行取值分析，通过 F-measure 值的变化来选取α的最佳值。如图所示。

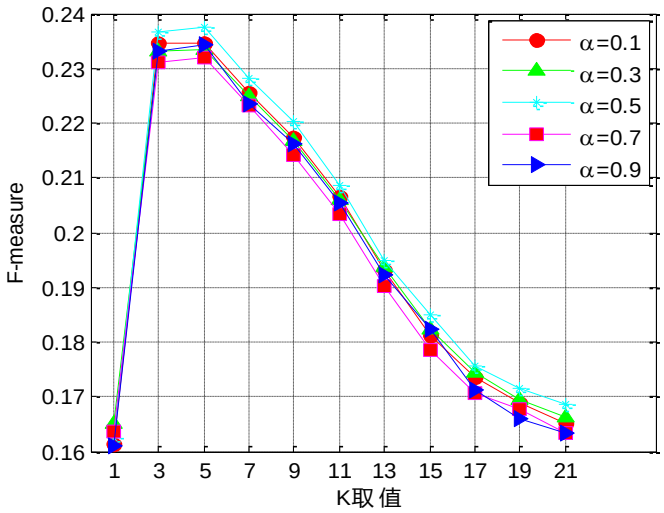


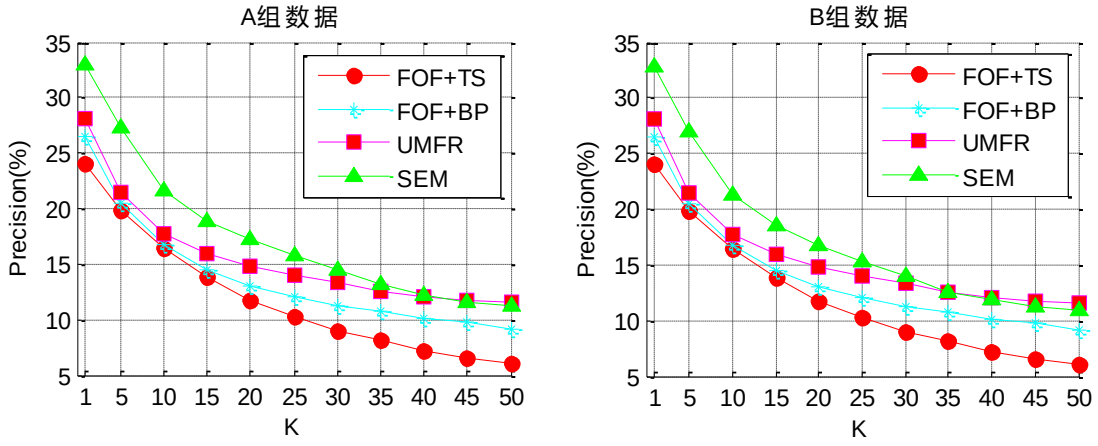
图 3.10 不同α下的 F-measure 值变化情况

从图中可以得出选取不同α值时，F-measure 值的变化都是先增大，达到顶峰之后逐渐减小，而在α=0.5时，好友推荐结果 F-measure 值均高于其它值。所以在好友情感程度分析时，对公式 3.3 进行修正，修正后的公式如 3.5。

$$Sim(U_i,U_j)=\sum_{k=1}^{m}\left(wf_k\times\frac{1}{1+0.5\times\left|E_i-E_j\right|}\right) \tag{3.5}$$

1. 推荐算法在 A 和 B 数据集上的准确率对比效果

融合时间因素之后，首先对数据集 A 和数据集 B 进行准确率的对比，在准确率的线形图中，如图 3.11，绿色的线条表示基于文本语义和情感程度的好友推荐算法(SEM)，其它算法如图中标签所示。

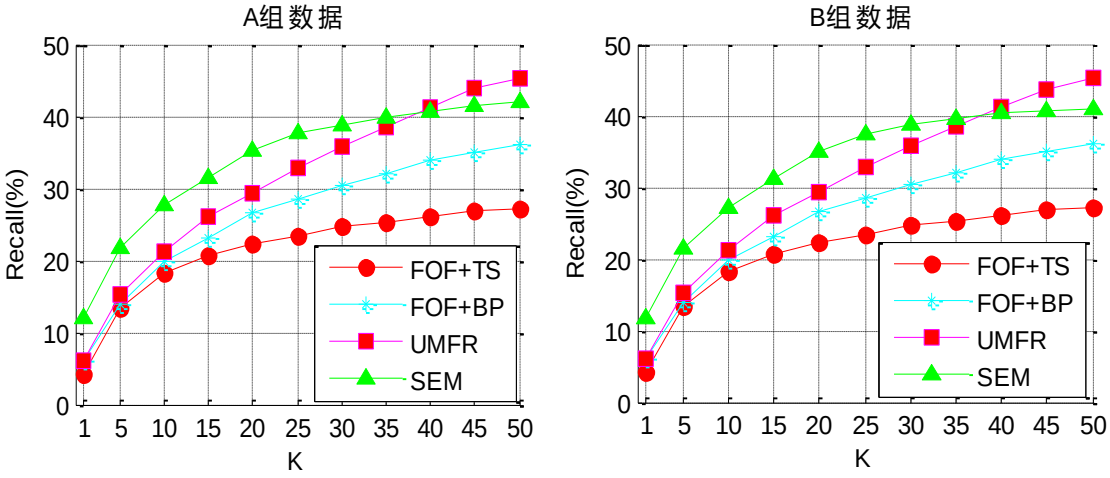


(a) A 组数据　　(b) B 组数据

图 3.11 准确率在不同数据上的对比图

从图 3.11 中可以得出以下结论：在对用户的准确率(Precision)分析时，不管是传统的好友推荐算法还是本文提出的融合时间的文本语义和情感程度的算法，随着推荐列表的增长，准确率反而降低，可以推断好友推荐列表的长度和准确率成反比。在推荐列表长度小于 10 时，准确率的值较高，符合推荐的特征。

2. 推荐算法在 A 和 B 数据集上的召回率对比效果

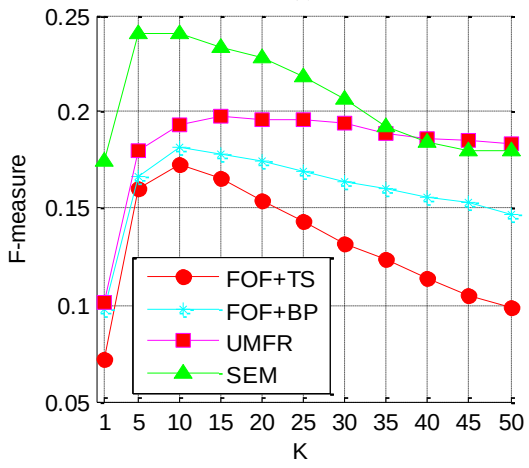


(a) A 组数据　　(b) B 组数据

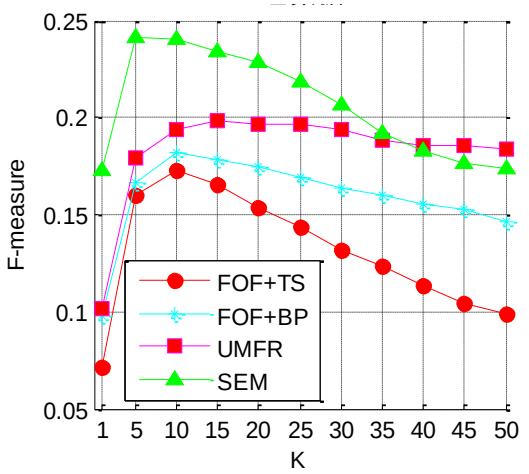
图 3.12 召回率在不同数据上的对比图

从图 3.12 可以得出，本文提到的算法和传统的算法的推荐效果和推荐长度成正比，和准确率的变化成反比。同时在第二章中谈到的准确率和召回率的关系是成反比关系得到验证。在推荐长度为 45 的时候达到最大值，增长趋于平稳。由于这两种评价指标是互为矛盾，则又对平衡这两种指标的 F-measure 值进行评价。

3. 推荐算法在 A 和 B 数据集上的 F-measure 值对比效果



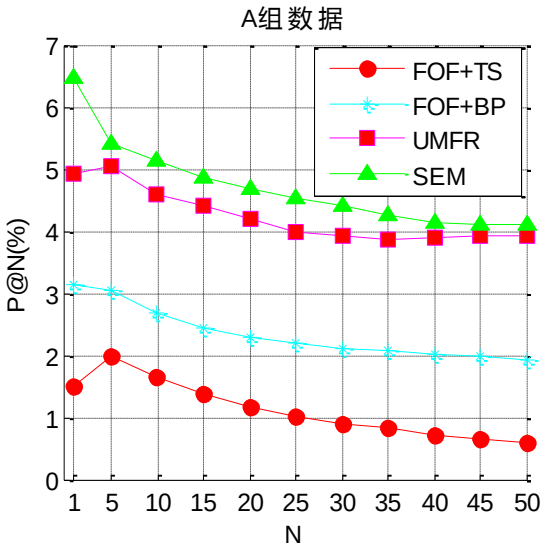
(c) A 组数据



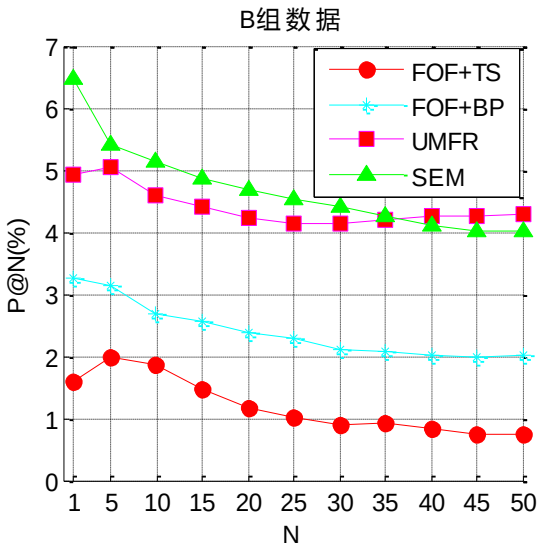
(d) B 组数据

图 3.13 F-measure 值在不同数据集上的对比图

图 3.13 表示 A、B 两组数据下的 F-measure 值对比图，从图中可以看出，本文提出的算法在推荐个数少于 40 时，能保持很好的结果，加入情感因素同样能够提高召回率。



(a) A 组数据



(b) B 组数据

图 3.14 P@N 值在不同数据上的对比图

图 3.14 是表示 A、B 两组数据的 P@N 的计算结果，从图中可以看出，本文提出的算法在得出的 P@N 值较其它算法效果更好，随着推荐好友个数的增加，总的

结果值较为平稳。

通过上述实验，可以得出对用户的文本语义和情感程度的分析中，在一定程度上可以提高推荐的准确率，但是本文提出的推荐模型会随着推荐用户的增多而有所下滑。然而在在线社交网络中，用户能够记住的好友数目符合“邓巴数字”这一规律，用户在交友过程中，真实交流的个数并不很多，提供过多的好友反而造成用户的负担，这说明本文提出的推荐模型能够充分提高用户的接受度和满意度，能够帮助用户找到志同道合的朋友，扩大自己的社交圈。

3.6 本章小结

本章首先介绍了传统的好友推荐算法在对用户的推荐时，考虑因素的不足，没有对用户的文本内容进行利用，提出了在好友推荐中加入用户文本语义和情感程度，并融合时间因素，分析在好友推荐推荐中的作用和影响。然后对基于用户文本语义分析和情感程度的好友推荐的思想和具体实现步骤进行阐述，最后对本章使用的数据集，评测指标进行了分析阐述，并通过实验在爬取的数据集上进行验证，得出该算法可以提高推荐质量，并在一定程度上可以提高用户的满意度。虽然本章节提出的算法与其它算法比较有一定的性能提升，但是对于用户而言，不能只通过简单的余弦相似度来对文本语义进行计算，同时情感程度对用户进行简单的定位分析也较为欠缺，用户的情感在社交网络以及推荐中起着关键作用。下一章节内容将重点改进文本相似度计算和深度考虑用户情感的好友推荐算法研究。

第4章 基于交叉文本相似性和情感词典的好友推荐

4.1 问题描述

新浪微博是在线社交网络中发展迅速的社交平台，是一种以用户为主体而建立的平台，经常被用来分享、发布和传播信息。用户常常在社交平台上做的主要动作就是通过转发来进行消息和新闻的传播，充当了信息传播的媒介。由于时间因素对用户相似度的结果会造成影响，当对用户进行文本相似度计算时，只计算用户同一天的微博信息的相似性。用户在转发微博时，可能看到一则消息的时间先后不同，所以应该使用交叉的比较方法。情感分析是表示用户在说话或者文本表示时，对说话者的态度和意见进行判断评估。在微博文本中对用户的情感分析，可以提高用户之间的匹配度，给用户推荐兴趣相同和情感相似的用户。

在第三章节中，考虑用户的情感时，只对用户微博文本中所包含的程度副词进行考虑，采用二阶段推荐方法。当提取关键词时，由于采用 TF-IDF 方法只能简单的对用户的文本进行关注点的定位，无法从关键词中得出用户的态度以及对关注点所表现情绪的强烈程度，所以对用户的程度副词进行了考虑，利用建立的程度副词词典来对用户的关注点进行情感过滤。用户的微博中富含用户对某个事件的情感，其中包含了对表达内容的积极或消极，正面或反面的看法。对用户的情感进行很好的分析可以挖掘群体行为规律，提高在好友推荐中的用户满意度，帮助用户找到有共同话题的好友。

4.2 引入情感词典的情感分析

在文本聊天环境下，通过文本分析技术可以很容易的发现用户的正面和负面情绪[46]，用户通常会在整个文档中频繁的使用感叹来表达积极情绪，表达消极情绪时会使用更多表达消极的词汇。情感分析分为基于词典和基于机器学习的方法，中文情感分析常用的词典是知网(HowNet)[47]的情感分析用词语集。在文本中经常会包含三类词语，正面情感、负面情感和中性情感词语，对三种词语的分析可以判断每条微博的大致情感趋向。程度副词在文本情感中起着关键作用，含有程度副词

的微博，所表现出的情感更加浓厚，所以计算相似性时必须要考虑程度副词所带来的影响。程度副词可分为客观程度副词和主观程度副词，可以分极度、甚度、递度和微度四个程度级别[48]，按照不同的级别分配权重。程度副词的权重分配如表 4.1 所示。

表 4.1 程度副词的权值分配

程度级别	词语示例	Weight
极度	最、顶、绝顶、极其、极、完全	1.2
甚度	很、非常、特别、十分、分外、相当	0.9
递度	更、更加、比较、较为、愈发	0.6
微度	稍、稍微、稍略、有点儿、有些	0.3

在中文文本处理中，否定词也是情感分析中的重要导向因素，当一个否定词出现在一个表现为正面情感的句子中，则该句子的情感趋向会直接反向变为负面情感，但同时出现两个否定词时，就变为了正面情感，即否定词可以改变文本情感的倾向。根据这一规律，在对否定词进行极性判断时，可以将句子中的所有否定词的个数作为判断条件，当个数为奇数时取反，反之不变。

4.3 融合文本语义和情感分析的好友推荐

对文本的语义分析和情感分析进行了融合，提出了 ESEM 模型。对用户的相似度计算时，综合考虑文本语义和文本情感两部分。在对两者进行分析时，语义分析和情感分析两者所占比重相同，直接进行相加得出用户的总体特征计算公式。具体计算公式如 4.1。

$$sim(U_{ik},U_{jk}) = sim(T_i,T_j) + sim(E_i,E_j) \tag{4.1}$$

其中 $sim(T_i,T_j)$ 为用户文本的相似度， $sim(E_i,E_j)$ 为微博文本的情感相似度。对提出的融合后的 ESEM 模型的算法流程如图 4.1 所示。

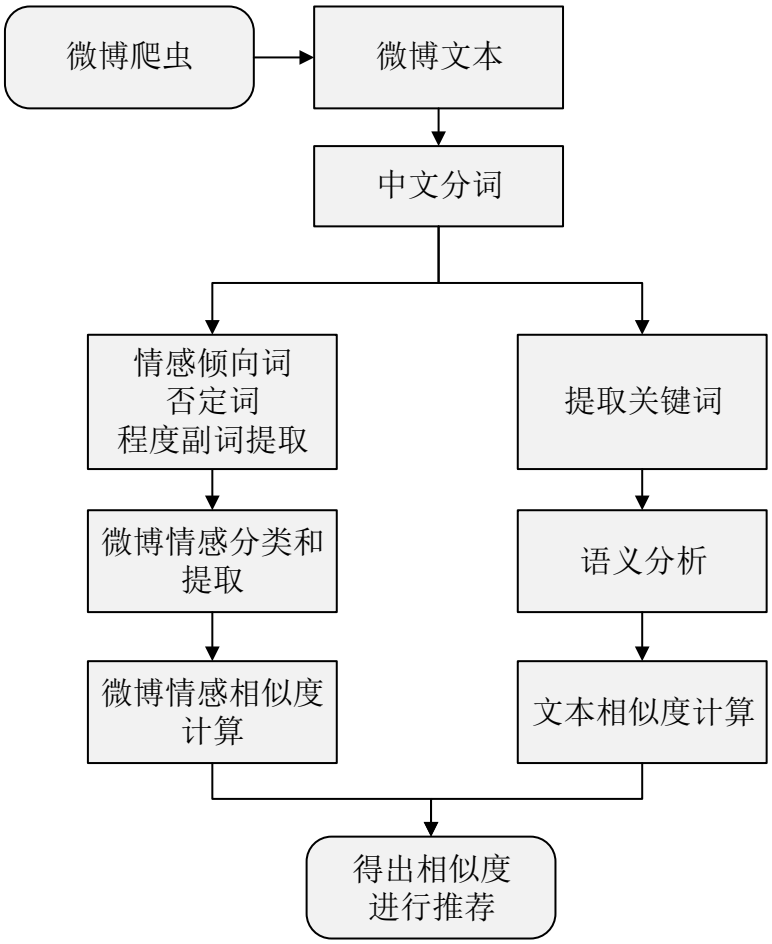


图 4.1 融合后的 ESEM 模型流程图

为了减少时间因素对相似度计算带来的影响，在文本相似度计算时采用交叉的相似度计算方法，修正公式 3.1 后如 4.2 所示。

$$sim(U_{ik},U_{jk})=\begin{cases}\sum_{m=1}^{k}\sum_{n=1}^{k}\left(wf_m\cdot wf_n\cdot\dfrac{\left|N(wb_{im})\cap N(wb_{jn})\right|}{\left|N(wb_{im})\bigcup N(wb_{jn})\right|}\right)\\ 0\end{cases} \tag{4.2}$$

具体详细的交叉余弦相似度计算方法如下，设经过层析分析法得出权值为 wf_1=0.7，wf_2=0.3，例子如表 4.2 所示：

表 4.2 用户的关键词信息

关键词 / 用户	*Key*1	*Key*2
U_i	A|B	A|C
U_j	B|C	A|B

根据公式 4.2，则用户U_i和U_j的相似度计算方式为：

$$\begin{aligned} sim(U_{ik},U_{jk}) &= \sum_{m=1}^{k}\sum_{n=1}^{k}\left(wf_m \cdot wf_n \cdot \frac{\left|N(wb_{im}) \cap N(wb_{jn})\right|}{\left|N(wb_{im}) \cup N(wb_{jn})\right|} \right) \\ &= 0.7\cdot 0.7\cdot \frac{1}{3} + 0.7\cdot 0.3\cdot \frac{1}{1} + 0.3\cdot 0.7\cdot \frac{1}{3} + 0.3\cdot 0.3\cdot \frac{1}{3} \\ &= 0.473 \end{aligned} \tag{4.3}$$

经过计算得出用户之间的相似度，融合时间因素之后的计算方法得出的相似度结果更为精准，符合用户的兴趣。

情感分析主要参考知网(HowNet)中文情感词典，通过计算词语情感的倾向性来计算微博文本的情感相似度。为了得到词语的情感倾向必须知道词语的极性，正面情感或者负面情感。假设有两句话$word_1$和$word_2$，经过分词之后会分成很多的词语，中文中的词语按照情感划分可以分为正面情感、负面情感和中性情感。则$word_1$可能会存在 n 个词语：v_1，v_2，...，v_n，$word_2$有 m 个词语：w_1，w_2，...，w_n，则计算$word_1$和$word_2$的情感相似度时，需要对文本中中的每个词进行判断。若词性为正面情感或者正面评价则用户 A 的情感倾向加 1，否则减 1，最后得出用户的总情感倾向。

根据否定词出现在情感词之前的个数，可以生成一个否定系数值。其计算公式如 4.4 所示。

$$C_{negation} = \begin{cases} 1, & Count(negation) = even \\ -1, & Count(negation) = odd \end{cases} \tag{4.4}$$

则用户之间的情感相似度计算公式如 4.5。

$$sim(E_i, E_j) = \frac{1}{1+\alpha\left|(d_i - d_j)\right|} \tag{4.5}$$

其中d_i表示用户i 的情感倾向，则情感词的倾向计算如公式 4.6 所示。

$$d = \left| C^{p}_{negation} \sum_{k=1}^{n} W_{kpos} - C^{n}_{negation} \sum_{k=1}^{m} W_{kneg} \right| \tag{4.6}$$

其中W_{kpos}表示用户文本集合中正面情感词的数量，W_{kneg}表示用户文本集合中负面情感词的数量，$C^{p}_{negation}$表示出现在正面情感词前面的否定系数。$C^{n}_{negation}$表示出现

在负面情感词前面得出的否定系数。当出现的次数为偶数的时候，则表示用户的情感保持不变，若次数为奇数，则用户的情感要进行反转，即当用户的情感为负面情绪时，如果情感词前有奇数个否定词，则用户的情感变为正面情绪，反之则仍为负面情绪。

加入程度副词带来的影响后，需要对正面情感和负面情感进行加权处理。则对公式 4.6 修正之后的计算公式为

$$d = \left| W_{adv} C^{p}_{negation} \sum_{k=1}^{n} W_{kpos} - W_{adv} C^{n}_{negation} \sum_{k=1}^{m} W_{kneg} \right| \tag{4.7}$$

4.4 实验及性能分析

4.4.1 数据集

为了验证本章提出的算法，同时为了弥补数据集小可能带来的问题，本文使用编写的爬虫程序进行了大数据规模的数据爬取。采集从 2015 年 2 月 1 日至 2016 年 1 月 3 日之间的微博数据，以周鸿祎为头结点，分别爬取 3 度节点的信息，即周鸿祎的朋友的朋友的朋友，其中爬取到的信息包含 15086 个用户，2966116 条微博信息和 1986600 条关系数据。为了更好的评价算法性能，本文分别进行 3 次试验，取 3 次结果的平均值作为为最终的评价指标值。在使用数据之前，需要对数据进行长尾分布的验证，用户的出度和入度分布如图 4.2 所示，

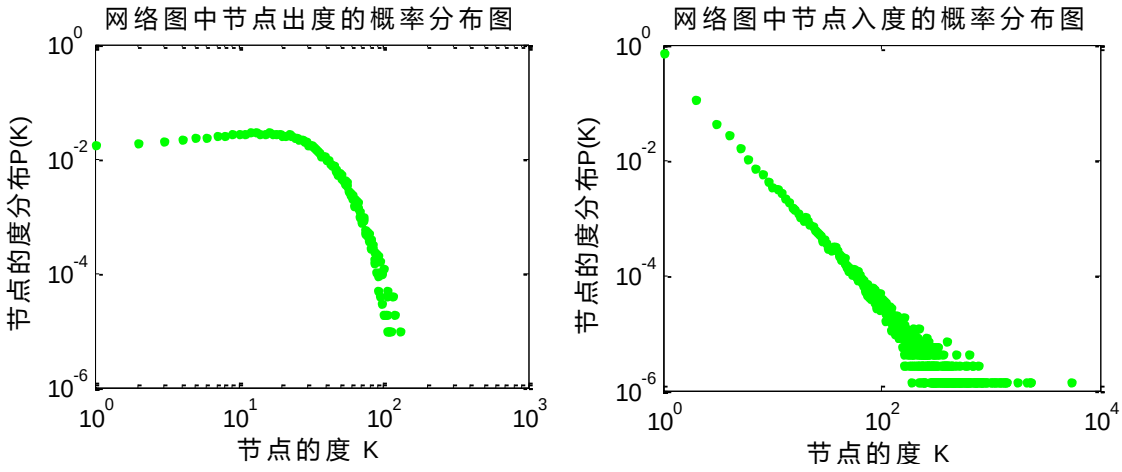


图 4.2 数据集中用户的出度和入度概率分布图

通过图中可以得出，用户的出度和入度都满足长尾分布和幂律分布。表明该组数据的网络是无标度网络，符合实际情况。

4.4.2 评价指标

本节采用的推荐算法评价指标为准确率、召回率和 F-measure 值来综合评价，具体在第二章节已经做了阐述。

4.4.3 实验分析

本章所采用的算法思想和第三章不同，本章没有采用二阶段方式进行推荐，而是采用融合后的计算方式来进行好友推荐。本章节选用的对比实验描述如下：

1. ACR-FoF (algebraic connectivity regularized friends-of-friends)，该方法认为大多数现有的好友推荐方法旨在提高推荐成功率，而没有考虑到在社交网络中信息的传播更为重要，引入了一个连接网络的代数连通度来估计其传播内容的能力[49]。

2. SEM 算法，采用二阶段式的好友推荐方法，考虑用户的文本语义和情感程度，并引入时间因素带来的影响，综合的进行好友推荐。

本章提出的 ESEM 算法和 ACR-FoF 算法以及上一章节提出的 SEM 算法准确度的对比如图 4.3 所示。分别选取推荐个数为 1、5、10、15、20、25、30、35、40、45、50 在三种算法下的准确率值。

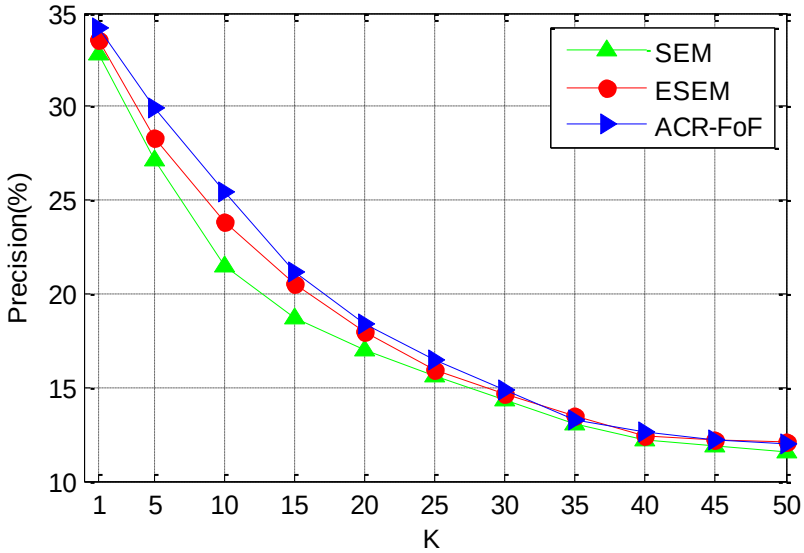


图 4.3 SEM、ESEM 和 ACR-FoF 算法的准确度对比图

本章提出的 ESEM 算法与 SEM 算法在召回率的对比时，分别选取推荐个数为 1、5、10、15、20、25、30、35、40、45、50 在两种算法下的召回率值。算法的结果如图 4.4 所示，

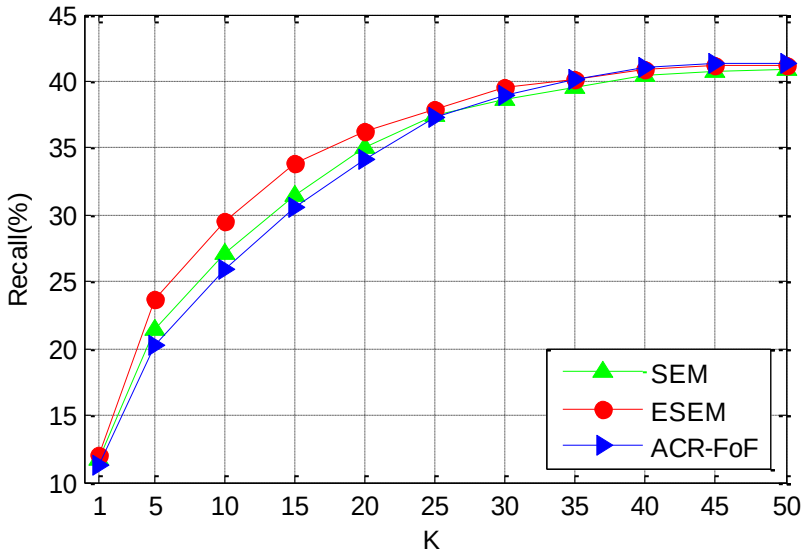


图 4.4 SEM、ESEM 和 ACR-FoF 算法的召回率对比图

在本章提出的 ESEM 算法、ACR-FOF 算法与 SEM 算法的 F-measure 值对比时，分别选取推荐个数为 1、5、10、15、20、25、30、35、40、45、50 在三种算法下的 F-measure 值。在计算时，准确率和召回率总是相反的，当准确率高的时候，召回率就会变低，准确率低的时候，召回率就会变高。所以引入 F-measure 值来平衡准确率和召回率，算法的对比结果如图 4.5 所示，

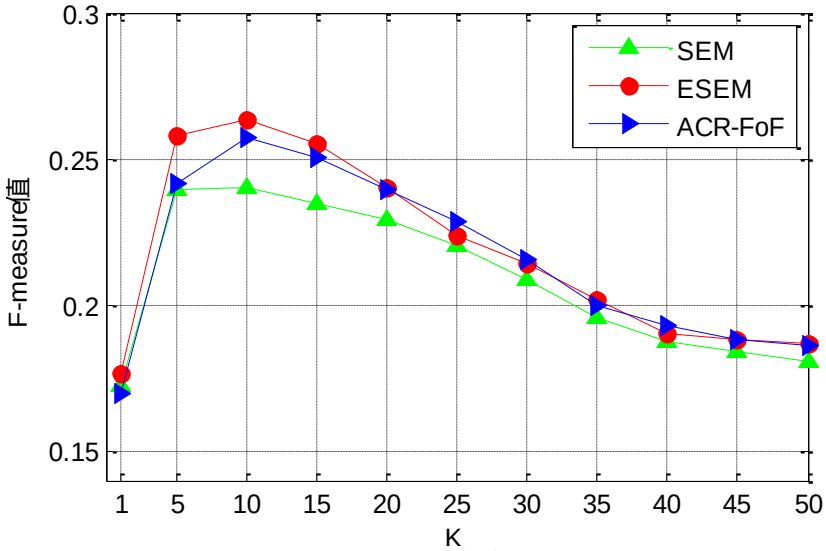


图 4.5 SEM、ESEM 和 ACR-FoF 算法的 F-measure 值对比图

通过图 4.5 可知在推荐个数为 5-10 左右时，所得到的算法处于最高点，相对于 SEM 算法有明显提高，同时随着推荐个数的逐步增加，准确率逐步下降。

综合分析得出，在微博中文本能够反映出用户的兴趣，同时在微博文本中蕴含的文本语义和情感都可以成为用户的特征标签，用户每天的文本信息都是动态变化，可以表征用户的心情和关注点的变化。当采用一些用户标签信息进行推荐时，具有一定的滞后性和不准确性，效果不理想，结果不新颖。融合用户的文本语义和情感分析的好友推荐方法在对比试验中，表现出的效果更好。

4.5 本章小结

本章主要介绍了文本内容对计算用户相似度的重要性，分析了传统相似度计算所存在的问题，为了减少单纯相似度比较时所带来的计算不准确性问题，通过对文本内容的分析，融合时间因素，并采取交叉相似度计算的方法来减少计算不准确性的影响。在对文本情感分析时，又引入了情感词典的分析方法，考虑了情感词、程度副词和否定词对文本情感所带来的影响。提出了融合时间因素和两者之后的推荐算法，通过在更大的真实数据集中进行多组对比实验，证明提出的算法具有更好的推荐效果。

第5章 融合文本语义和情感分析的好友推荐系统

5.1 问题描述

社交网络用户的信息多种多样，能够用来进行好友推荐的信息很多，比如用户在注册社交网站和社交工具时所填写的静态信息，这些信息的真实性较低，且随着时间的变化，用户对注册信息时所填写的内容更改的可能性小。同时包括一些用户的兴趣变更的动态信息，这些数据一般随着社会活动而在社交网站上产生的信息。通常情况下，用户的静态信息包括用户住址、性别、所属地、学校和职位等信息，动态信息包括用户每天通过社交网站发表的文本信息、GPS 信息、收藏和点赞信息等。

进行好友推荐时，最适合最便利的信息就是用户的一些静态信息。相对于动态信息，静态信息的获取比较方便，一般是用户注册时填写的信息，已经存在于本地现有的数据库中。通常情况，有些用户在注册时存在应付心理，同时为了保证自己的隐私问题，他们所填写的注册信息存在虚假信息，如有些用户为了方便在网上聊天，更改自己的年龄，甚至有些用户更改自己的性别和所在地。这些不真实的用户信息给推荐的准确性带来了困难。而进行个性化好友推荐时，采用的用户文本内容都是动态信息。在微博中，用户通过发表、转发一些文本内容来表达自己的观点和想法，基于这些文本内容可以很直观的推荐出最近时间与用户相同兴趣和爱好的好友。但是文本的相似度往往缺乏对用户情感的判断，而在用户的微博文本内容中也会存在用户的一些情感信息，包括正面和负面，积极和消极等。对这些信息可以对用户进行准确的定位，动态的进行好友推荐。

由于不同的用户在不同的时间，所关注的内容和方向不同，因此考虑到对不同类型用户的推荐，设计出了融合时间因素的用户文本语义的情感分析的好友推荐系统(SETRS，Personal Recommendation System with Users' text Sentiment and Emotion)，首先对用户的文本内容进行分析，针对不同时间下，用户对好友的需求可能不同的特点，对不同时间段的关键词分配不同的权重。

5.2 特征提取及需求分析

针对文本内容进行好友推荐的前提是对用户的微博内容进行预处理，提取用户的个性化标签。微博文本内容的预处理包括去除特殊符号、分词和去除停留词。采用中科院的分词接口对文本内容进行分词。当提取用户的个性化标签时，采用距离当前时间最近的 5 天内的文本内容进行提取处理。关键词的提取采用 TF-IDF 方法。TF-IDF 方法是统计学中一种常用的统计方法，用来评估词语在文本中的重要性。当一些字词重复的出现时，就认为该词汇比较重要，如果在语料库中重复出现，则重要性就下降。所以就采用了 TF-IDF 加权的方式来平衡在文件中出现的次数和语料库中出现的次数。

在现实生活中，在线社交网络中的用户，静态信息的改动非常少，而常见的好友推荐系统中对常常针对用户的静态信息进行推荐，这些信息只能过时的反映用户的情况，所以这种方法得出的推荐结果就会造成偏差。随着时间的变化，用户的特征和关注点都会有大大小小的变化，因此用户最近的信息能更好的描述用户特征。以 Facebook，Tweeter 和微博为代表的在线社交网络，用户可以发表短文本信息，这些微博文本信息能实时反映用户兴趣所在和情感性格特征。

在在线社交网络中不同的用户在通过微博进行交互时，能相互影响情绪。通常微博信息中会包含两类信息，即文本信息和用户情感信息，在传统的好友推荐算法中，用户的情感因素往往没有被考虑进去。融合时间因素、语义分析和情感分析技术对用户进行好友推荐可以提高用户的满意度。

5.3 好友推荐系统的设计与实现

5.3.1 总体设计

好友推荐系统的目标是能够根据用户的静态和动态信息，挖掘出用户的兴趣和爱好，产生使用户满意的推荐结果。整个推荐系统主要分为以下几个模块：推荐模块、搜索模块、数据存储模块和用户交互模块，如图 5.1 所示。

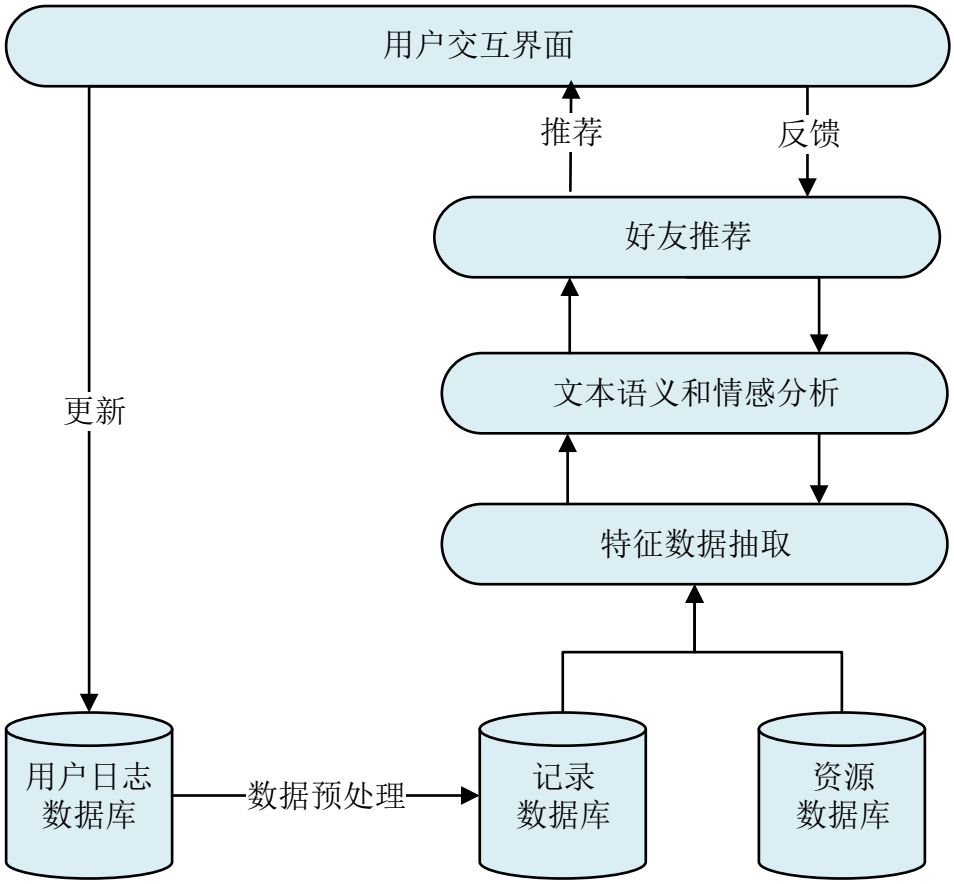


图 5.1 好友推荐系统架构图

用户交互模块主要是用户交互界面，是人与计算机或系统之间进行信息交互的通道，用户通过交互界面进行操作和输入信息，系统则通过用户交互界面向用户提供信息，供用户查看、分析和判断。

特征数据抽取模块，对用户的微博内容数据进行分析，提取关键词和情感词，特征抽取模块为上层模块提供服务，对用户的兴趣建立模型和标签。

数据储存模块包括用户日志数据库、记录数据库和资源数据库，其中，用户的日志数据库主要记录了用户的动态信息和静态信息以及所有的信息数据。由于日志数据库中的数据较为庞大，繁杂信息较多，记录数据库就是对其进行了数据清理，价值较高，可供好友推荐计算的数据信息。资源数据库保存了用户的注册信息和网络中的资源信息。

搜索模块可以通过用户的 ID 来查看用户的微博内容，主动发现兴趣相同的好友，同时采用分页功能，来方便用户的查看。

5.3.2 好友推荐模块

好友推荐系统的核心就是好友推荐模块，在好友推荐模块中，可以对模块进行细分，分为建立用户的特征标签和好友推荐。根据用户的历史微博数据进行关键词的提取作为用户的特征标签。好友推荐的功能根据用户的特征标签、时间因素带来的影响，对不同时间的特征标签分配不同的权重，与其他用户进行相似度计算、排序，得出最后的推荐结果，交付给用户交互模块。推荐系统中采用的是用户文本语义分析和情感分析。SETRS 推荐过程描述如下：

Step1：根据时间提取用户过去 5 天的文本内容数据；

Step2：文本内容预处理，通过 TF-IDF 算法提取关键词，构建特征标签之前，需要对用户的特征标签进行语义分析，中文中通常含有过多的同义词汇，影响用户的相似度计算，需转换之后构建用户特征标签；

Step3：计算目标用户与待推荐用户之间的相似性，从每天的微博中提取的关键词集合与其他用户的关键词集合进行余弦相似度的计算。为了避免只比较相同时间段的相似性而带来的计算不准确性，采用交叉相似度的计算方法 $sim(U_{ik},U_{jk})=\sum_{m=1}^{k}\sum_{n=1}^{k}\left(wf_m\cdot wf_n\cdot\frac{\left|N(wb_{im})\cap N(wb_{jn})\right|}{\left|N(wb_{im})\cup N(wb_{jn})\right|}\right)$。由于时间因素给文本带来的影响，利用层次分析法对每天进行分配权重，得到的权重分别为 0.51，0.26，0.13，0.07 和 0.03，距离当前时间越近，所得到的权重越大；

Step4：对用户的文本内容进行情感分析，采用情感词典的方法进行分析。得出具体的情感得分，根据情感得分得出用户的情感相似性；

Step5：对用户的文本语义相似度值和情感相似得分进行融合考虑，得出最终的相似度值；

Step6：根据得出的相似度值选出 TOP-N 用户推荐给用户；

5.3.3 界面显示模块

界面显示模块采用网页方法，根据分析和计算，将得出相似度较高的 TOP-N 用户显示在推荐页面中。本模块采用 Python 中的 Django 框架实现，使用账户和密码登录进入好友推荐系统，用户登录界面如图 5.2 所示。

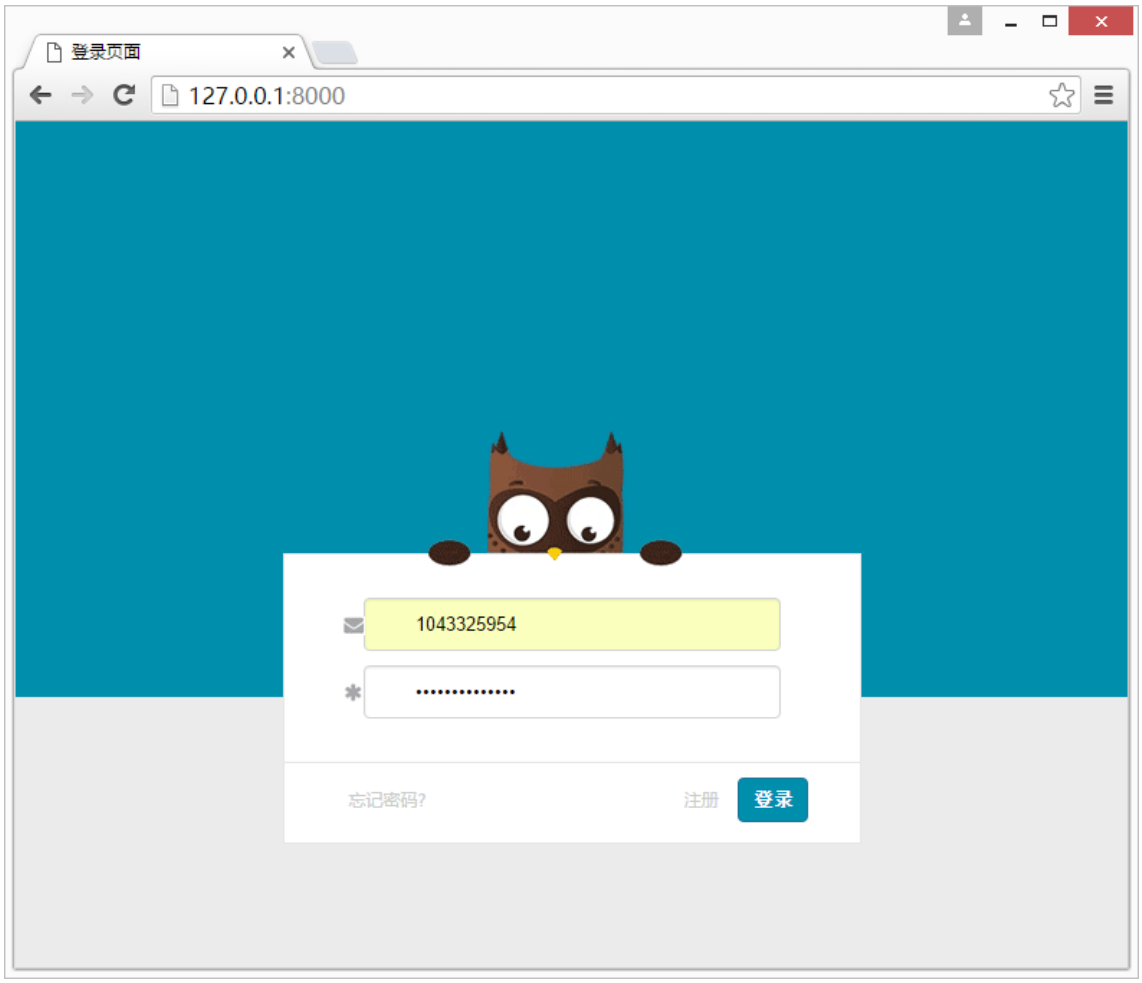


图 5.2 用户登录界面

用户登录成功后，将会出现用户的微博内容显示界面和好友推荐界面，显示界面如 5.3 所示，

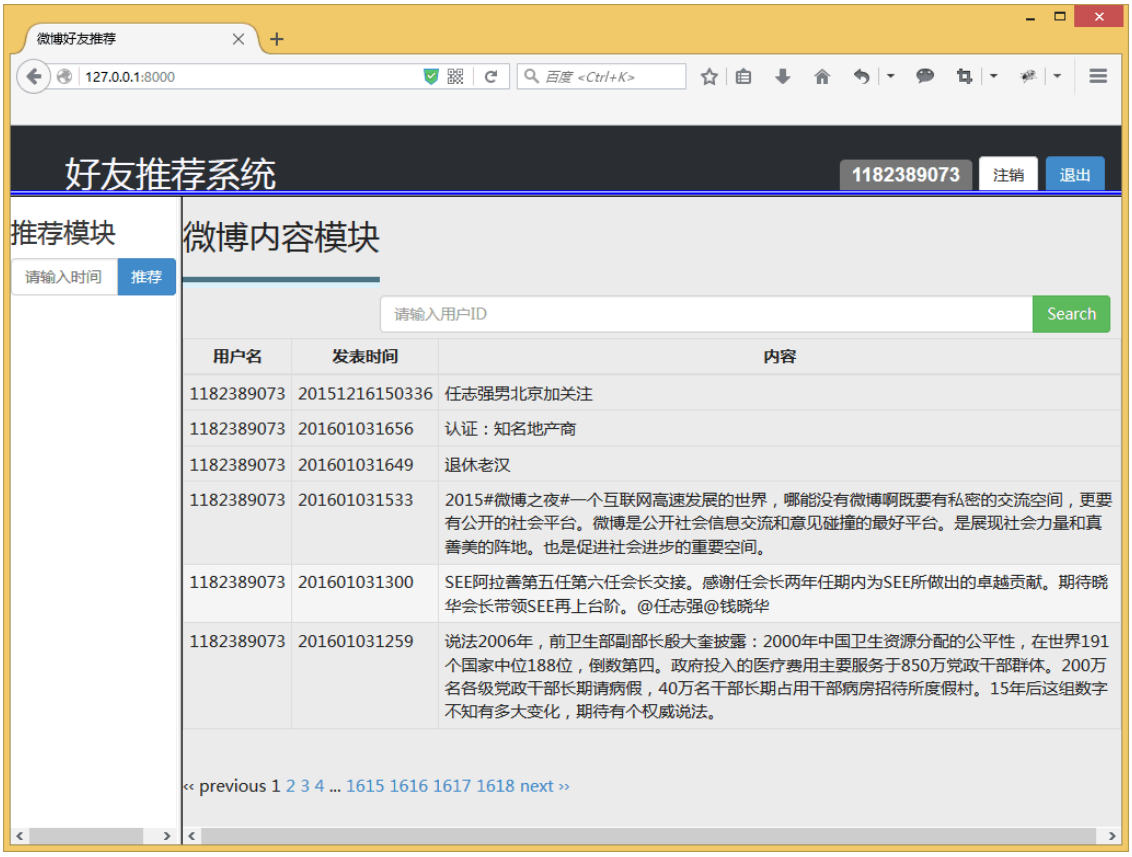


图 5.3 微博内容显示界面

图 5.3 是用户的微博内容显示界面，采用的数据集是爬虫程序抓取的 A 组数据集，在该页面中，用户可以通过用户的 ID 查看用户的微博内容，用户在微博中，通过发表微博内容来获得关注、添加好友、扩大自己的交际圈。整个界面显示了融合时间的好友推荐界面，根据用户的微博文本内容产生推荐结果。

为了更好的展示本论文提出的算法，加入了时间输入框，当选取不同的时间时，用户的关注点不同，得出的推荐结果也不同。图 5.4 是对用户 1043325954 选取特定时间 2015 年 9 月 26 日的好友推荐结果，图 5.5 是选取特定时间 2015 年 9 月 30 日的好友推荐结果。

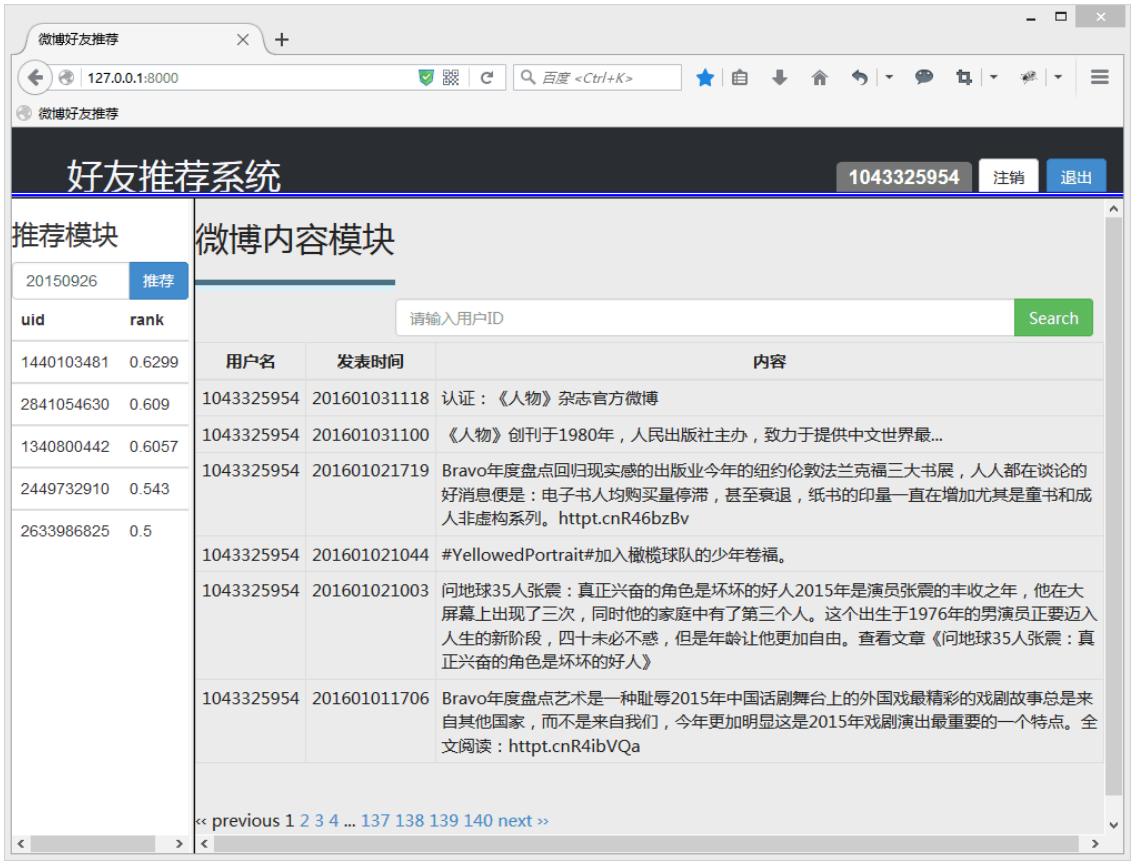


图 5.4 选定时间 2015 年 9 月 26 日的推荐结果

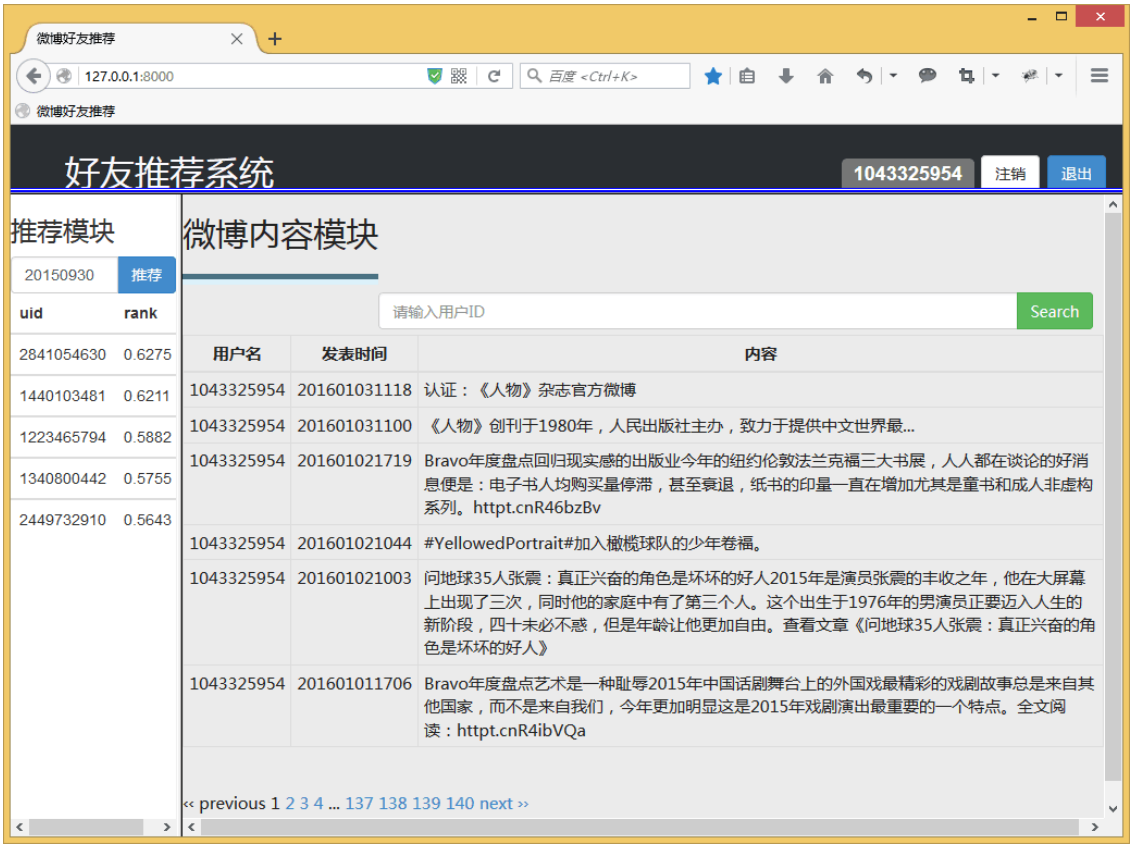


图 5.5 选定时间 2015 年 9 月 30 日的推荐结果

通过观察图 5.4 和 5.5，可知在不同的时间段，计算出的好友相似度值是不同的，推荐结果也不同，由于选择的时间较近，从推荐结果可以看出，推荐结果有一定的相似性。

5.4 本章小结

本章首先对社交网站中用户的行为和信息进行分析，根据用户的需求来对好友进行推荐。阐述了整个推荐系统的设计和实现，以及各个模块之间的关系，并对整个推荐算法的流程进行了详细的描述。最后通过对具体页面的展示效果，来阐明系统的具体推荐结果。为了满足不同时间段内不同用户的差异性需求，本章设计了基于时间的融合文本语义和情感分析的好友推荐系统。

第6章 总结和展望

微博已经成为一种流行的信息交流、传播和分享的新媒介，用户可以轻松的获取网络信息，信息在微博上具有传播速度快和范围广的特点。利用这种新的媒介，可以帮助用户建立庞大的社交圈，更便利的获取信息。本章对全文的研究重点进行总结，并分析了未来的研究计划和展望。

6.1 工作总结

微博这种新型媒介的大规模使用，给广大用户带来了极好的使用体验，用户能随时随地的获取信息。微博上存储着海量的信息资源，为了对繁杂信息的过滤，我们在使用微博时，只能看到关注好友发的微博消息，这是海量数据过滤的一种手段，也是推荐中的一种应用，这种方法能够解决信息暴增带来的问题。如何有效利用用户的个人信息给用户进行好友推荐，并以此扩大用户的交际圈，增强用户和社交网站的黏性是面临的难题。

本文对用户的微博信息进行研究，利用用户的微博文本富含的信息进行深入研究并通过实验进行了验证，主要的工作总结如下：

1. 对常用推荐系统的工作架构进行分析，综述了在好友推荐领域常用的推荐方法以及所存在的问题，并指出了推荐技术面临的机遇和挑战。

2. 对用户在微博中的行为进行分析，深入研究了用户微博文本的特点。在好友推荐领域，主题元素就是用户，在推荐时，所围绕的中心就应该是用户的个人信息。用户的文本中通常会存在文本语义相似性和情感相似性的特点。

3. 结合文本语义和中文中的程度副词提出了 SEM 模型的好友推荐方法，采用二阶段的推荐算法，并结合时间因素，根据 AHP 方法为时间分配权值。根据用户的文本特征计算用户之间的相似性，并得出待推荐列表集合。在第二阶段，对待推荐集合进行情感程度的相似度分析。用户的文本相似可以证明用户所关注的方向相似，根据程度副词的相似度可以得出用户所关注内容的重要程度，最终得出和用户最相似的用户。

4. 用户获取信息和分享内容有时间先后顺序的影响，在对文本计算相似度时必须考虑时间所带来的影响，针对这一问题，本文提出了交叉的相似度计算方法，来消除时间对用户获取信息先后顺序的影响。同时用户的情感表达，包括了用户的情感词倾向，程度副词和否定词，这些词都可以影响用户的情感。鉴于此，通过改进提出的 SEM 模型，本文又提出了 ESEM 模型，对常用的程度副词进行分配权重，同时根据情感词前否定词的个数得出否定系数，共同加入到用户的情感相似度计算中。在 ESEM 模型中，对文本相似性和情感相似性进行了融合。

5. 为了验证本文提出的算法模型，采用大众的社交网站新浪微博数据进行了爬取。针对新浪公开的 API 接口限制过多，爬取数据不全的情况，本文通过自行编写的爬虫程序对微博数据进行爬取，通过预处理，得出用户的特征标签。最后通过分析微博用户群体的行为，发现微博用户的出度和入度分布均符合幂律分布的特点，具有无标独特性，说明了数据的真实性和可用性。

6.2 不足与展望

随着科技的快速发展，各种新技术的诞生，许多针对好友推荐方面的难题将会逐渐得到解决，同时好友推荐方面的研究也是一个长期的科研任务。虽然本文提出的算法和模型在一定程度上能够提高性能，并取得了一些研究成果，但本文的研究工作还有一定的不足，在未来更进一步的研究中，我将考虑以下问题：

1. 本文中的数据集主要来源于新浪微博，虽然获取数据的数量达到了一定级别，但是没有在其它的社交平台上的数据集上进行实验，同时相对于现实中的海量计算和信息量来说，研究的实用价值有待提高，在未来的研究中，可以更全面的获取其它数据集的来源，如微信和各种社交软件等。在更多的数据集上进行分析、实验和研究。

2. 本文提出的算法主要是针对文本内容进行研究，分析其文本语义和情感。分析文本的语义相似度时，对提取的关键词进行同义词的处理，得出用户的特征标签。采用情感词典的方法对文本情感进行分析，对文本中含有的情感词、程度副词和否定词进行考虑计算，忽略了中文文本中，文本表达意思的复杂性。在未来的研究中，将会考虑采用机器学习的方法来对文本进行情感分析，在机器学习中可以采用有监

督和无监督的学习方法，来对用户进行分析，并可以结合深度学习的方法，来学习和挖掘出用户的特征，预测用户的行为，对用户进行精准的推荐。

3. 本文提出的算法都是采用单机模式，没有考虑并行计算等方式，随着技术的发展，云计算和大数据技术的普及，这些新技术能够很好的解决海量数据的处理和分析中出现的瓶颈问题。在未来的研究中，我将考虑结合大数据和并行计算技术对现有算法进行改进，提高计算的效率和推荐的性能。

参考文献

[1] Milgram S. The Small World Problem[J]. Psychology Today, 1967, 2(1): 185-195.

[2] Yin Zhijun, Gupta M, Weninger T, et al. LINKREC: a unified framework for link recommendation with user attributes and graph structure[C]//Proceedings of the 19th international conference on World wide web. North Carolina: ACM, 2010: 1211-1212.

[3] Hamid M N, Naser M A, Hasan M K, et al. A cohesion-based friend- recommendation system[J]. Social Network Analysis and Mining, 2014, 4(1): 1-11.

[4] Li Chengxin, Wu Huimin, Jin Qin. Emotion classification of chinese microblog text via fusion of bow and evector feature representations[M]. Natural Language Processing and Chinese Computing, Berlin: Springer, 2014: 217-228.

[5] Liben-Nowell D, Kleinberg J. The link-prediction problem for social networks[J]. Journal of the American society for information science and technology, 2007, 58(7): 1019-1031.

[6] Gou Liang, You Fang, Guo Jun, et al. SFViz: Interest-based friends exploration and recommendation in social networks[C]//Proceedings of the 2011 Visual Information Communication - International Symposium. HongKong: ACM, 2011: 1-10.

[7] Chin A, Xu Bin, Wang Hao. Who should I add as a friend?: A study of friend recommendations using proximity and homophily[C]//Proceedings of the 4th International Workshop on Modeling Social Media. Paris: ACM, 2013: 1-7.

[8] Yang Tan, Cui Yidong, Jin Yuehui. BPR-UserRec: a personalized user recommendation method in social tagging systems[J]. Journal of China Universities of Posts & Telecommunications, 2013, 20(1): 122-128.

[9] Chu Chenghao, Wu Wanchuen, Wang Chengchi, et al. Friend recommendation for location-based mobile social networks[C]//2013 Seventh International Conference on Innovative Mobile and Internet Services in Ubiquitous Computing (IMIS). Taiwan: IEEE, 2013: 365-370.

[10] Silva N B, Tsang I R, Cavalcanti G D C, et al. A graph-based friend recommendation system using genetic algorithm[C]//2010 IEEE Congress on Evolutionary Computation (CEC). Barcelona, Spain: IEEE, 2010: 1-7.

[11] Huang Shangrong, Zhang Jian, Lu Shiyang, et al. Social Friend Recommendation Based on Network Correlation and Feature Co-Clustering[C]//Proceedings of the 5th

ACM on International Conference on Multimedia Retrieval. Shanghai: ACM, 2015: 315-322.

[12] Wang Zhibo, Liao Jilong, Cao Qing, et al. Friendbook: A Semantic-based Friend Recommendation System for Social Networks[J]. IEEE Transactions on Mobile Computing, 2015, 14(3): 538-551.

[13] Koren Y. Factor in the neighbors: Scalable and accurate collaborative filtering[J]. ACM Transactions on Knowledge Discovery from Data (TKDD), 2010, 4(1): 1-9.

[14] Koren Y. Collaborative filtering with temporal dynamics[J]. Communications of the ACM, 2010, 53(4): 89-97.

[15] 项亮. 推荐系统实践[M]. 北京: 人民邮电出版社, 2012: 1-197.

[16] Dey A K. Understanding and Using Context Personal and Ubiquitous Computing Journal[J]. Personal & Ubiquitous Computing, 2001, 5(1): 4-7.

[17] McNee S M, Riedl J, Konstan J A. Being accurate is not enough: how accuracy metrics have hurt recommender systems[C]//CHI'06 extended abstracts on Human factors in computing systems. Canada: ACM, 2006: 1097-1101.

[18] 朱郁筱, 吕琳媛. 推荐系统评价指标综述[J]. 电子科技大学学报, 2012, 41(2): 163-175.

[19] 项亮. 动态推荐系统关键技术研究[D]. 北京: 中国科学院研究生院, 2011.

[20] Shani G, Gunawardana A. Evaluating recommendation systems[M]//Recommender systems handbook. US: Springer, 2011: 257-297.

[21] Resnick P, Iacovou N, Suchak M, et al. GroupLens: an open architecture for collaborative filtering of netnews[C]//Proceedings of the 1994 ACM conference on Computer supported cooperative work. North Carolina: ACM, 1994: 175-186.

[22] Linden G, Smith B, York J. Amazon.com Recommendations: Item-to-Item Collaborative Filtering[J]. IEEE Internet Computing, 2010, 7(1): 76-80.

[23] Herlocker J L, Konstan J A, Borchers A, et al. An algorithmic framework for performing collaborative filtering[C]//Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval. California: ACM, 1999: 230-237.

[24] 曾春, 邢春晓, 周立柱. 个性化服务技术综述[J]. 软件学报, 2002, 13(10): 1952-1961.

[25] 曾春, 邢春晓, 周立柱. 基于内容过滤的个性化搜索算法[J]. 软件学报, 2003, 14(5): 999-1004.

[26] 孙胜平, 张真继. 中文微博客热点话题检测与跟踪技术研究[D]. 北京: 北京交通大学, 2011.

[27] Yang Changchun, Zhou Meng, YE S, et al. An Improved Hot Topic Detection Method for Microblog Based On CURE Algorithm[J]. Computer Simulation, 2013, 11: 087-098.

[28] Ekman P. Biological and cultural contributions to body and facial movement[J]. 1977, 1977: 34-84.

[29] Picard R W, Picard R. Affective computing[M]. Cambridge: MIT press, 1997.

[30] Salton G, Wong A, Yang C S. A vector space model for automatic indexing[J]. Communications of the ACM, 1975, 18(11): 613-620.

[31] Lei Shi, Han Yingjie, Ding Xiaoguang, et al. An SPN-based integrated model for Web prefetching and caching[J]. Journal of Computer Science and Technology, 2006, 21(4): 482-489.

[32] Billsus D, Pazzani M J. Learning Collaborative Information Filters[C]//Proceedings of the Fifteenth International Conference on Machine Learning. Wisconsin: ICML, 1998: 46-54.

[33] Basu C, Hirsh H, Cohen W. Recommendation as Classification: Using Social and Content-Based Information in Recommendation[C]//Fifteenth National Conference on Artificial Intelligence. California: AAAI/IAAI, 1998: 714-720.

[34] Sarwar B, Karypis G, Konstan J, et al. Analysis of recommendation algorithms for e-commerce[C]// Proceedings of the 2nd ACM conference on Electronic commerce. Minneapolis: ACM, 2000: 158-167.

[35] Sarwar B, Karypis G, Konstan J, et al. Application of dimensionality reduction in recommender system-a case study[R]. Minnesota Univ Minneapolis Dept of Computer Science, 2000.

[36] Pazzani M, Billsus D. Learning and revising user profiles: The identification of interesting web sites[J]. Machine learning, 1997, 27(3): 313-331.

[37] Cheong M, Lee V. Dissecting Twitter: A Review on Current Microblogging Research and Lessons from Related Fields[M]. Springer Vienna, 2010.

[38] Xiong Xiaobing, Zhou Gang, Huang Yongzhong, et al. Dynamic evolution of collective emotions in social networks: a case study of Sina weibo[J]. Science China Information Sciences, 2013, 56(7): 1-18.

[39] 赵妍妍, 秦兵, 刘挺. 文本情感分析[J]. 软件学报, 2010, 21(8): 1834-1848.

[40] Bao Shenghua, Xu Shengliang, Zhang Li, et al. Joint emotion-topic modeling for social affective text mining[C]//Processings of the Ninth IEEE International Conference on Data Mining. Florida: IEEE, 2009: 699-704.

[41] da Silva N F F, Hruschka E R, Hruschka E R. Tweet sentiment analysis with classifier ensembles[J]. Decision Support Systems, 2014, 66: 170-179.

[42] Barabasi A L, Albert R. Emergence of Scaling in Random Networks[J]. Science, 1999, 286(5439): 509-512.

[43] ME N, M. G. Finding and evaluating community structure in networks[J]. Physical Review. E, Statistical, Nonlinear, and Soft Matter Physics, 2004, 69(2): 292-313.

[44] Kamishima T, Akaho S, Asoh H, et al. Efficiency Improvement of Neutrality-Enhanced Recommendation[C]//Proceedings of the 3rd Workshop on Human Decision Making in Recommender Systems in conjunction with the 7th ACM Conference on Recommender Systems (RecSys 2013), Hong Kong: ACM, 2013: 1-8.

[45] Feng Shi, Zhang Le, Wang Daling, et al. A Unified Microblog User Similarity Model for Online Friend Recommendation[M]//Natural Language Processing and Chinese Computing. Springer Berlin Heidelberg, 2014: 286-298.

[46] Hancock J T, Landrigan C, Silver C. Expressing emotion in text-based communication[C]//Proceedings of the SIGCHI conference on Human factors in computing systems. Atlanta: ACM, 2007: 929-932.

[47] Dong Zhendong, Dong Qiang. Hownet and the Computation of Meaning[M]//Hownet And the Computation of Meaning. World Scientific Publishing Co., Inc., 2006: 316.

[48] 陈颖. 简论程度副词的程度等级[J]. 牡丹江师范学院学报, 2008(1): 59-62.

[49] Yu Zhi, Wang Can, Bu Jiajun, et al. Friend recommendation with content spread enhancement in social networks[J]. Information Sciences, 2015, 309: 102–118.

致谢

值此论文完成之际，意味着三年的研究生生涯到了终点，同时表明我的几十年的学生生涯终于要告一段落。蓦然回顾三年的种种，重邮给我留下了深刻的印象，三年来，有所付出、有所收获、有得有失，在此我要对这段时间给予我支持和帮助的老师、同学、朋友以及家人们表示最衷心的谢意。

首先，我要感谢的是我的导师刘群教授，刘老师开启了我的研究生生涯，同时也是我研究生生涯的一盏长明灯，一直伴随我走过三年。刘老师在我写论文和发表论文期间能够耐心的批阅，并进行多次修改，不厌其烦的帮我解答问题和学术上的疑惑，总是在我迷途时给予我方向。刘老师拥有谦逊温婉的性格，严谨的学术态度是我学习的榜样。特别是能够为学生的个人发展着想，在我论文成稿之际，在我提出实习要求时，刘老师在实习工作中给予建议和帮助，这些都给我留下了深刻的印象。当我在科研道路和找工作路途遇到困难和迷茫时，刘老师能够通过各种例子进行讲解，毕业找工作不是一成不变的，第一份工作只是工作生涯的一部分。对找工作提出合理的建议。刘老师不仅是我的导师，也是我学习和生活中的榜样。

其次要感谢我的师兄弟和师姐妹们，感谢他们三年的陪伴。感谢李晓冰师兄等师兄师姐们在研究生生涯中耐心的指导，当我初入科研路途时，教我如何查找文献和确定研究方向。感谢同门的张振、刘荣鑫和时煜斌、齐会敏、刘秋霞、匡荣和戴大祥等师弟妹们，感谢你们在我论文成稿之际给予的帮助，和你们在实验室一起学习奋斗使我的研究生生涯多姿多彩，充实和美好。感谢我的室友曲省卫、王永超和冉欢在生活中和学习上的帮助。

特别感谢我的家人，一直在背后默默的付出，正是你们的付出和鼓励，我才能顺利完成学业，你们的付出我无以回报，特别是我的爱人和儿子，你们永远是我前进道路上的动力源泉，我永远爱你们。

感谢国家自然科学基金项目(61075019)，重庆市自然科学基金项目(CSTC2014jcyjA40047)等项目对本文的资助。

最后，感谢各位专家老师、教授在百忙中能抽出时间对本文进行评审。

攻读硕士学位期间从事的科研工作及取得的成果

参与科研项目：

[1] 国家自然科学基金项目(61075019)，项目名称：三支决策聚类理论与方法研究，项目参与年限：2014.1-2017.12

[2] 重庆市自然科学基金项目(CSTC2014jcyjA40047)；项目名称：多智能体网络分组一致动力学行为分析与研究，项目参与年限：2014.07-2017.06

[3] 重庆市教委研究项目(KJ1400403)；项目名称：多智能体复杂网络的牵制一致性研究；项目参与年限：2014.1-2015.12

[4] 重庆邮电大学博士启动项目(A2014-20)；项目名称：多智能体网络分组一致性问题研究；项目参与年限：2014.1-2017.9

发表及完成论文：

[1] 刘群，**孙红涛**，纪良浩.一种融合文本语义和情感分析的好友推荐方法研究[J].(二类期刊，已录用)

[2] 刘荣鑫，**孙红涛**，李晓冰，杨鸿滟，刘群，一种基于LBS的微信用户行为的交友方式[P].(已申请)

获奖：

[1] **孙红涛**，刘荣鑫，李晓冰，杨鸿滟.第四届“华为杯”全国大学生智能设计竞赛国家级二等奖。

[2] **孙红涛**，刘荣鑫，李晓冰，杨鸿滟.第二届“腾讯-重邮”专利创意大赛二等奖。