各位答辩专家晚上好，我的论文题目是：融合用户文本语义和情感分析的好友推荐方法，我的导师是刘群教授，我的研究方向是智能信息处理。

我的内容主要分为以下6个部分，其中第三部分，第四部分和第五部分是我做的主要工作。

首先是研究背景、研究意义和研究内容

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，各种社交网站也层出不穷，它们给人们的生活带来了便利同时也正在逐渐的改变人们的生活方式和社交方式。

在社交网络和互联网信息传播中的主导因素是人与人的关系。好友推荐能提升社交网站中用户的满意度，可以增强用户与社交网站之间的黏性。这对于社交网站有着重要意义，同时好友推荐功能可以应用在各种应用中，具有很强的实用性和巨大的商业价值和研究价值。

我的研究内容，首先是数据获取阶段，本文采用自行编写的爬虫程序，以迭代的方式对新浪微博数据进行爬取，并分析微博用户的行为，通过度分布图验证了数据的真实性和有效性。

构建模型阶段。本文根据微博信息的特点，进行文本语义分析，并提取文本中含有的情感程度词进行两步考虑，构建SEM好友推荐模型，并引入时间因素，综合进行好友推荐。

改进模型阶段。本文进一步对微博内容分析，采用交叉的文本相似度计算方法，并结合情感词典对用户进行情感分析，并综合考虑时间因素，提出了ESEM模型，在更大的数据集上进行验证。

最后是开发了简易版好友推荐系统。对用户的文本和情感进行分析，设计开发了融合时间因素的基于用户文本语义和情感分析的好友推荐系统，采用Django框架进行开发。

第二部分是好友推荐的相关理论基础

1是推荐系统和算法。

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。达到信息过滤的目的。

常见的推荐系统框架如图。

首先是获取用户的偏好，根据用户的偏好来构建用户模型。再根据推荐算法，从推荐对象模型中选出符合推荐的用户或物品列表，推荐给用户。其中提供个人偏好和采集个人偏好是个相互的过程。

和本文相关的好友推荐算法。

协同过滤算法和基于内容的好友推荐算法

基于协同过滤的好友推荐思想是根据用户行为预测用户的兴趣点，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。

基于协同过滤的推荐算法包括三个步骤

1. 首先是构建评分矩阵，第i行j列，表示用户i对物品j的评分或是对j的关注关系。
2. 第二步是相似度的计算。其中用户和用户在评分矩阵中的评分可以看作向量和，则可以用余弦相似度来计算两个的相似度。
3. 第三步是得出推荐结果。通常采用TOP-N和相关阈值的方法。其中N和阈值都是指定参数或者通过多次实验得出。

基于内容的好友推荐最早出现在信息检索和信息过滤领域，是一种根据用户的文本内容进行推荐的方法。一般分为三个步骤：

* 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。
* 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。
* 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。

其他的理论基础主要包括。

关键词的空间向量模型文本语义和情感分析和层次分析技术（AHP）

* 关键词的空间向量模型。这一步主要是为了提取关键词

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进行提取。

* 文本语义和情感分析

文本语义是对文本材料进行综合分析，提取关键词并利用同义词词林进行转义，并分析蕴含在用户的文本中的情感。并引入时间时间，综合考虑。

* 层次分析技术AHP

用AHP解决多指标决策问题的步骤是：

1) 问题分解，构建AHP层次；

2) 构建比较矩阵M；

3)用极差比较法构建判断矩阵C；

4)一致性验证

评价指标

本文采用的评价指标包括以下四种。从表中可以看到。当我们进行推荐时会出现以下四种情况。

则推荐指标的计算公式如下：

第三部分是基于用户文本语义和情感程度的好友推荐

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。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。

特点：

1、字数限制在140个字符以内。

2、微博的数据量很大，包含高价值的信息。

3、这些消息是用户描述近况、发表评论或抒发情感

同时微博文本短小，数据量，且格式繁杂，给预处理带来一定的困难，本文采用图中的流程进行预处理。

其中分词本文采用ICTCLAS（中科院分词系统）开放的接口进行分词。

提出的SEM推荐算法

主要分为以下四个部分。

首先是融合时间因素

然后是推荐算法的整体流程，采用二阶段式的好友推荐模型。对整体进行控制。

其中wfk是加入的时间因素控制参数，通过AHP方法计算得出。

实验部分。实验数据的采集一般主要分为两种方法，调用官方公开的API。和爬取分析网页。

本文主要采用第二种方法，因为公开的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。第二种方法较为复杂，但用户可以自己编写程序来定制爬取自己所需数据。

实验数据主要A和B两部分。在微博网络中。由于存在……..

所以对数据进行了长尾验证。

用户的出度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。

接下来就是在四个评价指标下的对比图。

……

……

……

……

第四部分是基于交叉文本相似性和情感词典的好友推荐

….

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。