**面向电影领域的微信聊天机器人**

**马晶义**

院 （系）： 计算机科学与技术学院 专业：计算机科学与技术

学　　号： 1130310723 指导教师：杨沐昀

**2017年6月**



**毕业设计（论文）**

题 目 面向电影领域的微信聊天机器人

专 业 计算机科学与技术

学　　 号 1130310723

学 生 马 晶 义

指 导 教 师 杨 沐 昀

答 辩 日 期 2017年6月

**摘 要**

目前聊天机器人是自然语言处理研究最活跃的方向之一，不论是学术界还是工业界，对聊天机器人的热衷不下于自然语言处理的其他方向。随着微信影响逐渐扩大，结合微信在电影领域做聊天机器人是一个较有趣的研究课题。

聊天机器人的实现主要有基于检索的技术和基于生成的技术，本系统采用的是基于检索的技术来设计微信聊天机器人。系统实现的第一步是数据获取，这里我们选择了三个数据源：豆瓣影评、猫眼影评和微博电影，通过网络爬虫及相应的数据处理得到系统需要的数据。本系统主要使用的检索模型有BM25模型、布尔模型和语言模型中的狄里克雷平滑方法及线性插值平滑方法。在实验中对BM25模型和线性插值模型在数据集上进行调参处理，而后在这四种模型上作了性能对比，结果显示BM25模型在本系统中应用性能最好。本系统将BM25模型应用到检索模块，与微信模块一起构建成微信聊天机器人。

关键词**：**检索模型；聊天机器人；语言模型；自然语言处理

# Abstract

At present, Chatbot is one of the most active directions of Natural Language Processing research. Whether it is in academic or industrial, the Chatbot is no more interested in other directions of Natural Language Processing. With the gradual expansion of WeChat's influence, it is an interesting research topic to chat in WeChat about the films on releasing.

There are two main methods to design Chatbot which are retrieval-based technology and generation-based technology.Our system is based on retrieval technology to achieve WeChat Chatbot. The first step in the implementation of the system is data acquisition, where we select three data sources: film reviews in douban, film comment in maoyan and Sina Weibo movies. We have gained the data needed by the system through web crawler and corresponding data processing. The main use of the system retrieval model includes BM25 model, Boolean model and language model which has Jelinek-Mercer smoothing method and Dirichlet smoothing method.In our experiment,we adjusted the parameter of BM25 model and Jelinek-Mercer smoothing method to make retrieval system perform better based on our data set.Then we made the performance comparison in the four models, and the results showed that the BM25 model in this system got the application of the best performance. This system applied the BM25 model to the retrieval module, and constructed the WeChat robot with the WeChat module.

Keywords**:**Retrieval Model,Chatbot,Language Model,Natural Language Processing

# 目 录

[**摘 要** I](#_Toc484807613)

[Abstract II](#_Toc484807614)I

[目 录 III](#_Toc484807615)

[第1章 绪 论 1](#_Toc484807616)

[1.1 课题背景及研究的目的和意义 1](#_Toc484807617)

[1.2 国内外研究现状与分析 1](#_Toc484807618)

[1.2.1 问答系统和聊天机器人的研究现状 1](#_Toc484807619)

[1.2.2 电影领域的问答系统研究现状 2](#_Toc484807620)

[1.3 主要研究内容 3](#_Toc484807621)

[第2章 聊天数据的获取加工 6](#_Toc484807622)

[2.1 聊天数据获取 6](#_Toc484807623)

[2.2 评价数据获取 8](#_Toc484807624)

[2.3 数据加工过程 9](#_Toc484807625)

[2.3.1 检索数据处理 9](#_Toc484807626)

[2.3.2 评价数据处理 10](#_Toc484807627)

[2.3 本章小结 10](#_Toc484807628)

[第3章 检索模型性能比较 12](#_Toc484807629)

[3.1 检索模型相关介绍 12](#_Toc484807630)

[3.1.1 BM25模型 12](#_Toc484807631)

[3.1.2 布尔检索模型 13](#_Toc484807632)

[3.1.3 线性插值（JelinekMercer）平滑方法 13](#_Toc484807633)

[3.1.4 狄里克雷（Dirichlet）平滑方法 15](#_Toc484807634)

[3.2 检索模型在Lucene中评分算法 15](#_Toc484807635)

[3.2.1 BM25模型的评分算法 16](#_Toc484807636)

[3.2.2 语言模型在Lucene中的评分算法 17](#_Toc484807637)

[3.3 检索模型在本系统的性能对比与分析 17](#_Toc484807638)

[3.3.1 检索模型性能的评价方法 17](#_Toc484807639)

[3.3.2 模型调参 19](#_Toc484807640)

[3.3.3 检索模型在系统性能上的比较分析 24](#_Toc484807641)

[3.4 本章小结 26](#_Toc484807642)

[第4章 整体系统设计与实现 27](#_Toc484807643)

[4.1 本系统的结构与模块 27](#_Toc484807644)

[4.2 检索模块的实现 27](#_Toc484807645)

[4.2.1 构建索引过程 28](#_Toc484807646)

[4.2.2 检索查询实现 28](#_Toc484807647)

[4.3 微信机器人模块实现 28](#_Toc484807648)

[4.3.1 与微信用户间聊天 29](#_Toc484807649)

[4.3.2 微信群中聊天 29](#_Toc484807650)

[4.4 模块连接方法 31](#_Toc484807651)

[4.5 系统演示 31](#_Toc484807652)

[4.7 本章小结 32](#_Toc484807653)

[结论 33](#_Toc484807654)

[参考文献 34](#_Toc484807655)

[哈尔滨工业大学本科毕业设计（论文）原创性声明 36](#_Toc484807656)

[致谢 37](#_Toc484807657)

[附录 38](#_Toc484807658)

# 绪 论

* 1. 课题背景及研究的目的和意义

在刚刚过去的人机大战中，AlphaGo以3:0的战绩完胜世界第一棋士柯洁证明了人工智能的强大之处。随着人工智能的逐渐发展，机器变得越来越“智能”了。近期QQ小冰在QQ群中非常火热，也证明了智能机器人在聊天软件中深受用户的喜爱。QQ小冰能够在群中通过与用户互相艾特进行交流，也能够单独与用户进行聊天。QQ小冰作为聊天的工具是足够的，但若是让小冰准确回答某一领域的问题，小冰就无法做到了，这是小冰的一大缺陷。现在微信的影响力非常巨大，有腾讯官方公布的微信数据显示微信的月活跃用户已达到8亿人次，而微信公众号作为媒体、企业及公众人物在微信平台上推送内容、建立联系的渠道，已经成为影响人们生活不可忽视的力量，因而聊天机器人在这些社交平台的作用也越来越显得重要。

近些年电影在国内的影响越来越大，特别是最近热映的电影，大都是大家茶余饭后的话题。更多的电影爱好者通过微信、QQ群加入讨论，发表各自的看法。随之而来的就会有很多用户，希望能够得到关于最新电影的信息，了解电影的动态，因此在微信上做面向电影领域的聊天机器人具有一定的研究意义和价值。聊天机器人作为自然语言处理研究最活跃的方向之一，在每年的文本检索会议TREC上，一直是一个非常热门的主题。聊天机器人不仅在学术研究上非常有价值，而且在工业界还有非常好的商业前景，例如客服机器人等。

* 1. 国内外研究现状与分析

### 1.2.1 问答系统和聊天机器人的研究现状

国外在问答系统技术上的研究起步比较早，目前在工业界和学术界已经有比较成熟的问答系统，例如 MIT的Start系统，谷歌的Google Now及cornata等。聊天机器人是自动问答技术的一个方向，之前聊天机器人的研究不受重视，但近年来智能手机普及以及电子商务的发展，聊天机器人在客户服务及社交媒体上显得很有前途，从而聊天机器人的研究火热了起来，聊天机器人中技术较为成熟的是微软小冰。

Start是由MIT大学计算机科学系人工智能实验室的研究员Boris Katz和他的同事们一起开发的，是世界上第一个基于web的问答系统。Start和信息检索系统不同，它的目标是给用户返回恰到好处的信息。Start包含两个知识库（”START KB”、”Internet Public Library”）以及一个搜索引擎。如果能够通过知识库回答用户的问题，系统就会给出准确的回答，否则返回搜索的网页链接，供用户选择。

Google Now和Cornata也都是基于搜索引擎的问答系统。其中Google Now基于用户过往的搜索习惯，预测用户可能需要的信息；Cornata是以个人数字助理为目标的，注重于帮助用户的日程安排、问题回答等。它们的核心技术都是基于搜索引擎的深度挖掘、大数据技术。

微软小冰是由微软亚洲研究院研发的聊天机器人，自2014年发布以来，获得数千万的用户量。小冰具有海量的知识库，基于搜索引擎（必应）、文本挖掘、自然语言处理结合深度学习等技术，其核心技术是情感计算框架。小冰还可以结合历史对话和用户的历史情感，来和用户进行对话聊天。

### 1.2.2 电影领域的问答系统研究现状

电影领域的问答系统，较新的是2016年MIT开发的关于电影的MovieQA[1]问答评测系统，系统采用了408部电影的视频、情节、剧本、字幕以及DVS等作为数据，经过形式化描述，通过CNN[2]模型整合为所有的文本表示。MovieQA针对每个问题，设置五个答案，其中只有一个答案是正确的，其他四个是极具有难度的干扰选项。据介绍此系统的评测性能要优于余弦相似度[3]的方法。

国内由浙大计算机学院院长陈纯及其学生构建的支持评价类问题与电影智能搜索的问答系统，采用基于聚合评论信息的灵活查询策略。允许用户灵活、自由输入对电影的描述，给用户智能推荐匹配的电影；同时依托基于tag信息查询扩展的个性化搜索技术，挖掘用户潜意识需求，不断完善电影的搜索结果。此系统针对评价类问题，设计了由语义匹配、情感极性匹配、回答句式筛选与赋权、答案的去重与重新组织组成的解决模式，完成对评价类问题的回答。其核心内容是对用户提出的问句，利用分词器进行词性分析与标注，提取核心关键词；通过核心关键词在语料库中进行检索，选取前k个回答，对这k个回答计算是否与问句的情感极性相符；利用公式计算每个回答的分数进行排序，选取最终的答案集返回给用户。

* 1. 主要研究内容

本文的主要研究内容是面向电影的微信聊天机器人设计与实现。本课题将从近期热映电影的影评入手，构建较为准确的电影领域的检索系统，并与微信进行结合，在微信中利用机器人实现与用户之间的自动聊天。整个系统的架构如图1-1所示

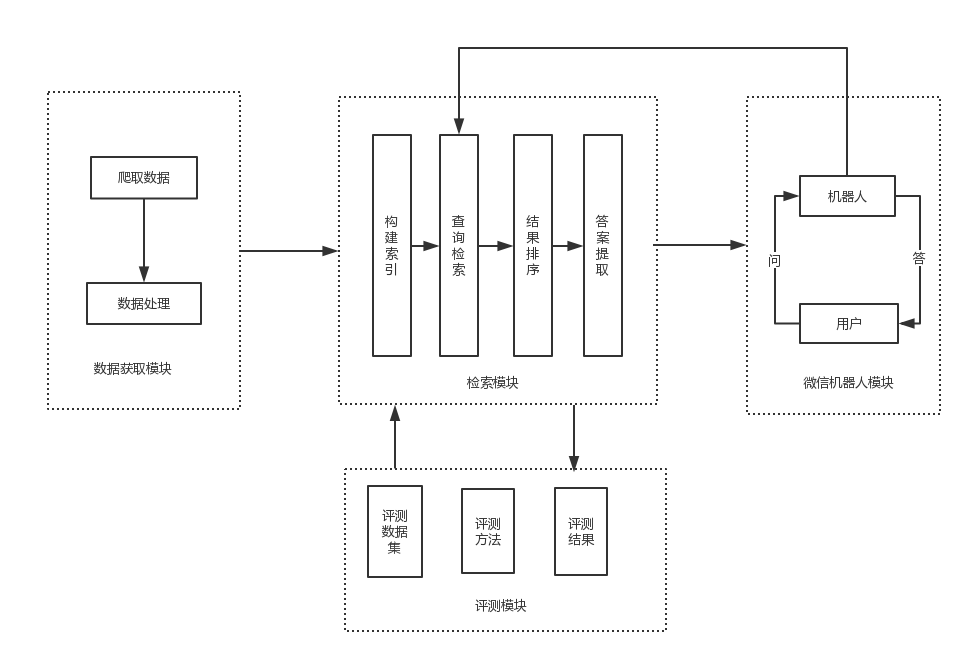


图 1-1 系统架构图

本文的主要研究内容：

1. 影评数据获取与数据处理 由于要获取的数据是最近热映电影的影评，没有现有的数据可使用，因此只能采用网络爬虫在对应的网站上爬取影评数据。选取的数据源有豆瓣影评、猫眼影评和微博电影，分别针对这三个网站进行爬取数据即可。爬虫爬取的数据，不能直接给答案检索模型使用，需要进行数据去重，生成作为检索用的文档。
2. 答案检索研究与模块设计 答案检索模块又包括四部分，分别是：构建索引、查询检索、结果排序及答案提取。在经过数据处理得到的文档，在此文档上通过Lucene分析文档提取关键字，建立倒排索引。此时获取用户的查询语句，经过词法分析、语法分析及语言处理，生成语法树，再在索引上进行查找，即可得到检索的结果。但是检索的结果一般会有很多，如何获取最相关的结果就需要对检索的结果进行排序，采用适当的排序策略获取最优的结果。但是我们获取到检索的文档结果，还不是我们最终想要的句子语言，在建立索引时设置的有各个域，这里就可以通过域来获取文档中所需要的信息，最终取得检索得到的句子，返回给用户。
3. 微信机器人模块的实现 在整个系统中，微信是直接与用户进行交流的工具，也是整个系统的应用场景。因此，利用热映电影的影评构成的知识库完成基于电影的微信机器人系统设计，将数据处理、答案检索、答案提取扩展融合到系统中去，实现能够针对用户的自然语言问题，给出较合理的回答，并且通过微信这一聊天工具与用户进行对话交流。
4. 检索结果的评测 在实现面向电影领域的微信聊天机器人系统之后，还需要做的工作是对系统性能的评测。系统的核心在于答案检索的结果是否契合用户的问题，因此整个系统的评测即对检索系统的评测。

# 聊天数据的获取加工

2.1 聊天数据获取

聊天数据获取是整个课题的首要步骤，因为无论是在答案检索模块还是在系统评测模块，都需要数据的支撑。聊天数据获取源主要是豆瓣影评、猫眼影评和微博电影，采用网络爬虫爬取三个数据源上的电影数据，然后经过针对检索和评测分别作数据处理。爬虫的结构如图2-1，本部分主要介绍数据获取的详细过程。

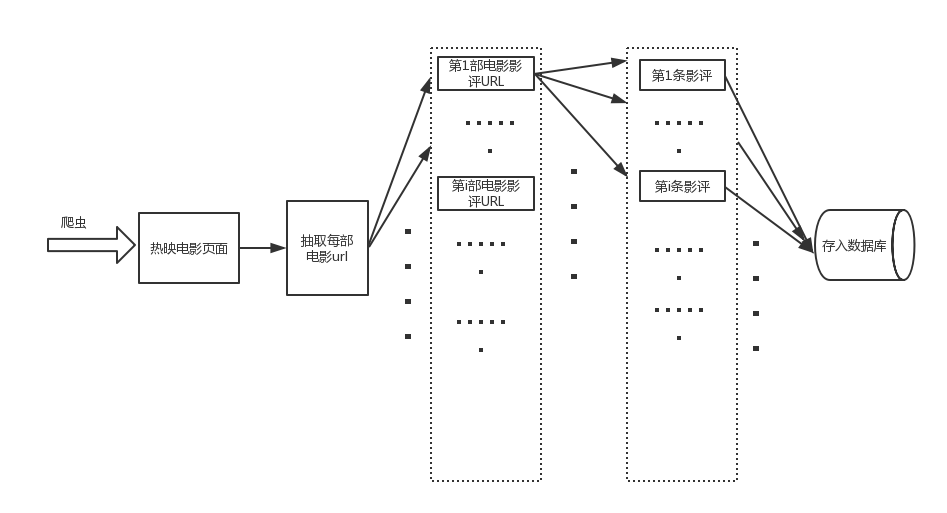


图 2-1网络爬虫结构图

针对上述三个数据源，分别设计了三个网络爬虫，用来分别在对应的网站上爬取数据。这里爬虫主要分为四个步骤：

1. 网站的模拟登陆 由于大多数网站数据量较多，很多数据需要用户登陆之后才能查看。这里的三个数据源，只要猫眼网站不需要模拟登陆。豆瓣和微博的网站都需要模拟登陆。而网站的模拟登陆的技术已经较为成熟，但其中个一个难点是验证码的问题。在模拟登录时，需要用户输入验证码，程序就需要提早下载好验证码，识别出其中的文字信息作为登陆的一部分数据。由于验证码的自动识别有一定的难度，本系统的核心也不在这个地方，因此选择进行人工识别验证码输入，然后传递参数进行模拟登陆。而有的网站登陆的参数经过了加密算法加密，例如新浪微博的登陆过程就对验证码进行了加密，因此需要知道网站的加密算法并用相同的方法进行加密，新浪微博的加密方法是base64，我们采用Python中base64方法进行加密即可。
2. 获取每部电影的链接 在模拟登陆网站之后，能够获取电影首页的页面，针对HTML网页，可以解析成复杂的树形结构，然后通过抽取网页中的标签信息从而获得标签中的内容。而对于一些无法通过标签方便解析的，也可以采用正则表达式进行提取数据。对于豆瓣和猫眼的网站，这样做是完全可以的，但由于我们爬取微博的手机网页端，获取的不再是HTML数据而是json数据，从而问题的关键是json数据的解析，只要通过Python在json数据中解析出所需的数据即可。
3. 获取影评或微博的链接 在获取电影链接之后，访问即可得到电影页面的信息，从而提取出影评的链接或微博的链接。这里的问题不是如何提取，而是在大量的影评数据，如何能够连续的获取数据。网站对于影评数据及微博都作了分页处理，因此我们总是需要获取进入下一页的方法，在电影网站上可以通过提取网站中的下一页的链接获得。而对于微博，下一页的链接需要一个sinceid的参数来组合而成，这里的sinceid则需要从上一页的json数据中提取。针对微博的评论，下一页的数据可以只用页码组合成的链接中获取，而在验证没有下一页时，只需提取json中的ok参数，当参数为1代表还有下一页，而为0时则代表没有下一页评论数据。
4. 数据存储 由于三个数据源的数据都不相同，故采用了三个数据库进行存储。在保存数据的过程中考虑到Python对数据编码处理的问题，统一设置“utf-8”编码，Python对sqlite3的支持较好，故作为获取数据的数据库，数据库存储说明如下表2-1所示。

表 2-1 获取的数据存储结构

|  |  |  |
| --- | --- | --- |
| 数据源 | 数据库表 | 说明 |
| 豆瓣 | better\_than | 存储电影的好于“百分比+电影类型” |
| comments | 存储豆瓣电影的影评数据 |
| movies | 存储电影的信息（名字、演员、简介等） |
| score\_box | 存储豆瓣电影的各种评分及评分人数 |
| 猫眼 | box | 存储电影票房数据 |
| celebrities | 存储电影的导演、编剧、演员信息 |
| comments | 存储电影的影评 |
| movies | 存储电影的详细信息（国家、上映日期等） |
| 微博 | movies | 存储电影的信息（评分、别名等） |
| movies\_detail | 存储电影的详细信息（导演、演员等） |
| weiblog | 存储微博的信息（内容、时间、用户信息） |
| weicomment | 存储评论的信息（内容、微博ID、时间等） |

## 2.2 评价数据获取

评价数据的来源主要是上一步聊天数据的一部分，因为作为评价数据要求与检索的数据有一定的相关性。这里的相关性是在同一部分电影的数据，且数据来源是一样的，而且要保证不会与检索数据重复。由于微博数据爬取时是按微博的类别爬取的，经过抽样观察数据库中的数据发现同一微博可能属于多个类别，因此在数据库中微博是会有重复的。评价数据另一个问题是需要提取出一对一对的对话，作为聊天机器人，不论是基于检索还是基于生成的，都需要这种成对的对话来评价聊天机器人的性能。

1. 数据去重及数据选取 对于数据去重，根据数据库中的微博与评论之间的主键外键关系，来达到去重的目的。微博数据中有微博ID，在新浪微博中微博ID是可以唯一识别一条微博的，因此作为微博的主键；而微博所对应的评论，也是根据微博ID知道所属于哪条微博，也就是微博ID作为评论的外键。方法是构建一个新的数据库，定义微博表的主键和评论表的外键为微博ID，以达到数据去重的目的。在获取到的微博数据中，微博有五种类型，考虑到测试数据与评价数据之间的二八原则，选择一种类型的微博作为评价数据，这里选择的是“好评微博”这一类。
2. 微博中对话抽取 微博中对话抽取思想借鉴于Hao Wang[4]提出的一种抽取方式，即对每条微博选择它评论中与之最相关的一条构成对话。本课题中计算句子的相关性采用的是bleu[5]值评价。因此，本课题中通过计算微博与其每一条评论的belu值，进行排序获取belu值最高的评论。为了保证微博与数据相关性的质量，在此设置bleu阈值对数据进行了部分筛选，才作为评价数据。

2.3 数据加工过程

### 2.3.1 检索数据处理

检索数据处理主要是将数据库中的数据提取出来，生成作为检索的文档。对于豆瓣影评和猫眼影评，数据主要分为三部分：电影信息、电影评分及电影影评。对于影评数据，直接提取影评的内容，去除内容中的一些HTML标签内容（评论内容会有URL链接），保留纯文本信息生成txt文档。对于电影信息和电影评分，通过提取数据库中的项，在项与项之间加入合理的语言描述，生成txt文档。例如数据库中对于电影《摔跤吧！爸爸》的评分信息只有8.8和6834两个数据，我们需要加工成“摔跤吧！爸爸#豆瓣评分8.8分#6834人评”这种能够理解这两个数字意思的句子。对于微博数据，则分为电影信息、电影微博及微博评论。对于电影信息，处理方式和上述类似，对于微博和评论理论上处理成对话形式，即一条微博与一条评论构成对话，将对话存入txt文档中，作为检索数据源。但本系统数据量少及这样处理在检索中较难区分对话顺序，暂就一条评论一个文档，作为检索的数据文档。

### 2.3.2 评价数据处理

在2.2节中介绍评价数据获取是来自于微博数据中的五类中的一类，即在获取去重后的微博数据库时，提取微博数据设置微博的种类为“好评微博”。在获取一条微博数据时，提取出微博ID，利用微博ID是评论的外键来从数据库中提取评论数据。此时得到评论的列表，然后利用nltk自然语言处理工具包中的bleu评分方法，计算微博与每条评论的bleu值，并将bleu值加入到评论的列表之中。通过简单的排序算法，本系统用的冒泡排序，对bleu值排序获得评论与微博的相关性序列。本系统利用设置bleu阈值来筛选与微博相关的评论，在后续模型评价中有详细的例子说明，经过筛选得到的微博与评论构成参与评价的数据。

## 2.3 本章小结

本章主要介绍聊天数据的获取与处理，首先对数据源的数据分别编写网络爬虫进行爬取数据，获取的数据量结果如表2-2。爬取的数据存入对应的数据库中，然后经过数据去重，筛选得到检索数据和评价数据，经过上述的数据处理得到本系统检索模块要使用的数据文档和作为评价数据的对话。其中得到作为检索模块的数据有174500个文档，作为评价数据的对话有1544对。未设置bleu阈值的对话是1544对，因为在评价模型方法时，筛选数据的bleu阈值和要和搜索结果与候选结果的bleu阈值一样的，在实验中进行调整阈值。综上，本章的主要工作是从数据源获取数据，并为检索模块和评价模块提供可用的数据，分别是174500个文档和1544对对话。

表2-2 数据获取结果

|  |  |  |
| --- | --- | --- |
| 数据源 | 数据量 | 说明 |
| 豆瓣影评 | 21012条影评 | 12部电影的长短影评数据 |
| 猫眼影评 | 11854条影评 | 12部电影的猫眼影评数据 |
| 微博电影 | 16779条微博83799条评论 | 23部热映电影的微博及对应的评论 |

# 检索模型性能比较

## 3.1 检索模型相关介绍

本课题采用的检索模型主要有BM25[6]模型、布尔模型、及语言模型中的线性插值（JelinekMercer）平滑方法和狄里克雷（Dirichlet）平滑方法[7]。

### 3.1.1 BM25模型

BM25模型是基于二元独立模型的概率模型，是目前检索上最有效和流行的排序算法之一。BM25模型通过加入文档权值和查询项权值，拓展了二元独立模型的得分函数，这种拓展是基于概率论和实验验证的，并不是一个正式的模型。BM25最普遍的得分函数如(3-1)所示，其中K值在(3-2)中定义：

 (3-1)

 (3-2)

公式(3-1)中 是包含词项i的文档数目，N是整个数据集中文档的数目， 是包含i的相关文档数目，R是和这个查询Q相关的文档数目， 是词项i在文档中的频率， 是词项i在查询项中的频率，、和 K都是经验设定的参数。公式(3-2)中b是一个经验参数，dl是文档长度，avdl是数据集中文档的平均长度[8]。

在BM25模型的计算中，对于查询Q中出现的每个查询词，依次计算每个词在文档D中的分值，累加后就是文档D与查询Q的相关性得分。公式（3-1）可以拆成三部分，第一部分组成就是二元独立模型的计算得分，设置默认值的话等价于IDF因子的作用；第二部分代表的是查询词i在文档D中的权值；第三部分则代表查询词在查询Q的权值。由此公式可以看出BM25模型是在二元独立模型的基础上，整合了查询词在文档的权值和查询词在查询中的权值，得到计算相关性的公式。BM25中包含三个自由调节参数，在TREC试验中调节因子b设置为0.75被证明是有效的；参数决定增加时 部分的词项权值如何变化，即当=0，词项频率部分将被忽略，只有词项的存在是否会有所影响，当很大时，词项权重部分会随着线性增长，在TREC实验中=1.2是有效的，因此词频的影响是很大的，由于log函数的影响，在词出现次数越多时，影响越低。参数 在查询项权重中有类似的作用，参数范围是0指1000较大的范围，值较大很多，代表系统性能对的敏感性较低。之所以如此设置是因为查询一般较短，查询词的频率比较低，查询词之间的频率差异很小，较大的调节参数范围数值可以对这种差异放大。

### 3.1.2 布尔检索模型

布尔检索模型用在最早的搜索引擎之中并沿用至今。因为被检索的文档都能精确匹配检索需求，不满足的文档不会被检索到，故又称精确匹配检索[9]。布尔检索模型假设在检索到集合中，所有文档的相关性都是等价的且是二元的。在检索评价中只有两种输出结果(TRUE和FALSE)，并且查询项往往被描述为布尔逻辑操作符。

布尔检索有很多优点，模型的结果很容易推断并向用户解释。在布尔查询项的运算域不仅仅是词语，而可以是任何文档特征，因此可能在检索规范中融入元数据。从实现的角度看，由于文档在评分过程中更容易快速剔除，布尔检索往往比排序检索更加有效。

### 3.1.3 线性插值（JelinekMercer）平滑方法

由于在多项式分布中，如果查询项中的任意一个词语没有在文档中出现的，那么查询似然模型给的分值就会是0，不适用于较长的查询。平滑技术用于避免这种估计问题以及数据稀疏问题，这意味着我们不需要使用大量文本来估计语言模型的概率。平滑技术一般的方法是降低文档文本中出现词语的估计概率，并对文本中未出现的词语赋给估计的“剩余”概率。未出现词语的概率通常都是基于整个文档数据集中词语的出现频率来进行估计的。如果 是文档集合C的数据集语言模型中词语i的出现概率，那么文档中未出现词语的估计概率为 ，其中 是控制未赋予未见词语概率的系数。一般来说， 依赖于文档。为了保证概率值和为1，文档中一个出现过的词语概率被估计为 。

的赋值方式导致不同估计形式，选择设定为常数得到公式(3-3) 。集合语言模型概率估计中，估计词语 的概率为 ，其中  是文档数据集中查询词出现的次数，|C|是集合中所有词语出现次数的总和。

 (3-3)

这种形式称为Jelinek-Mercer方法。在查询似然模型的文档得分中替换掉这个估计，得到

 (3-4)

由于在连乘较小值会导致精确率的问题，采用取对数的方法将得分转换为等价排序求和得到

 (3-5)

 较小造成较小的平滑，查询会更像布尔操作AND，因为任何查询的缺失都会在本质上惩罚这个得分，更进一步极大似然估计得到的词语的相关性权值对确定分值是重要的。如果 越接近1，相关性权值会变得越不重要，此时查询会变得更像布尔操作OR或协调水平匹配。TREC评测中，实验证明短查询参数 接近0.1效果较好；长查询参数 接近0.7效果较好。短查询倾向于只包含显著性的词语，较低的值会偏爱包含所有查询词的文档，在较长的查询中，缺少一个词语的影响较小，较高的值会更加强调包含许多高概率词语的文档。

### 3.1.4 狄里克雷（Dirichlet）平滑方法

上述中说明，一种不同的更加有效的估计形式，来自于使用依赖于文档长度的变量。若对作公式 (3-6)设置，即为狄里克雷（Dirichlet）平滑方法。

 (3-6)

其中 是经验参数。在中替换掉得到

 (3-7)

从而推导得到如下的文档得分函数：

 (3-8)

参数较小的数值对词语的相关性权值赋予了更大的重要性，较大的数值偏向于匹配上的词项数目。在TREC实验中，达到最佳效果的的数值范围在1000到2000之间。一般的，狄里克雷平滑方法优于线性插值平滑方法，尤其是在短查询的搜索中。

狄里克雷分布在多项式分布的概率估计中考虑先验知识的自然方式。贝叶斯估计确定估计概率的过程就是基于这种先验知识和观察文本的。最终的概率估计可以看成是结合了文本中的实际词项计数以及来自狄里克雷分布的伪计数。如果没有文本，词项 的概率估计可以是 。拥有更长的文档，先验知识就具有越小的影响。

## 3.2 检索模型在Lucene中评分算法

输入用户的查询用户之后，在经过词法分析、语法分析及语言处理，生成语法树在索引上检索出结果后，结果的排序是非常重要的。我们总是希望与问题最相关最契合的答案排在前面，因此结果的评分算法尤为重要，公式(3-9)是Lucene[10]的相似度评分公式，用来衡量查询语句和对应匹配文档之间的相似度。本课题在Lucene中使用了BM25算法、布尔模型、线性插值平滑和狄里克雷平滑方法作为评分算法，其中布尔模型较为简单，以下不再介绍。

 (3-9)

表 3-1 Lucene评分公式中因子

|  |  |
| --- | --- |
| 评分因子 | 描述 |
|  | 协调因子（Coordination factor），基于文档中包含查询项的个数，该因子会对包含更多搜索项的文档进行类似AND的加权 |
|  | 每个查询的归一化值，指每个查询项权重的平方和 |
|  | 项频率因子­--文档(d)中出现项(t)的频率 |
|  | 项(t)在倒排文档中出现的频率：被用来衡量项的“唯一”性，出现频率较高的term具有较低的idf；出现较少的term具有较高的idf |
|  | 域和文档的加权，在索引期间设置，可以通过此项对某个域或文档进行静态加权 |
|  | 域的归一化(Normalization)值，表示域中包含的项数量，该值构建索引期间计算，并保存在索引norm之中。对于该因子，更短的域（或更少的语汇单元）能获得更大的加权 |

### 3.2.1 BM25模型的评分算法

由于在典型的情况下，没有相关信息，即公式（3-1）中的参数R和 设置为0，得到如下公式(3-10)

 (3-10)

而在Lucene中评分集中体现在idf和tf上，因此它的实现过程就由如下两个公式构成

 (3-11)

 (3-12)

从而得到评分函数 。在检索应用中使用BM25评分算法，将BM25算法实现的类封装，在Lucene构建索引或检索的过程中通过setSimilarity方法将评分算法设置为BM25模型即可，Lucene中BM25模型参数默认设置k1=1.2，b=0.75，实验中可以通过调参来验证。

### 3.2.2 语言模型在Lucene中的评分算法

Lucene中使用语言模型来评分的有线性插值（JelinekMercer）平滑方法和狄里克雷（Dirichlet）平滑方法。在Lucene中应用的方法和应用BM25算法类似，均是通过setSimilarity方法设置。JelinkMercer平滑参数 可以通过在调用LMJelinekMercerSimilarity类方法设置，从3.1节中说明短文本查询的一般设置为0.1，实验中通过调参来验证。Dirichlet平滑的参数 默认设置为2000，在3.1节中参数正常范围内，无法修改，故实验中直接使用默认值。

## 3.3 检索模型在本系统的性能对比与分析

### 3.3.1 检索模型性能的评价方法

本系统的评价主要是评价聊天机器人给出的回答与问题答案的相关性，通过准备好的评价数据——微博中提取的对话，来评价检索系统的性能。由于对话是由微博和评论构成一对一对的，故本实验拿微博作为查询语句，拿检索的结果与构成对话的微博对应的评论计算相关性。本实验中采用了bleu值来评价检索结果与评论之间的相关性，通过在实验中采集检索的样本分析知，bleu值较高的，检索的结果与评论的相关性较高，如表3-2所示；bleu值较低的，检索的结果与评论的相关性较低，如表3-3所示。

表 3-2 bleu值高的样本展示

|  |  |  |
| --- | --- | --- |
| 类别 | 内容 | Bleu值 |
| 微博 | 摔跤吧！爸爸[半星]，男尊女卑 如此自私的父亲何来伟大之处 也是印度这个社会的悲哀 | - |
| 评论 | 一些缺乏基础知识的人总是觉得自己站在真理的一端，并且喜欢用自己观点任意judge别人的观点，心疼楼主十秒钟 | 0.5542089483371553 |
| 检索结果 | 每个人都有自己的观点，胖迪的粉丝不要骂博主。我们自己喜欢便好，不要没素质 | 0.5313102542584737 |

表 3-3 bleu值低的样本展示

|  |  |  |
| --- | --- | --- |
| 类别 | 内容 | Bleu值 |
| 微博 | 神秘家族[星星][星星][星星][星星]，约！！！@陈晓 @电影神秘家族 | - |
| 评论 | 窝窝《爱微吧君》爱#神秘家族#更爱#陈晓# | 0.08807755569397381 |
| 检索结果 | 神秘家族中陈晓演树树 | 0.042277929617726875 |

检索模型中最常用的评价方式是召回率(recall)和准确率(precision)，在此评价中我们设置Bleu阈值作为筛选规则。由于我们选取的评价数据是通过计算微博与评论之间的bleu值，取其中bleu值最大的评论作为结果，与微博构成一对对话。在这里设置bleu阈值，本课题认为微博的排序最靠前的评论的bleu值达不到所设的阈值表明此微博和评论无法构成对话，不能作为评价的数据。在检索出的结果，计算结果与对应评论之间的bleu值，若bleu值高于之前设置的阈值，表明在此数据集上检索的结果与评论是相关的，若无法达到即为不相关。

召回率 = 检索结果的bleu值达到阈值的数目/没有设阈值的数据集数目

准确率 = 检索结果的bleu值达到阈值的数目/同样bleu阈值的数据集数目

召回率和准确率的计算规则如上规定，由于设置的bleu阈值对召回率的影响非常大，严重会影响系统的评价性能，这里设定准确率作为系统的主要评价指标。

### 3.3.2 模型调参

#### 3.3.2.1 BM25模型调参

在3.2.1节中介绍BM25模型在Lucene中的评分公式中有两个经验参数和b，根据TREC实验验证=1.2,b=0.75效果最好，本实验是验证在本系统的数据集上这两个经验参数如何设置效果最好。已知=1.2是较好的默认值，因此我们选取的调参范围就在1.2的周围。如下表3-4所示为当bleu阈值设置为0.4时，BM25模型在经验参数变化的准确率的情况。图3-1所示可以看出，在本系统上BM25模型的经验参数=1.4的性能相对较好。

表 3-4 BM25模型参数k1的调参

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| bleu值 | k1=1.0 | k1=1.1 | k1=1.2 | k1=1.3 | k1=1.4 | k1=1.5 | k1=1.6 |
| 0.4 | 0.3154 | 0.3190 | 0.3272 | 0.3454 | 0.3646 | 0.3636 | 0.3272 |

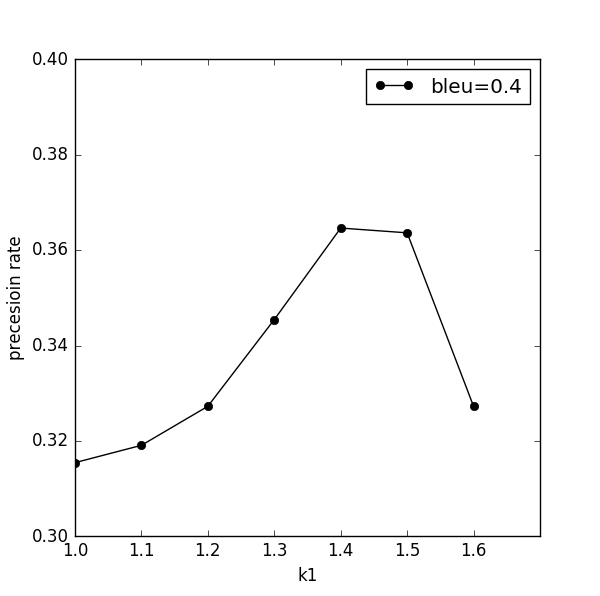


图 3-1 BM25模型准确率随k1变化情况

已知BM25模型的经验参数b=0.75时在TREC实验中性能较好，实验中设置0.55到0.95之间每隔0.1测试一次系统的性能，发现准确率在b=0.65最好，接着加入b=0.6和b=0.7来进行验证，如表3-5所示为设置b值所对应的系统的准确率。由图3-2可以看出准确率曲线在b=0.7后开始下降，在之前是线性增长的。因此b=0.7作为本系统中BM25模型的经验参数相对较好。

表 3-5 BM25模型的参数b调参

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| bleu值 | b=0.55 | b=0.6 | b=0.65 | b=0.7 | b=0.75 | b=0.85 | b=0.95 |
| 0.4 | 0.3636 | 0.3616 | 0.3818 | 9.3818 | 0.3272 | 0.3272 | 0.1818 |

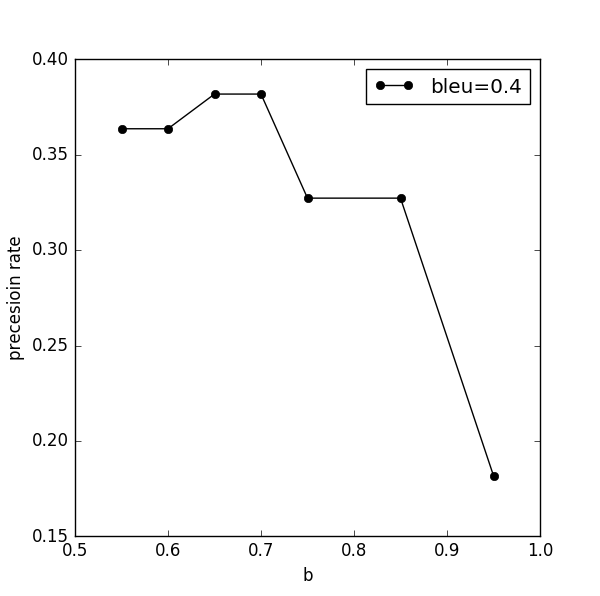


图 3-2 BM25模型准确率随参数b的变化情况

BM25调参中设置bleu阈值为0.4是因为在本数据集中0.4相对较好的衡量相关性还能保证数据量不会因为少量变化较大影响准确率的问题。经过此次调参，本系统中BM25模型设置经验参数=1.4、b=0.7相对较好，故作为以下内容的BM25模型的默认参数。

#### 3.3.2.2 JelinekMercer平滑方法调参

由3.2节中所述，JelinekMercer的参数 一般设置为0.1。表3-6所示为设置为0.1、0.3、0.5、0.7、0.9的结果，图3-3即是根据此数据作出的，图中lamda即为，可以形象显示随着bleu阈值设置越高，系统的准确率越低，可以理解当bleu阈值较高，能够达到要求的评价数据集越小，从而准确率受影响越大；还可以看出随着参数越接近1，系统的准确率越低，值设为0.1可以是较好的选择，进一步验证了3.1.3中TREC实验的验证结果，在短文本查询中值为0.1效果较好，越接近1则相关性权值越不重要。

表 3-6 JelinekMercer检索微博参数设置

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Bleu阈值 | =0.1 | =0.3 | =0.5 | =0.7 | =0.9 |
| 0.01 | 0.8469 | 0.8239 | 0.8048 | 0.7767 | 0.6658 |
| 0.1 | 0.7522 | 0.6978 | 0.6676 | 0.6012 | 0.4138 |
| 0.2 | 0.4892 | 0.4731 | 0.4301 | 0.3978 | 0.2580 |
| 0.3 | 0.3173 | 0.2884 | 0.2596 | 0.2307 | 0.1153 |
| 0.4 | 0.2 | 0.1818 | 0.1454 | 0.1454 | 0.0545 |
| 0.5 | 0.0384 | 0.0384 | 0 | 0.0 | 0.0 |
| 0.6 | 0.0 | 0.0 | 0 | 0.0 | 0.0 |

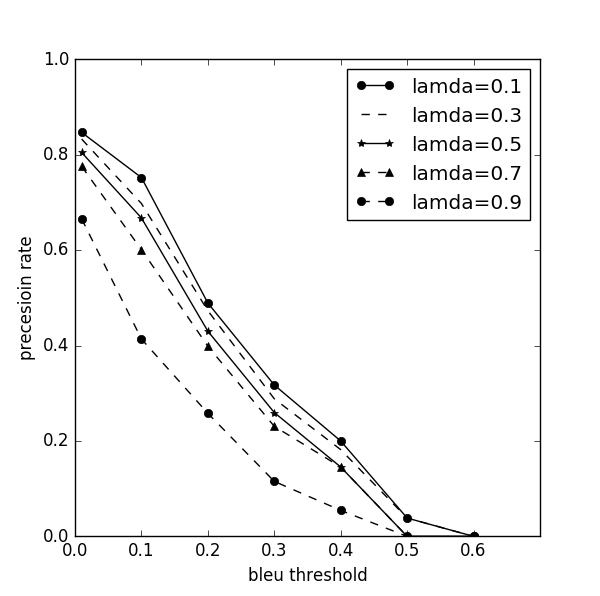


图 3-3 JelinekMercer检索微博中参数影响

在实验中，对话是由微博和评论构成的，理论上微博在前评论在后，即对话中微博作为第一句，评论作为第二句。这里选择检索评论，计算结果与微博的bleu值，评价系统的准确率。由图3-4中可以看出，参数设置为0.1的优势不是很大，图中lamda即为。

表 3-7 JelinekMercer检索评论

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Bleu阈值 | =0.1 | =0.3 | =0.5 | =0.7 | =0.9 |
| 0.01 | 0.8367 | 0.8316 | 0.8188 | 0.7729 | 0.6352 |
| 0.1 | 0.6646 | 0.6737 | 0.6465 | 0.6314 | 0.4864 |
| 0.2 | 0.4462 | 0.4408 | 0.4354 | 0.4569 | 0.3978 |
| 0.3 | 0.3557 | 0.375 | 0.375 | 0.4038 | 0.375 |
| 0.4 | 0.2363 | 0.2727 | 0.2545 | 0.2909 | 0.2727 |
| 0.5 | 0.1153 | 0.1538 | 0.1153 | 0.1538 | 0.1153 |
| 0.6 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 |

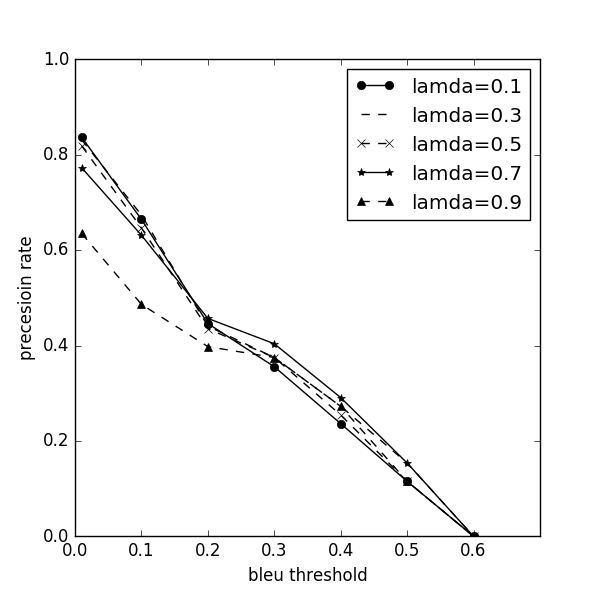


图 3-4 JelinekMerce检索评论中参数影响

图3-5对在JelinekMerce检索微博和检索评论的性能上有一个对比，从图中可以看出检索微博的系统性能要略优于检索评论的系统性能。

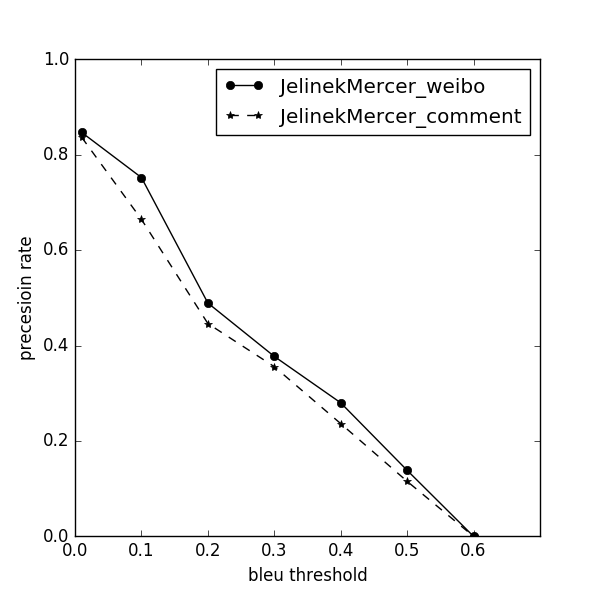


图 3-5 JelinekMercer检索微博和评论的差异

### 3.3.3 检索模型在系统性能上的比较分析

检索模型在系统性能上的评价方法在3.2节中已有描述，表格3-8为课题中所用到的检索模型在检索微博上的准确率。图3-6是其直接表现形式，从图中看出虽然在Bleu阈值较低时,BM25算法性能还没有优势，但在bleu值较低时说明相关性也较低，对系统的性能评测帮助不大；随着bleu阈值增大，可以看出BM25算法的性能的优势较明显，而且通过对文件结果采样得到bleu阈值在0.5以上可以看出有明显的相关性。

表 3-8 检索模型准确率对比

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Bleu阈值 | BM25 | BOOL | Dirichlet | JelinekMercer |
| 0.01 | 0.8163 | 0.8992 | 0.8163 | 0.8469 |
| 0.1 | 0.7552 | 0.7945 | 0.6948 | 0.7522 |
| 0.2 | 0.5752 | 0.4892 | 0.4516 | 0.4892 |
| 0.3 | 0.4326 | 0.375 | 0.3557 | 0.3173 |
| 0.4 | 0.3272 | 0.2909 | 0.2909 | 0.2 |
| 0.5 | 0.1538 | 0.1538 | 0.1154 | 0.0384 |
| 0.6 | 0 | 0 | 0 | 0 |

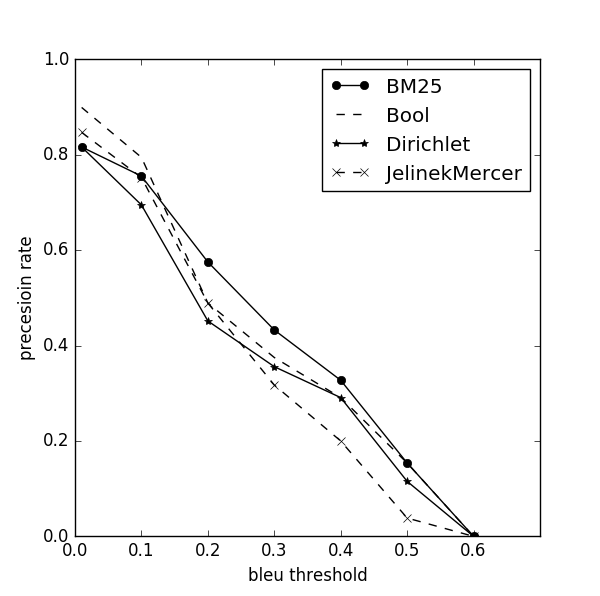


图 3-6 模型在检索微博中的比较

为了进一步说明BM25算法的优越性，实验中还进行检索评论，计算检索结果与微博的相关性，虽然没有检索微博的性能好，但从图3-7中也能明显看出BM25相对其他检索模型的优势。故整体来看，在BM25模型、布尔模型和语言模型中的狄里克雷平滑及线性插值平滑方法中，BM25模型在本系统中有明显的优势。

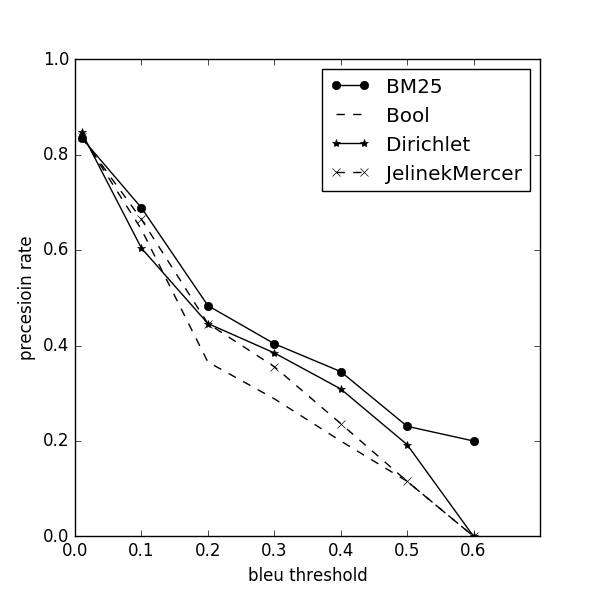


图 3-7 模型在检索评论中的比较

## 3.4 本章小结

本章中介绍了本课题中要使用的检索模型及其原理，并设定一种评价方法来评价系统的性能。在模型调参的结果中验证了TREC实验中有效的经验参数，并对实验结果进行对比。在对所用的检索模型性能评价中，通过两种方式验证了BM25模型在本系统中具有的优越性。因此，在本课题的微信聊天机器人中设置的检索模型为BM25模型，并且经验参数 设置为1.4，参数b设置为0.7,以求本系统达到相对最好的性能。

# 整体系统设计与实现

## 4.1 本系统的结构与模块

该系统设计分为两个主要模块：后台端（检索模块）和微信端（聊天机器人模块）。如图4-1所示，后台端与微信端之间的通信。检索模块本课题使用Lucene检索工具采用java语言实现，微信端使用Python3的wxpy包进行实现，因为是不同的程序语言，采用了Jnius包来实现跨语言调用类，即在Python中调用java类，实现这两个模块的连接。

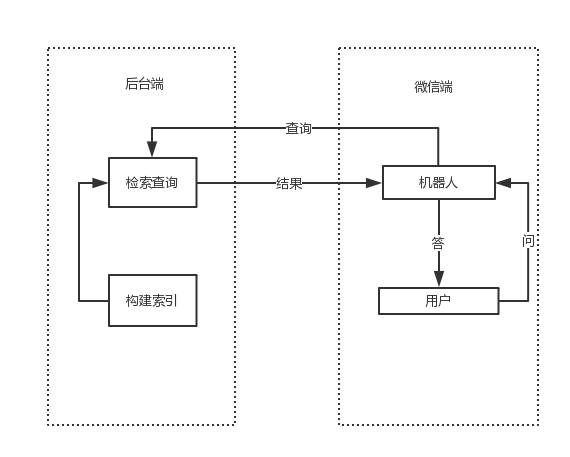


图 4-1 系统框架

## 4.2 检索模块的实现

检索模块主要有两个实现步骤：构建索引和检索查询。在第2章内容中通过对获取的数据加工，生成供检索模块使用的文档数据，在这里对这些文档数据建立Lucene中的倒排索引，然后针对用户的问题生成对应的查询树，在索引中检索查询，提取出检索的结果。

### 4.2.1 构建索引过程

本部分主要说明对第2章中得到的检索文档进行构建索引的过程，在Lucene中索引构建主要有一下四个步骤：

1. 将文档传给分词组件，分词组件会将文档内容分成一个一个字（对于中文）,去除标点符号，去除停用词，得到词元
2. 将词元传给语言处理组件，语言组件对英文单词变成小写以及将单词转成词根形式。对于中文则不会有此过程
3. 将得到的词传给索引组件[12]，将词创建为一个字典，并排序，合并相同的词生成倒排链表
4. 通过索引存储将索引写入影评

### 4.2.2 检索查询实现

搜索的过程针对输入的语句，在上一步构建的索引中搜索结果，然后对结果进行排序，获取排序最靠前的结果返回。这里的排序算法在第3章已有详解，本部分主要说明这里搜索的过程，主要有以下五个步骤：

1. 对查询语句经过语法分析和语言分析得到一系列的词
2. 通过语法分析得到一个查询树
3. 通过索引存储将索引读入内存
4. 利用查询树搜索索引，进而得到每个词的文档链表，对文档链表进行并交差运算得到结果文档
5. 将搜索得到的结果文档对查询的相关性进行排序，取排序靠前的结果文档

## 4.3 微信机器人模块实现

微信机器人的账号设置和普通用户的账号设置一样，因此机器人和普通用户一样可以单独和好友聊天，也可以通过在微信群中与群成员互动。在与微信好友进行聊天时，机器人有自动接受好友请求、自动拉好友进电影讨论群和自动回复好友消息（文字、语音和图片）三个功能。在微信群中进行聊天时，机器人有自动回复艾特机器人的消息（仅包含文字）、剔出不遵守群规则的成员以及在无法回答问题时进行转发问题搜集答案三个功能。

### 4.3.1 与微信用户间聊天

微信机器人与用户之间进行交流，实现了三个功能：自动接受好友请求、自动拉好友进群及自动回复好友消息（文字、语音和图片）。

1. 自动接受好友请求功能 用户加机器人微信为好友，机器人识别系统消息为加好友消息，自动接受好友请求。实现方式是通过wxpy中的消息注册机制，注册好友消息，然后识别是否是好友请求消息。在成功加用户为好友时，机器人会自动回复“我是小影机器人，我们来聊天吧！”来通知用户并邀请用户与机器人聊天。
2. 自动拉好友进群功能 本系统的目的是为了机器人参与到用户与用户之间的讨论之中，因此实现拉取好友进群讨论是一个必备功能。本系统在添加用户为好友时会给用户发一条消息“欢迎入群讨论热映电影啦！请回复：我要加群”，这里的“我要加群”即为入群的“暗号”，机器人在获取消息中解析出此“暗号”，即可通过add\_member方法将用户添加入群。当然，在入群前，会对群成员上限检测（本系统设置群成员上限）以及用户是否已经加过此群进行检测。
3. 自动回复好友消息功能 自动回复好友消息功能是机器人的主要功能，因为此功能与用户之间进行互动才能体现机器人的作用。自动回复好友的消息有三种：文本、语音和图片。文本消息是本系统的核心，获取用户的文本消息进行加工传送给检索模块，通过检索模块获取检索的结果，经过机器人加工之后发送给用户。对于语音消息，本系统暂未做语音识别功能，因此对用户的语音消息给出“本机器人暂不能语音，请发文字消息”的提示消息。对于图片消息，在微信上主要是为了表情包斗图，本系统未涉及图像识别领域，因此每次在表情包库中随机选择一张表情图片返回给用户。

### 4.3.2 微信群中聊天

如图4-2给出微信机器人关于微信群中聊天的设计流程图，展示了微信机器人在群聊中的两个核心功能：自动回复群中艾特机器人的消息和搜集机器人无法回答问题的答案。

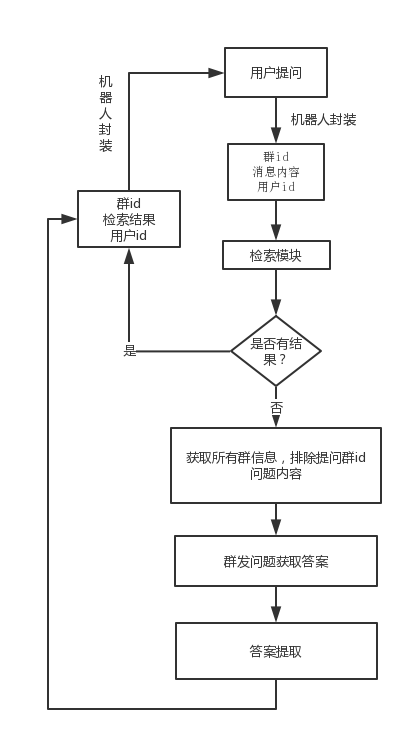


图 4-2 机器人设计流程图

自动回复群中的消息的实现机制和自动回复好友消息的机制相同，只是群中是通过艾特消息来提醒是对机器人的提问，而艾特消息只能是文本消息，消息类别就比较单一。搜集机器人无法回答问题的答案功能实现主要是依靠众包的思想，本系统是基于检索的聊天机器人，本系统的检索数据集是有限的，那就必然存在问题无法检索到结果的情况，此时我们将这种问题群发到非提问者的群中，交由其他用户提供答案，机器人获取其他用户的回答后整合成微信消息返回到提问者的群中。在这里设置了数据库用来存储搜集的问题和答案，暂未对问题获取的答案进行筛选，而是直接使用。群聊机器人的另一个功能是剔出不遵守规则的用户，这里我们在群中设置管理员，管理员在群中发现有用户不遵守规则的用户，通过发送“踢出@用户名”消息，机器人在群中删除此用户。

## 4.4 模块连接方法

由于微信机器人设计采用的是Python语言而检索模块的设计则为Java语言，需要将两个模块进行连接，即用Python调用Java语言的类包，实现跨语言调用的功能。本课题采用Jnius包来实现此功能。

首先将Java实现的检索类封装成jar包，并对其中调用的第三方jar包进行封装与在MINIDEST设置对应的路径。在Python代码中设置jvm的路径和所占用的空间，本系统设置为1024兆的空间。初始化调用的类，之后使用java的类方法传值给Python即可。

## 4.5 系统演示

如图4-3是微信聊天机器人在微信群中回答用户的问题的过程演示，如图所示用户通过在群中艾特机器人小影，小影机器人会立刻拿到检索的结果返回给用户，并艾特用户给予提醒。

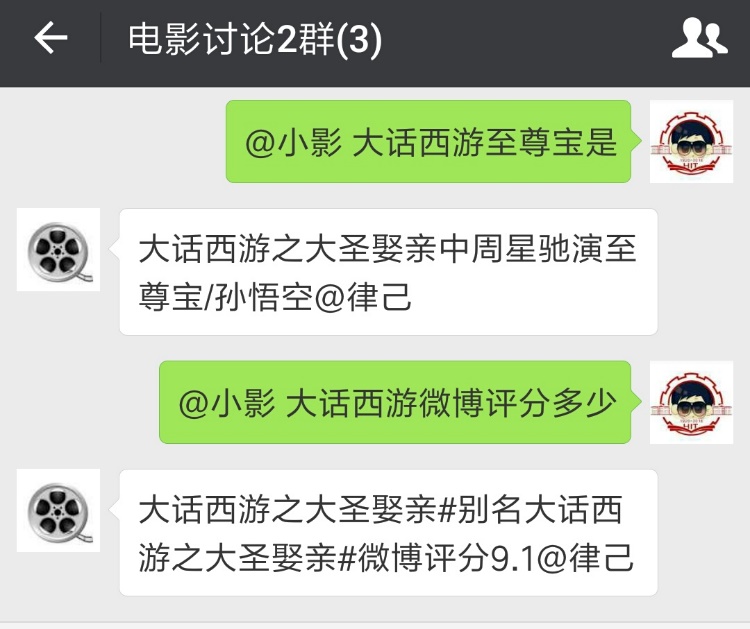


图 4-3 聊天机器人演示

## 4.7 本章小结

本章主要介绍了微信聊天机器人的系统结构以及各个模块的设计方法。首先介绍了核心的检索模块中如何利用Lucene构建索引及检索查询，而后介绍了系统应用的微信模块，详细说明了各个功能的设计过程；然而两个模块之间是跨语言开发的，因此又简单介绍了模块之间的连接方法，并最后给出了系统的功能演示。

# 结论

本课题是研究面向电影领域的微信聊天机器人的设计与实现，目的是能够在利用微信与自然语言处理相关知识实现聊天机器人，并让其参与到用户关于电影的讨论中。聊天机器人实现目前有基于检索和基于生成的技术可以采用，本课题采用的是基于检索的方法实现的，因此整个机器人系统的核心是检索模块的设计。在本课题的实验中，通过bleu值的设置来计算检索模型的准确率，用来评价检索系统的性能。在实验中，我们通过对相关模型进行调参，通过准确率的评价选择在检索系统中性能最好的模型参数作为本系统的模型参数。本系统中用到的四种检索模型：BM25模型、布尔模型和语言模型中的狄里克雷平滑方法及线性插值平滑方法，在系统的评测实验中，我们对这四种模型进行了对比，结果显示BM25模型在本系统中的效果最好。因此本系统的聊天机器人的检索模型设置为BM25模型。

在评测系统性能时，可以看出随着bleu阈值设置的越高，系统的准确率在急速下降，但只有bleu值较高时，查询与检索结果的相关性才更能体现出来。而在评价句子相关性上，bleu值作为评测标准是不够充分的，未来我们可以尝试用向量空间模型或编辑距离[13]来计算句子的相关性。由于Lucene的检索模型使用的分词对中文支持不友好，我们还可以实现效果更好的中文分词，结合到Lucene中去，以提高检索模型的性能。本系统的检索数据不是很充足，因此在提高系统的性能上，我们还需要获取更多有用的数据。

# 参考文献

1. Wang Z, Zhang Y. A Text Information Retrieval Method by Integrating Global and Local Textual Information[C] IEEE, Computer Software and Applications Conference. IEEE Computer Society, 2016:504-505.
2. Tian Y, Stewart C M. Framing the SARS crisis: A computer-assisted text analysis of CNN and BBC online news reports of SARS[J]. Asian Journal of Communication, 2005, 15(3): 289-301.
3. Shum S, Dehak N, Dehak R, et al. Unsupervised Speaker Adaptation based on the Cosine Similarity for Text-Independent Speaker Verification[C]//Odyssey. 2010: 16.
4. wang Hao,Lu Zhengdong,Li Hang,Chen Enhong.A Dataset for Research on Short-Text Conversation[A].In Proc.of the 2013 Conference on Empirical Methods on Natural Language Processing[C].Seattle,AMERICA,2013,935-945.
5. Papineni K, Roukos S, Ward T, et al. BLEU: a method for automatic evaluation of machine translation[C]//Proceedings of the 40th annual meeting on association for computational linguistics. Association for Computational Linguistics, 2002: 311-318.
6. Robertson S, Zaragoza H. The probabilistic relevance framework: BM25 and beyond[J]. Foundations and Trends® in Information Retrieval, 2009, 3(4): 333-389.
7. Zhai C, Lafferty J. A study of smoothing methods for language models applied to Ad Hoc information retrieval[C]// International ACM SIGIR Conference on Research and Development in Information Retrieval. ACM, 2001:334-342.
8. Brucecroft W, DonaldMetzler, TrevorStrohman, et al. Search engines:Information retrieval in practice[M]. 机械工业出版社, 2009:277-286.
9. Salton G, Fox E A, Wu H. Extended Boolean information retrieval[J]. Communications of the ACM, 1983, 26(11): 1022-1036.
10. Gennaro C, Amato G, Bolettieri P, et al. An approach to content-based image retrieval based on the Lucene search engine library[C]//International Conference on Theory and Practice of Digital Libraries. Springer Berlin Heidelberg, 2010: 55-66.
11. Wan J, Pan S. Performance evaluation of compressed inverted index in lucene[C]//Research Challenges in Computer Science, 2009. ICRCCS'09. International Conference on. IEEE, 2009: 178-181.
12. Bennett G, Scholer F, Uitdenbogerd A. A comparative study of probabilistic and language models for information retrieval[C]//Proceedings of the nineteenth conference on Australasian database-Volume 75. Australian Computer Society, Inc., 2008: 65-74.
13. Che W, Liu T, Qin B, et al. Similar Chinese sentence retrieval based on improved edit-distance[J]. High technology letters, 2004, 14(7): 15-20.

# 哈尔滨工业大学本科毕业设计（论文）原创性声明

本人郑重声明：在哈尔滨工业大学攻读学士学位期间，所提交的毕业设计（论文）《面向电影领域的微信聊天机器人》，是本人在导师指导下独立进行研究工作所取得的成果。对本文的研究工作做出重要贡献的个人和集体，均已在文中以明确方式注明，其它未注明部分不包含他人已发表或撰写过的研究成果，不存在购买、由他人代写、剽窃和伪造数据等作假行为。

本人愿为此声明承担法律责任。

作者签名： 日期： 年 月 日

# 致谢

本论文是本人在导师杨沐昀副教授的悉心指导下完成的。从毕设选题到论文的完成，杨老师给了我很多建议和指导，花费了杨老师很多宝贵的时间与精力。在每次遇到问题和困难时，杨老师都会给予及时的帮助与支持。在与杨老师相处过程中，也让我明白了很多人生道理而受益匪浅。在此谨向杨老师致以诚挚的谢意和崇高的敬意。

衷心的感谢实验室的张越师兄。张越师兄在信息检索方面经验丰富，在我的毕业设计中给予了很多的指导和建议。感谢实验室的杨艳师姐和赵晶晶师姐，在我完成论文中提供了很多的帮助。

感谢陪伴我的朋友和同学，在我的生活中给予了很多帮助，给我带来丰富多彩的人生。

最后感谢我的父母，在我身后一直无条件的支持着我，为我付出。

# 附录