****

**毕业设计（论文）开题报告**

**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 生 马 晶 义**

**学 号 110310723**

**指导教师 杨 沐 昀**

**日 期 2017.3.30**

**哈尔滨工业大学教务处制**

1. 课题来源及研究目的和意义

近期QQ小冰在QQ群里火了起来，QQ小冰能够在群里与用户通过互相艾特进行交流，用户也很乐意向小冰提问，不过小冰不能够准确回答某一领域的问题，大部分是在闲聊；小冰作为陪聊的工具是足够的，但是用户获取不到想要的信息，这是小冰的一大缺陷。后来偶然通过一个微信公众号paperweekly加入了一个自然语言处理相关的群，群里有一个论坛机器人，可以转发用户的消息到其他群中，进行“头脑风暴”；同时可以将用户之间讨论较热烈的话题，进行统计通过发帖的形式保存下来，此公众号的用户非常之多，群里的用户也是非常积极讨论问题。由微信公众号到微信群聊，不知不觉中我们已经深受微信的影响。有数据显示，微信的用户已经超过6亿人次，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，而聊天机器人在这些社交平台的作用也越来越显得重要，给我毕设提供了一个很好的方向。

QQ小冰不能够提供某一领域的相关问题的准确答案，paperweekly的机器人不够智能，因此结合微信公众号做一个某一领域的微信机器人，不仅能够弥补上述两个机器人的缺陷，还在自然语言处理的问答系统方向有一定的研究价值。

电影在国内影响越来越大，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，了解电影甚至选择购票看电影。而自己对电影也有浓厚的兴趣，因而做面向电影领域的聊天机器人应该是毕设的一个不错的选择。

聊天机器人作为自然语言处理方面的一个重要应用，可以结合自然语言处理很多方面的前沿知识，特别是问答系统模块；在一年一次的文本检索会议TREC上，问答系统一直是一个非常热门的主题，而且国内对问答系统的研究也是如火如荼。

2.国内外在该方向的研究现状及分析



## 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[1]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商的发展，聊天机器人在客服服务及社交媒体上显得很有前途，从而聊天机器人的研究热了起来，聊天机器人其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”）以及一个搜素引擎。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基于用户过往的搜素习惯，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。

## 电影领域的问答系统研究现状

电影领域的问答系统，较新的是目前MIT开发的关于电影的MovieQA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[2]等作为数据，经过形式化描述，通过CNN[3]模型整合为所有的文本表示。MovieQA[4]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[5]的方法。

国内由浙大计算机学院院长陈纯及学生王北斗等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[6]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r个回答，对这r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Score进行排序，选取最终的答案集返回给用户。

Score(i)=Num(i)\*TFIDF(i)\*Pa(i) (2-1)

3．主要研究内容

本课题的主要研究内容是面向电影领域的微信聊天机器人研究与实现。本课题将从近期热映电影的影评入手，构建电影的知识库，力图实现较为准确的电影领域的问答系统，并与微信进行结合，在微信公众号和微信群中实现与用户之间的交互问答。本课题的研究的整体系统框架如图Figure 1。主要研究内容有：

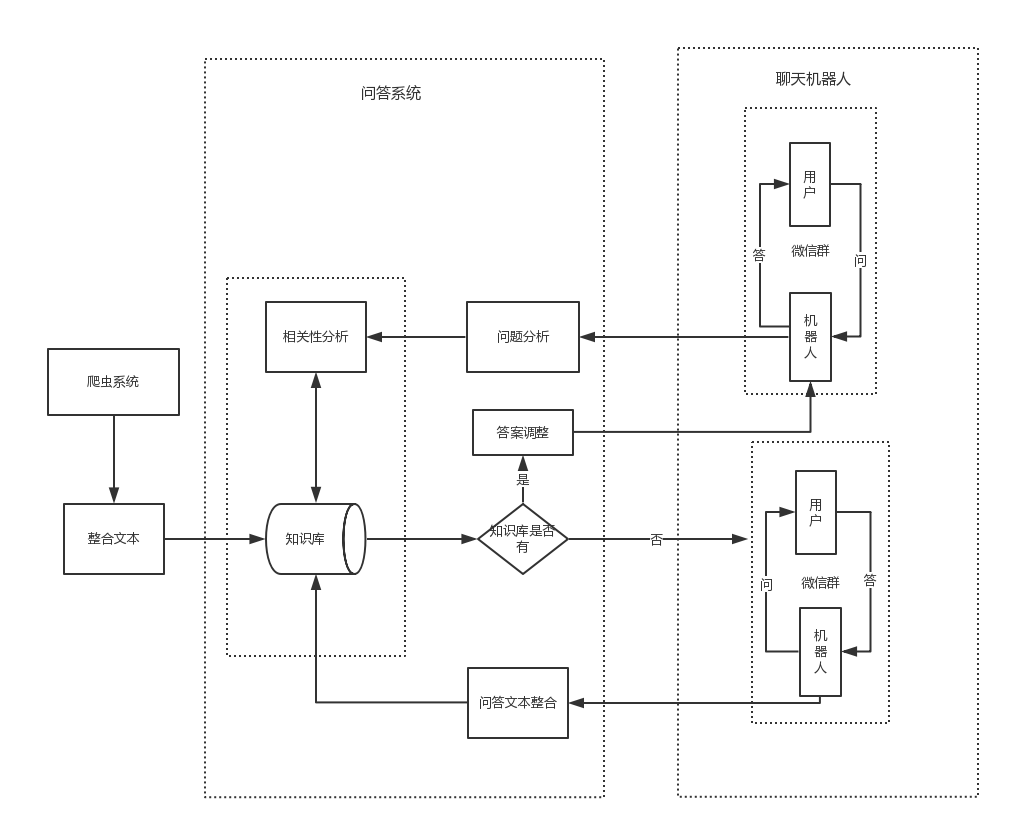


Figure 1 系统架构图

1）影评数据获取与数据训练、测试。由于影评的获取主要是通过爬虫在豆瓣影评上爬取，该阶段主要编写爬虫的代码进行爬取即可；而在数据训练与测试中，该阶段则需要选取合适的训练测试方案进行调整系统的参数，具体方案参照研究方案4.2。

2）问答系统的研究与设计。问答系统又包括三个组成部分：问题理解、答案检索和答案抽取。则这部分的主要研究内容为：

问题的语义理解与分析。在聊天机器人设计中，用户的问题是通过自然语言来表述的，因此必须要将自然语言表示的问题进行语义理解与分析，提取用户的意图，为问答系统检索答案奠定基础。

问题答案的检索。问题的答案需要在知识库中进行检索得到，因此优秀的检索方案是问答系统中必不可少的。问答检索的方法有LDA主题模型方法和Word2Vec词向量模型方法，本系统采用的是佐治亚州立大学的Zhibo Wang和Yanqing Zhang在2016年提出的整合LDA模型和Word2Vec模型方法，其性能的F值能达到85.6%，优于LDA模型（68.0%）和Word2Vec模型（80.75%）。

问题答案的抽取。在知识库中检索得到的答案句子有多条，需要进行相关性排序，并提取出合理的答案。本课题采用BM25[7]算法对检索的结果进行排序，提取靠前的答案作为对用户的回答。

3）微信机器人系统的实现。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，将问题分类、语义解析、答案检索、答案提取扩展融合到问答系统中去，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. 研究方案



## 爬虫系统设计

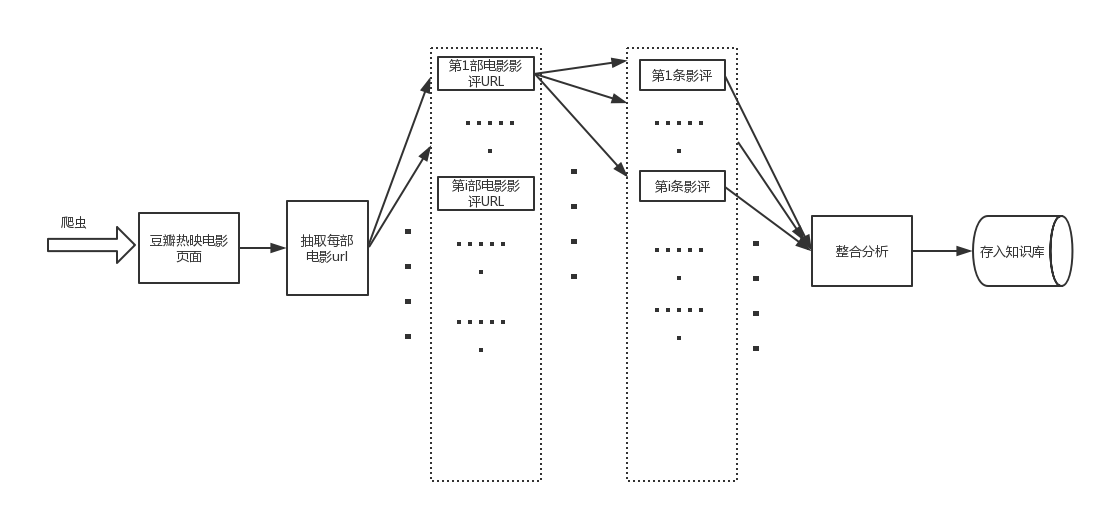


Figure 2 爬虫系统设计

爬虫系统暂以爬取豆瓣影评为主，采取如下爬取策略。

（1）爬取豆瓣热映电影首页，抽取每部电影的url

（2）访问每部电影的url，爬取网页，提取每页的影评文本

（3）影评文本进行句子划分，分析整合为三元组，存入数据库

## 数据训练、测试方案

人工设置1000条关于最新电影的问题及答案，其中80%作为系统的训练集，剩余20%作为系统的测试集。

训练方案：训练集800条问题，作为系统的输入，人工对系统给出的回答作判断（分为两类，其中一类是回答正确，另一类为回答错误或是无关回答），通过计算精确率（Precision）和召回率（Recall）从而计算F值，根据F值调整系统的参数。

测试方案：测试集200条问题，作为系统的测试输入，通过计算得到的F值来表示系统的性能。

这里的精确率是系统回答问题被判为正确的条数与输入问题的条数的比率；召回率是系统回答问题被判为正确的条数与问题总条数的比率。F值=精确率\*召回率\*2/（精确率+召回率）。

## 问答系统设计

问答系统一般包含三个主要组成部分：问题理解、信息检索和答案抽取。依据处理数据的格式，问答系统可以划分为三类：基于知识库的问答系统、基于大规模文本的问答系统和基于问答对的问答系统[9]。本系统只针对电影领域进行问答，因此选择构建基于知识库的问答系统。

### 问题理解

问题理解是问答系统的首要任务，接收用户的自然语言问题，从中提取用户的意图。问题理解分为问题类型分类、关键词提取和限定词提取。

（1）关于电影问题分类，本系统暂定分为：评价类、情节类、数据类、影片信息类及其他类。

电影评价类问题：美女与野兽这部电影怎么样？——显然用户希望得到关于《美女与野兽》这部电影的评价答案，因此很容易归类于评价类。

电影情节类问题：贝尔是谁？野兽最终的结局是什么？——显然贝尔和野兽都是电影《美女与野兽》中的角色，问电影中的角色以及角色的结局都是电影的情节类问题。

电影数据类问题：美女与野兽的票房达到多少了？——这就是一个最明显的数据类问题，用户希望得到的是《美女与野兽》这部电影票房的具体数据，因为要保证这个数据的准确性；而不是用“很多”这种模糊词代替。

影片信息类问题：《美女与野兽》中野兽谁演的？——这类问题就需要获取演员表这类数据，这种数据只需要在百科中获取即可。

其他类型的问题：有没有美女与野兽的免费资源啊？——这种问题就与上述的电影评价、情节及数据就没关系，暂归类到其他类问题。这种问题，在库中无法找到，在下述微信机器人设计中有解决方法。通过“众包”的方法，从用户那获取答案。

（2）问题关键词提取

问题关键词不仅影响答案检索的准确性，同时影响计算答案与问题关键字的相似度。关键字抽取一般首先对问题进行分词和词性标注，然后根据“停用词表”去除问题中没有价值的词，保留下来的作为关键词。分词和词性标注可以采用哈工大的语言技术平台（LTP[9]）中词性标注的模块，停用词表可以选用哈工大版停用词表。

（3）限定词提取

限定词对问题的答案有一定的修饰限定作用。可以利用哈工大语言技术平台进行句法分析，得出对应的主谓关系（SBV）和动宾关系(VOB)[9]。由于大部分的简单问句所期望的答案都是问句中核心谓语动词的发出者或接受者，故设定限定词就在主谓之间、动宾之间。

### 答案检索

关于文本检索，比较常用的是LDA主题模型方法及google的Word2Vec开源工具。LDA采用分级贝叶斯模型，根据主题分布的概率展示全局关系。LDA注重于高频词而忽略低频词，因此会有意义的低频词不能在模型中起作用，一定程度上降低了模型的准确率。Word2Vec根据语义关系，将词划分为词向量表示，展示局部关系。通过目标词所在句子的周围的词向量计算目标词的词向量。近期Zhibo Wang[10]和Yanqing Zhang[10]提出结合LDA和Word2Vec的方法，并且取得较好的效果。

整合的方法主要分为三个步骤：

1）生成word-to-topic向量

采用LDA的思想，用向量的每个维对应每一个topic在文档中的概率。

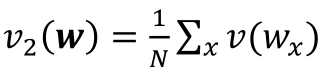
 （4-1）

公式（4-1）中**w**代表词的集合，**a**是潜在分布的参数向量，可以生成；C:\Users\马晶义\Desktop\paper\CodeCogsEqn (4).gif代表在文档中选出主题的可能性，是的多项式函数；是整个语料库生成矩阵，其中每一项代表一个词在主题中的可能性。Word-to-topic向量。

2）派生word-to-word向量

利用K维的词向量代表一个词，利用词与词之间哈弗慢树编码的关系，使用逻辑回归的方法更新向量。利用Word2Vec的思想，由x词周围的词向量生成x的词向量。

 （4-2）

计算文本向量，其中N代表文本中词数目。

3）构建强文本向量

由于LDA和Word2Vec分布考虑全局与局部的关系，因此整合这两个模型计算的词向量[11]就都考虑了全局关系与局部关系。

 （4-3）

其中和C:\Users\马晶义\Desktop\CodeCogsEqn (1).gif分别由上两步计算得出。

### 答案抽取

通过检索得到的候选句子，需要计算与问题的相关性来进行评分排序。选用比较常用的检索相关性评分算法BM25算法。这里BM25算法的主要思想：对问句Q进行分词，然后计算每个词语与句子文本T的相关性得分；最后将所有词语的相关性得分进行加权求和，从而得到问句Q与句子文本T的相关性得分。

C:\Users\马晶义\Desktop\paper\CodeCogsEqn (3).gif （4-4）

C:\Users\马晶义\Desktop\paper\CodeCogsEqn (1).gif （4-5）

其中，score(D,T)就是问句Q和T的相关度评分；q为句子Q中的语素，即分词后的词语；f(C:\Users\马晶义\Desktop\paper\CodeCogsEqn (2).gif,T)为T中q出现的频率；|T|为T的长度，avgdl为所有T的平均长度；k1和b是调整参数。N为T的总数，n(C:\Users\马晶义\Desktop\paper\CodeCogsEqn (2).gif)为所有包含C:\Users\马晶义\Desktop\paper\CodeCogsEqn (2).gif的总数。

## 微信机器人设计

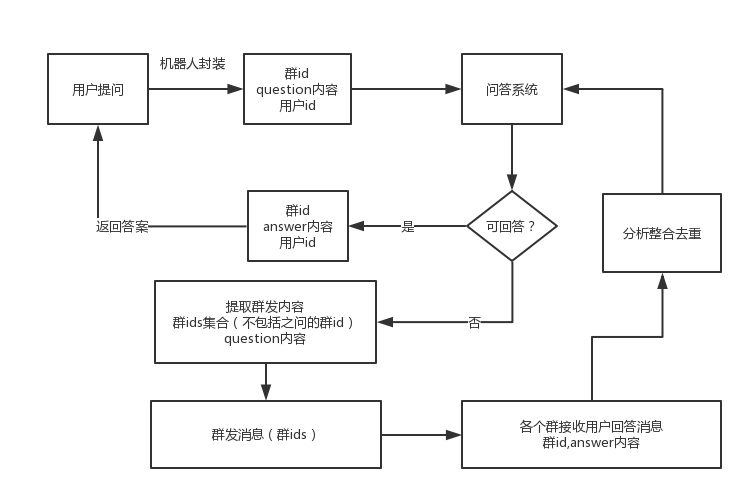


Figure 3 聊天机器人设计

机器人设计步骤：

（1）用户在群中通过艾特机器人，提出问题

（2）机器人获取问题内容、用户当前所在群id、用户id，进行封装，通过itchat接口发送后台

（3）问答系统分析，是否能返回答案；是则进入步骤（4），否则进入步骤（5）

（4）返回内容进行封装为群id、回答内容、用户id，发送到群id对应的群，通过艾特用户名（用户id对应）加答案内容，结束

（5）将问答系统获取的消息，进行修改封装为群id集合（去掉用户提问的群id）、问题内容，发送给所有群

（6）获取每个群中用户给机器人的回答，进行分析整理，返给问答系统；进入步骤（3）

## 评价指标

由设置的关于最近热映电影的1000个问题，由系统回答出每个问题的答案，进行人工判别。对每个问题的答案，人工判别为两类：其中一类是回答正确，另一类为回答错误或是无关回答。从而计算系统的精确率（Precision）和召回率（Recall），最终计算F值，通过F值来评价系统的性能。

5．进度安排，预期达到的目标



## 课题进度及工作安排

Table 1 课题进度和工作安排

|  |  |  |
| --- | --- | --- |
| 起始时间 | 完成时间 | 计划工作内容 |
| 2017.4.1 | 2017.4.8 | 爬虫系统设计 |
| 2017.4.9 | 2017.5.1 | 知识库的构建及问答系统相关模型搭建 |
| 2017.5.2 | 2017.5.16 | 微信机器人设计 |
| 2017.5.17 | 2017.6.1 | 微信与知识库连接，数据训练、测试 |
| 2017.6.2 | 2017.6.20 | 测试系统性能，撰写毕业论文 |

## 预期达到的目标

1) 微信机器人层面，界面友好，与用户之间通过自然语言交流

2）问答系统层面，预期准确率达到90%，由于答案检索部分的模型算法的F值最高为85.6%，预期本系统的F值达到85%

3) 作为聊天机器人，当在知识库中能够找到答案，答案能立即返回给用户（延时1秒以内）；当需要通过在其他群询问用户答案，则返回答案的时间较长，预计延时1分钟以内

4）整体系统而言，用户问聊天机器人问题，预期机器人能够给出80%的令用户满意的答案

6．课题已具备和所需的条件、经费



## 课题已具备的条件

课题主要研究面向电影领域的微信聊天机器人，目前已获取微信接口机器人itchat源码。现有哈工大语言技术云平台（LTP）可作分词、词性标注[12]、句法依赖，可以选择调用。同时已阅读部分有关问答、文本检索、文本相似度的文献，对系统的流程有基本的了解。

## 所需要的条件、经费

条件、经费暂无需求。

7．研究过程中可能遇到的困难和问题，解决的措施

问题1：因为是以近期热映的电影为方向，数据来源不容易寻找，数据信息不够准确。

解决方案：通过爬虫爬取豆瓣影评上近期热映电影的影评，以句子为单位划分存储，解决数据来源问题。通过对句子进行整合统计去重，得到更准确的信息。

问题2：用户讨论比较多的是情节，然而热映电影的情节仅仅靠影评来抽取，是完全不够的，数据就不够完备，不足以回答用户的某些问题

解决方案：机器人参与用户讨论是通过微信群的方式，机器人通过管理多个微信群，可以跨群进行转发用户的问题，由其他群的用户进行回答，得到问题的答案进行分析整合丰富语料库，然后转发给发问的用户。

问题3：用户的问题表达不清楚，或者缺少核心的意图，从而机器人无法判别用户的意图不能给出回答

解决方案：尝试在遇到机器人无法确定用户的意图时，通过反问用户一些问题，丰富用户的问题，从而得到较清晰的用户意图，给出回答。

问题4：问题数据集如何获取？问答系统给出的答案是否正确，如何判别？

解决方案：问题数据集没有具体的来源，因此选择人工针对热映电影进行编写。对系统回答的答案，与问题进行整合成问答对的形式，做成判断题，作为问卷让周围人进行判别，最后进行统计结果。

8．主要参考文献

[1]. Varian H R. Beyond Big Data[J]. Business Economics, 2014, 49(1):27-31.

[2].Deng L, Wang S, Jin C, et al. Test Structure with Variable Test Data Width of IP Cores for Multicore SoCs with DVS and Multiple Voltage Islands[C]. Sixth International Conference on Instrumentation & Measurement, Computer, Communication and Control. 2016:665-670.

[3].Nicolosi L, Tetzlaff R, Abt F, et al. Cellular Neural Network (CNN) based control algorithms for omnidirectional laser welding processes: Experimental results[C]. International Workshop on Cellular Nanoscale Networks and Their Applications. IEEE, 2010:1-6.

[4].Tapaswi M, Zhu Y, Stiefelhagen R, et al. MovieQA: Understanding Stories in Movies through Question-Answering[J]. IEEE Signal Processing Letters.2016:4631-4640.

[5]. Jun Ye, Cosine similarity measures for intuitionistic fuzzy sets and their applications, Mathematical and Computer Modelling, Volume 53, Issues 1–2, January 2011, Pages 91-97

[6].Chien-Hsing Chen. Improved TFIDF in big news retrieval: An empirical study[J]. Pattern Recognition Letters,2016

[7]. Robertson S, Zaragoza H, Taylor M. Simple BM25 extension to multiple weighted fields[C] Thirteenth ACM International Conference on Information and Knowledge Management. ACM, 2004:42-49.

[8].王宝鑫, 郑德权, 王晓雪, 赵珊珊， 赵铁军.基于文本蕴含的选择类问题解答技术研究[J]. 北京大学学报,2016，52(1): 1-2.

[9].Che W, Li Z, Liu T. LTP: A Chinese Language Technology Platform.[C] COLING 2010, International Conference on Computational Linguistics, Demonstrations Volume, 23-27 August 2010, Beijing, China. DBLP, 2010:13-16.

[10].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.

[11].Faruqui M, Dyer C. Non-distributional Word Vector Representations[J]. Computer Science, 2015.

[12]. Haitao Liu, The complexity of Chinese syntactic dependency networks, Physica A: Statistical Mechanics and its Applications, Volume 387, Issue 12, 1 May 2008, Pages 3048-3058