[摘要 1](#_Toc514276140)

[Abstract 2](#_Toc514276141)

[第一章 概述 3](#_Toc514276142)

[1.1引言 3](#_Toc514276143)

[1.2 课题的主要任务和目标 3](#_Toc514276144)

[1.3本文的组织 3](#_Toc514276145)

[第二章 相关技术背景 4](#_Toc514276146)

[2.1 知识图谱 4](#_Toc514276147)

[2.2文本情感分析 4](#_Toc514276148)

[第三章 数据整理和算法设计 6](#_Toc514276149)

[3.1数据的清洗 6](#_Toc514276150)

[3.2中文分词 7](#_Toc514276151)

[3.3停用词处理 8](#_Toc514276152)

[3.4知识图谱提取设计 9](#_Toc514276153)

[3.5本章小结 9](#_Toc514276154)

[实验设计、分析和评估 10](#_Toc514276155)

[4.1 实验设计(实验过程、工具等) 10](#_Toc514276156)

[4.2实验结果分析 12](#_Toc514276157)

[4.3实验评估 12](#_Toc514276158)

[4.4本章小结 12](#_Toc514276159)

[总结和展望 13](#_Toc514276160)

[5.1课题总结 13](#_Toc514276161)

[5.2课题展望 13](#_Toc514276162)

[参考文献 14](#_Toc514276163)

摘要

自从google公司推出旗下的产品Knowledge Graph以来，知识图谱这个概念越来越受到学术与工业界的关注。如何以质量层次不齐的网页数据作为原始数据源，构建知识图谱已经成为了一个热门的研究课题。

本文首先介绍本课题的主要任务与目标；然后介绍本课题相关的技术背景；接着介绍实验的数据处理和构建知识图谱的算法设计；随后我们介绍具体的实验过程，给出实验结果和实验评估；最后我们对本课题的工作做一个总结和展望。

关键词：知识图谱 情感分析

作 者：徐双奇

指导老师：朱晓旭

Abstract

Since google company introduced its product Knowledge Graph, the concept of knowledge graph has been increasingly concerned by academic and industrial circles. How to use the Web page data with uneven quality as the original data source, building a knowledge map has become a hot research topic.

This paper first introduces the main tasks and objectives of this topic; then introduces the relevant technical background of this topic; then introduces the experimental data processing and the algorithm design to construct the knowledge map; then we introduce the specific experimental process, and give experimental results and experimental evaluation; In the end, we will make a summary and outlook on the work of this topic.

Keywords: knowledge map sentiment analysis

Author: Shuangqi Xu

Instructor: Xiaoxu Zhu

第一章 概述

1.1引言

互联网的发展导致了“数据大爆炸"。但正所谓"乱花渐欲迷人眼”，大量的数据反而导致人们很难从中获得有效信息。于是诸如谷歌百度之类的搜索平台在上世纪九十年代应运而生,他们一般采用向量空间模型(VSM)。然而这些基于关键词匹配的搜索技术目前存在一些缺陷:搜索引擎缺乏互联网网页中人，物，事，之间的联系，因此无法完成精确搜索。

于是知识图谱这一概念应运而生。它将人物，事物作为实体，并构建他们之间的关系。数据不再是孤立的个体，相反充满了联系。

目前，面向特定领域的中文知识图谱的项目比较缺乏。因此，构建一个特定领域的知识图谱具有重要现实意义。本文的工作关注用户投诉这一特定领域，以投诉文本作为原始数据，构建基于投诉文本的知识图谱。

1.2 课题的主要任务和目标

本课题的主要内容是（1）对投诉文本进行分词，从而为之后提取实体以及他们之间的关系提供基础。（2）从投诉文本中提取实体概念。（3）从投诉文本中提取实体概念之间的关系。（4）用实体概念以及他们之间的关系构建知识图谱。本课题的主要任务是基于用户投诉文本进行知识图谱的构建，即输入一个关键词输出与该关键词相关联的信息。并以视觉(图片)的方式呈现这些信息。

1.3本文的组织

本文的第一章是概述，主要阐述课题背景以及课题的主要任务与目标。本文第二章介绍相关技术背景:一是知识图谱技术背景，二是文本情感分析技术背景。第三章介绍数据整理和算法的设计:首先是数据清洗的方法，接着是中文分词，然后是停用词处理，最后是知识图谱的提取。第三章介绍具体实验设计，分析，和评估方法。第五章对本课题做一个总结并且对课题做进一步展望。最后是参考文献和致谢。

第二章 相关技术背景

2.1 知识图谱

2012年谷歌搜索发布了一款重磅产品“知识图谱”（Knowledge Graph）。知识图谱可以将谷歌搜索的结果进行知识系统化，用户所输入的任何一个关键词都能获取完整的知识体系。比如你输入特朗普，谷歌搜索的侧栏就会告诉你特朗普的相关信息，比如他的身份，配偶，以及其子女信息等等。换句话说在知识图谱的概念里面，一切都是一个实体，而非简单的，没有联系性的字符串[1]。事实上知识图谱是大数据时代的必然产物，互联网上爆炸的数据使得人们迫切需要一个将这些数据关联起来的工具。

知识图谱本质上是一个语义网[2]。语义网络最早可以追溯的上个世纪六十年代，它是一个有向图，这个有向图由顶点和边构成。顶点代表实体或概念，有向图中的边代表实体或概念之间的关系。在早期，语义网络的构建一般由手工构建。知识图谱的数据源是互联网上的网页数据，包括结构化，半结构化，以及非结构化数据。其中大部分数据都是非结构化的。这些数据需要通过数据清洗等手段进行数据的预处理，因为这些原始数据里面充满了噪声比如歧义等等。

大体上，知识图谱的构建可以分为以下几个部分：一是实体或概念的获取；二是关系的提取。三是实体或概念与关系的关联，去除伪信息以及互斥信息，构建知识图谱。常见的实体或概念与关系的识别方法有两种:一是流水线方法[3]，二是端到端识别方法。关于实体关系的去除伪信息以及去除互斥信息。最近提出了一种可行的方法是使用概率软逻辑[4]，它既可以捕获事实之间的概率依赖关系又在实体和其关系之间执行全局约束。根据应用场景，实体或概念与关系的存储可以采用关系型数据库以及NoSQL。如果使用场景需要高性能计算，则需要配合Hadoop或者Spark使用。

2.2文本情感分析

文本情感分析又被称为意见挖掘[5]，是指对存在主观情感倾向的文本进行情感分析和提取的过程。文本情感分析技术一般被用来做网络营销、企业舆情监控、政府舆论监控等等用途。一般来讲，文本情感分析技术可以划分成几个子任务，首先是提取给定文本的主题。接着是确定陈述者主体，然后是提取该主体的陈述的主观性语句，过滤那些客观性语句。最后是对他的情感倾向进行判断分析。根据给定文本的粒度不同，可以将文本情感分析划分为词汇级，语句级，篇章级，以及海量数据级。针对海量数据级，随着互联网的发展，网络的各个平台都存在“网络水军”，他们所陈述的意见往往都是不真实的，可能会对最终结果造成较大影响。因此需要首先根据噪声的特征进行数据清洗。

常见的文本情感分析方法有：（1）基于机器学习的技术，比如支持向量机等等。（2）基于词典的方法，即对一个新出现的词，根据词典中相似词对其进行情感倾向推理。目前，中文的文本情感分析还存在一些挑战，具体在:一，效果良好的英文文本情感分析技术无法在中文文本上获得相同效果。二，不同语境下的情感分析技术效果不佳，等等。随着智能手机的火热以及诸如微博这类信息发布和分享平台的发展，短文本的情感分析将会越来越受到人们的重视。

第三章 数据整理和算法设计

3.1数据的清洗

本课题的数据是“苏州阳光便民12345”网站的投诉网页,格式是HTM，如图3.1。

“苏州阳光便民12345”是由苏州市人民政府组织创建的网站。它成立于2005年，网站的建站目的是通过整合政府部门和有关行业单位的公共服务资源，设立统一的服务热线平台，365天×24小时为市民提供方便、快捷、优质、规范的服务。

市便民服务中心为市政务办下属、副处级、公益一类事业单位。主要职责：一是负责拟订全市便民服务工作的标准、规范，并组织实施；二是负责市“12345”政府公共服务热线的运行管理，承担与其他市级公共服务热线和服务平台的对接、联动、合作；三是负责市“12345”政府公共服务热线的诉求受理、答复，事项的交办、跟踪、督办和回复回访，协助做好承办单位的业务协调、服务质量和服务效果的监督考核工作；四是依托“12345”服务平台，开展社情民意调研、舆情分析，向公众提供相关的政务信息服务；五是参与文明创建、社会信用、城市管理、城市综合服务标准化、政风行风、作风效能、绩效管理等公共事项的管理；六是参与防汛防旱、防台风、突发水污染、突发固体废物污染、食品安全等突发事件的应急管理；七是指导各市、区便民服务业务工作。



图3.1.苏州阳光便民12345网页。

我们可以从图3.1看到该HTM网页中的我们所需要的投诉文本：（1）投诉标题；（2）投诉正文（3）网友针对投诉内容的跟帖与回复。

为了提取这三类文本，我们需要观察该HTM的源码。三类文本都在HTM网页固定的标签之中：对于投诉标题，如图2所示，它总是位于H1标签中最里面的位置。而投诉正文和跟帖回复则位于属性的开头为“postman”的TD标签之中，如图3以及图4。

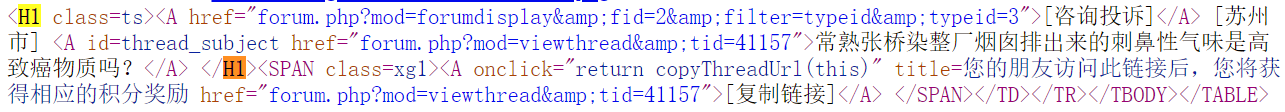


图2.投诉标题位于H1标签最里面

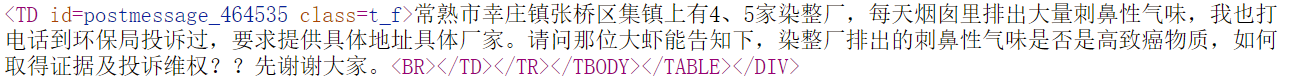


图3.投诉正文位于属性的开头为“postman”的TD标签之中



图4.投诉回复也位于属性的开头为“postman”的TD标签之中

从HTM格式的数据中提取特定标签内容的工具有很多，我们使用比较常见的工具：Beautiful Soup。利用该工具提取出投诉文本之后我们需要对其做进一步处理。

3.2中文分词

为了从文本中提取实体或概念与关系，我们首先得对投诉文本进行中文分词。所谓分词就是将连续的字符串或序列按照一定的规范重新组合成词序列的过程。在英文的行文中，单词之间是以空格作为自然分界符的，因此，在词理解上就比较直观。而汉语中词与词之间没用分界符，因此需要我们进行分词。

分词的算法有很多，总体可以分为两类：一类是无词典的分词方法，另一类是有词典的分词方法。所谓无词典就是根据汉字串在汉字语料中出现的概率来推断出该汉字串是否为词。有词典的分词方法是根据现有词典进行分词的方法。常见的有词典分词算法有：（1）正向最大匹配算法；（2）逆向最大匹配算法；（3）双向匹配法；（4）邻近匹配算法；（5）最短路径匹配算法；（6）基于统计的最短路径分词算法。

正向最大匹配算法的思想是以中文字符串的首字为起点，找到在中文字符串中出现的最长的词。这即是该字符串的第一个分词，剩余的字符串按照相同的方法处理。

逆向最大匹配算法的做法和上面的正向对最大匹配算法相反，它以汉字字符串的尾部为起点，找到最长的词汇，然后切分。并对剩余的汉字字符串做相同的动作。

双向匹配算法是将正向最大匹配算法和逆向最大匹配算法结合起来，构成新的算法。

邻近匹配算法是一种改进的最大匹配算法，它改进了了正向最大匹配算法的效率。

最短路径匹配算法是将汉字字符串进行全切分，接着构建汉字字符串切分的有向无环图。每一个汉字分词对应一条边，每条边上有一个权值。找出从有向图的起点到终点的最短路径，该路径包含的词就是汉字字符串的切分结果。

基于统计的最短路径分词算法是对最短路径匹配算法的一种改进，它根据词频给词典里的词分配不同的权重，具体的权重值可以通过大规模语料库获得。

常用的分词工具有很多，我们选择比较常见的jieba分词工具。Jieba分词工具所采用的分词算法便是上述的基于统计的最短路径分词算法。在有向无环图中求最短路径，把分词问题变成用动态规划解决的最短路径问题。对于未登录词，jieba采用了基于汉字成词能力的 HMM模型，使用维特比算法。

3.3停用词处理

经过上一步的分词处理，投诉文本变成了汉字词汇序列。但是这些词汇序列中还存在着停用词，我们需要对其进行过滤。停用词就是停止处理的词语，因为这些词语没有实际的含义。却占用存储空间，并且会对结果产生影响。停用词的过滤一般有以下几种方法：（1）将数据集中的高频与低频词汇找出，人工提取有用的词汇，剩余的就是停用词。（2）参考互联网上的常用中文文本停用词表。

我们将使上述两种方法结合起来，先参考哈工大停用词表，再对高低频词汇人工筛选。

3.4知识图谱提取设计

知识图谱中的结点代表实体，边代表实体之间的关系。接下来我们需要做的是实体识别和关系抽取。

在本章的前面几小节里面，我们对投诉文本进行分词，停用词过滤。一组中文词汇集合。为了进行实体识别，我们对该词汇集合中的所有词按照其词频从大到小进行排列，选取词频较高的词语，把他们当作实体。关于关系抽取，常见的主要有两种方法。一种是流水线方法[6]，另一种方法是基于神经网络的联合学习[7]。

流水线方法的思想是：第一步，将实体与实体之间两两组合。第二步，进行关系分类。

与流水线方法不同，联合学习对实体识别和关系抽取使用单一模型，比如使用LSTM模型等等。

本文使用实现简单的流水线方法，首先进行实体识别，在进行关系抽取。

3.5本章小结

本章具体讨论了构建基于投诉文本的中文知识图谱的步骤以及各个步骤所使用的算法：首先从HTM格式网页提取投诉文本。我们使用Beautiful Soup来提取这些文本。接着对投诉文本进行中文分词。这一步我们使用的工具是jieba分词。随后我们对词汇集进行停用词过滤。针对不同的主题，停用词也不尽相同。我们采用了参照停用词表与人工筛选相结合的方法，对停用词过滤。最后我们进行提取知识图谱。我们采用的是流水线方法，将问题分成两个子任务：实体识别和关系抽取。对于实体识别，我们采用两两组合配对的方法。

第四章 实验设计分析和评估

4.1实验设计(实验过程、工具等)

如图4.1所示，本文的实验设计可以划分为以下几个步骤：（1）从网页中提取投诉文本；（2）对投诉文本进行分词；（3）停用词过滤；（4）提取高频词作为实体；（5）计算实体相关性；（6）构建知识图谱。

首先我们使用Beautiful Soup提取HTM文件中投诉文本，投诉文本又可以分成投诉标题，投诉正文，以及投诉回复三个部分。我们分别编写不同的代码提取之。

对于投诉标题，由于它位于H1标签中最里面的位置。因此我们可以用以下主要代码提取它：

1. **from** bs4 **import** BeautifulSoup #导入包
2. **import** re
3. **import** os
4. list=os.listdir("../html")
5. f=open("../text/投诉文本.txt","w",encoding='utf-8')#utf-8方式打开
6. **for** i **in** range(0,len(list)):
7. path = os.path.join("../html", list[i])
8. **if** os.path.isfile(path):
9. soup = BeautifulSoup(open(path, 'rb'))  # 以rb的方式打开文件
10. headlist = soup.find\_all('h1')
11. **for** head **in** headlist:
12. alist = head.find\_all('a')
13. **for** a **in** alist:
14. **if** ("[咨询投诉]" **not** **in** a.string):
15. f.write(a.string)
16. f.write("\n")

在上面代码的第14行，我们过滤了“咨询”,“投诉”这两个停用词。原因是短语“咨询投诉”在每个HTM网页中固定位置都会出现，因此失去了实际的意义。

对于投诉正文和投诉回复，它们都位于属性的开头为“postman”的TD标签之中，因此可以用下面的代码提取：

1. tdlist = soup.find\_all('td', attrs={'id': re.compile("postmessage")})
2. **for** td **in** tdlist:
3. text = td.get\_text()
4. f.write(text)
5. f.write("\n")

我们通过上述两份主要的代码从HTM网页中提取投诉文本，并将投诉文本写入txt格式文件中。

接下来，我们从这个txt格式文件中读取投诉文本，使用jieba分词工具对其进行分词。并用停用词表进行停用词过滤。最后按照词频从高到低对词语排序，选取前1500个词语。主要代码如下：

1. **import** jieba
2. **from** collections **import** Counter
3. f=open("../text/投诉文本.txt",'r', encoding='UTF-8')#只读，注意编码方式
4. of=open("../text/高频词.txt",'w', encoding='UTF-8')
5. #加载停用词表
6. stopWord = [line.strip()**for** line **in** open('../text/停用词.txt').readlines() ]
7. line=f.read() #读取文件一行代码，有可能读的是空行
8. **if** line:
9. seg\_list = jieba.cut(line, cut\_all=False)
10. c = Counter()
11. **for** x **in** seg\_list:
12. **if** len(x) > 1 **and** x != '\r\n'**and** x **not** **in** stopWord:
13. c[x] += 1
14. **for** i **in** c.most\_common(1500):
15. of.write(str(i[0])+' '+str(i[1])+'\n')
16. **print**(i)
17. f.close()
18. of.close()

经过观察，jieba对于极个别未登录词的分词效果不好，例如词语“寒山闻钟”被划分成了“寒山闻”。因此我们需要在jieba中导入自定义词：

1. jieba.load\_userdict('../text/自定义词.txt')#导入自定义词

同时，高频词中存在诸如“&#”，英文字符串，数字，等等无意义的词语，这些都没有被过滤掉。我们可以在在停用词表里添加停用词，并且配合正则表达式进行过滤：

1. **for** x **in** seg\_list:
2. **if** len(x) > 1 **and** x != '\r\n'**and** (x **not** **in** stopWord )**and** **not** x.isdigit():
3. **if**(**not** re.match(r'[a-zA-Z]+',x)):#正则表达式过滤英文
4. c[x] += 1

接下来，本文将进行实体识别和关系抽取。为了探究实体之间的相关性，我们将之前的前1500个高频中文分词两两组合，并且计算这两个实体在同一个HTM格式的投诉网页上出现的频率。经过实验，单独按照某两个实体在网页中一同出现的频率降序排列并不能准确描述实体与实体之间的关联性。本文采用了如下公式：

公式4.1

上面公式4.1中的*P(A)*表示实体A在数据集中出现的概率，P(B)表示实体B在数据集中出现的概率。而P(A,B)则表示实体A和实体B同时出现在某个投诉网页上的概率。本文将计算K值并降序排列得出关联性最强的实体，

4.2实验结果分析

4.3实验评估

4.4本章小结

第五章 总结和展望

5.1课题总结

5.2课题展望

参考文献

* 1. Singhal A. Introducing the knowledge graph: things, not strings[J]. Official google blog, 2012.
  2. sowa J F. Principles of semantic networks: Exploration in the representation of Knowledge[J]. Frame Problem in Artificial Intelligence, 1991(2-3):135–157.
  3. Young D M. Pipelined method and apparatus for processing communication metering data: U.S. Patent 6,377,939[P]. 2002-4-23.
  4. Pujara J, Miao H, Getoor L, et al. Knowledge graph identification[C]//International Semantic Web Conference. Springer, Berlin, Heidelberg, 2013: 542-557.
  5. Pang B, Lee L. Opinion mining and sentiment analysis[J]. Foundations and Trends® in Information Retrieval, 2008, 2(1–2): 1-135.
  6. Young D M. Pipelined method and apparatus for processing communication metering data: U.S. Patent 6,377,939[P]. 2002-4-23.
  7. Zheng, Suncong, et al. "Joint entity and relation extraction based on a hybrid neural network." Neurocomputing 257 (2017): 59-66.

致谢

我要感谢我的指导老师朱晓旭老师，他所讲授的《中文信息处理》课程生动活泼。在课堂上，他深入浅出的授课风格给我留下了深刻的印象，激发了我对自然语言处理的兴趣。他对学生认真负责，每周都会组织一场例会，与我们探讨当前工作进展，并指出下一步的方向。在朱老师身上，我看到了很多老师所缺失的师风师德。他是我学习的榜样，向他致敬！