基于领域知识的评价对象抽取研究

杨森

2015 年 6 月

?1994-2018 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

中图分类号：TP391
UDC 分类号：004.8

基于领域知识的评价对象抽取研究

作 者 姓 名　　杨森
学 院 名 称　　计算机学院
指 导 教 师　　冯冲 副研究员
答辩委员会主席　　李侃 教授
申请学位级别　　工程硕士
学 科 专 业　　计算机技术
学位授予单位　　北京理工大学
论文答辩日期　　2015 年 6 月

?1994-2018 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

The study on domain-based opinion target extraction

Candidate Name:	Sen Yang
School or Department:	Computer Engineering
Faculty Mentor:	Associate Fellow Chong Feng
Chair, Thesis Committee:	Prof. Kan Li
Degree Applied:	Master of Engineering
Major:	Computer technology
Degree by:	Beijing Institute of Technology
The Date of Defence:	June，2015

?1994-2018 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

研究成果声明

本人郑重声明：所提交的学位论文是我本人在指导教师的指导下进行的研究工作获得的研究成果。尽我所知，文中除特别标注和致谢的地方外，学位论文中不包含其他人已经发表或撰写过的研究成果，也不包含为获得北京理工大学或其它教育机构的学位或证书所使用过的材料。与我一同工作的合作者对此研究工作所做的任何贡献均已在学位论文中作了明确的说明并表示了谢意。

特此申明。

签　　名：　　日期：

?1994-2018 China Academic Journal Electronic Publishing House. All rights reserved.　http://www.cnki.net

摘 要

Web2.0 时代，用户不仅仅是网络内容的消费者，也同时是互联网内容的生产者。网络上产生了大量的用户评价信息，这些评价信息不仅能给消费者提供对商品的参考，而且能给生产者反馈产品信息，了解自己产品的不足并加以改进。大数据时代人工来做这些工作是不现实的。如何通过使用自然语言处理技术，从大量数据中自动提取出评价对象是当前研究的热点。

评价对象抽取是在有观点性的句子中抽取出所讨论的评价对象。中文倾向性评测 COAE 从 2008 年开始举办，每年都会有评价对象抽取方面的任务。评价对象的抽取和分析方法主要分为有监督、半监督、无监督的机器学习方法。其中有监督的方法往往能取得较高的准确率，但需要一定的人工标注结果。通用的评价对象抽取研究由于范围广，往往难以取得较高的准确率和召回率。而对特定领域进行抽取工作，往往可以通过利用领域知识提高抽取的效果，使其有较高的实用价值。

本文以中文汽车领域为代表，研究了基于领域知识的评价对象抽取技术。本文主要工作和创新点如下：

1. 完成了对中文汽车领域代表性网站（www.autohome.com.cn）大量的半结构化数据的抓取；设计并构建了汽车知识库，以多元组方式存储；在大量的数据上训练汽车领域 word embedding 词向量。

2. 利用条件随机场的方法，把评价对象抽取转化为序列标注问题，提出了一种融合所构建的领域知识库的新方法，结合词、词性、句法等特征，对结果产生了较大的提升；同时，在 CRF 抽取的基础上，进一步利用 word embedding 技术设计一些策略方法对抽取的评价对象结果进行扩展，较大程度的提高了系统的召回率。

3. 实现了基于 Web 的汽车领域评价对象抽取验证平台。基于 Linux 平台构建；Web 服务器使用 Nginx；使用 python Django 经典框架；前端展示层通过 Socket 方式与后端层算法进行通信，减少了与算法之间的耦合，提高了系统的实用性。

关键词：评价对象；汽车领域；知识库；条件随机场；word embedding

?1994-2018 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

ABSTRACT

With the further development of the Internet in China, more and more people acquire information through the Internet. In the era of Web2.0, the user is not only the consumer of network, but also the producer of the Internet. A large number of users' evaluation information is produced on the Internet, it can not only provide references of goods for consumers, but also offer feedback information of products to the producers and so as to make them understand the shortcomings of their products and perform some improvements. However, it is not realistic in the big data era with so much information and how to use the technology of natural language processing to perform opinion targets extraction is a hotspot of current researches.

The task of pinion targets extraction is extracting the objects which people make comments in the opinion sentence. And the methods of opinion targets extraction is mainly divided into three categories: supervised, semi-supervised and unsupervised. Although the supervised method can often obtain higher precision score, its requirements for manual annotation can not be overlooked. Chinese orientation analysis evaluation is held from 2008, and task on opinion targets extraction is contained every year. Considering the difficulties to obtain high precision and recall on domain-ignored opinion targets extraction, this paper made practical researches on domain-specific opinion targets extraction and used domain knowledge to reach a higher precision, recall and F-measure score.

This paper mainly studied the opinion targets extraction in Chinese automotive field. This paper mainly focus on the automotive-field opinion targets extraction and the main contributes are as follows:

1) The automobile-field knowledge base is constructed through grasping semi-structured data from the website of http://www.autohome.com.cn and stored in form of multiple-groups. And then the automotive-specific word vector is produced through training on a large amount of automotive-field datasets.

?1994-2018 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

2) In this paper, CRF-based method is adopted and the task of opinion targets extraction is transformed to a sequence labeling problem, in which words, part-of-speech, syntax and the constructed automobile-field knowledge base are selected as features of CRF to improve the extraction results. Then on the basis of prediction with CRF, some designed strategies and methods with word embedding are adopted for opinion targets expansion to reach a higher recall score in the experiments.

3) This paper also builds an automobile-field opinion targets extracting and verifying platform. Nginx web server, Python Django framework and socket communicate form are adopted in this platform and the whole system are established on the Linux platform. The algorithm performs communicate with the platform through the socket interface and so as to reduce the coupling. Consequently, the Internet web solutions provided a application for our studied algorithm of opinion targets extraction.

Key Words: opinion targets; automobile-field; knowledge base; conditional random field; word embedding

?1994-2018 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

目录

?1994-2018 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

?1994-2018 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

第 1 章 绪论

1.1 研究背景及意义

中国互联网络信息中心（CNNIC）在京发布第 35 次《中国互联网络发展状况统计报告》（以下简称《报告》）。《报告》显示，截至 2014 年 12 月，中国网民规模达 6.49 亿，互联网普及率达到 47.9%。网上信息资源的增长速度非常迅猛，从网页内容上看，文本占到绝大多数 80%以上。在这样的信息爆炸中人们如何从中高效的获取对自己有用的信息，过滤掉无效信息成为越来越重视的问题。搜索引擎的出现很好的帮助了人们查找到相关信息，但是也存在一些不足，搜索引擎只做了信息筛选和排序，而对数据的进一步处理与分析搜索引擎便无能为力了。这就需要我们在更高级层面上进行对数据的进一步处理与分析。评价对象的抽取研究便应运而生。

在网络中，各式各样的产品评论信息出现在各大论坛、门户网站及专业网站。这些产品的信息往往有着重要的价值：一方面，用户在购买产品前会对产品进行了解，而网络上的这些评价信息往往能决定着用户对其的偏好。另一方面，公司的一款产品，公司想得到用户对其产品的反馈意见，以对自己的产品进行升级。而在浩瀚的信息网络中，人工的去进行筛选会耗费较大的人力物力，为了解决这方面的问题，评价对象的抽取研究便应运而生，这是因为评价对象在人类语言中既是杂乱无章的，又会呈现出一定的规律性。在评价对象抽取方面，如果对不限领域对任何文本内容都适用往往在准确率方面难以提高，而对特定领域的研究更能达到较高的准确率，越高的准确率往往代表着更高的价值。本文也正是针对特定领域，汽车领域，产品评价对象做了一点尝试。

评价对象抽取为实现实际可用的评论信息智能系统，使计算机能自动的对网络信息进行处理分析，给决策者提供有用的结论性信息，帮助企业了解用户的消费习惯，帮助用户了解对于某种产品更加全面的信息，分析热点时间舆情和发展趋势，给企业、政府等机构提供真实数据层面的决策依据。情感分析的处理技术涉及到中文分词，数据获取，数据挖掘，机器学习等多个领域的相关技术。对网络评价对象抽取技术的研究，具有重要的学术价值和实用价值。

?1994-2018 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

此外评价对象的研究还可以作为情感分析的基础应用到，网络评论信息分析处理、网络舆情监控、信息预测等领域。

汽车与人们的生活息息相关，我们曾经或未来都可能会面对买车租车等问题，如何才能认识到汽车的好与坏，不能仅仅听一家之言，而应听别人对该车的评价，甚至说更细节的，比如操控性、油耗等。而对于汽车生产商，想要改进自己的产品，就需要收集用户对自己汽车产品的反馈，了解用户对自己产品的态度，以辅助决策，达到最大的收益。如此，对汽车领域的评价对象抽取研究有很好的实用价值。

综上所述，评价对象抽取是近几年信息处理、自然语言理解领域的一个新的研究方向，已经成为学术界和工业界所关注的焦点。目前，在中文评价对象抽取领域研究还并不是很深入，摆在我们面前的仍有很多亟待解决的难题与挑战。因此，对于评价对象抽取以及情感分析方面的技术，不仅具有重要的学术意义，而且具有重要的实用价值。

1.2 国内外研究现状

1.2.1 观点挖掘概述

互联网中包含着大量的非结构化文本信息，这些文本信息中含有大量的价值。观点挖掘就是研究文本中包含的观点信息。是当今研究的热点内容。应用领域包括有舆情分析，客户关系管理、产品信誉度分析等等。想要利用网络中大量的文本信息，一般通过以下三个步骤：首先是观点提取，从网络文本信息中提取具有观点信息的文本。其次是观点的极性分析，判断文本中所包含观点的极性，一般分为三类，正向、负向和中性。最后是观点总结，通过对观点的整合，总结出对某一对象或领域的分析结果。

观点挖掘的相关主要任务包括：情感分类、基于特征的观点挖掘、比较语句和关系挖掘、观点搜索、欺诈性观点识别。由于本文主要研究评价对象抽取，与情感分类密切相关，下面对情感分类做一个简要概述。

情感分类，就是判别说话人对某一主题的态度，积极的、还是消极的。是情感分析的一个重要问题。根据处理的粒度不同，可以分为词语级、句子级、篇章级、领域集几种层次。现在的研究工作主要集中在篇章级和句子级。而最近的研究热点转向了词语级。因为，往往一个句子中包含了多个观点，有时相同有时不同。把粒度转为词

?1994-2018 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

语级就显得有必要了，也使情感分类的更加精确。而情感要素抽取则是基础性工作。情感要素抽取是指从句子中抽取关键要素。一般地，我们所说的情感信息包括：评论实体、评价对象、评价词、观点持有者。在不同的任务中往往侧重点不同，在 COAE 中文倾向性分析评测中，此任务定义为从所识别的微博观点句中抽取相应的评价对象，需要抽取出被评价的产品名、被评价的产品属性。

1.2.2 评价对象抽取概述

评价对象抽取是指在一条评论性的自然语句中，分析抽取出带有情感色彩的词所评价的对象。具体表现是观点句子中观点词语所修饰的对象。在产品领域的评价对象抽取上有过很多学者研究，但是商业系统往往还是在实用基本的规则方式匹配名词或词组，准确率不高。基于统计学习的方法是近年来的热门，许多学者把机器学习的相关技术应用到评价对象抽取上来产生了较大提升。

对评价对象抽取一般通过以下三个指标进行评价，准确率、召回率和 F1 值。准确率指预测正确的结果占预测结果对总数占比例；召回率是由系统预测正确的结果占文档中所有评价对象的总数的比例表示;F1 值是综合了准确率和召回率的指标,一般的，准确率和召回率是相辅相成的，不能只追求准确率的提高，而忽略召回率，用 F1 值进行比较相对有意义。

评价对象抽取往往为情感分析进行服务。通常的情感分析任务是对一句话或一篇文章分析。而评价对象抽取的情感粒度更小，对细粒度抽取的准确性决定了上层结果的好坏。评价对象应用包括且不限于观点问答系统、推荐系统、汇总统计系统。这些系统都需要准确的找到具有观点性句子中评价词评价的对象。本文就是通过设计改善算法，抽取汽车领域的评价对象。如：“方向盘指向很准确，底盘非常扎实，高速时稳定性也让我很放心。”句子中，方向盘、底盘、稳定性即为评价对象。准确、扎实、放心是对应的评价词。

1.2.3 评价对象抽取研究现状

评价对象的提取与分析是自然语言处理领域中的一个重要的应用方向,近些年来，无论是国内还是国际都受到了领域学者的广泛关注。多次的中文倾向性分析评测（Chinese Opinion Analysis Evaluation, COAE）和 NTCIR（NII Test Collection for IR

?1994-2018 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

Systems）[2]中，都将评价对象的抽取与分析作为一个重要的评测内容。从评测与具体应用来看，评价对象的抽取工作目前还离成熟较远，有相当大的提升空间。NTCIR 是公认的最权威的倾向性分析大赛，历届的比赛，我们可以发现，国外相对于国内更加成熟一些。在评价对象的提取与分析对于英文语料的处理中，准确率和召回率取得了较高的水平，在具体实践上已经有了初步的应用，日本富士通公司推出的评价分析系统已经被一些国外的网站使用，并作为参考反馈给产品的使用者。但也都是一些基于某个或几个特定领域的应用。而国内，中文倾向性分析评测[1]在 2008 年起开始举办，task3 任务是关于评价对象抽取的评测。在实际应用中，国内成熟的产品相对还比较少。

从研究手段上说，评价对象的抽取很多技术源于明明实体识别技术。命名实体识别主要指语言中特定的指代如人名、地名、机构名等。可以看出，评价对象与命名实体识别有很大的关联。

评价对象抽取的方法大致分为两大类：基于规则的方法和基于统计机器学习的方法，基于统计的方法又包含监督学习（Supervised Machine Learning method）、无监督学习（Unsupervised Machine Learning method）。

基于规则的方法有，Hu 和 Liu[11]利用关联规则挖掘评价对象，认为高频词更有可能是评价对象。Li 和 Zhou[12]利用情感词典和主题词词典筛选出<情感词，评价 对象>的组合，使用该二元组中的评价词与评价对象之间的关系进行抽取。Popescu 等[13]利用互信息算法抽取特性。刘鸿宇等[14]就是利用句法分析获取候选评价对象集，然后利用一定的过滤算法提取评价对象。随着近年来中文信息评测(COAE)的举办，基于规则的方法也得到了很好的应用和发展。

监督学习方法是基于人工加入的语言上下文模式，通过训练语料（Training Data）生成规则抽取模型或字符序列的标注模型，完成识别工作。其本质是通过机器学习的学习函数通过分析大量的训练数据的特征。监督的算法在评价对象抽取任务上近年发展起来。其经典方法有：决策树（Decision Trees）[3]、隐马尔科夫模型（Hidden Markov Models）[4]、最大熵（Maximum Entropy）[5]、支持向量机（Support Vector Machines）[6]、条件随机场（Conditional Random Fields）[7]。Zhuang 等[8]提出了一个意见述评价对象序偶的提取算法。Kessler 等[9]提出了利用机器学习算法对句子进行分类。Jakob

?1994-2018 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

等[10]首次把评价对象抽取转化成序列标注问题，然后利用条件随机场模型来标注。在本文中，我们使用条件随机场作为主要的方法。监督学习方法是最常用的方法，效果也相对较好，在 COAE 的比赛中也是使用最多的方法。但是，监督学习对语料具有较高的依赖性，有时需要较多的人工参与，成本较高。当在训练语料不足的时候，我们可以使用半监督和无监督的学习方法。

1.2.4 评价对象抽取发展趋势

从目前评价对象抽取现状来看，评价对象抽取的主要发展趋势有以下几个方面：

1. 随着大数据时代的到来，有大量的未标注数据，通过半监督或无监督的方法从数据中挖掘出信息，是新的机遇和挑战。
2. 领域无关的评价对象抽取往往难以取得较高的准确率和召回率，难以产生较高的应用价值，而分领域的抽取，往往能在其基础上产生较大的提高，便可以在其基础上产生应用价值。
3. 在抽取评价对象后难以直接得以利用，需要后续的对评价词的抽取以判断其倾向性。如今，已经有人提出了一些方法，基于评价对象和评价词的联合抽取工作，这也是未来研究的热点之一。

1.3 论文主要研究内容

本文综合了前人关于评价对象抽取的研究基础，针对来自互联网汽车领域的句子级语料信息，对中文评价对象抽取进行了深入系统的研究。主要研究内容包括：

1. 在领域评价对象抽取中，领域知识的利用是非常重要的。基于这点，本文构建了汽车领域的多元组知识库。通过爬取汽车之家网站（http://www.autohome.com.cn）上大量半结构化信息，包括并不限于：品牌、厂商、车系、车型、评测、百科、口碑、论坛等数据，对数据进行规范提取形成汽车领域知识库。并利用 word embedding 算法生成领域词向量。

2. 条件随机场（CRF）是解决序列标注问题的有效和常用办法，本文也是基于 CRF 进行评价对象的识别工作。特征主要包括：词特征、词性特征、依存特征、语义角色特征。并通过引入知识库特征，融合了各特征进行进一步的实验。

?1994-2018 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

3. 利用条件随机场识别评价对象时，往往具有较高的准确率，但是召回率较低。如何提高召回率是函待解决的问题。本文提出利用领域 word embedding 数据，通过设计算法规则，对评价对象进行扩展。

4. 通过调研和学习业界流行的互联网技术，搭建一套整体的基于 web 方式的验证平台。

1.4 论文组织结构

论文共分五章：

首先，介绍了本文的研究背景，国内外的研究现状、研究内容和主要共工作，以及论文的主要结构。

本文第 2 章研究了汽车领域知识采集与表示。介绍互联网数据采集的设计与实现，并通过观察网页规律，抽取出高质量的网页正文内容。并在采集的半结构化数据上设计和构建了汽车领域四元组知识库。最后，介绍 word embedding 方法，并使用 word2vec 生成领域词向量，作为领域知识的表示。

本文第 3 章研究了基于 CRF 的领域评价对象抽取。首先通过对比各统计机器学习的方法阐述为何选择 CRF。其次介绍前人研究的特征，提出了如何利用构建的领域知识库与 CRF 方法的结合。为了弥补 CRF 的不足，提出了如何通过利用领域 word embedding 信息，对评价对象进一步扩展，提高召回率，最后通过实验分析算法效果。

本文第 4 章介绍我们设计并实现的基于 web 的汽车领域评价对象抽取验证平台。首先通过介绍业界流行的 web 技术引入我们为何选择这些作为我们的系统的设计方案，然后介绍平台设计与架构、以及系统平台的展示与使用。

最后，对本论文的工作进行总结与展望，总结论文的内容和主要工作，客观分析算法与系统的优缺点，并提出下一步可以继续研究的内容。

?1994-2018 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

第 2 章 汽车领域知识采集与表示

2.1 概述

随着大数据时代的到来，数据具有了以下的 4V 特点：Volume（大量）、Velocity（高速）、Variety（多样）、Value（价值）。数据是有价值的，数据是很庞大的。庞大的数据存在于互联网中，如何获取这些数据使之为我们利用起来，这就是数据采集技术，即网络爬虫技术，第 2 小结将介绍我们是如何高效爬取汽车之家网站内容。而我们获取到的这些数据并非结构化数据，而混乱的数据我们是无法利用的，如何从非结构化数据或半结构化数据基础之上建立起结构化的数据，即为知识库构建技术，应用在特定领域中即为领域知识库，在第 3 小节中我们将介绍如何设计和构建汽车领域四元组知识库。词向量作为一种新的词表示方法，能在大量数据中自动的学习，第 4 小结介绍我们如何构建出汽车领域的词向量。

2.2 领域数据采集

数据采集技术是通过使用爬虫从海量的互联网数据中获取所需要的数据的技术。爬虫是一个自动的从互联网上获取内容的程序，是搜索引擎的重要组成部分和检索内容的生成者。爬虫从一个种子 URL 开始进行网页爬区，在爬取到的新网页中的 URL 把符合我们要求的置入带爬队列中。抓取方式是通过 HTTP 协议向站点发送访问请求，下载页面内容，并对内容按规则解析，把所有待爬队列中的 URL 全部抓完即为停止。

2.2.1 抓取工具和内容

本论文使用的是开源的抓取工具 Scrapy。Scrapy 是由 Python 语言设计开发的 web 抓取的框架。提供多线程的抓取功能，并且可以抽取结构化信息。Scrapy 经常应用于数据挖掘、爬取、数据清洗等。功能非常强大，提供了包括 BaseSpider、sitemap 等功能。。

抓取内容主要来源于汽车之家（http://www.autohome.com.cn）网站。包括其中如下四大块的数据：

评测：http://www.autohome.com.cn/bestauto/

百科：http://car.autohome.com.cn/shuyu/index.html

?1994-2018 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

口碑：http://k.autohome.com.cn/

论坛：http://club.autohome.com.cn/

其中，评测和百科用于第 3 节的自动化构建四元组知识库。口碑和论坛数据用于第 4 节的词向量的训练。

2.2.2 抓取数据的正文抽取

抓取到的数据为 html 格式数据，如何从 html 格式数据中抽取出正文内容而没有杂乱的标签和无关内容是一项非常重要的工作。传统的提取方式都是通过简单的去除 html 标签，但是这样会存在一些问题。比如一篇奥迪的文章，我们需要的只是文章的正文，但是如果仅仅去除标签，我们会得到不仅仅是文章，更是所有的网页汉字，这对质量要求较高的语料是不满足的。本文通过对爬区数据的细分，如开始所述，我们爬取了四个方面的内容，评测、百科、口碑、论坛。在具体的工作中，这四项结果是分开进行处理的，由于每一主题下面的网页都符合其规律性，我们通过观察其规律，得到了完整的所需的网页正文内容，较高程度的去除了干扰。

我们使用的是 Beautiful Soup 开源 XML 解析工具。 Beautiful Soup 由 Python 语言编写的 HTML/XML 解析器，它处理不规范的标签。提供简单强大的功能，可以大大节省编程时间，提高工作效率。

例如，源网页的片段如下：

```
<!--end 标题 start 内容区-->
<div class="conmain clearfix endcon" id="termbox">
    <!--start 左侧-->
    <div class="conleft">
        <p><font style="FONT-FAMILY: 宋体; FONT-SIZE: 14px">　　从前些年到现在，不少中国企业一直想在中型车层面有所作为，于是出来了长安睿骋，比亚迪思锐，奔腾 B70 等车型，但都叫好不叫座。很多人骂车企没诚意，国产车也定那么高的价格。但据笔者观察，中国中型车价格普遍只是合资紧凑型的价格，不可谓没有诚意，但如同之家以前分析过的一样，中国中型车卖得不
```


?1994-2018 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

好的深层次原因，其实是中型车用户对于品牌的成见，以及“面子”这个特殊需求。</p>

我们通过观察网页规律性结构最终提取到的内容如下：

从前些年到现在，不少中国企业一直想在中型车层面有所作为，于是出来了长安睿骋，比亚迪思锐，奔腾 B70 等车型，但都叫好不叫座。很多人骂车企没诚意，国产车也定那么高的价格。但据笔者观察，中国中型车价格普遍只是合资紧凑型的价格，不可谓没有诚意，但如同之家以前分析过的一样，中国中型车卖得不好的深层次原因，其实是中型车用户对于品牌的成见，以及“面子”这个特殊需求。

源文链接：http://car.autohome.com.cn/shuyu/detail_11_12_1099.html 。

2.2.3 网络数据采集与正文抽取结果

表 2-1 网络数据抓取结果

	百科	口碑	评测	论坛	合计
网页文件数量	1543	30742	1079	31983	65347
网页文件大小	81M	11G	140M	2.8G	14G
利用文本大小	12M	674M	5.1M	139M	830M

表 2-1 显示了我们获取到的互联网数据，网页数达 65347 个，总网页大小 14G，抽取获得高质量内容 830M。网页数据中口碑数据包括 160 个品牌和 1057 种车系数据，均为 UGC 内容，基本覆盖了市场上的所有的品牌和车系。

图 2-1 为口碑数据的目录展示，品牌为一个文件夹，文件夹下面不同的车系的口碑数据作为一个文件。

?1994-2018 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

图 2-1 口碑数据文件结构

?1994-2018 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

2.3 领域知识库构建

随着语义网技术的不断发展，越来越多的知识库[15]被建立起来。在自然语言处理中，为了使计算机在处理问题时具有较高的智能性，让机器拥有先验知识往往是关键所在。在知识的基础上，可以让机器做一些简单的推理。整体的世界知识是一个非常庞大的工程，其复杂性是难以想象的。但是，我们在处理问题的时候，往往问题是针对某一领域的，建立起领域的知识库显得尤为重要。本节将会介绍我们是如何建立起汽车领域知识库。

2.3.1 知识库概述

知识是人类活动积累的，经过分类、归纳、综合处理过的信息。人类认识世界就是通过习得这些信息，从而更好的理解世界。在人工智能领域，希望通过利用把这些知识结构化用逻辑语言表示，使机器利用起来。本体理论就是用来表达知识和其关系的理论。知识库是事实、规则和概念的集合，用特定的方式表达和存储形成知识库。整个世界知识是非常庞大的，而且一直产生和消亡，这么庞大的信息通过人为的是无法构建和维护的。所以，专家和学者们希望能通过限定领域范围进行构建知识库，这样既避免了知识库太庞大无法构建的问题，又能构造出一个高质量的领域知识。但是损失是知识库的普适性。知识库的设计有很多种方式，最常用的是本体库，描述本体库的语言有 OWL 等。多元组也是构造知识库的常用办法，多元组因表达清晰，简单易用得到广泛使用。本文就是采用多元组的方式进行知识库的构建。

2.3.2 汽车领域四元组概念关系构建

本文提出一种用于汽车领域的四元组[16]知识库，其结构为：(concept1, concept2, relation, label)。四元组整体作为唯一。Concept1 为概念 1，concept2 为概念 2，relation 为概念 1 与概念 2 的关系，label 是对概念的分类。

元组中概念一般为领域词，通过构建词之间的关系构成知识库。

概念（concept）：元组中概念一般为领域词。

关系（relation）：三种主要的关系如下：

?1994-2018 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

1. 种属关系（P1< SubclassOf >P2）指概念 P1 与概念 P2 是继承关系，子类可以继承父类的所有属性，可选择某一合适的标准划分下位类。本文使用了 OWL 语言中对种属关系的定义。常见的种属关系有：car_serie<SubclassOf>car_brand，car_factory<SubclassOf>car_brand 等。

2. 属性关系（P1< AttributeOf >P2）指概念 P1 是概念 P2 的属性，指 P2 的某一方面。如下文样例第四行，“可靠性”<AttributeOf>汽车，表示可靠性是汽车的一个属性。

3. 成员关系（P1< MemberOf >P2）指概念 P1 是概念 P2 的实体成员，主要是用于描述汽车领域中上下位类间的构成关系。如下文 2.3.4 样例第 5 行，“车厢”< MemberOf >“零件”，表示车厢是零件的一个成员。

类别（label）：类别是对当前元组概念 1 的标注。主要分为六大类：car_brand（品牌）、car_factory（厂商）、car_serie（车系）、car_part（部件）、car_behavior（车的动力行为）、car_opinion_target（评价点）。

其中类别主要有以下几种：

car_brand：大众、宝马、奔驰、马自达等。

car_factory：郑州日产、通用雪佛兰、一汽大众、上海大众等。

car_serie：途观、英菲尼迪 QX80、科鲁兹等。

car_opinion_target：设计、工艺、加速性、安全性、性能等。

car_behavior：制动、操控、传动、焊接、减震等

2.3.3 领域知识库构建结果

最终我们生成的的知识库中一共有 **3456** 个四元组。

四元组样例：

大众	品牌	MemberOf	car_brand
高尔夫	大众	SubclassOf	car_serie
郑州日产	日产	SubclassOf	car_factory
车厢	部件	MemberOf	car_part

?1994-2018 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

注：其中第一行，“大众”是一个概念，与概念 2 “品牌”之间的关系是“MemberOf”，表示是成员之一，其 label 为“car_brand”，表示为汽车品牌。其他的含义以此类推。

在构建四元组的过程中，我们通过发掘网站规则，从半结构化中找出其中的结构，然后进行规范化，其中主要使用了一下三种方法：

1. 商与车系数据是通过解析汽车之家口碑数据标准化分类得到的结果。其中品牌 160 个、厂商 220 个、车系类别 1057 种。其中也包括品牌-厂商-车系的树形结构关系。

2. 部件主要通过解析汽车之家百科数据，通过超链接的方式得到汽车零件。如下面一段话：四冲程汽车发动机主要有气缸、活塞、活塞连杆、曲轴、配气机构（气门、凸轮轴等）、火花塞（汽油机）。带有超链接的词均为汽车之家具有百科解释的词，也往往是汽车的零部件名称。

3. 合网络开源的汽车领域词典对知识库进行完善，外加一部分的人工编辑。

2.4 基于 word2vec 领域词向量

2.4.1 词向量介绍

用机器学习的方法进行自然语言处理任务，一般首先要做的就是将文本转化为可计算的形式。词向量就是一种将文字转化为数字的方式。自然语言处理中最经典的词表示方法是 One-hot Representation 方式，该方法用一个长度为词表大小的向量表示一个词，词对应维度为 1，代表当前词，其它全为 0。如，一个语料中包含了 10 个单词，那么每个单词都要表示为一个 10 维的向量。对于其中的两个词“汽车”、“性能”，用 One-hot 方式表示如下:

汽车 [0 0 1 0 0 0 0 0 0 0]

性能 [0 0 0 0 1 0 0 0 0 0]

这种表示方法比较简洁清晰，在自然语言处理任务中得到了广泛的应用。这种表示方法一个最大的问题是无法捕捉词与词之间的相似度，就算是近义词也无法从词向量中看出任何关系。即这种方法存在“词汇鸿沟”问题，用这种方式表示的任意两个词都是孤立的，通过两个向量无法表示任何两个词之间的关系。即使是两个词义非常

?1994-2018 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

相近的两个词，也无法通过词向量看出两个词之间是否有关系。同时在处理大规模语料时，由于词表规模通常比较大导致 One-hot 的方法表示的向量维度会非常高。通过这种表示方式构建语言模型时往往会面临维数灾难，其计算复杂度会变的无法接受。

词向量作为一种新的词表示方法正在吸引越来越多的学者关注。Distributed representation 最早由 Hinton 提出，后来有学者借鉴这种思想在自然语言处理领域来表示词，将词典中的每个词表示成一个 N 维向量，向量的每一个维度表示一个隐含的特征，向量的维数可以更根据应用场景灵活调整，这种方法的优点在于避免了 One-hot 方法的维数灾难和数据稀疏问题，同时可以通过向量的距离表示词的相似度。用 Distributed representation 表示词通常称为“Word Embedding”，为了标书方便如无特殊说明后文中提到的词向量都是指代这种词表示方法。

Distributed representation 基本思想是通过训练将每个词映射成 K 维实数向量（K 一般为模型中的超参数），通过词之间的距离（如余弦相似度、欧氏距离等）来判断他们之间的语义相似度。而 word2vec 是用的就是这种 Distributed representation 的词向量表示方式。

2.4.2 word2vec 介绍

Word2vec[17]是 Google 在 2013 年年中开源的一款将词表征为实数值向量的高效工具，采用的模型有 CBOW（Continuous Bag-Of-Words，即连续词袋模型）和 Skip-Gram 两种。Word2vec 代码连结地址：https://code.google.com/p/word2vec/，遵循 Apache License 2.0 开源协议[18]。

Word2vec 通过训练，可以把对文本内容的处理转化为 K 维向量空间中的向量运算，而向量之间的相似度表征词与词之间的语义距离。Word2vec 输出的词向量经常用很多方面，如聚类、同义词、词性分析等。向量之间可以进行线性运算如：

vector('Paris') - vector('France') + vector('Italy') ≈ vector('Rome')

vector('king') - vector('man') + vector('woman') ≈ vector('queen')

神经网络语言模型 NNLM（Neural Network Language Model）采用的是 Distributed Representation，也就是每个词是一个向量，模型如下图 2-2：

?1994-2018 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

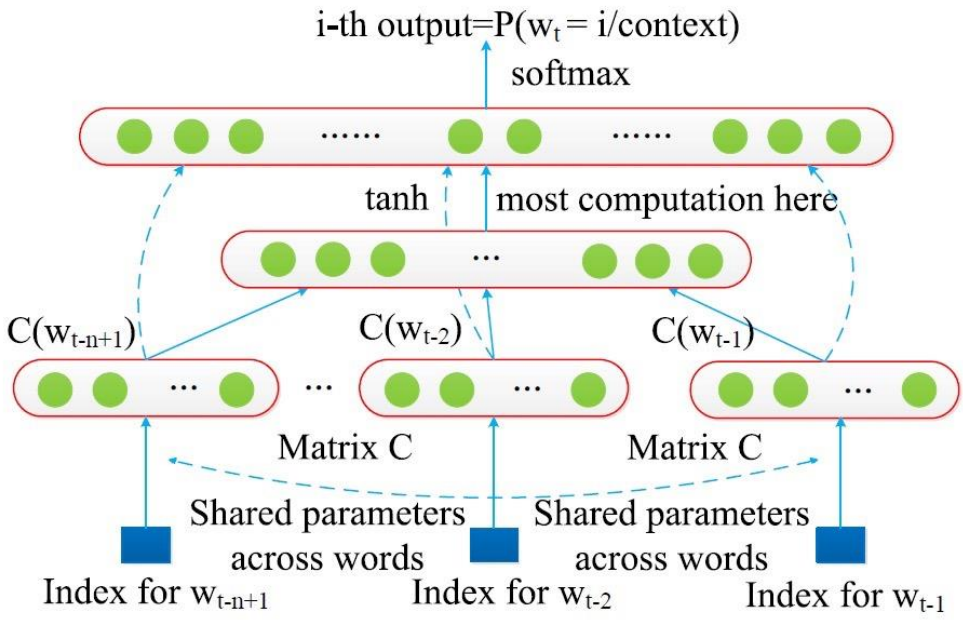

图 2-2 神经网络词向量模型

我们的目标是学习以下函数：

$$f(w_t, w_{t-1}, \dots, w_{t-n+2}, w_{t-n+1}) = p(w_t | w_1^{t-1}) \quad (2\text{-}1)$$

需要满足连个约束：

$$f(w_t, w_{t-1}, \dots, w_{t-n+2}, w_{t-n+1}) > 0 \quad (2\text{-}2)$$

$$\sum_{i=1}^{|V|} f(i, w_{t-1}, \dots, w_{t-n+2}, w_{t-n+1}) = 1 \quad (2\text{-}3)$$

Word2vec 包括 CBOW 和 Skip-gram 两种模型，一般地 Skip-gram 效果会比 CBOW 稍好，本文实验采用的就是 Skip-gram 方式进行训练的。下面简要介绍 Skip-gram 原理。

?1994-2018 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

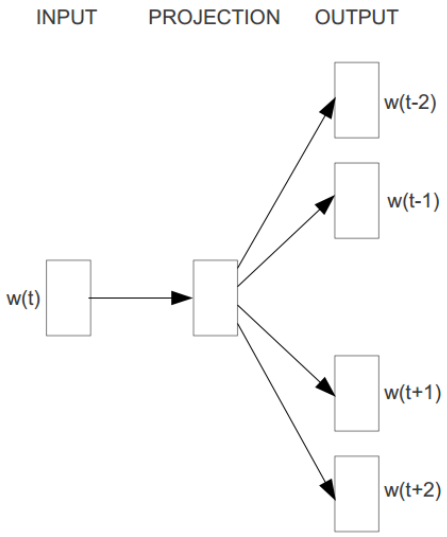

图 2-3 Skip-Gram 模型

如图 2-3，假设存在一个 w1、w2、w3、..、w_T 的词组序列，Skip-Gram 的目标是最大化：

$$\frac{1}{T}\sum_{t=1}^{T}\sum_{-c\le j\le c, j\neq 0}\log p\left(w_{t+j}|w_t\right) \qquad (2\text{-}4)$$

其中 c 决定上下文的大小，c 越大则需要考虑的 pair 就越多，一般能够带来更精确的结果，但是训练时间也会增大。基本的 Skip-Gram 模型$p(w_o|w_I)$定义为：

$$p(w_o|w_I)=\frac{exp\left({v'_{w_O}}^{T}v_{w_I}\right)}{\sum_{w=1}^{W}exp\left({v'_w}^{T}v_{w_I}\right)} \qquad (2\text{-}5)$$

其中，v_w和v'_w分别表示词 w 的输入和输出向量，W 表示词典中词的总量。由于模型训练采用的数据集比较大，为了提高训练速度，Mikolov 等给出了几种近似的计算方法，这里就不再一一介绍了。

2.4.3 领域词向量训练

在 2.2.3 节我们爬取的口碑和论坛数据基础上，使用哈工大分词工具 LTP，然后利用使用 word2vec 进行训练，K 设置为 200 维，进行训练。

?1994-2018 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

表 2-2 领域词向量与中文维基对比

	汽车领域		中文维基	
奥迪	Position in vocabulary: 1151		Position in vocabulary: 14533	
	奔驰	0.792071	艾斯唐	0.627918
	宝马	0.639865	纳塔莱	0.597354
	雷克萨斯	0.613132	集团型	0.576904
	保时捷	0.594880	英退	0.575320
	BMW	0.587699	保时捷	0.569228
	路虎	0.564869	纳塔莱迪	0.565170
斗气	Position in vocabulary: 11602		Position in vocabulary: 46981	
	赌气	0.778031	我爱俏	0.536747
	飞车	0.619090	吴娇	0.531686
	飚车	0.590431	何正	0.531519
	不跟人	0.579447	欢喜冤家	0.524515
	急车	0.572532	扑水	0.523444
	开赛	0.571441	井乔	0.509808
习近平	Position in vocabulary: 97840		Position in vocabulary: 17239	
	一汽红旗	0.678832	胡锦涛	0.823996
	复兴	0.670190	江泽民	0.757793
	造福人类	0.643405	李克强	0.738144
	来求	0.642307	温家宝	0.716654
	管理者	0.641855	贾庆林	0.710127
	改革开放	0.641195	曾庆红	0.704690

从表 2-2 中可以看出，word2vec 对语料的依赖性还是很强的。比如：“奥迪”在我们手机的汽车领域语料上把“奔驰”、“宝马”等其他品牌全都找出来了，而在维基百科中文语料中第五名是“保时捷”汽车品牌，其他的几个我们可能都很难发现他们有什么语义关联。“斗气”也是如此。相反的，我们查询国家主席“习近平”在我们手机的汽车语料中很难看出关联，倒是也能发现点什么比如“复习”、“改革开放”，而在维基百科中文语料中发现最相似的全是中国国家领导人的姓名。以此可以说明，把在使用 word2vec 时语料的选择是非常重要的。

在此我们建立了我们领域的词向量，后续我们会使用这里的数据进行改善我们评价对象抽取工作。

?1994-2018 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

2.5 本章小结

本章首先提出了如何在庞大的互联网中采集到我们所需要的数据，利用网络爬虫爬去和超文本标签解析器使我们得到高质量的网络内容。然后介绍了领域知识库的相关知识，提出了我们在汽车领域知识库的设计，通过四元组的方式解决领域概念之间的关系和分类问题。最后，我们通过在领域语料中计算词向量，以提高词向量结果的质量，通过观察实验结果，我们的领域词向量在领域词方面要优于在中文维基百科上的结果的。

?1994-2018 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

第 3 章 基于 CRF 的领域评价对象抽取

评价对象抽取方法可以分为两大类方法，一类是传统的基于规则的方法。姚天昉等[19]开发了一个汉语汽车评论信息分析系统，采用领域词典和语法关系树来挖掘。基于规则的方法有一定的局限性，一般只能处理简单、评价对象和评价词距离比较近的情况。另一类是利用统计机器学习的方法。统计的方法只需要标注数据并让机器进行学习，省去了繁琐的规则制定，也往往具有较高的准确率，本文主要采用的是此类方法。统计的方法一般有隐马尔科夫模型、条件随机场模型等。本文使用的的是 CRF 条件对随机场，下面对 CRF 方法做一些简要介绍。

3.1 算法模型介绍

3.1.1 条件随机场理论

条件随机场（Conditional Random Fields，CRF）是 Lafferty 等在 2001 年在隐马尔科夫模型和最大熵模型的基础上提出的一种判别式模型。CRF 克服了隐马尔科夫模型中严格独立假设问题，能够表达更多的上下文信息，特征选择也更加灵活，较好的克服了标记偏置问题。条件随机场已经在分词、命名实体识别等许多领域得到了广泛的应用。下面介绍一下 Lafferty 给出的 CRF 的定义。

设$G=(V,E)$是一个无向图，$Y=(Y_v)_{v\in V}$中元素与 G 众的顶点一一对应。在给定 X 的情况下，如果任意一个随机变量Y_v的条件概率服从图的马尔科夫属性：$p(Y_v|X,Y_w,w\neq v)=p(Y_v|X,Y_w,w\sim v)$，其中 w~v 表示 w，v 是图 G 的相邻顶点，则称(X,Y)为一个条件随机场。如图 3-1 所示为条件随机场模型。

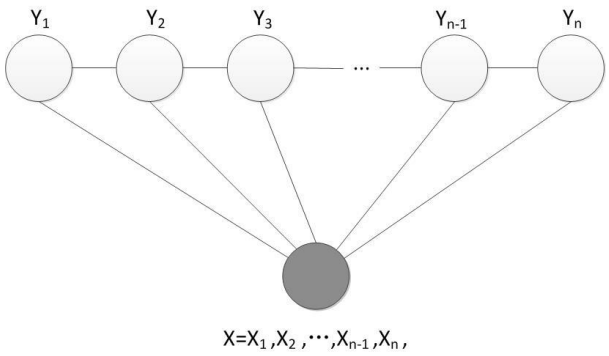

图 3-1 条件随机场模型

?1994-2018 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

根据 Hammersley-Clifford 定理[21]可知，当给定观察序列 $X=\{x_1, x_2, \ldots, x_n\}$时，标记序列 $Y=\{y_1, y_2, \ldots, y_n\}$的后验概率$p(Y|X)$服从 Gibbs 分布。对于线性条件随机场，令参数为 $\Lambda=\{\lambda_1, \lambda_2, \ldots, \lambda_n\}$，则状态序列的条件概率为：

$$P(Y|X) = \frac{1}{Z(x)} exp(\sum_{i=1}^{N} \sum_{k=1}^{K} \lambda_k f_k(y_{i-1}, y_i, x, i)) \quad (3\text{-}1)$$

$$Z(x) = \sum exp(\sum_{i=1}^{N} \sum_{k=1}^{K} \lambda_k f_k(y_{i-1}, y_i, x, i)) \quad (3\text{-}2)$$

其中，$Z(x)$是归一化因子，N 表示输入序列的长度，$f_k(y_{i-1}, y_i, x, i)$是特征函数，λ_k是第 k 个特征函数的权重。其中特征函数 f 是状态特征函数$S_{y,x}(y_i, x_i)$和转移特征函数$t_{y',y}(y_{i-1}, y_i, x, i)$的统一行使，定义如下：

$$f(y_{i-1}, y_i, x, i) = \mu t_{y',y}(y_{i-1}, y_i, x, i) + \zeta S_{y,x}(y_i, x_i) \quad (3\text{-}3)$$

其中μ和ζ分别是$t_{y',y}(y_{i-1}, y_i, x, i)$和$S_{y,x}(y_i, x_i)$的权重。一下两个公式分别为转移特征函数和状态特征函数的定义。

$$t_{y',y}(y_{i-1}, y_i, x, i) = \begin{cases} 1 & if\ y_{i-1} = y\ and\ y_i = y \\ 0 & otherwise \end{cases} \quad (3\text{-}4)$$

$$S_{y,x}(y_i, x_i) = \begin{cases} 1 & if\ y_i = y\ and\ x_i = x \\ 0 & otherwise \end{cases} \quad (3\text{-}5)$$

3.1.2 基于 CRF 的评价对象抽取流程

实现 CRF 的常用工具有 CRF++、CRF Mallet tookit 等，本文选用 CRF++。

用 CRF 作为评价对象抽取的流程如下图 3-2。

?1994-2018 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

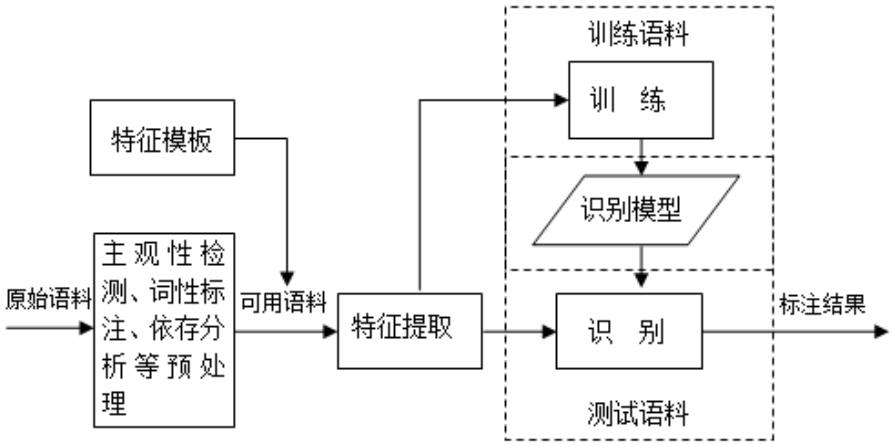

图 3-2 CRF 方法抽取流程

1. 对原始语料进行处理，包括预处理、分词等。
2. 设计要提取的特征，如词性、句法等。
3. 设计特征模板，即设计影响关系。
4. 使用标注数据进行训练。
5. 对测试语料进行识别。

3.2 融合领域知识的特征设计

3.2.1 特征选择

在 CRF 抽取工作中，特征选择[22][23][24]是非常关键的，经典的做法是考虑词、词性特征，最近 Chun 等[25]提出了结合句法和语义相关特征的方法，得到了较好的效果。主要应用的特征有：

1. 词特征

词特征即选择当前词为特征，例如前文提及的“**动力性能**不错”，动力性能词本身能对评价对象识别提供较大的信息，在领域抽取时，往往词本身是非常重要的特征。词特征的缺点是不具有通用性。

2. 词性特征

?1994-2018 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

即选择词的词性作为特征，在评价对象抽取中，评价对象往往是以名词的形式出现，而动词和形容词作为评价对象往往会较少的出现。并且，评价对象词的上下文词性同样纳入为特征，即表示，评价对象往往是一个词性序列中的某一个词。比如“动力性能不错”就是 n+n+adj 其评价对象就是该组合中的名词。

3. 依存结构[26,27]

当我们向一个产品发表评论的时候，常常需要一些评价词来表达观点，而这些评价词往往是和评价对象有着很强的语义联系的。它的依存分析结构下图所示。从图 3.1 中可以看到，我们可以从评价词“不错”与“性能”的 SBV 关系，得出“性能”为评价对象，从“大”与“扭矩”的 SBV 关系，得出“扭矩”为评价对象。他们之间的这些句法关系，往往能帮助我们找到所需要的评价对象。

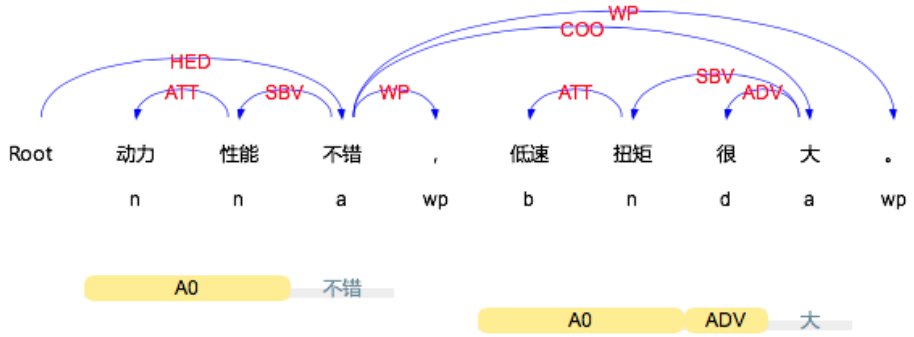

图 3-3 LTP 结果示例

4. 语义角色[28,29]

作为浅层语义分析[30,31]的一部分，语义角色标注在词汇语义分析中占据着非常重要的地位。还是如图 3.1 所示，在观点句中，人们通常通过评价词来表述观点，而形容词和动词则为评价词的两种主要形式。通过观察，我们发现，当评价词为形容词的时候，施事者 A0 即为评价对象。

3.2.2 融合了领域知识库

从本文的第二章我们已经介绍了我们是如何抽取出汽车领域的四元组知识库[32,33]。如何利用这些知识库是个值得研究的问题。本文提出利用领域词的 label 结果作为 CRF 新增特征进行抽取工作。

?1994-2018 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

首先，我们认为不同的类型的词作为评价对象时，其上下文关系是不一样的，会显示出某一类的特征。这种区分为评价对象的抽取提供了较准确的先验信息。

其次，本文还提出了利用了情感词的分类信息。情感词典使用的是中文情感词汇本体库[37,38]，是大连理工大学信息检索研究室在林鸿飞教授的指导下经过全体教研室成员的努力整理和标注的一个中文本体资源。该资源从不同角度描述一个中文词汇或者短语，包括词语词性种类、情感类别、情感强度及极性等信息。一共本体中的情感共分为 7 大类 21 小类。具体类别如表 3-1。

表 3-1 情感词汇本体库情感词分类含义

编号	情感大类	情感类	例词
1	乐	快乐(PA)	喜悦、欢喜、笑眯眯、欢天喜地
2		安心(PE)	踏实、宽心、定心丸、问心无愧
3	好	尊敬(PD)	恭敬、敬爱、毕恭毕敬、肃然起敬
4		赞扬(PH)	英俊、优秀、通情达理、实事求是
5		相信(PG)	信任、信赖、可靠、毋庸置疑
6		喜爱(PB)	倾慕、宝贝、一见钟情、爱不释手
7		祝愿(PK)	渴望、保佑、福寿绵长、万寿无疆
8	怒	愤怒(NA)	气愤、恼火、大发雷霆、七窍生烟
9	哀	悲伤(NB)	忧伤、悲苦、心如刀割、悲痛欲绝
10		失望(NJ)	憾事、绝望、灰心丧气、心灰意冷
11		疚(NH)	内疚、忏悔、过意不去、问心有愧
12		思(PF)	思念、相思、牵肠挂肚、朝思暮想
13	惧	慌(NI)	慌张、心慌、不知所措、手忙脚乱
14		恐惧(NC)	胆怯、害怕、担惊受怕、胆颤心惊
15		羞(NG)	害羞、害臊、面红耳赤、无地自容
16	恶	烦闷(NE)	憋闷、烦躁、心烦意乱、自寻烦恼
17		憎恶(ND)	反感、可耻、恨之入骨、深恶痛绝
18		贬责(NN)	呆板、虚荣、杂乱无章、心狠手辣
19		妒忌(NK)	眼红、吃醋、醋坛子、嫉贤妒能
20		怀疑(NL)	多心、生疑、将信将疑、疑神疑鬼
21	惊	惊奇(PC)	奇怪、奇迹、大吃一惊、瞠目结舌

例如：“车厢内部整体**空间**_令人满意_”，“具有_良好_的**稳定性**”。这两句话两句话的评价对象的词性模板前后是不一样的，而他们在知识库中的 label 分别是：“空间”是 car_part、“稳定性”是 car_opinion_target。此时利用知识库已经在一定程度上识别出来了评价对象。而这两个评价对象上下文的情感词为“令人满意”和“良好”，他们在情感词本体库中分别标注为 PH 和 PB，由上表可以发现，PA 为快乐之意，而 PB

?1994-2018 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

为喜爱之意。两者虽均为正向情感，但细分起来还是可以有更多的信息可以利用。经过试验，利用这些细分的情感信息能一定程度的提高系统抽取的效果

我们的算法设计如下：

算法 1：评价对象抽取算法

输入：分词结果　输出：特征值

对分词结果中的每一个词：

　　若词在情感词典集合中：

　　　返回情感词典的情感分类结果

　　若词在领域知识库集合中：

　　　返回四元组的 label 结果

　　其他：

　　　返回 NULL 字符串

最后我们每一句话生成如下特征结果，例如“音响和空调按键设计清晰易用。”：

音响	n	SBV	car_part	B
和	c	LAD	NULL	O
空调	n	COO	car_part	B
按键	d	ADV	car_part	O
设计	v	COO	NULL	O
清晰	a	ATT	dlut_PH	O
易	a	ADV	NULL	O
用	v	VOB	NULL	O

3.2.3 特征模板

特征模板[35,36]的设计直接影响着抽取结果的性能。本文通过特征模板进行特征的筛选，充分考虑语言特点和领域特性。CRF++使用特征模板文件，可以让使用者综合考虑上下文特征和外部知识。

特征模板是为特征函数提供统一的模板格式。特征模板主要包含当前项的位置信息和属性信息。位置信息表示用于确定当前观察单元的标注结果的窗口大小。属性信

?1994-2018 China Academic Journal Electronic Publishing House. All rights reserved.　http://www.cnki.net

息是指当前单元在当前特征下的特征值。当前的研究，一般选取 2 的窗口大小（-2，-1，0，1，2）。

特征模板主要包括原子特征、复合特征和二元特征及其他们之间的组合。每个原子特征只考虑一种位置信息或者属性信息，复合特征则是考虑两个或两个以上的属性信息和位置信息。

例如：

表 3-2 特征结构样例

音响	n	SBV	Car_part	
和	c	LAD	NULL	
空调	n	COO	Car_part	<=当前观察
按键	d	ADV	Car_part	
设计	v	COO	NULL	

表 3-3 模板文件语法样例解释

U01:%x[-1,0]	和
U12:%x[0,1]	n
U16:%x[-1,1]/%x[0,1]	c/n
U53:%x[1,3]	adv

原子特征，如：U01:%x[-1,0]，U01 代表该模板编号用于区分，中括号中-1 表示当前词的上一个词，0 表示特征的第 0 列，即为和。表示当前的词的上一个词对当前词的抽取有影响。

复合特征，如：U16:%x[-1,1]/%x[0,1]，考虑了当前词上一个词的词性与当前词的词性联合对抽取结果的影响。

3.3 领域词向量扩展评价对象

用 CRF 模型融合各个特征进行序列标注，进行评价对象抽取的时候，结果往往能取得较高的准确率 0.65 左右，而召回率则较低 0.3 左右，召回率的较低，会影响整体的效果。本文提出了一种方法，通过利用领域词向量对 CRF 抽取结果进行扩充，可以较大程度的提高召回率，不明显的降低了准确率。

在 word2vec 中，我们会得到词向量，一般地，我们使用余弦相似度[34]作为两个词之间的语义相关度。我们认为，评价词往往是对一些特定的对象进行评价，他们之

?1994-2018 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

间的距离也往往较近。而如果这个评价词较少的与某个评价对象在一起，则他们的语义相似度相对较低。

如下表 3-4 是对一些样本的测试：

表 3-4 情感词与评价对象距离对比

	技术	宝马	底盘	舒适性	操控	马自达	后备箱
提升	0.134	-0.085	0.082	0.157	0.173	0.040	-0.071
更好	-0.037	-0.003	0.117	0.139	0.102	0.033	-0.022
舒适	-0.074	-0.078	0.077	0.350	0.272	-0.124	0.167

从结果中也可以发现，“提升”情感词与“操控”评价对象相似度最高，“更好”情感词与“舒适性”最高，即舒适性更好。“舒适”排除舒适性后，与“操控”也是最高，即常说的操控舒适。

本文提出一个假设，当判断某一个词是否为评价对象时，应该找与其周围评价词相似的最接近的词，也通常是惯用搭配。当一个评价词旁有多个候选评价对象时，我们通过对其进行相似度量扩展，并设置阈值的方式，找出未被 CRF 预测出来的词，最终扩展我们的评价对象的结果抽取。

算法的具体流程如下：

算法 2：评价对象扩展算法

```
输入：CRF 预测结果      输出：扩展后的结果
按顺序遍历文档中的词：
    若词在情感词典集合中：
        判断前后 k 个词是否已有 CRF 结果：
            若有：
                继续循环
            若无：
                找出前后 k 个词中在知识库中的词，取得与当前情感词相似度最大的词
                若相似度 sim 超过阈值 S，则增加标注其为评价对象
```

其中 k 为窗口大小。S 为认为其为评价对象的相似度大小的阈值，一般为 0.2~0.5。sim 为两个词的词向量余弦距离。

?1994-2018 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

3.4 实验与分析

3.4.1 中文倾向性分析评测

情感分析是近些年来在自然语言处理中的越来越热门。各种国际国内的顶级会议上，针对这一问题的文章有很多；针对倾向性分析的国际评测有 TREC[20] BlogTrack 以及 NTCIR 等。在国内针对汉语的倾向性问题的研究还处于起步阶段。中文倾向性评测是探索针对中文倾向性分析的新技术、新方法的一个组织。

中文倾向性分析评测 2008（COAE2008）共设置 6 个任务，其中那个任务 3 为评价对象抽取。语料主要来源于商品评论，所以评价的对象主要针对于商品本身和商品的属性。这些都可以成为评价对象，即人的自然语言中所评价的事物。

表 3-5 评价的对象的例子

	需要抽取的对象	例句
商品的属性	价格	**价格**方面很“厚道”
商品的附属物	发动机	**发动机**其实没什么可说的
商品附属物的属性	油门反应	**油门反应**也很直接
商品的相关概念	耗油	可能比福克斯还**耗油**
商品相关概念的属性	车内的空气质量	我首先要批评的是**车内的空气质量**

任务 3 包含包含任务所需的语料共计 494 篇文档，共涉及两个领域，汽车领域以及电子产品领域。因为本文主要做汽车领域评价对象识别研究，所做的实验在汽车领域语料中。汽车领域语料共包含 158 篇文章，每篇文章评价对象约为 20-30 个左右，大约 4000 个评价对象的识别。

例如就是期望在下面一段话中找出评价对象（红字）：

斯柯达欧雅虽然已经并入大众，但由于初进中国市场，**价格**方面很“厚道”，在当时可以算是原装车中**性价比**最高的了，于是决定买了这辆斯柯达欧雅 2.0 自动豪华车。 **发动机**成熟，**性能**较好欧雅的这台 **2.0 发动机**其实没什么可说的，现在已经装配了大众很多车型：甲壳虫、帕萨特、高尔夫用的都是这款。从媒体的一些测评来看**发动机** 没什么特别优秀的地方，但是很成熟可靠。**动力性能**不错，**低速扭矩**很大，**油门反应**也很直接，在市内行车非常舒适。尤其是赶上堵车或者插队的时候往往能够先人一步。**制动**更是“一点就有”，典型的德国血统。

?1994-2018 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

3.4.2 不同特征组合结果对比

我们把词特征称作 w、词性特征称为 pos、依存特征称为 dp、和本文新提出的利用领域知识的新特征称为 Domain-Knowledge-Feature，简称 DKF。利用准确率、召回率和 F1 值作为评价结果。

表 3-6 不同特征预测结果

	准确率	召回率	F1 值
w	0.552	0.368	0.441
w+pos	0.604	0.387	0.471
w+pos+dp	0.637	0.393	0.486
w+pos+dp+DKF	**0.664**	**0.412**	**0.508**

通过对比可以发现，首先，当我们只使用词特征的时候，已经能预测出大部分的结果了。在加入 pos 词性特征后，结果有了较大的提升，达到 0.47，这说明，词性特征对评价对象的抽取是相当有用的。然后加入了依存句法特征，结果又有一定程度的提升，说明句法不同词在句子中使用的句法特点是不一样的。最后，当加入本文新提出的领域知识特征后，结果有所提升，这也是符合预期的，因为我们增加了知识。

3.4.3 word2vec 扩展不同相似度阈值结果对比

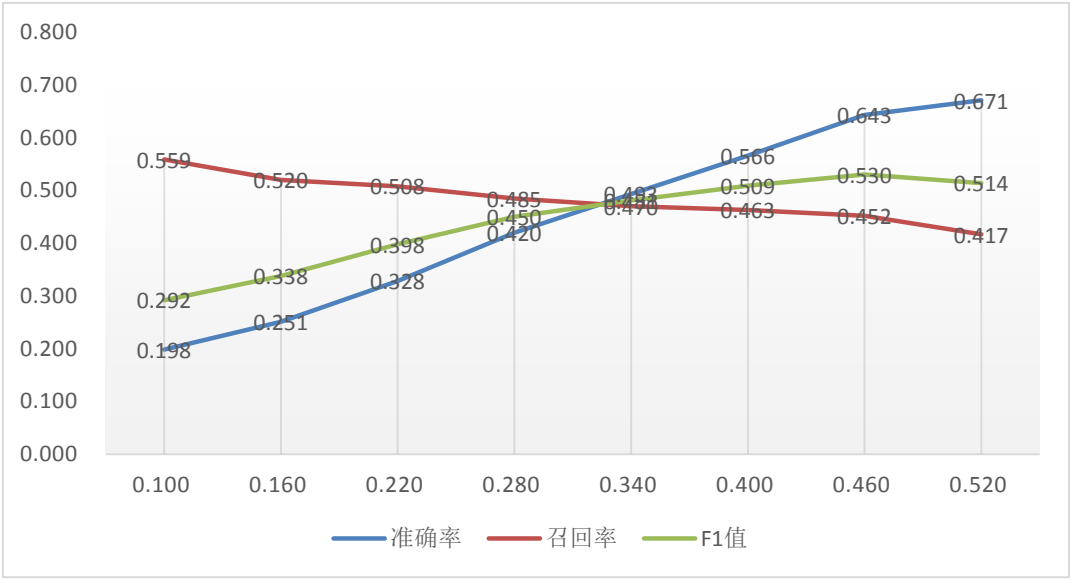

图 3-4 不同相似度阈值对结果的影响

注： X 轴表示在不同的相似度阈值 S 的数据。

?1994-2018 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

从图 3-4 的结果可以看出，随着 S 值的增大，准确率逐渐提高，召回率逐渐降低，F1 值在 S=0.460 的时候达到最高 0.53。

随着继续增加 S 值，准确率有所提升，但是召回率基本回归到了 CRF 结果的水平上，因为当相似度阈值过大的时候，我们在扩展的时候，很难有符合条件的词加入，导致结果回归的 CRF 预测的结果之上。

而我们阈值较小额时候，又会因为加入的词较多，虽然使召回率得以提升，但是准确率会大大折扣。与 CRF 最高的 F1 值 0.508 相比，在 S=0.460 时提高了 0.022。

3.5 本章小结

本章主要介绍了本文提出的两个办法，以提高和扩展评价对象抽取结果。首先，在 CRF 特征选择的时候，利用了构建的领域知识库信息，提高了系统的准确率和召回率。其次，通过使用领域词向量信息，在 CRF 抽取结果的基础上，进行对评价对象的扩展。这样克服了 CRF 结果的准确率较好，但召回率偏低的问题。通过实验可以发现，当相似度阈值设置在 0.460 的时候，系统结果达到了最优 0.53。

?1994-2018 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

第 4 章 基于 web 的汽车领域评价对象抽取验证平台

本章主要介绍依托于评价对象抽取算法的验证平台。考虑到算法与实际应用的结合，我们构建了验证算法结果的凭台，主要是用户通过访问 web 界面，输入文本就可以获取算法的结果，并通过观察验证算法的有效性。

4.1 平台架构设计

4.1.1 系统环境及技术点简介

操作系统 Linux、语言 Python、Web 服务器 Nginx[39]、后端框架 Django[40]、前端 html/css/javascript、交互 ajax[41]。

1. **Linux** 是一套开源的，广泛使用的类 Unix 操作系统，基于 POSIX 和 UNIX。支持 32 位和 64 位 CPU。Linux 性能稳定的网络操作系统。Linux 在国内工业界的非常普及，特别是互联网行业。由于 Linux 的开源与免费，包括在其上的应用软件大多数都是免费开源的，可以方面我们的很多工作。

2. **Python** 作为最近几年特别火的解释型语言，特别是在数据科学领域，天生的对处理文本、数据运算非常方便。特别是在其上的扩展包更是极大的让 python 大展身手，包括 Numpy、Scipy、Sklearn、matplotlib 等。本文算法设计主要是通过 python 实现的。考虑到平台与算法交互的统一性，也选择使用 python 作为 web 平台的开发语言。

3. **Nginx** 是一个高性能的 HTTP 和反向代理服务器，占用内存较小，事实上 nginx 的并发能力确实在同类型的网页服务器中表现较好。有渐渐取代 Apache 服务器的趋势。越来越多的公司选择 Nginx 作为 web 服务器。其主要功能包括 web 服务器、反向代理、负载均衡。特别是在处理高并发、大访问量网站性能卓越。本文使用的是 Nginx 的反向代理功能。

4. **Django** 是开源的 Python Web 框架。Django 是 python 下最流行的框架，主要因为：用于创建模型的对象关系映射、为最终用户设计的完美管理界面、一流的 URL 设计、设计友好的模板语言、缓存系统等。特别是在 URL 与模板语言方面，大大方便了

?1994-2018 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

我的工作。并且可以用他轻松定制 HTTP 协议服务接口，方便程序调用。采用了 MVC 设计模式。

5. **html/css/javascript** 这三种技术是做 web 前端必须的。Html 是超文本标记语言，我们平常浏览的几乎所有网页都是以他为主体构建出来的。Css 是级联样式表，html 语言本身只能定义简单的样式，主要是提供内容，而 Css 则是专门用来定义样式与结构的，是网页显的美观大方。Javascript 浏览器解析的脚本语言，是定义页面动态效果的，web2.0 时代在前端可以说是 javascript 时代，他使我们的网页不再是静的，而是动态的，甚至在网页上做各种各样的功能。

6. **Ajax**（Asynchronous Javascript And XML）是一种可以创建交来回交互的应用的网页技术。通过前端浏览器内部执行 http 请求，与服务器进行通信，AJAX 使网页能够实现异步更新功能。这意味着可以在不重新加载整个网页的情况下，对网页的某部分进行更新。而传统的网页（不使用 AJAX）如果需要更新内容，必须重载整个网页页面。那样会造成，我们每点一个东西，整个页面会刷新一次，重新建立访问，首先造成用户体验的不好，其次是造成资源的浪费，每处理一个页面都是经过服务器各种运算的。现在的 web 应用基本都会使用 ajax 技术，包括 weibo、百度等。往往我们看到的非常炫酷的网站，根本感觉不到页面刷新，就是使用 javacript 和 ajax 设计构建的。本文用以处理数据与服务器交互，而不需要刷新页面，增加用户体验。

4.1.2 平台架构设计

如下图 4-1 所示，整个系统是建立在 Linux 系统平台之上。LTP 服务、算法、模型训练、CRF++等是算法服务的基础，抽取算法是依托于这些算法之上，通过调用这些算法的接口，实现抽取的功能。而抽取算法又为更上层的 web 平台提供服务，算法本身是内容的输入和输出，对用户使用不够友好，需要专业技术才能使用，而在其上封装一层 web 服务，便能够很好的解决问题，并且我们可以方面的通过 web 界面验证我们算法的结果。通过使用 Django 框架调用算法接口，获取抽取结果，把结果返回到浏览器，展示给使用者。

?1994-2018 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

图 4-1 系统分层架构

4.1.3 MVC 开发模式

MVC 模式[42,43,44,45]是“Model-View-Controller”，中文意思是模型-视图-控制器。MVC 由三个部分组成。视图是交互、模型是业务逻辑、控制器用来控制模型和视图的。MVC Web 应用程序被分成 3 个核心部件：数据模型（Model—M）、视图（View—V）、控制器（Controller—C）。MVC 是一种软件开发的设计模式，一个好的 MVC 设计，可以使模型、视图、控制器独立的较低耦合的完成各自的任务，使它们较好的配合，形成 Web 应用程序。

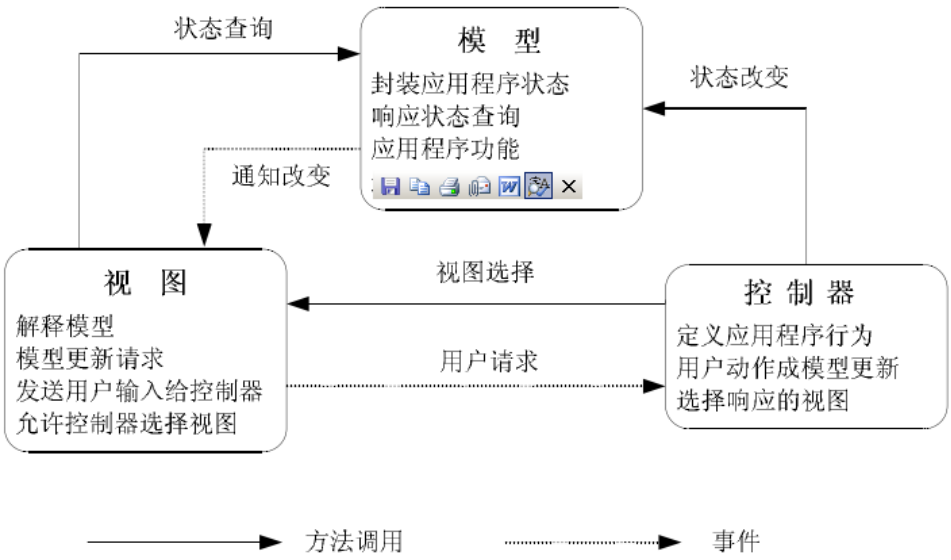

图 4-2 MVC 模型内部关系

?1994-2018 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

（1）视图层

视图是用户看到并与之交互的界面。在 Web 程序中，视图就是前端标签。一个 web 应用有很多的视图界面。数据获取和解析，如模板解析等。把后端控制器的数据填充到相应的位置。视图被传入的数据是由模型层计算生成的，视图仅仅是数据的表示。数据模型与视图的关系是多对多的关系。如“商品”数据模型中，会有视图用户购买的商品，还会有商品列表用于查看商品。这两个视图同时访问“商品”这一数据模型。

（2）模型层

模型是业务流程/状态的处理和业务规则的实现。业务流程可视为黑箱操作，模型接受视图请求的数据，并返回结果。其内部结构对视图是透明的。业务处理的过程对于其他是透明的，或者说黑盒的。总得来说，模型通过接受输入数据，并返数据。对于后端开发者，更需要考虑业务流程。我们在构建 web 应用的时候，往往把各个模型作为独立的模块，这样可以降低系统直接的耦合。对于模型方面，应该做好的是，每一个模块的可用性，正确性和安全性。

（3）控制层

控制层相当于总控制，用来对模型和视图进行组合与调度，完成某一任务，达到较高的复用效果。划分控制层的作用就是作为一个分发起，我们用他来选择模型和视图。控制器是用来调度系统的各个模块。控制器的前方是 url 机制，正常的就是根据不同的 url 来进入不同的控制器，进行不同的业务逻辑的处理。每一个控制器都可以说是一个 http 服务接口，接受输入并返回输出。只不过，当某个控制器是界面时，我们看到了实际的页面。而有的控制器是一层服务，这层服务不返回具体的标签，只返回数据，由 ajax 对前端界面进行修改。

MVC 开发模式具有很多优点：

1）耦合性低、重用性高。视图层和业务层分离，允许更改视图而不需要改变控制器和业务的代码。MVC 允许使用各种不同样式的视图来访问同一个服务器端的代码，因为多个视图能共享一个模型。

?1994-2018 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

2）生命周期成本低、部署快。MVC 使开发和维护用户接口的技术含量降低。而且大大缩短开发时间。

3）分工明确。系统将不同的模块分成不同部分，在程序设计中能够有效的实现合理的分工。使后端程序员专注于业务逻辑，前端程序员专注于表现形式。

4.2 系统详细设计

4.2.1 基于 Socket 的算法间数据传递

评价对象抽取算法作为服务进行提供，而我们的 web 平台是另一个进程，如何使两个程序之间交换数据，这就是进程间通讯。进程间通信主要包括三种方式，管道[46,47]、系统 IPC[48,49]（Inter-Process Communication）、套接字（Socket）。

Socket[50,51,52,53]发展到现在，并不为 Linux 所专有，在所有提供了 TCP/IP 协议栈的操作系统中几乎都提供了 Socket，而所有这样操作系统，对套接字的编程方法几乎是完全一样的。本文便是采用 Socket 方式使 web 平台与算法服务之间进行通信。

Socket 是应用层与 TCP/IP 协议族通信的中间软件抽象层，可以称之为中间件，是一组接口。Socket 是一个门面模式，把 TCP/IP 协议族隐藏在 Socket 下面，而用户只需要调用 Socket 接口就可以了，剩余的由 Socket 去组织，以符合协议的规范。结构如下图 3-4：

?1994-2018 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

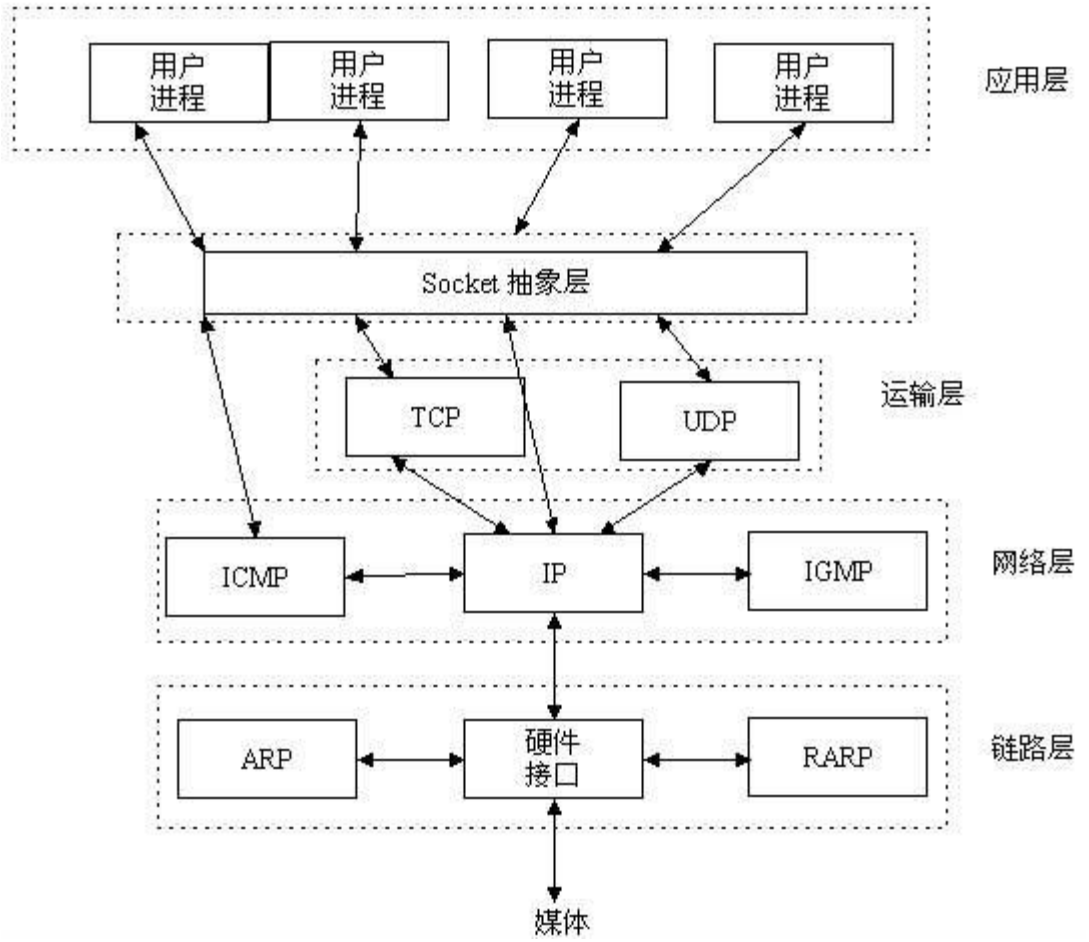

图 4.3 计算机网络协议分层关系

Socket 通信目前已有方便的第三方模块可直接使用，其原理与生活场景中两人间的交流类似：首先你得知道对方的地址，或者说手机号，对方需要手机开机，手机开机就是监听着自己的手机。这时你的电话打来，对方收到，选择接收即建立连接，你们可以自由的通话。通话结束，挂断电话，关闭连接。Socket 一次通信的流程大致是这样的。

Socket 通信具体的流程如下图所示。首先，服务端先初始化 Socket，然后与端口绑定（bind），绑定意味的进程就不能再使用该端口了。然后进行监听（listen），调用 accept 阻塞，等待客户端的连接。在这个时候，如果有一个客户端初始化了一个 Socket，然后连接服务器监听的地址，如果连接成功，客户端便与服务端建立了连接。然后客户端发送数据请求，服务器接受请求并处理请求，然后把响应数据发送给客户端，客户端读取数据，最后关闭连接，一次交互就结束了。

?1994-2018 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

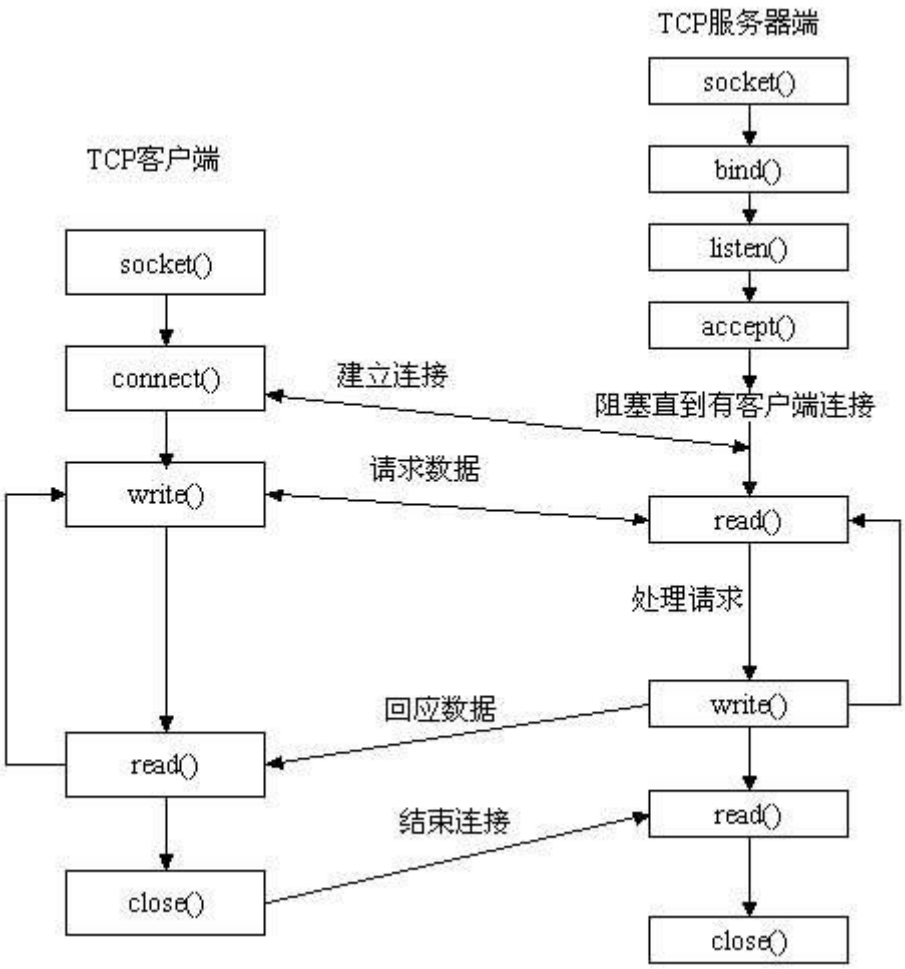

图 4-4 Socket 通信过程

在我们的系统中，评价对象抽取算法被封装成为 Socket 服务，并监听 2015（默认）端口。只要接受到文本数据传入，便进行评价对象算法的抽取，并以 json 字符串的形式返回给客户端数据。

4.2.2 一次抽取详细流程

到这里，我们不仅了解了我们算法设计与实验，而且了解了我们对外的应用平台的技术架构及实现。

?1994-2018 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

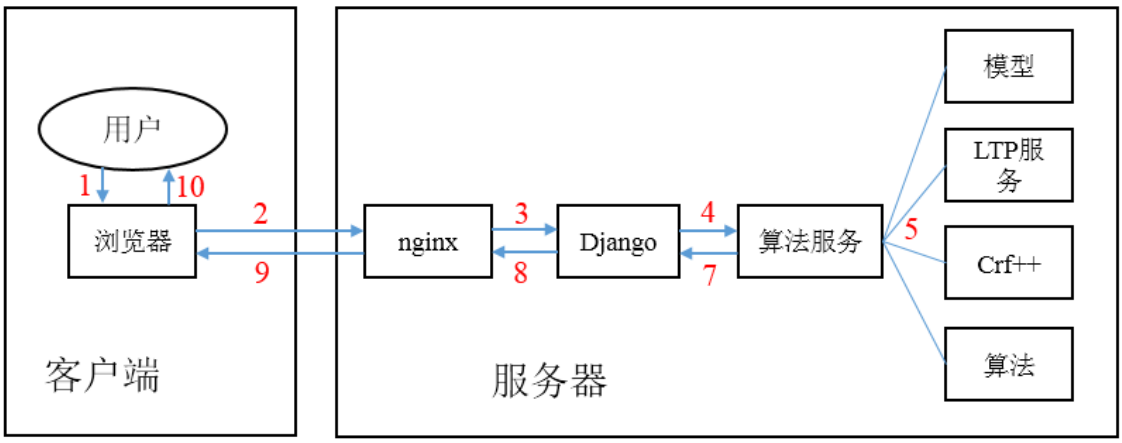

图 4-5 信息在系统中传输过程

1. 首先用户通过浏览器浏览我们网站的地址，在左侧输入想要抽取的文本。
2. 点击“分析”按钮，向服务器发起 ajax 请求，内容为文本内容。
3. 请求经过互联网到达 web 服务器 Nginx，即服务器的 80 端口，Nginx 什么也不做，只对 http 请求做一次转发，转发到 Django 服务器 8000 端口。
4. Django web 收到请求，首先对请求 url 进行解析，发现是要抽取文本评价对象，于是封装好数据，与服务端建立 Socket 连接，发送要抽取的数据。
5. 算法服务，具体工作在第三章，这里就不赘述。我们可以把它看成黑盒，输入为文本内容，输出为 json 格式的带标注结果的内容。是评价词的标注为 B。
6. Django web 收到算法服务结果，json 字符串格式。
7. 把结果传递给 Nginx。
8. Nginx 再转发给浏览器。
9. 浏览器收到 json 格式结果后，对结果进行解析，比如，把标注为评价对象的词字体颜色设置为红色。把结果填充到右侧文本框中。
10. 用户看到返回的结果。

?1994-2018 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

4.3 系统展示

4.3.1 算法抽取样例

如下表 4-1 展示了文本的抽取结果，其中第一列表示内容来源，主要是选取的互联网上的文本信息。第二列是进行分析的文本。第三列为分析结果，空格表示分词，加粗表示抽取出的评价对象。第四列展示抽取出的该段文本的评价对象。

表 4-1 抽取结果样例表

内容来源	待分析文本	分析结果	评价对象
http://www.autohome.com.cn/drive/201505/873434.html	相比锐界的淡定和稳重，汉兰达的表现则要显得更轻快一些。转向的阻尼较小，转动起来比锐界更加轻松。抛开较高的坐姿和视野，锐界的驾驶感受能让人随时意识到自己驾驶的是一台 SUV 的话，那么当你握住汉兰达的方向盘时，你会有一种置身在舒适的轿车中的错觉。	相比 锐界 的 淡定 和 稳重 ， **汉兰达** 的 表现 则 要 显得 更 轻快 一些 。转向 的 阻尼 较 小 ，转动 起来 比 锐界 更加 轻松 。 抛开 较高 的 坐姿 和 **视野** ， 锐界 的 **驾驶感受** 能 让 人 随时 意识 到 自己 驾驶 的 是 一台 SUV 的话 ， 那么 当 你 握住 汉兰达 的 方向盘 时 ， 你 会 有 一种 置身 在 舒适 的 轿车 中 的 错觉	**汉兰达、视野、驾驶感受**
http://www.autohome.com.cn/drive/201505/873499-all.html	最后我总结一下，这台顶配的 C3-XR 性价比很一般，在同级别中配置较低而且售价太高，所以对于讲究实用性的用户来说这台车并不是一个最佳的选择。如果你喜欢雪铁龙 C3-XR 这台车外观的话，那不妨多考虑考虑低配版本。	最后 我 总结 一下 ， 这 台 顶配 的 C3 - XR **性价比** 很 一般 ， 在 同级别 中 配置 较 低 而且 **售价** 太 高 ，所以 对于 讲究 实用性 的 用户 来 说 这 台车 并 不是 一个 最佳 的 选择 。如果 你 喜欢 雪铁龙 C3 - XR 这 台车 外观 的话 ， 那 不妨 多 考虑 考虑 低 配 版本	**性价比、售价**
http://www.autohome.com.cn/drive/201505/871535-all.html	全系最低配的 20T 两驱领先型没有强大的动力和智能的四驱系统，放弃了性能方面的追求，转而走起务实路线。从结果上看，它的各项测试都得到了令人满意的成绩，价格也确实便宜不少，唯一最大的不足是 7 挡双离合变速箱在平顺性和油门响应方面表现实在一般，想买这款车的消费者一定要做好充分的心理准备。	全系 最低 配 的 20 T 两驱 领先型 没有 强大 的 **动力** 和 智能 的 四驱 系统 ， 放弃 了 性能 方面 的 追求 ，转而 走 起 务实 路线 。从 结果 上看 ， 它 的 各项 测试 都 得到 了 令人满意 的 成绩 ， 价格 也 确实 便宜 不少 ， 唯一 最大 的 不足 是 7 挡双离合 变速箱 在 **平顺性** 和 **油门响应** 方面 表现 实在 一般 ， 想 买 这 款车 的 消费者 一定 要 做好 充分 的 心理准备	**动力、平顺性、油门响应**

?1994-2018 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

4.3.2 系统界面

如下图 4-6 所示，从用户角度来看，我们的系统实现了如下页面的应用。用户通过左侧输入框，输入想被抽取的文本，点击“按钮”，会在右侧展示算法分析的结果。标红色的字体为抽取的结果。

图 4-6 系统用户界面

?1994-2018 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

4.3.3 代码文件结构

Web 平台的代码目录结构（只展示到 3 级结构）如下：

```
platform/
├── README.md
├── service     #抽取算法服务目录
│   ├── coae2008-car
│   │   ├── car_target_daemon.py
│   │   ├── client_daemon.py
│   │   ├── configure.py
│   │   ├── crf.py
│   │   ├── func0317.py
│   │   ├── main.py
│   │   ├── output/
│   │   ├── preprocess.py
│   │   ├── source/
│   │   ├── template/
│   │   ├── test.py
│   │   └── test.result
│   └── nohup.out
└── web       #web 平台目录
    ├── data
    │   └── crf_data
    ├── db.sqlite3
    ├── manage.py
    └── web_platform
        ├── __init__.py
        ├── analysis/
        ├── car_extraction/
        ├── settings.py
        ├── urls.py
        └── wsgi.py
```


?1994-2018 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

第 5 章 总结与展望

本文针对汽车领域的评价对象抽取问题，开展了较为系统的研究工作，包括从开始的领域数据获取和知识表示，再到评价对象抽取算法的分析和设计，最后实现了一个 web 方式的抽取验证平台。在评价对象抽取方面，提出了结合领域知识库信息的方式，提高了系统的准确率。同时，利用词向量的办法扩展了评价对象，从而弥补了 CRF 抽取评价对象召回率较低的问题。最后，通过构建基于 Web 的验证平台，不仅完成了算法工作的展示环境，而且可以更为直观的验证算法的有效性。

本文的主要工作和创新主要包括如下几个方面：

1）通过获取并分析汽车领域网站的半结构化信息，抽取出有结构的部分，构建了汽车领域知识库，以四元组（概念 1，概念 2，关系，类别）的形式存储并得以利用。最后一共生成领域 3456 个四元组。并把这些概念分为了品牌、车系、部件等类别。然后，通过抽取汽车领域正文内容，得到近 1G 的纯文本，通过 word2vec 训练出领域词向量。通过对比发现词向量在领域范围内要好于维基百科中文语料。

2）在评价对象抽取方面，我们依托于 COAE2008 评测的 Task3 汽车语料，进行了算法的实验。提出了结合领域知识库信息融合到 CRF 中，提升了系统的效果。由于 CRF 准确率较高，但召回率较低的原因，我们通过利用领域词向量，定义相似度阈值，并对比不同阈值对实验结果的影响，发现，当阈值定位 0.46 时，F1 值达到了最高值 0.53。

3）实现了基于 Web 的评价对象抽取验证平台，使用了互联网行业当前的主流技术，包括 Nginx、Django、Ajax、Socket 通信等，保证了系统的技术先进性、可扩展性和较好的时间性能。

后续工作可以考虑以下两个方面的改进。

首先，在知识库利用方面，现在我们虽然建立了四元组知识库，但是在利用方面并没有把概念之间的关系利用起来，而只是使用了词典的分类信息，未来考虑如何把领域词之间的关系利用起来，这是一块相当有用的信息。

?1994-2018 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

其次，在抽取验证平台方面，现在我们主要是用来把它当做快速展现算法结果和验证平台，在未来，我们可以考虑开发给普通用户，并可以利用用户的反馈信息以提升系统结果。比如，可以在我们抽取结果的页面下面，增加一个判断标识，当用户看到我们结果的时候，可以表示对抽取结果的态度，比如好或中或差。然后通过利用这些反馈信息，使我们的抽取结果更加的准确。

?1994-2018 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

参考文献

[1] 赵军, 许洪波, 黄萱菁. 第一届中文倾向性分析评测情况介绍[C]. //中国中文信息学会成立二十七周年学术会议. 2008.

[2] Gey F, Larson R, Kando N, et al. NTCIR-GeoTime overview: Evaluating geographic and temporal search[C]//NTCIR. 2010, 10: 147-153.

[3] Quinlan J R. Induction of decision trees[J]. Machine learning, 1986, 1(1): 81-106.

[4] Eddy S R. Hidden markov models[J]. Current opinion in structural biology, 1996, 6(3): 361-365.

[5] 李荣陆, 王建会, 陈晓云等. 使用最大熵模型进行中文文本分类[J]. 计算机研究与发展, 2005, 42(1):94-101.

[6] 祁亨年. 支持向量机及其应用研究综述[J]. 计算机工程, 2004, 30(10):6-9. DOI:10.3969/j.issn.1000-3428.2004.10.003.

[7] Lafferty J, McCallum A, Pereira F C N. Conditional random fields: Probabilistic models for segmenting and labeling sequence data[J]. 2001.

[8] Zhuang L, Jing F, Zhu X Y. Movie review mining and summarization[C]//Proceedings of the 15th ACM international conference on Information and knowledge management. ACM, 2006: 43-50.

[9] Kessler J S, Nicolov N. Targeting Sentiment Expressions through Supervised Ranking of Linguistic Configurations[C]//ICWSM. 2009.

[10] Jakob N, Gurevych I. Extracting opinion targets in a single-and cross-domain setting with conditional random fields[C]//Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics, 2010: 1035-1045.

[11] Hu M, Liu B. Mining and summarizing customer reviews[C]//Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining. ACM, 2004: 168-177.

[12] Li B, Zhou L, Feng S, et al. A unified graph model for sentence-based opinion retrieval[C]//Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics. Association for Computational Linguistics, 2010: 1367-1375.

[13] Popescu A M, Etzioni O. Extracting product features and opinions from reviews[M]//Natural language processing and text mining. Springer London, 2007: 9-28.

[14] 刘鸿宇, 赵妍妍, 秦兵等. 评价对象抽取及其倾向性分析[J]. 中文信息学报, 2010, 24(1):84-88. DOI:10.3969/j.issn.1003-0077.2010.01.015.

[15] 俞士汶, 段慧明, 朱学锋等. 综合型语言知识库的建设与利用[J]. 中文信息学报, 2004, 18(5):1-10. DOI:10.3969/j.issn.1003-0077.2004.05.001.

?1994-2018 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

[16] 许德山, 张运良, 李芳. 中文本体三元组的单字索引与更新方法研究[J]. 图书情报工作, 2014, 58(22). DOI:10.13266/j.issn.0252-3116.2014.22.018.

[17] Mikolov T, Chen K, Corrado G, et al. Efficient estimation of word representations in vector space[J]. arXiv preprint arXiv:1301.3781, 2013.

[18] Apache. The Apache Software Foundation[J]. Proceedings of the Apachecon Las Vegas, 2009, 45(10):8-9.

[19] 姚天昉, 聂青阳, 李建超等. 一个用于汉语汽车评论的意见挖掘系统[C]. //中文信息处理前沿进展——中国中文信息学会二十五周年学术会议. 2006.

[20] Voorhees E, Harman D. TREC : experiment and evaluation in information retrieval[J]. Clientrd Com, 2005, 53(1):45--39.

[21] Chandgotia N. Generalisation of the Hammersley-Clifford Theorem on Bipartite Graphs[J]. Eprint Arxiv, 2014.

[22] 徐冰, 赵铁军, 王山雨等. 基于浅层句法特征的评价对象抽取研究[J]. 自动化学报, 2011, 37(10):1241-1247. DOI:10.3724/SP.J.1004.2011.01241.

[23] 王荣洋, 鞠久朋, 李寿山等. 基于 CRFs 的评价对象抽取特征研究[J]. 中文信息学报, 2012, 26(2):56-61. DOI:10.3969/j.issn.1003-0077.2012.02.011.

[24] 郑敏洁, 雷志城, 廖祥文等. 中文句子评价对象抽取的特征分析研究[J]. 福州大学学报：自然科学版, 2012, (5):584-590.

[25] Liao C, Feng C, Yang S, et al. A Hybrid Method of Domain Lexicon Construction for Opinion Targets Extraction Using Syntax and Semantics[M]//Social Media Processing. Springer Berlin Heidelberg, 2014: 108-116.

[26] 王步康, 王红玲, 袁晓虹等. 基于依存句法分析的中文语义角色标注[J]. 中文信息学报, 2010, 24(1):25-29. DOI:10.3969/j.issn.1003-0077.2010.01.005.

[27] 刘挺, 王开铸. 基于篇章多级依存结构的自动文摘研究[J]. 计算机研究与发展, 1999, 36(4):479-488. DOI:doi:10.1007/BF02946505.

[28] 刘挺, 车万翔, 李生. 基于最大熵分类器的语义角色标注[J]. 软件学报, 2007, 18(3):565-573.

[29] 刘怀军, 车万翔, 刘挺. 中文语义角色标注的特征工程[J]. 中文信息学报, 2007, (1):79-84.

[30] 连乐新, 胡仁龙, 杨翠丽,等. 基于中文宾州树库的浅层语义分析[J]. 计算机应用研究, 2008, 25(3):674-676. DOI:doi:10.3969/j.issn.1001-3695.2008.03.008.

[31] 陈耀东, 王挺, 陈火旺. 浅层语义分析研究[C]. //计算机研究与发展. 2007:321-325.

[32] 梁邦勇, 唐杰, 李涓子,等. 语义 Web 下知识正确性检查的研究[J]. 计算机集成制造系统, 2005, 11(3):446-450. DOI:doi:10.3969/j.issn.1006-5911.2005.03.026.

?1994-2018 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

[33] 杨靖. 领域本体自动构建的关键技术研究[D]. 哈尔滨工业大学, 2008.

[34] 张振亚, 王进, 程红梅等. 基于余弦相似度的文本空间索引方法研究[J]. 计算机科学, 2005, 32(9):160-163. DOI:10.3969/j.issn.1002-137X.2005.09.041.

[35] 鞠久朋. 评价对象抽取研究[D]. 苏州大学, 2011. DOI:10.7666/d.y1991179.

[36] 王荣洋. 评价对象抽取关键技术研究[D]. 苏州大学, 2012. DOI:10.7666/d.y2120828.

[37] 陈建美. 中文情感词汇本体的构建及其应用[D]. 大连理工大学, 2008. DOI:doi:10.7666/d.y1417690.

[38] 徐琳宏, 林鸿飞, 潘宇等. 情感词汇本体的构造[J]. 情报学报, 2008, 27(2):180-185. DOI:10.3969/j.issn.1000-0135.2008.02.004.

[39] 凌质亿, 刘哲星, 曹蕾. 高并发环境下 Apache 与 Nginx 的 I/O 性能比较[J]. 计算机系统应用, 2013, (6):204-208. DOI:10.3969/j.issn.1003-3254.2013.06.048.

[40] 高峰, 杨连贺. Flex 技术与 Django 开发框架的整合研究[J]. 计算机与数字工程, 2010, (1):94-96. DOI:10.3969/j.issn.1672-9722.2010.01.027.

[41] 王星, 潘郁. 基于 AJAX 技术的 Web 模型在网站开发中的应用研究[J]. 微计算机信息, 2006, 22(27):206-207. DOI:10.3969/j.issn.1008-0570.2006.27.072.

[42] 任中方, 张华, 闫明松等. MVC 模式研究的综述[J]. 计算机应用研究, 2004, 21(10):1-4. DOI:10.3969/j.issn.1001-3695.2004.10.001.

[43] 黎永良, 崔杜武. MVC 设计模式的改进与应用[J]. 计算机工程, 2005, 31(9):96-97. DOI:doi:10.3969/j.issn.1000-3428.2005.09.036.

[44] 戴朝晖, 吴敏. 基于 MVC 模式的 Web 管理信息系统分析与设计[C]// 第 14 届中国过程控制会议暨第 3 届全国技术过程故障诊断与安全学术会议. 2003:413-415.

[45] 赖英旭, 刘增辉, 李毛毛. MVC 模式在 B/S 系统开发中的应用研究[J]. 微计算机信息, 2006, 22(30):62-64. DOI:doi:10.3969/j.issn.1008-0570.2006.30.020.

[46] 钱培德, 张元道. UNIX 管道通信机构及其程序设计[J]. 计算机世界月刊, 1992, (2):63-65.

[47] 刘悦, 杨波. UNIX 下进程间的消息传递与同步[J]. 中国经济和信息化, 1997, (34):50-50.

[48] 李新明, 李艺. UNIX 系统中 IPC 机制综述[J]. 指挥技术学院学报, 1996, (1).

[49] 陈霖. UNIX 内部进程协作机制应用研究[J]. 电脑知识与技术：学术交流, 2007, 1(6). DOI:doi:10.3969/j.issn.1009-3044.2007.06.105.

[50] 孙钦龙, 邵惠鹤. Socket 套接字在工业数据通信中的应用[J]. 控制工程, 2006, 13(3):274-277. DOI:10.3969/j.issn.1671-7848.2006.03.026.

?1994-2018 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

[51] 何进, 谢松巍. 基于 Socket 的 TCP/IP 网络通讯模式研究[J]. 计算机应用研究, 2001, 18(8):134-135. DOI:doi:10.3969/j.issn.1001-3695.2001.08.045.

[52] 王丰锦, 邵新宇, 喻道远,等. 基于 SOCKET 和多线程的应用程序间通信技术的研究[J]. 计算机应用, 2000, 20(6):65-67. DOI:doi:10.1007/BF02948846.

[53] 王晓鹏. TCP/IP 下的 Socket 及 Winsock 通信机制[J]. 航空计算技术, 2004, 34(2):126-128. DOI:doi:10.3969/j.issn.1671-654X.2004.02.037.

?1994-2018 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

攻读硕士学位期间发表的论文

[1] Liao C, Feng C, Yang S, et al. A Hybrid Method of Domain Lexicon Construction for Opinion Targets Extraction Using Syntax and Semantics[M]//Social Media Processing. Springer Berlin Heidelberg, 2014: 108-116.

?1994-2018 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

致谢

时光荏苒，岁月如梭，两年的研究生生涯就快要结束了。不禁感慨时间过的真快，在学校里，不仅收获了专业能力的增长，而且还结识了一群最可爱的人。

感谢我的导师冯冲老师。冯老师和蔼可亲、幽默风趣、能设身处地的考虑学生的问题。冯老师在学习上给了我很多的帮助，特别的，工作实践上让我得到了充分的锻炼，这对我未来踏入工作岗位有非常大的帮助。同时感谢冯老师对我论文的悉心指导。

感谢我的爸爸妈妈。

感谢院长黄河燕老师，实验室辛欣老师、鉴萍老师、史树敏老师。黄老师对学生无微不至的关怀，还组织了每周大组会，让我们互相学习，增强了相互交流。特别的，感谢辛老师机器学习课程对模型的仔细推导，让本对公式特别恐惧的我，只想说这一次就这一次我真的懂了些。

感谢人工智能课程刘峡壁老师、感谢面向对象技术课程金旭亮老师、感谢英语课程石艳老师。铭记刘老师在结课时给我们的嘱托“做不臣之人、建不臣之制度”。感谢石老师的课程，每节课都提问我，让我感受到了课堂上被关注的感觉和演讲的锻炼。

感谢廖纯，同在一个屋檐下一起学习研究很快乐和对我论文的帮助。感谢胡燕林，本文中的一个重要创新点就是在与你一起头脑风暴时产生的。感谢付佳。感谢刘全超博士、魏骁驰博士，与你们交流学术为何总会让我茅塞顿开。由于 NLP2013 小伙伴们声势浩大，并且欠我很多顿饭，需要先请再谢。

?1994-2018 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net