# 摘要

# Abstract

# 目录

# 第1章 绪论

## 1.1 研究背景及意义

当前的社会是一个高度网络化的社会，随着近年来网络的发展，人们的社交有相当大的一部分逐渐转移到网络上。近年来，随着一些网络社交平台如国内的新浪微博、腾讯微博，国外如Twitter等的出现，大批的人群加入到网络社会这个虚拟的社会中，而网络社会里的信息也呈爆炸式的增长。这些社交媒体上包含了大量的有用的信息，我们从中可以发掘出有指导意义的内容。

新浪微博是一个可以把文字、图片等信息分享至自己的关注者的平台，它基于关注/取消关注来建立/解除好友关系。它的一大特点是发布的微博最多只能是140个字符，这种限制能够让用户把信息说明的更加清晰，因此，选择微博作为本文获取信息的来源。

要让计算机程序能够代替人力从微博中获取到有用的信息，就必须把微博文本转换成计算机能够识别的格式，并且相应的信息短能够用一些特定的标记符标记，这一过程通常被称作自然语言处理（Natural Language Processing，简称NLP）。

一些社会事件常常第一时间在微博平台上曝光出来，因此，微博上的信息具有及时性强的特点。然而，由于微博用户群体的庞大，虽然数据量会很大，但其稳定性却很难保证，因此，我们的微博数据来源主要是一些政府、相关部门，相关民间权威人士的微博，这样的数据更有真实性。

但微博这种类型的社交媒体属于新兴媒体，与传统媒体不同，它所携带的信息具有很多冗余，其中有很多都是无用信息。因此，如何从海量的信息中发掘出有用的信息点成了人们关注的焦点。

## 1.2 国内外研究现状

本小节将从自然语言处理和社交媒体事件发掘两个方面来阐述国内外的研究现状。

### 1.2.1 自然语言处理

自然语言处理是实现人与计算机之间直接通信的重要手段，自上个世纪90年代以来，国内外关于自然语言处理的研究日益增多，在特定的应用领域，一些商业化的应用已经开始出现，例如多语种数据库和专家系统的自然语言接口、各种机器翻译系统、全文检索系统等。

基于中文自然语言处理的难点目前主要在两个方面：一是当前的语言处理分析都限定在一个孤立的语句上，对于上下文环境的研究少之又少；二是由于自然语言的数据量庞大，无法完全存储到计算机中，当前的研究只能建立在一个有限的数据基础范围内。

### 1.2.2 社交媒体事件发掘

社交媒体事件发掘，就是从社交媒体平台上，海量用户发布的信息中，发掘出有价值的信息。随着网络的普及，社交媒体事件发掘的研究也越来越多，一些产品化的应用也开始出现，例如百度舆情等。

## 1.3 本文主要研究内容

本文主要研究如何从微博中发掘出有关的交通新闻类事件。主要包括：微博数据的抓取、过滤以及预处理部分、对微博预料的自然语言处理部分、事件挖掘的网站控制台部分以及事件挖掘的模型建立部分。

其中，微博数据的抓取研究了如何从微博平台抓取数据以及其标准化过程；对微博语料的自然语言处理部分研究了如何对微博语料进行自然语言处理；网站控制台部分研究了如何搭建一个可以查看并维护事件发掘系统的网站；模型建立部分研究了如何建立并完善一个微博事件主题模型。

本文所实现的事件发掘系统的整体结构及流程如诶图1.1所示。首先从微博获取数据经过相关过滤、标准化处理之后存入数据库，然后将数据库中的微博预料通过自然语言处理模块进行相关处理，同时，在网站上对事件进行人工标注相关元素，将人工标注后的结果和自然语言处理后的结果进行合并后转换成CRF训练输入标准格式的文件，通过CRF训练模块进行训练，最后得到事件发掘的结果。



图1.1 事件发掘系统流程图

经过本系统的训练处理，最后事件发掘的结果召回率为76.98%，精确度为83.62%，准确率为93.92%，综合F指数为80.17%。

## 1.4 本文主要结构

本文的主要结构如下：

1. 为绪论部分，论述本文研究的课题、项目背景、国内外研究现状以及本文所设计系统的结构、流程。
2. 论述了微博抓取及预处理模块的实现方案以及运行情况，讨论了相关方案的选择原因及性能。
3. 论述了自然语言处理也就是语言特征标注部分的实现过程及理论依据，描述了当前流行的一些自然语言处理方案的优缺点。
4. 论述了本系统网站控制台部分的技术选型以及实现过程，该网站控制台是系统操作的主要方式，本章还讨论了从网站完全控制系统运行的可能性。
5. 论述了条件随机场算法的定义以及核心思想，讨论了目前研究领域对CRF算法的应用以及CRF工具的选择。并论述了利用CRF工具完成事件发掘核心模块的实现方案。本章还给出了提高结果可靠性的方案及测试过程，展示了最优训练参数的选取过程，是本文的核心章节之一。
6. 描述了整个系统的运行情况，展示了事件发掘的结果。
7. 分析了系统运行结果，对系统的不足提出了改进方案。

# 第2章 微博数据获取及预处理

## 2.1 数据需求简述及分析

## 2.2 抓取方案选择

### 2.2.1 Python爬虫框架Scrapy

### 2.2.2 自写爬虫

### 2.2.3 新浪微博API

### 2.2.4 综合方案选择

## 2.3 微博的数据清洗及标准化

### 2.3.1 微博主题筛选

### 2.3.2 无关字符过滤

### 2.3.3 主体内容截取

## 2.4 数据存储入库

## 2.5 本章小结

# 语言特征标注

## 3.1 自然语言处理概述

## 3.2 NLP工具选择

## 3.3 中文分词

## 3.4 词性标注

## 3.5 命名实体识别

## 3.6 依存句法分析

## 3.7 人工事件标记

## 3.8 本章小结

# 后台管理网站搭建

## 4.1 网站需求分析与说明

## 4.2 网站技术栈确定

### 4.2.1 前端技术选型

### 4.2.1 后台技术选型

## 4.3 本章小结

# CRF算法与事件挖掘的实现

## 5.1 事件发掘的模型选取

## 5.2 条件随机场（CRF）

### 5.2.1 CRF定义

### 5.2.2 CRF优势

## 5.3 CRF工具

### 5.3.1 CRF工具选择

### 5.3.2 CRF++训练套件

## 5.4 CRF训练

### 5.4.1 训练数据的获取

### 5.4.2 训练模板的设计

### 5.4.3 训练结果测试

## 5.5 最优训练参数筛选

### 5.5.1 评测体系

### 5.5.2 训练参数选择

### 5.5.3 最优参数确定

## 5.6 本章小结

# 系统运行结果

## 6.1 系统整体结构

## 6.2 微博抓取模块结果

## 6.3 网站运行情况

## 6.4 语言特征标注结果

## 6.5 CRF训练结果测试

# 结论

## 7.1 本文工作总结

## 7.2 研究与展望

# 参考文献

# 附录

# 致谢