

SHANGHAI JIAO TONG UNIVERSITY

**学士学位论文**

THESIS OF BACHELOR



论文题目：基于CRF的中文社交媒体的事件发掘研究

学生姓名: 祁思蒙

学生学号: 5100369001

专 业: 信息安全

指导教师: 黄征

学院(系): 信息安全

**基于CRF的中文社交媒体的事件发掘研究**

摘要

网络社会和现实社会处于同一时空维度下，与现实社会息息相关。随着网络社会的发展，它能越来越真实地反映出现实社会的当前关注点、指定方向关注点等社会信息，甚至对现实社会正在施以越来越大的影响力。社交媒体（Social Media）承载了大量的信息，从中可以发掘出很多有指导、启发意义的内容。

由于中文的特殊性，例如表意方式多种、理解方式多种、缩略语及新生词较多等，中文社交媒体事件发掘中很多语言处理问题，例如分词、词性标注等相对英文来说具有更大的挑战性。进一步，同样基于中文表意的复杂性和不确定性，中文事件的标注也面临较大的困难，标注方式多样化，给工具的训练过程带来一定的不可预知性。

基于以上，本课题将中文自然语言处理技术与社交媒体，特别是微博的自身特性相结合，进行中文社交媒体的事件发掘研究。经过微博抓取、事件标注、命名实体识别（NER, name entity recognition）、分词（Seg, segmentation）、词性标注（POS, part of speech）等阶段，手动建立事件标注集，建立模板文件，设计并实现基于CRF的中文社交媒体事件发掘工具。

关键词：中文，社交媒体，事件，发掘，CRF

**EVENT EXTRACTION FROM CHINESE**

**SOCIAL MEDIA BASED ON CRF**

**ABSTRACT**

Network society exists at the same time and space dimensions with our traditional realistic society, and is closely related to it. Along with the development of the network society, it can more truly reflect the social reality with social information such as the current concerns, or the very concern we want to know about. Even more, the network society may have greater influence on the real society lately. Social media carry a large amount of information, and from them, we can extract meaningful content to help us building a better understanding of our life.

However, due to the particularity of Chinese, such as a variety ways of understanding, a variety ways of expression, abbreviations, and more new words, Chinese social media event extraction has a lot of language processing problems, such as word segmentation and part-of-speech tagging. They are relatively more challenging compared with English. Furthermore, also based on the complexity and uncertainty of Chinese ideographic, annotation also faces great difficulties. The diversification way for marking brings certain unpredictability to our training tools.

This project combines Chinese natural language processing technology with the characteristics of social media, using Weibo for example, and explore research on Chinese social media event extraction. There are five steps for natural language processing: Weibo extraction, Event labeling, NER (name entity recognition), Seg (segmentation), and POS (part-of-speech tagging). The event set and the template for CRF is manually generated. Based on above, this project conducts the aim of event extraction from Chinese social media based on CRF.

**Key words:** Chinese, social media, event, extraction, CRF

目录

[第一章 绪论 1](#_Toc389121945)

[1.1 项目背景 1](#_Toc389121946)

[1.2 项目意义 1](#_Toc389121947)

[1.3 本文结构 2](#_Toc389121948)

[第二章 研究现状及难点分析 3](#_Toc389121949)

[2.1微博信息提取策略 3](#_Toc389121950)

[2.2自然语言处理发展 3](#_Toc389121951)

[2.3中文自然语言处理 4](#_Toc389121952)

[第三章 课题题目内容解读及思考 6](#_Toc389121953)

[3.1 社交媒体 6](#_Toc389121954)

[3.2 事件发掘 6](#_Toc389121955)

[3.3 中文 6](#_Toc389121956)

[3.4 CRF 7](#_Toc389121957)

[3.4.1 CRF定义 7](#_Toc389121958)

[3.4.2 CRF优势 9](#_Toc389121959)

[3.4.3 CRF工具 9](#_Toc389121960)

[第四章 事件发掘系统结构及实现 10](#_Toc389121961)

[4.1 微博抓取 12](#_Toc389121962)

[4.2事件标注 14](#_Toc389121963)

[4.3 命名实体识别 16](#_Toc389121964)

[4.4 分词 17](#_Toc389121965)

[4.5 词性标注 18](#_Toc389121966)

[4.6 训练 19](#_Toc389121967)

[4.6.1 建立训练模板 19](#_Toc389121968)

[4.6.2 训练模型 20](#_Toc389121969)

[4.6.3 测试 21](#_Toc389121970)

[第五章 结论 22](#_Toc389121971)

[5.1 项目总结 22](#_Toc389121972)

[5.2 项目不足 22](#_Toc389121973)

[5.3 进一步研究方向 23](#_Toc389121974)

[谢辞 26](#_Toc389121975)

**第一章 绪论**

网络社会和现实社会处于同一时间和空间维度下，与现实社会有着千丝万缕的联系。随着网络社会的发展，它能越来越真实地反映出现实社会的当前关注点、指定方向关注点等社会信息，甚至对现实社会正在施以越来越大的影响力。社交媒体（Social Media）承载了丰富的信息，从中可以发掘出很多有指导、启发意义的内容。

1.1 项目背景

当前社会手机、pad等移动终端迅速发展，成为继电脑之后又一重要交流媒介。由于它们的方便、易携性，人们逐渐适应了使用移动终端的社交媒体。这部分社交媒体具有简短、即时等特性，其中的重要代表之一即为“微博（Microblog）”。与人人等支持博客类的社交网站相比，微博的最多140字的字数要求使提取信息更加简单、精确，微博的关注公共主页功能使提取信息更加具有针对性。在中国大陆的微博平台中，新浪微博人数最多影响力最大，2014年3月27日新浪微博改名为“微博”，成为中国最主流、发展最完善的微博平台。

微博基于用户关系信息，是一个通过关注-被关注的机制去分享他人实时消息的平台。用户可以使用电脑、手机、pad等各种客户端登陆微博，一方面发送自己的信息，包括文字、图片、音频、视频、网页等形式，梳理自己当时的见闻及感受；另一方面关注其他人或公共主页，获得好友的信息或针对某各领域的信息。用户之间通过关注、评论、转发等形式来进行互动。它具有很强的实效性和随意性，从宏观角度看事实上形成了一个个和现实社会拥有很大相似点的个性化社交网络群，并且信息在群中方便快捷地产生并传播。

除了传统意义上朋友间的互相关注外，微博中有很多名人认证账户和公共主页账户，这些账户更加趋向于单向关注，即普通微博用户将它作为单纯的信息来源，交流互动较朋友间要少一些。近几年出现的名词“大V”指的即是名人身份获得认证的在微博领域的意见领袖。这些名人主要是一些学者、明星等，拥有巨大的粉丝（关注者）团体，在微博上具有一定的影响力，时时引导者互联网上的言论和话题，带给普通民众新的关注点。公共主页账户则是提供某一个特定领域的信息，例如 “环球时报”、“上海交通大学”、“上海交通广播”等。在这种特定领域的公共主页中，博主会以该领域的微博为主，为关注者提供比较纯粹、更有针对性的信息。

除了娱乐交流、分享实时信息等作用外，微博在政治领域也在逐渐被认可。据新京报报道，截止2013年6月，政务微博认证账号也已经超过了24万个。近年来，“反腐微博”等工作更是在微博与政治间搭了一座桥梁。

1.2 项目意义

基于以上背景内容我们可以看到，网络社会和现实社会处于同一时空维度下，与现实社会息息相关，具有便捷性、背对脸、原创性、草根性等鲜明的特点。不论是众多民众个体生活的点点滴滴、感想与琐事，还是爆发性的新闻，亦或是大批人同时关注的全球重大事件，微博已经成为众多网民分享生活、发表意见、关注事件的一个重要的渠道。随着微博的迅速发展，它能越来越真实地反映出现实社会的当前关注点、指定方向关注点等社会信息，以新兴媒体的身份融入到了现实社会中，甚至对现实社会正在施以越来越大的影响力。因此，在这个平台进行事件发掘，对于政府舆情分析、网络舆情分析、事件监控等领域都有着广阔的应用空间。在本设计中，主要以微博作为社交媒体进行研究，以新浪微博作为微博的代表进行信息抓取与分析。

1.3 本文结构

本文结构如下：

第一章：介绍项目的研究背景，即微博的发展和壮大及给社会生活带来的重大影响。分析项目的意义，即微博事件发掘能帮助人们通过微博获得的信息。

第二章：研究现状及难点分析。主要从微博信息提取策略、自然语言处理发展和中文自然语言处理这三方面分析同一领域研究项目的研究现状和研究难点。

第三章：课题题目内容解读及思考。根据课题题目展开，第三章分别就“社交媒体”、“事件发掘”、“中文”、“CRF”四个关键词进行说明，对各个关键词在本项目中的现实意义进行分析与解读。着重介绍了CRF这一数学模型，并确定针对各个关键步骤的研究思路和方法。

第四章：事件发掘系统结构及实现。主要说明了从项目的起步到最后得出实验结果的详细步骤。从微博抓取、事件标注、命名实体识别、分词、词性标注这几个方面详细说明了每一步骤的内容，以及每一模块中间的关联。最终得到实验结果。

第五章：对项目进行总结，思考不足之处，并就不足之处和未尽之处进行接下来工作的展望。

**第二章 研究现状及难点分析**

在微博的事件发掘框架中，微博具有文本松散、语言随意、信息量低、数据量大等特点，这些都为事件发掘带来了负面影响。因此，如何在一个有如此大影响力的社交媒体中，提取出具有指导、启发意义的事件并呈现在大家面前，是目前中文社交媒体处理领域具有较大发展空间的研究课题。

2.1微博信息提取策略

对于微博的事件发掘，主要指在海量的微博中，针对每条微博信息的140字之内，确定信息的主体/客体/时间/地点/事件等信息，判断其是否有潜在用途或是否是感兴趣的信息，并将有用信息格式化地呈现出来，即概括为事件。

相比中国的新浪微博，美国开发的Twitter在世界上具有更大的关注度和用户量，相应的事件发掘研究工作也更加完善。2011年，Jui-Yu Weng等人提出了对微博信息进行“疑问”、“讨论”、“分享”、“聊天”四种信息分类，进行信息的简化呈现[1]；2012年，Alan Ritter等人提出了一套完整的基于开放域Twitter事件发掘系统，虽然处理的是英文语料，但是系统思路对本项目的研究具有很大的启发意义[2]。

在国内，2011年，张晨逸等人提出了基于MB-LDA模型的微博主题挖掘方法[3]；2012年，郑斐然等人提出了一种微博数据处理的新方法以及一种新闻话题检测算法[4]。

2.2自然语言处理发展

自然语言是一种自然地随文明发展而演化的语言，如汉语、英语、阿拉伯语等，目的是为了人与人之间的通信。从信息论的角度来讲，我们可以认为，语言其实就是对信息进行编码，编码和解码的规则就叫做“语法”。在实际交流、使用自然语言的过程中，我们把大脑中要表达的意思，通过语言的语句表达出来，告诉其他人或记录下来，本质上就是用这种作者使用的语言特有的编码方式，对作者大脑中想要表达的信息编码成一段文字。而对方“会说”某种语言，或能阅读记录的文字，则本质上就是用该种语言独特的解码方法获得文字中表达的信息。用计算机进行自然语言处理，其实并非是很长时间以来大家理解的用电脑来模拟人脑，而是凭借数学模型去理解自然语言。[5]

传统的自然语言处理采用基于规则方法，该方法是理性主义，主要是由语言学家预先针对各种不同的语言，编写相应的字典和语义规则表等规则文件。然后计算机科学家使用计算计算法去描述这些规则，得到计算机程序能够解释的机器语言，从而利用计算机科学家写出的程序执行各种规则来处理输入的文字，从而达到“理解”这些输入的文字的目的。但是在应用方面面临着两个重要的问题，并且几乎无法解决。第一，对于规则数量的要求极其庞大，基本上无法完成。在真实生活中，人脑对句子的理解并不局限于规则，而一个新的、人类很容易理解的句子在计算机中或许就要匹配一个新的规则。另外，即使在相同的文法规则下，一个句子在不同环境中也会呈现不同的语义。为了获得句子的正确信息，便需要加以额外说明。这种基于特定的使用环境的规则也不太可能被一一列举。因此，这种方法语言知识覆盖率并不高，同时也没有一个统一的机制来彻底解决各种语言知识之间的冲突。第二，计算及解析语法规则是困难的，因为描述自然语言的文法和计算机程序的问法不同。主要问题在于自然语言的文法是上下文有关文法，而程序语言多设计成上下文无关文法。

20世纪70年代后，基于统计学的自然语言处理方法开始出现。这种方法是经验主义，使用机器学习的方法来获得语言知识。同样是语言学家建立文件库。不同的是，基于统计学概念下，语言学家建立的是具有抽样代表性的语料库。然后计算机科学家则去建立统计模型，基于数学知识，利用前面建立的语料库去训练统计模型的参数，最后利用训练好的数学模型归纳出算法，进行自然语言的处理。在这一阶段，词汇库的构建获得了普遍的关注。基于统计学的自然语言处理方法有很多，比较主流的模型有隐马尔科夫模型(HMM, Hidden Markov Model)、最大熵马尔科夫模型(MEMM, Maximum Entropy Markov Model)、条件随机域(CRF, Conditional Random Field)等。

John Lafferty等人提出[6]，相比隐马尔科夫模型和随机文法，CRF具有很多优势。前者是生成模型，要求参数被训练至训练集联合概率最大化，需要去列举所有的观察序列可能性。通常意义上，即需要一个表示方法，使得观察值都是合适的原子实体，例如词语。事实上，这种表示多个交互特点或者距离较远的依赖关系是不切合实际的，因为这种模型推断极难处理。基于以上问题，条件模型具有更好的效果。对于一个观察序列，条件模型可以为可能的标记序列指定概率值。因此，它在观察值上并不增大建模工作量。进一步，标记序列的条件概率可以依赖观察序列中任意的、有重叠性的、非独立的特征，而不必要求模型来解释这些依赖关系的分布。同时，条件随机域模型可以任意地区添加外部特征，从而提供了增大模型信息量的条件。所选特征可能会表示同一观察值不同粒度的属性，例如文字和单词。标记之间的转移概率可能不仅取决于当前的观测值，也取决于过去和未来的观测值，即可以充分利用待识别文本中的上下文信息。相比最大熵马尔科夫模型和其他基于图形有识别力的马尔科夫模型，CRF同样具有一些优势。前者存在着标记偏差问题，即状态间转移时，转移概率的大小仅仅是下一状态的相对概率大小，并且对概率进行归一化，并不考虑整个模型中的其他状态转移。这种做法称为“得分权重守恒”，后继状态的概率值和必须为1。它可能带来不可避免的偏差，在极端情况下，如果某一状态只有一个后继状态，那么即使事实上可能性极低，但是转移概率还是为1。而条件随机域的概率模型范围是所有的状态，那么即使某一状态只有一个后继状态，他的转移概率也不必强制为1。

2.3中文自然语言处理

中文自然语言处理相比英文等语言，一个比较特殊的地方在于中文需要进行分词。2.2中提到的基于统计语言模型的自然语言处理工作中，语言模型是建立在词的基础上的，因为在大部分语言中，词语是表达特定语义的最小单位。对于西方的拼音语言来说，词语之间有天然且明确的分解符（例如英语中的空格）；而对于中文来说，自然语言中最小粒度则是汉字。因此，在进行中文自然语言处理时，首先要对句子进行词语切分后，才能进行经典的词性标注、命名实体识别、语义角色标注等工作。然而由于微博客户端中，手机等便携式终端占很大比重，在这类终端中用户的输入更加趋向与简洁自由，语言不规范性极大。尤其是一些非格式化的、由输入法提供的字符表情等内容的存在，更是给分词带来了极大的困难。

中文分词（Seg）工作有很多种方法。从历史发展看来，基于字符串匹配的机械分词方法，建立在语义规则上基于文法的分词方法，以及基于统计模型的统计分词方法，都在当时的年代获得了深入的研究。词性标注（PoS）是在分词的基础上，标记每个词语的词性。命名实体识别（NER）是指识别文本中具有某些特定含义的词语，例如人名、地名等等。在英文的命名实体识别中，往往由某些词语的特殊格式或天然的首字母大写即可基本上完成命名实体识别，而在中文中词语切分和命名实体识别则是互相影响的。在微博事件发掘中，更是很有可能出现命名实体识别中提取的人名、地名等信息即是事件关键字的情况，所以微博事件发掘中命名实体识别很重要，能为事件发掘提供很好的特征。语义角色标注，即任务中的标记事件，则是基于以上对文字的处理之后，对句子中词语成分在事件中的地位进行标记，例如时间、地点、实施体、受事体、事件等等。

**第三章 课题题目内容解读及思考**

3.1 社交媒体

本项目中的社交媒体主要以微博为代表，着重进行金融和交通信息方面的事件发掘。

微博基于用户关系信息，是一个通过关注-被关注的机制去分享他人实时消息的平台。用户可以使用电脑、手机、pad等各种客户端登陆微博，一方面发送自己的信息，包括文字、图片、音频、视频、网页等形式，梳理自己当时的见闻及感受；另一方面关注其他人或公共主页，获得好友的信息或针对某各领域的信息。用户之间通过关注、评论、转发等形式来进行互动。它具有很强的实效性和随意性，从宏观角度看事实上形成了一个个和现实社会拥有很大相似点的个性化社交网络群，并且信息在群中方便快捷地产生并传播。

微博的来源是基于新浪微博官方API的微博内容抓取。新浪微博提供了微博开放平台（Weibo Open Platform），它向用户提供各种丰富应用和完全服务。一般来说，微博开放平台为手机、电脑、pad等终端的应用程序提供账号支持、链接/发布功能等，便于应用程序进行产品的推广。它开发者提供了200多个免费开放的公共API，用户注册开发者账号后可以在一定的限制内调用这些API，来获得与微博有关的信息，例如微博内容、评论、用户、话题、关系等。它提供了多种SDK，包括C++、JAVA、Action Script、PHP、Python、iOS、Android、WP7、JS等[7]。

在本项目中需要调用开放平台的Python接口，实现信息读取，获得所关注的人的最新微博，API为statuses/friends\_timeline，其返回得到的微博的JSON串。在程序中获得JSON串中“text”等关键字内容，并储存在PostgreSQL数据库中，作为最基础的语料。

3.2 事件发掘

本项目中的“事件发掘”指在文本条目中提取出感兴趣的信息。项目中设置兴趣点为时间(When)、地点(Where)、事件(Trigger)、施事体(Who)和受事体(Whom)。

训练集为手工标注获得。在3.1中的语料数据库的基础上，使用Django框架搭建网页，进行手工事件标注后，写回数据库中。

事件发掘的目标为输入微博，输出标记过的事件，包括时间和/或地点和/或事件和/或施事体和/或受事体。

3.3 中文

本项目中，使用到的中文自然语言处理步骤包括命名实体识别（NER, name entity recognition）、分词（Seg, segmentation）、词性标注（POS, part of speech）、事件标注四个部分。其中命名实体识别和词性标注这两部分工作研究已经颇具成果，已有模型和工具已经能比较理想地完成预期任务；分词系统则采用上海交通大学毕业设计中，信息安全工程学院姚雨石同学的毕业设计成果，即针对社交媒体的分词工具。事件标注部分主要是人工构建微博数据训练集，来训练出适合于中文微博的CRF模型。

值得提出的是，由于中文的特殊性，导致命名实体识别同样比在英文系统中更难处理。命名实体识别和分词的处理结果可能会出现冲突。例如，“上海交大学生”作为输入，命名实体识别系统可能识别“上海交大”为一个命名实体，而分词系统可能识别“大学生”为一个词语。为了避免这种冲突，同时考虑微博词汇语言具有很强的随意性和口语化特点，微博中命名实体，尤其是未登录词出现频率较大而且往往占据相当重要的地位，本项目提高命名实体识别的优先级。即将句子首先输入命名实体识别系统中，将识别出的命名实体前后人为添加空格，防止分词得到的词语截断命名实体识别结果。在这种前提下，例如某段短语为ABCDEFGH，命名实体识别识别出CDEFG为一个命名实体，则在分词处理之后，BC不可能在同一个此种，GH不可能在同一个词中，但CDEFG可能被分为多个词语，例如CD和EFG。在这种情况下，CD和EFG的命名实体识别标记与CDEFG标记相同。

3.4 CRF

2001年，John Lafferty等人提出了条件随机域模型。CRF是一个为了对序列数据进行分段和标注而构建的条件模型概率框架，模型思想主要源自最大熵模型，可以看成是一个马尔科夫随机场，或者是无向图模型。它不是在当前状态下去定义下一状态，而是在具有观察序列的基础上去获得标记序列联合概率。

3.4.1 CRF定义

在实验中，记X为数据序列中待标注的随机变量，记**Y**为对应的标注序列变量。所有**Y**的元素都在一个有限的集合中。例如，**X**可以是自然语言句子，而相应的**Y**可以相应地是词性标注标记，而**Y**的元素则分布在所有可能得词性标注标记。我们构建一个条件模型，和一个隐含的边际概率模型。

下面是CRF的定义：

在图中，有，**Y**中元素与*G*中的顶点相对应。则在条件**X**下，随机变量服从马尔科夫属性：

 (1)

其中，表示*w*和*v*是G中相邻的两个顶点，即是G的边。此时，成为一个条件随机域。

CRF是在观察值**X**下的全局随机域。我们不妨假设图G是固定的。在最简单也是最重要的序列模型中，可以认为G是一个简单链式或者线式模型：。事实上**X**可以具有一个自然图结构，但是通常情况下，我们并不一定必须假设**X**和**Y**具有相同的图结构，甚至可以认为**X**并没有任何图结构。在这里，我们可以以，为例去研究分析。

如果图下，**Y**是一棵树。则有

 (2)

式中**x**是数据序列，**y**是标记序列，是与子图S中的定点相关联的**y**的组分集。

我们可以假设，特征和是指定的。例如，词是大写字母并且标记是专有名字的时候，特征作为一个布尔顶点，值就可以为真。

参数估计的问题事实上就是根据具有经验分布的训练集去估计参数。Lafferty等人给出了一个迭代算法来最大化对数似然函数。

 (3)

作为一个特例，我们可以通过为每一个状态对和每一个状态观测对分别定义一个特征，来构造一个类似于HMM的CRF。

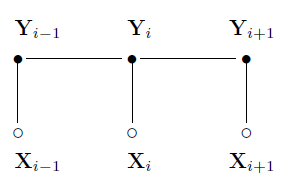
 (4)

 (5)

参数和相互对应，事实上作用就类似于通常HMM中和的作用。玻尔兹曼链模型也有一个相似的形式，但是它使用的是一个单独的归一化的常量来描述联合分布，而CRF是用的是基于观测值的归一化函数来构建条件分布。

虽然说CRF可以包含类似于HMM的模型，但是事实上条件随机域更具有表达力，因为条件随机域允许观测序列有任意的依赖关系。另外，CRF中的特征并非必须完全指定一个状态或者观察值，所以它建立模型可能会需要更少的训练数据。并且，CRF的损失函数是上凸的，它具有上凸模型的普遍优点。

在分析过程中，假设在X条件下Y的依赖关系形成一个连式结构。为了简化一些表达式，我们添加特殊的开始和停止状态，。由此，我们可以使用图1中的图结构。



**图1 CRF链式图结构**

对于链式结构来说，标记序列的条件概率可以简洁地使用矩阵形式表示，它对于描述参数估计和推断算法有着巨大的作用。假设是（1）中表示的CRF，在观察序列**x**中的位置***i***，定义的变量为：

 (6)

 (7)

其中，是标签的的边，是标记的顶点。

归一化函数（配分函数）就是以下矩阵(start, stop)路径。

 (8)

标记序列y的条件概率为：

 (9)

其中，。

3.4.2 CRF优势

相比隐马尔科夫模型和随机文法，CRF具有很多优势。前者是生成模型（generative models），要求参数被训练至训练集联合概率最大化，需要去列举所有的观察序列可能性。通常意义上，即需要一个表示方法，使得观察值都是合适的原子实体，例如词语。事实上，这种表示多个交互特点或者距离较远的依赖关系是不切合实际的，因为这种模型推断极难处理。基于以上问题，条件模型具有更好的效果。对于一个观察序列，条件模型可以为可能的标记序列指定概率值。因此，它在观察值上并不增大建模工作量。进一步，标记序列的条件概率可以依赖观察序列中任意的、有重叠性的、非独立的特征，而不必要求模型来解释这些依赖关系的分布。同时，条件随机域模型可以任意地区添加外部特征，从而提供了增大模型信息量的条件。所选特征可能会表示同一观察值不同粒度的属性，例如文字和单词。标记之间的转移概率可能不仅取决于当前的观测值，也取决于过去和未来的观测值，即可以充分利用待识别文本中的上下文信息。

相比最大熵马尔科夫模型和其他基于图形有识别力的马尔科夫模型(discriminative Markov model)，CRF同样具有一些优势。前者存在着标记偏差问题，即状态间转移时，转移概率的大小仅仅是下一状态的相对概率大小，并且对概率进行归一化，并不考虑整个模型中的其他状态转移。这种做法称为“得分权重守恒（conservation of score mass）”，后继状态的概率值和必须为1。它可能带来不可避免的偏差，在极端情况下，如果某一状态只有一个后继状态，那么即使事实上可能性极低，但是转移概率还是为1。而条件随机域的概率模型范围是所有的状态，那么即使某一状态只有一个后继状态，他的转移概率也不必强制为1。这样就解决了后继状态较少时出现的标记偏差问题。

3.4.3 CRF工具

本项目使用的CRF工具为上海交通大学信息安全工程学院黄征老师开发的PCRF工具。PCRF是开源实现的条件随机域工具，它能完成对序列数据进行分段、标签等工作。PCRF基于Python2.7.5，可以应用为各种各样的NLP任务重，例如命名实体识别、信息提取和中文分词、事件发掘等等[8]。

PCRF相比其他CRF工具，提供了一些很重要的特性。包括完全基于python并使用numpy、scipy等库、可以重新定义特征集、基于LBFGS的快速训练、训练和测试都使用更少的内存、编码解码均在实用时间内、完全开源、训练文件和特征模板文件同热门CRF工具CRF++通用、多线程工作等。基于这些特性，本项目使用的工具为PCRF。

**第四章 事件发掘系统结构及实现**

系统流程如下：



**图2 获得数据集流程**



**图3 训练及测试流程（十倍交叉验证）**

4.1 微博抓取

新浪微博提供了微博开放平台（Weibo Open Platform），它向用户提供各种丰富应用和完全服务。一般来说，微博开放平台为手机、电脑、pad等终端的应用程序提供账号支持、链接/发布功能等，便于应用程序进行产品的推广。它开发者提供了200多个免费开放的公共API，用户注册开发者账号后可以在一定的限制内调用这些API，来获得与微博有关的信息，例如微博内容、评论、用户、话题、关系等。它提供了多种SDK，包括C++、JAVA、Action Script、PHP、Python、iOS、Android、WP7、JS等。在本项目中需要调用开放平台的Python接口，实现信息读取。

首先，注册开发者账号。一个完备的开发者账号需要姓名、电话等各项信息，但是考虑到本项目需要的微博条目并非庞大，对API次数要求不高等原因，不需要进行开发者身份认证审核，只需提供基本信息即可。由于不是一个真正的网页/客户端应用程序去调用API，在应用程序控制中，授权回调页和取消授权回调页都使用微博开发平台自身的Oauth2.0默认页面https://api.weibo.com/oauth2/default.html即可。

其次，关注一些特定的微博。在本项目中，关心的是例如金融、交通等方向的微博。需要注意的是，在本项目中，要分别构建两个领域的事件集。为了方便起见，两个领域的事件集时间不重合，即分别关注两个领域的微博，省略了分辨事件类型的步骤。

在本项目中，希望获取的是当前登录用户及其所关注用户的最新微博，对应的API为statuses/friends\_timeline。该API支持格式为JSON，HTTP请求方式为GET，访问级别为普通接口，需要登录授权，有频次限制。我的账号具有测试授权，调用总限制为150次/小时[9]。返回的JSON结果示例如下（部分数据，仅供格式示例，下同）：

{

"statuses": [

{

"created\_at": "Tue May 25 17:46:55 +0800 2014",

"id": 11488058555,

"text": "求关注~！小伙伴们一起来呀！"，

"source": "<a href="http://weibo.com" rel="nofollow">新浪微博</a>",

"favorited": true,

"truncated": false,

"in\_reply\_to\_status\_id": "",

"in\_reply\_to\_user\_id": "",

"in\_reply\_to\_screen\_name": "",

"geo": null,

"mid": "5612866880546515491",

"reposts\_count": 5,

"comments\_count": 12,

"annotations": [],

"user": {

"id": 1404846560,

"screen\_name": "meng",

"name": "meng",

"province": "16",

"city": "8",

"location": "北京 海淀区",

"description": "最喜欢第十任Doctor！ALLONS-Y！ ",

"url": "http://blog.sina.com.cn/ meng ",

"profile\_image\_url": "http://tp1.sinaimg.cn/1404334560/50/0/1",

"domain": " meng ",

"gender": "f",

"followers\_count": 8234,

"friends\_count": 456,

"statuses\_count": 9345,

"favourites\_count": 8,

"created\_at": "Fri Aug 26 00:00:00 +0800 2011",

"following": true,

"allow\_all\_act\_msg": true,

"remark": "",

"geo\_enabled": true,

"verified": false,

"allow\_all\_comment": true,

"avatar\_large": "http://tp1.sinaimg.cn/2304376580/180/0/1",

"verified\_reason": "",

"follow\_me": true,

"online\_status": 0,

"bi\_followers\_count": 300

}

},

...

],

"ad": [

{

"id": 4566614911582342,

"mark": "AD21321JYTJJH"

},

...

],

"previous\_cursor": 0,

"next\_cursor": 16688015866,

"total\_number": 82345

}

以上JSON串仅为示例，不保证符合真实情况。

在工程中，可以获得其“text”、“crated\_at”等属性，即微博信息内容、微博创建时间等。需要注意的是，该API需要登陆授权，当每天多次运行该API获得一天不同时段的微博时，需要多次进行登陆token获取工作。该部分需要编写程序自动获得token。

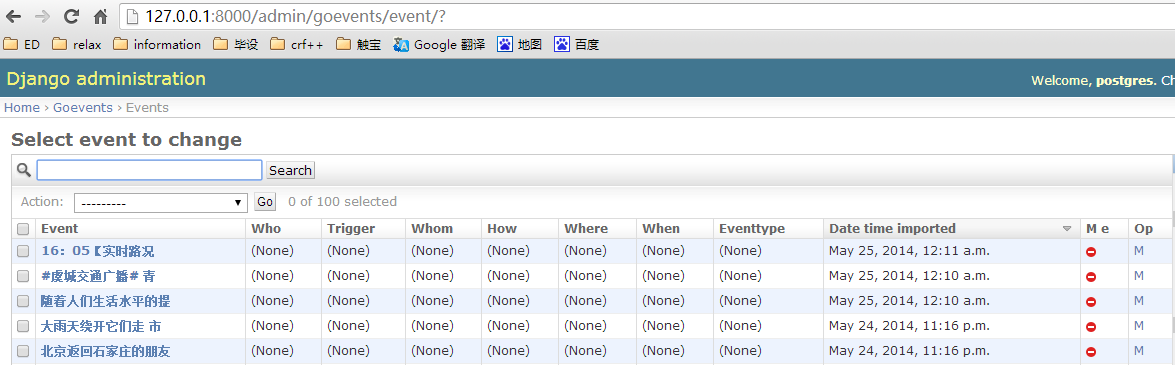
在写入数据库处理中，一条微博中可能包含多个句子。在多个句子中进行指定成分的挖掘是一件复杂且正确率较低的事情。所以针对每条微博，程序提取其中的一句话作为微博的代表。在当前程序中，如果微博具有形如“【xxx】”的格式，且方括号中的内容超过9个字，则认为方括号中为微博内容的提要，对他进行抓取。如果没有该格式的内容或内容字数太少，则舍弃方括号中的文字，直接搜索微博内容中以句号分隔的句子中数字最多的一句话作为微博内容的提要，将其保存在数据库中。

另外，在经过人工标注以后，未标注的条目被舍弃掉，它可能包括签到信息、古文诗词、购物链接等在本课题中视为噪音的内容。

通过window 8系统中的“任务计划”程序功能，我设置此微博抓取模块在每天8：00-23：00，每个整点自动运行。如果获取的微博有重复数据，则根据hash值比对，进行抛弃。

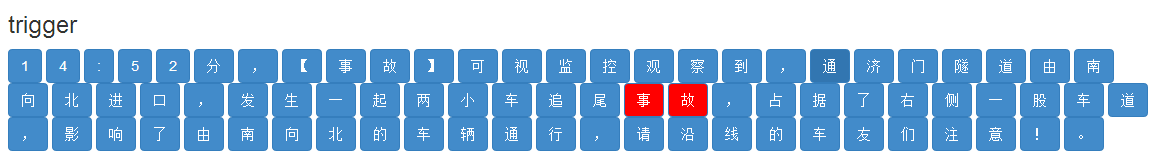
4.2事件标注

事件标注是本任务的重要环节，指在微博条目中提取出感兴趣的信息。项目中获得的微博为交通信息类微博，所以设置兴趣点为时间(When)、地点(Where)、事件(Trigger)、施事体(Who)和受事体(Whom)。事件标注的初始集合为5.1中得到的微博文本数据库。数据库使用PostgreSQL。训练集为手工标注获得。其中，金融方向的事件集是实验室同学共同标注的，交通方向的事件集是我一个人单独标注的。将文本存入PostgreSQL数据库中，使用Django框架搭建网页，进行事件标注后，写回数据库中。标记前网页如图4：



**图4 标记前网页**

标记页面如图5、6、7：（标记仅需点击选择词语的第一个字和最后一个字）



**图5 标记页面-trigger**



**图6 标记页面-who**



**图7 标记页面-when**

标记后网页如图8：



**图8 标记后页面**

上文列举的条目如图9：



**图9 标记后条目**

导出事件，获得的标注集合如下：

<row>

<id>174</id>

<url>3694897624551458</url>

<text>中铁信托拟转让30%股权引入战投交易金额或超65亿。</text>

<eventtype>产经商业信息</eventtype>

<who>中铁信托</who>

<whom>股权</whom>

<how></how>

<when></when>

<where></where>

<whywho></whywho>

<whyhow></whyhow>

<trigger>转让</trigger>

<swho>1</swho>

<ewho>4</ewho>

<swhom>11</swhom>

<ewhom>12</ewhom>

<show>100000</show>

<ehow>0</ehow>

<swhen>100000</swhen>

<ewhen>0</ewhen>

<swhere>100000</swhere>

<ewhere>0</ewhere>

<swhywho>100000</swhywho>

<ewhywho>0</ewhywho>

<swhyhow>100000</swhyhow>

<ewhyhow>0</ewhyhow>

<strigger>6</strigger>

<etrigger>7</etrigger>

<evaluser>sxx</evaluser>

<evaluation xsi:nil="true"/>

<gettime>2014-04-02T13:15:04+08:00</gettime>

<pubtime>2014-04-02T13:15:04+08:00</pubtime>

</row>

其中，“s..”和“e..”分别表示相应关键字在微博文本中出现的序号，100000和0表示未出现。需要说明的是，为了以后进一步的研究使用，在网站中使用了eventtype、trigger、who、whom、how、where、whyhow、whywho等多个关键字，在本设计中，为了简单起见，标注时只使用了who、trigger、whom、where、when等关键字，其他关键字并未进行标注。

对于获得的格式化微博，需要进行两个方向的处理。一是提取所有的微博句子，二是将格式化信息转化为PCRF能处理的语料的格式。在这里我编写了两个程序goeventsToSentence.py、goeventsToCorpus.py。goeventsToSentence.py输出结果为每条微博为一句话、占一行的txt文本。goeventsToCorpus.py输出结果格式如下：

美 bwho

国 who

R who

i who

b who

b who

e who

c who

k who

L who

a who

w who

律 who

所 ewho

， O

将 O

起 btrigger

诉 etrigger

波 bwhom

音 whom

和 whom

马 whom

航 ewhom

。 O

4.3 命名实体识别

命名实体识别（NER），是指识别文本中具有某些特定含义的词语，例如人名、地名等等。在英文的命名实体识别中，往往由某些词语的特殊格式或天然的首字母大写即可基本上完成命名实体识别，而在中文中词语切分和命名实体识别则是互相影响的。在微博事件发掘中，更是很有可能出现命名实体识别中提取的人名、地名等信息即是事件关键字的情况，所以微博事件发掘中命名实体识别很重要，能为事件发掘提供很好的特征。

值得提出的是，在本项目中，将命名实体识别放在了分词之前。原因是命名实体识别和中文分词系统出列结果可能会出现冲突。例如，“上海交大学生”作为输入，命名实体识别系统可能识别“上海交大”为一个命名实体，而分词系统可能识别“大学生”为一个词语（用例仅供说明，不表示一定会发生）。为了避免这种冲突，同时考虑微博词汇语言具有很强的随意性和口语化特点，微博中命名实体，尤其是未登录词出现频率较大而且往往占据相当重要的地位，本项目提高命名实体识别的优先级。即将句子首先输入命名实体识别系统中，将识别出的命名实体前后人为添加空格，防止分词得到的词语截断命名实体识别结果。在这种前提下，例如某段短语为ABCDEFGH，命名实体识别识别出CDEFG为一个命名实体，则在分词处理之后，BC不可能在同一个此种，GH不可能在同一个词中，但CDEFG可能被分为多个词语，例如CD和EFG。在这种情况下，CD和EFG的命名实体识别标记与CDEFG标记相同。

本任务使用的命名实体识别工具为上海交通大学电子信息与电气工程学院的赵海老师开发的baseNER工具，它是未切分中文文本的命名实体识别工具。它基于CRF++写成，使用n-gram特征设置进行训练。中文简体的训练语料为Bakeoff-3的命名实体训练语MSRA，提供在Bakeoff-3对应的测试集上0.8524的总体F值[10]。

命名实体识别过程的输入为goeventsToSentence.py输出的每条微博为一句话、占一行的txt文本。

例如，输入某一行为：美国RibbeckLaw律所，将起诉波音和马航。

则输出格式如下：

[美国]/LOC RibbeckLaw律所，将起诉 [波音]/ORG 和 [马航]/ORG 。

4.4 分词

分词是中文自然语言处理中不可避免的一个项目。基于统计语言模型的自然语言处理工作中，语言模型是建立在词的基础上的，因为在大部分语言中，词语是表达特定语义的最小单位。对于西方的拼音语言来说，词语之间有天然且明确的分解符（例如英语中的空格）；而对于中文来说，自然语言中最小粒度则是汉字。因此，在进行中文自然语言处理时，首先要对句子进行词语切分后，才能进行经典的词性标注、命名实体识别、语义角色标注等工作。

由于微博客户端手机等便携式终端占很大比重，在这类终端中用户的输入更加趋向与简洁自由，语言不规范性极大。尤其是一些非格式化的、由输入法提供的字符表情等内容的存在，更是给分词带来了极大的困难。

本项目使用的分词工具为统一课题组的姚雨石同学毕业设计的内容——中文社交媒体分词工具。此工具的模型使用CRF++进行训练。工具中在CRF++训练之前先利用MMSEG方法对于登录词的高识别性，使用MMSEG算法预先进行一次分词。介于特征向量在CRF算法中具有十分重要的地位，指定特征向量突出了CRF算法可扩展的特性，因此，将MMSEG分词的结果加入训练集作为CRF训练的一个特征，以待提升CRF模型的准确率。MMSEG算法有着比较强的去除歧义方法。它使用四个规则来消除歧义：最大匹配算法、最大平均词语长度、词语长度最小变化率，以及单词单子的语素自由度最大和。它的核心思想是把一个句子尽可能长、尽可能均匀地区分，看上去比较符合中文的语法习惯。但是这种过度抽象的模型忽略了词汇之间的关联性，因此在效果上表现较差，但是对于已登录词汇的高分辨率使得它非常适合来区分出社交媒体中的网络用语。

由4.3分析可以看出，对于本任务，命名实体识别在分词之前完成是很有必要的。对于命名实体识别的结果，我先进行了预处理（goeventsToSentence.py），即将识别出的命名实体前后人为添加空格，防止分词得到的词语截断命名实体识别结果。

例如4.3中提到的输出：

[美国]/LOC RibbeckLaw律所，将起诉 [波音]/ORG 和 [马航]/ORG 。

进行预处理为：

美国 RibbeckLaw律所，将起诉 波音 和 马航。

则分词输出结果为：

美国 RibbeckLaw 律所 ，将 起诉 波音 和 马 航。

注意“马”和“航”作为一个识别出的命名实体，可能会被分词系统分割。

4.5 词性标注

词性标注（PoS tagging或PoST）也叫做语法标记，或单词分类消歧，是一个标记单词在一段特定文本中词性的工具。项目中使用词性标注作为CRF可用的特征之一。

与分词相似，词性标注也是基于计算机语言学的，有着基于规则和基于统计两种方法。本任务使用的PoS工具为上海交通大学电子信息与电气工程学院的赵海老师开发的basePoS工具，它基于最大熵分类器写成。中文模型使用Bakeoff-4的CTB词性标注的训练语料，在Bakeoff-4的CTB测试语料上提供0.941的中文标注精度[10]。

词性标注的输入为分好词的句子。其中分词用空格隔开。输出为标记好词性的句子。例如4.4中提到的分词输出：

美国 RibbeckLaw 律所 ，将 起诉 波音 和 马 航。

经过词性标注过程后，输出为：

美国/NR RibbeckLaw/NR 律所/NN ，/PU 将/AD 起诉/VV 波音/NR 和/CC 马/NR 航/NN 。/PU

词性标注结束后，需要将它和命名实体识别结果进行合并。在之前对命名实体识别结果添加空格的预处理的前提下，制订规则，即例如某段短语为ABCDEFGH，命名实体识别识别出CDEFG为一个命名实体，分词将CDEFG分为CD和EFG，则CD和EFG的命名实体识别标记与CDEFG标记相同。我编写了getPoSNER.py，得到命名实体识别和词性标注合并的结果。结果格式如下：

美国 LOC NR

RibbeckLaw O NR

律所 O NN

， O PU

将 O AD

起诉 O VV

波音 ORG NR

和 O CC

马 ORG NR

航 ORG NN

。 O PU

将分词、命名实体识别、词性标注合并后，需要添加4.2中手动标记的事件。4.2中goeventsToCorpus.py得到的语料是以字为单位，需要和上面的以词为单位的语料进行合并。合并程序为combine.py。合并结果如下：

美国 LOC NR bwho

RibbeckLaw O NR who

律所 O NN ewho

， O PU O

将 O AD O

起诉 O VV etrigger

波音 ORG NR bwhom

和 O CC whom

马 ORG NR whom

航 ORG NN ewhom

。 O PU O

这也是训练集和测试集的格式。

4.6 训练

4.6.1 建立训练模板

PCRF的模板（template）定义如下：模板文件中，每一行代表一个模板（template），定义宏%x[row,col]来作为输入数据中的块（token）。其中，row表示相对于关注块的行数，col表示相对于关注块的列数。即以关注块为坐标（0,0），横向为y轴，向右为y轴正方向，纵向为x轴，向下为x轴正方向的坐标（x,y）[11]。

举例如下：

A a Ⅰ

B b Ⅱ

C c Ⅲ

D d Ⅳ

E e Ⅴ

假设关注块为C，那么特征模板形式为：

Template Expanded feature

%x[0,0] C

%x[0,1] c

%x[-1,0] B

%x[-2,1] Ⅰ

%x[0,0]/%x[0,1] C/c

ABC%x[0,1]123 ABCc123

在模板文件中，模板类型分为U和B两种。

U表示Unigram template，用来描述unigram特征。当模板中出现“U00:%x[0,1]”时，PCRF会自动生成一系列的特征函数，形如：

func1 = if (output = Ⅰ and feature="U01:a") return 1 else return 0

func2 = if (output = Ⅱ and feature="U01:a") return 1 else return 0

func3 = if (output = Ⅲ and feature="U01:a") return 1 else return 0

...

funcXX = if (output = Ⅰ and feature="U01:e") return 1 else return 0

funcXY = if (output = Ⅱ and feature="U01:e") return 1 else return 0

...

一个U模型生成的特征函数有L\*N个，其中L是输出成分的类别，N是该模板扩展出的特征函数的个数。

B表示Bigram template。用来描述bigram特征。在B模板下，系统会自动生成当前关注块和上一个关注块的组合，因此生成的特征函数有L\*L\*N个，L和N含义同上。B模板会产生更多的可区分的特征，相应地，会导致训练时间增长。

本项目中采用的模板文件如下：

# Unigram

U00:%x[-2,0]

U01:%x[-1,0]

U02:%x[0,0]

U03:%x[1,0]

U04:%x[2,0]

U05:%x[-2,0]/%x[-1,0]

U06:%x[-1,0]/%x[0,0]

U07:%x[0,0]/%x[1,0]

U08:%x[1,0]/%x[2,0]

U10:%x[-2,1]

U11:%x[-1,1]

U12:%x[0,1]

U13:%x[1,1]

U14:%x[2,1]

U15:%x[-2,1]/%x[-1,1]

U16:%x[-1,1]/%x[0,1]

U17:%x[0,1]/%x[1,1]

U18:%x[1,1]/%x[2,1]

U20:%x[-2,2]

U21:%x[-1,2]

U22:%x[0,2]

U23:%x[1,2]

U24:%x[2,2]

U25:%x[-2,2]/%x[-1,2]

U26:%x[-1,2]/%x[0,2]

U27:%x[0,2]/%x[1,2]

U28:%x[1,2]/%x[2,2]

U30:%x[-2,1]/%x[-1,1]/%x[0,1]

U31:%x[-1,1]/%x[0,1]/%x[1,1]

U32:%x[0,1]/%x[1,1]/%x[2,1]

U40:%x[-2,2]/%x[-1,2]/%x[0,2]

U41:%x[-1,2]/%x[0,2]/%x[1,2]

U42:%x[0,2]/%x[1,2]/%x[2,2]

# Bigram

B

此模板序列窗口大小为5，在特征方向上参考坐标0（词语本身）、1（命名实体识别）、2（词性标注）等特征，并进行相应的结合来得到模板文件，训练事件发掘模型。

4.6.2 训练模型

PCRF基于命令行的训练命令形如python pcrf-train.py train.txt template model。

得到model文件。

4.6.3 测试

PCRF的测试命令形如python pcrf-test test.py model test-info.txt。

得到test-info文件，文件内容形如：

华为 O NR who who

超越 O VV trigger trigger

行业 O NN whom whom

巨头 O NN whom whom

爱立信 O NN whom whom

。 O PU O O

新华社 ORG NR who who

记者 O NN who who

再次 O AD O O

实名 O VV how trigger

举报 O VV trigger whom

华润 ORG NR whom whom

集团 ORG NN whom whom

董事长 O NN whom whom

。 O PU O O

在自然语言处理中，召回率R（Recall）、精确度P（Precision）和准确率A（Accuracy）常被用来衡量系统的性能。召回率指系统能正确识别出的数目（TP）占标准结果（TP+FN）中数目的比例，精确度指系统正确识别出的数目(TP)占系统识别出的数目(TP+FP)的比例，准确率指系统判断正确的数目（TP+FN）占所有条目（TP+TN+FP+FN）的比例。将召回率和准确率进行几何加权平均，得到F1-Measure（F1指数），作为系统的整体评价。F1的计算公式如下：



在本系统中，认为具有事件信息量的词语即为待识别词语，没有事件信息量（即标记为O）的词语即为非待识别词语。

本项目中，金融信息和交通信息在不同时间段内获取，分别标记、训练、测试，得出召回率、精确度、准确率和F1指数值。训练和测试采用十倍交叉验证。

两个数据集测试结果如表1：

**表1 两种工具处理中文微博测试结果对比**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 事件类型 | 召回率(%) | 精确度(%) | 准确率(%) | F1指数(%) |
| 金融信息 | 42.80 | 72.90 | 80.25 | 53.94 |
| 交通信息 | 77.42 | 82.76 | 93.57 | 80.00 |

**第五章 结论**

5.1 项目总结

在本中文社交媒体事件发掘系统研究项目中，主要是按照管道模型，依次进行了微博抓取、事件标记、命名实体识别、分词、词性标注等过程，获得训练集。使用PCRF进行训练，得到训练结果，进行处理，展示从微博中获得的事件。

在这个模型下，具体的研究内容有以下几点：

第一，系统地分析了相对于英文来讲的中文，和相对于正式文本来讲的微博，在事件发掘系统中具有的困难因素，为接下来的研究工作做准备。

第二，基于以上困难，提出一定的解决方法。例如使用专门为社交媒体语言训练的中文分词工具、手工标注微博建立训练集等。

第三，分析并应用PCRF工具，选取合适的特征，并根据独特的特征和应用场所（微博），建立有针对性的模板。分析模板的特征选取方案，以及模板的运行原理。对于其他研究人员的研究具有一定的借鉴意义。

第四，建立一个完整的微博事件发掘系统，介绍各部分的工作原理，完成整个运行流程。

在不同的两个数据集上，交通信息比金融信息四个参数的数值均显著偏高，且稳定重现。分析原因有以下两点：

第一，金融信息的事件是实验室多名同学共同标注的，而交通信息的事件是我一个人标注的，后者的标注准则比前者要更加一致，得到的数据集更加统一。

第二，金融信息中句式繁多，金融领域事件类型、信息角度等比交通信息量大；而交通信息句式则比较单一，多数满足时间、地点、道路名、路况动词的格式，其中时间、地点可选。

5.2 项目不足

中文社交媒体事件发掘是一个很有意义的研究领域，能为生活带来一定的指导意义。但是从本项目的结果来看，这套系统还很不完善。

从对语料的处理来看，命名实体识别、分词、词性标注这三个系统的正确率虽然比正规语料的训练、测试结果低一些，但是相比其他部分来说已经具有足够的正确率。结果不够理想的主要问题出在训练集的构建和事件发掘系统的训练上。经过分析，本系统出现的主要不足有以下三点：

第一，微博抓取模块抓取关键句不准确。在将微博写入数据库处理中，一条微博中可能包含多个句子。在多个句子中进行指定成分的挖掘是一件复杂且正确率较低的事情。所以针对每条微博，程序提取其中的一句话作为微博的代表。在当前程序中，如果微博具有形如“【xxx】”的格式，且方括号中的内容超过9个字，则认为方括号中为微博内容的提要，对他进行抓取。如果没有该格式的内容或内容字数太少，则舍弃方括号中的文字，直接搜索微博内容中以句号分隔的句子中数字最多的一句话作为微博内容的提要，将其保存在数据库中。事实证明，这种抓取策略抓取关键句不是很准确，部分微博的关键句并非最长的一句，而是第一句或最后一句。最长的一句或几句反而起的是解释的作用。

第二，微博中未登录词造成识别不准确。例如在交通信息的微博中，“沪金高速公路春申路到剑川路路段”中，“沪金高速公路”、“春申路”、“剑川路”这三项并不能被模型很好地区分，从而很难单独将“沪金高速公路”正确地识别为施事体（who）。

第三，训练集太少且精度不够。微博中出现未登录词的概率本身便比正规文本大一些，一些时下流行的短语等也需要在训练集中多次出现才能稳定地正确理解。在标注中，不同标注者、或同一标注者的不同时间段的标注标准不一，造成训练model有偏差。

5.3 进一步研究方向

基于以上分析的项目不足之处，可以从以下四点来进行改进，进一步进行研究：

第一，继续研究针对微博等社交媒体的分词、命名实体识别、词性标注系统，增加对缩略语、新型词汇、表情词汇等内容的理解正确率。

第二，研究如何抓取微博的关键句，避免简单地使用最长句、“【】”括起来的句子、首句、尾句等基于统计得到的抓取方法，而是宏观地对整个微博文本进行分析、研究，得到关键句。

第三，在模板文件的基础上，增加规则，从而增加对类似于“沪金高速公路春申路到剑川路路段”等格式化短语的理解。

第四，建立更大的训练集，同时增强对语言学的了解和认识。制定更加完善的标注规则，增加标注事件的准确性，从而提高模型的精确度。

第五，将管道模型变换为更大的、基于整个微博文本分析的宏观概率模型，从全局的角度进行分析，避免管道模型系统带来的命名实体识别、分词、词性标注部分错误累加。

参考文献

[1] Weng J Y, Yang C L, Chen B N, et al. Imass: An intelligent microblog analysis and summarization system[C]//Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: Systems Demonstrations. Association for Computational Linguistics, 2011: 133-138.

[2] RITTER Alan，Mausam，ETZIONI Oren. Open Domain Event Extraction from Twitter[C]//Proceedings of KDD.[s.l.]：ACM，2012：1104-1112.

[3] 张晨逸, 孙建伶, 丁轶群. 基于 MB-LDA 模型的微博主题挖掘[J]. 计算机研究与发展, 2011, 48(10): 1795-1802.

[4] 郑斐然, 苗夺谦, 张志飞. 一种中文微博新闻话题检测的方法[J]. 计算机科学, 2012, 39(1): 138-141.

[5] 吴军, 谷歌. 数学之美[M]. 人民邮电出版社, 2012.

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

[7] 微博开放平台开发文档.概述[EB/OL].[2014-05-28]. http://open.weibo.com/wiki/概述.

[8] 黄征.Github. Python Linear CRF[EB/OL].[2014-05-28]. https://github.com/huangzhengsjtu/pcrf

[9] 微博开放平台开发文档.API[EB/OL].[2014-05-28]. http://open.weibo.com/wiki/2/statuses/friends\_timeline

[10] 赵海.homepage[EB/OL].[2014-05-28]. http://bcmi.sjtu.edu.cn/~zhaohai/

[11] CRF++: Yet Another CRF toolkit[EB/OL].[2014-05-28]. http://crfpp.googlecode.com/svn/trunk/doc/index.html

[12] Nakagawa T, Uchimoto K. A hybrid approach to word segmentation and pos tagging[C]//Proceedings of the 45th annual meeting of the ACL on interactive poster and demonstration sessions. Association for Computational Linguistics, 2007: 217-220.

[13] Li P, Zhou G, Zhu Q, et al. Employing compositional semantics and discourse consistency in Chinese event extraction[C]//Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning. Association for Computational Linguistics, 2012: 1006-1016.

[14] Sutton C, McCallum A, Rohanimanesh K. Dynamic conditional random fields: Factorized probabilistic models for labeling and segmenting sequence data[J]. The Journal of Machine Learning Research, 2007, 8: 693-723.

[15] Russell M A. Mining the Social Web: Data Mining Facebook, Twitter, LinkedIn, Google+, GitHub, and More[M]. " O'Reilly Media, Inc.", 2013.

[16] Sharifi B, Hutton M A, Kalita J K. Experiments in microblog summarization[C]//Social Computing (SocialCom), 2010 IEEE Second International Conference on. IEEE, 2010: 49-56.

[17] 高强,游宏梁. 事件抽取技术研究综述[J]. 情报理论与实践,2013,04:114-117+128.

[18] 李博. 网络热点事件挖掘及特征描述研究[D]. 国防科学技术大学, 2010.

[19] 文坤梅, 徐帅, 李瑞轩, 等. 微博及中文微博信息处理研究综述[J]. 中文信息学报, 2013, 26(6): 27-37.

[20] 丁荩,涂浩. 微博感知突发重大新闻事件的研究与分析[J]. 广西大学学报(自然科学版), 2011,S1:335-338.

[21] 李博. 网络热点事件挖掘及特征描述研究[D].国防科学技术大学,2010.

[22] 周晶. 基于条件随机域模型的中文实体关系抽取[J]. 计算机工程,2010,24:192-194.

谢辞

大学本科生活已经接近了尾声，毕业设计工作也即将结束。家乡的表弟将在6月份迎来高考，激动的他已经开始期待着自己的本科生活；而即将本科毕业的我则在目睹一批学生的轮回中思绪万千。

四年前我踏入交大，不知道离家一千多公里的代价能否让我让我获得足够的能力。幸运的是，回想这四年，母校交大带给我了很多很多。除了知识技能以外，上海、以及上海交大带给我的更多的是一种看世界的方法，一种新的生活方式，一种家乡城镇永远无法给我的高度。感谢交大，让我在这四年的时光内迅速成长。

一个学校、一个城市的灵魂不是钢筋水泥的建筑，而是活跃在建筑内的人。感谢交大的各位老师，让我从一个懵懂无知、初出家门的少年，成长为有了一定知识技能、懂得如何做人的交大毕业生。信息安全工程学院的老师们让我感受到了前沿的科技、得以一窥知识的殿堂。而我有幸参加过的其他选修课的老师们，则给我打开了另外一扇扇大门，通往自然科学、通往人文科学、通往社会科学。

特别地，我要感谢我的毕业设计指导老师黄征老师。从选题、到开题、到中期、到最后毕业论文的撰写工作，黄老师都一直给与我了很大的帮助和鼓励。当我对实验感到迷茫的时候，黄老师总能一针见血地指点出我的困惑，帮我指清前进的方向。在这里向黄老师表达我最诚挚的谢意！

信息安全工程学院是一个小院，每年的本科生只有100人左右。2014年，信安并回了电院，带给我了一个契机来让我意识到，我有那么喜欢信安的同学们，喜欢信安这个家庭。大一面对好多数学课程大家一起愁眉不展、全院去出游的欢声笑语、面对其他学院的同学说“我认识我们院所有的同学”的自豪，这些或忧伤、或欢笑，都是一份难得的回忆。感谢同学们四年来的相伴。

最后，毕业设计虽然即将结束，但是我的求学道路还在继续。毕业设计中的种种问题，我会在以后的研究生生涯中继续去探索。毕业，不是船儿归港，而是扬帆出航！