硕士学位论文

中文微博评论对象抽取研究

论文作者：梅　寒

指导教师：胡小华　教授

学科专业：计算机应用技术

研究方向：自然语言处理

华中师范大学计算机学院

2016 年 5 月

Research on Opionion Target Extraction in Chinese Microblog

A Thesis

Submitted in Partial Fulfillment of the Requirement

For the M.S. Degree in Computer Application Technology

By

Han Mei

Postgraduate Program

School of Computer

Central China Normal University

Supervisor: Xiaohua Hu

Academic Title:Professor

Signature

Approved

May. 2016

华中师范大学学位论文原创性声明和使用授权说明

原创性声明

本人郑重声明：所呈交的学位论文，是本人在导师指导下，独立进行研究工作所取得的研究成果。除文中已经标明引用的内容外，本论文不包含任何其他个人或集体已经发表或撰写过的研究成果。对本文的研究做出贡献的个人和集体，均已在文中以明确方式标明。本声明的法律结果由本人承担。

作者签名：杨寒　　日期：2016年6月1日

学位论文版权使用授权书

学位论文作者完全了解华中师范大学有关保留、使用学位论文的规定，即：研究生在校攻读学位期间论文工作的知识产权单位属华中师范大学。学校有权保留并向国家有关部门或机构送交论文的复印件和电子版，允许学位论文被查阅和借阅；学校可以公布学位论文的全部或部分内容，可以允许采用影印、缩印或其它复制手段保存、汇编学位论文。（保密的学位论文在解密后遵守此规定）

保密论文注释：本学位论文属于保密，在____年解密后适用本授权书。

非保密论文注释：本学位论文不属于保密范围，适用本授权书。

作者签名：杨寒　　导师签名：胡小华

日期：2016年6月1日　　日期：2016年6月1日

本人已经认真阅读“CALIS高校学位论文全文数据库发布章程”，同意将本人的学位论文提交“CALIS高校学位论文全文数据库”中全文发布，并可按“章程”中的规定享受相关权益。同意论文提交后滞后：□半年；□一年；□二年发布。

作者签名：杨寒　　导师签名：胡小华

日期：2016年6月1日　　日期：2016年6月1日

摘 要

随着自媒体技术的迅速发展，广大互联网用户逐渐从信息的被动接收者转变为信息的生产和分享者。微博平台的迅速崛起产生了海量的文本数据，其中蕴含的信息无论是对研究还是应用都具有非凡的价值。近年来，针对微博的情感分析发展迅速，而作为情感分析的关键任务之一，评论对象抽取由于在文本摘要抽取、舆情分析等多个领域具有极大的应用价值，逐渐受到研究人员的重视。然而目前面向中文微博的相关研究还不充分，由于微博文本具有缺乏语言规范、句子结构不清晰等内在特性，使得研究难度加大。因此，本文将微博作为研究对象，以评论对象抽取任务为研究内容，寻找更有效的中文微博评论对象抽取方法。本文将评论对象抽取任务分为候选词抽取与标准评论对象抽取两个主要步骤，针对不同步骤，对现有方法进行了改进优化。具体而言，本文工作主要有以下四点：

第一，在候选词抽取阶段，本文改进了现有方法中的基于话题标签分词和简易规则的评论对象候选词抽取方法。现有方法是基于 SCP 理论将话题标签分词然后将结果加入分词工具的词典中进而对微博文本进行分词，然后用基于规则的方法来抽取文本中的评论对象候选词。由于传统分词工具的词典具有很大的局限性无法对包含了大量网络流行语的微博文本进行良好的分词，并且使用的抽取规则过于简单和粗糙，从而导致候选词抽取效果不佳。本文通过收集中文输入法的细胞词库构建新的用户词典，然后对抽取规则进行优化和扩展，有效提高了候选词抽取效果。

第二，在标准评论对象抽取阶段，提出了一种改进的基于聚类的多图平行标签传播的评论对象抽取方法，提高了微博文本评论对象抽取效果。现有的基于标签传播的评论对象抽取方法是以话题为单位，将同一个话题下所有微博无差别的构建到一个无向图中然后实行标签传播算法以达到协同抽取评论对象的目的，这种方法简化了建模复杂度，但是忽略了同一个话题下的微博存在不同讨论主题的情况，而不同的讨论主题下的微博无论是在表达方式、遣词用句等方面都具有差别，无差别的构图方法将会在标签传播过程中产生错误的传播路径和效果，并且这种错误会随着传播的进行而不断积累。为了克服上述问题，本文选择将同一个话题下的所有微博通过相似度进行聚类分为多个类别，每个类别对应一个主题，在此基础上，为每一个主题下的微博构建一个无向图，然后平行地对多个无向图实行标签传播算法，从而避免了由于讨论主题不同而造成的传播过程中的可信度不平等化的问题，实验表明改进的算法较现有方法性能有明显提升。

第三，在标准评论对象抽取阶段，语句相似度计算是十分重要的一步。本文提出了一种改进的基于微博上下文与浅层词汇特性相结合的微博语句相似度计算方法。在标签传播算法中，节点相似度计算是极为重要的一个步骤，相似度计算的准确与精度直接影响着整个图的传播过程，进而影响到最后的抽取结果。现有的基于标签传播的抽取方法中，语句作为无向图中的节点，其相似度计算是直接采用标准向量空间表示下计算向量余弦值的方法，这种方法虽然简单却丢失了语句所处的上下文信息。对于微博这种结构松散的短文本来讲，单句的理解往往依赖于对上下文信息的理解，忽略上下文的单句表达能力十分有限，因此，在计算相似度时，除语句本身所具有的词汇特征等内在特性外，本文还将语句所在微博的上下文信息考虑在内，设计了融合上下文与浅层词汇特征的语句相似度计算方法。

第四，在基于标签传播算法的评论对象抽取方法中，候选词相似度也是影响传播过程的重要因素。本文改进了现有方法中的计算候选词相似度的方法。在现有方法中，候选词的相似度是计算两个词汇（短语）的杰卡德距离指数得到，通过共有字符数来衡量相似程度，然而这种仅仅考虑词形特征的方法是粗放的，很容易造成误传播，影响候选词可信度的排序结果从而影响到最后的抽取结果。本文在现有研究基础上，提出一种面向微博的基于同义词词林与词形特征相结合的 candidate similarity 计算方法，将词形与词义特征相结合以计算候选词的相似度。

关键词：标签分词；评论对象抽取；标签传播

Abstract

Along with the rapid development of technology of the We media, broad masses of netizens have been shifting gradually from passive recipients of information to the publisher and distributor. Microblog has vast amounts of text data, which contains a lot of information that has great value on research and application. In recent years, the research on microblog sentiment analysis has been developing rapidly and turning to be more in-depth and detailed. As one of the key tasks, opinion target extraction gradually catch the attention of researchers because of its great application value in text summarization extraction, public opinion analysis, etc. Yet the research on Chinese microblog is not enough, the inherent property of microblog make it hard to study well, such as lacking language standard and sentence structure not clear, etc. Therefore, this article takes Chinese microblog text as research object and opinion target extraction as research content, and strives to explore a more effective opinion target extraction method for Chinese microblog. In this paper, we divided the opinion target extraction into two steps, namely candidate extraction and standard target extraction and improved the two steps respectively. To be specific, the work mainly includes the following four points:

Firstly, in the candidates extraction step, we improved the hashtag segmentation and rule based opinion target candidate extraction method in existing research. Existing method segmented the hashtags based on the Symmetrical Conditional Probability, then put the segmentations into user dictionary to further segment the microblog text, then used rule based method to extraction candidates. Because the dictionary of traditional segmentation tools has significant limitations, this method cannot segment the microblog that contain a lot of network buzzwords well. Furthermore, the rule is too simple and rough to extract candidates well. In this paper, we collected several Cell Thesaurus of Chinese input software to build new user dictionaries. In addition, we optimized and extended the extraction rules. Experimental result showed the effect of the new method.

Secondly, in the standard target extraction step, we proposed an improved clustering based multiple graph parallel label propagation algorithm. Existing label propagation based opinion target extraction algorithm (LPA) uses all the messages in a topic indiscriminately to build undirected graph and then run LPA to collectively extract opinion target. This method ignored that there exist diffident discussion aspects in a topic. Different

discussion aspects have different expression styles and usage of words. Building graph indiscriminately will lead to wrong propagation path and effect, and the error will accumulate in the propagation process. To avoid the above problem, we choose to divide all the messages in one topic to several categories using clustering, then we build undirected graphs for each category, and then run LPA parallel. This method can avoid confidence inequality problem. The experimental result showed that the extraction effect of this improved method improved obviously.

Thirdly, in the standard target extraction step, sentence similarity calculation is an important point. In this paper, we proposed an improved context and shallow lexical feature based sentence similarity calculation method. Similarity calculation is a very important step in LPA. The similarity calculation accuracy directly affects the propagation process of the whole figure, and thus affects the final extraction effect. The LPA takes a sentence as a node in the undirected graph, and use cosine to indicate the similarity of two sentences under standard VSM. This similarity calculation method lost the context information. For loosely structured text like microblog, the understanding of sentences often rely on the context information, and expression ability of one single sentence is limited. Therefore, in addition to the inherent vocabulary characteristics of sentences, this article also took into consideration of the context information of a sentence and designed a method that integrates context information and shallow vocabulary features.

Forthly, in the LPA based opinion target extraction method, the candidate similarity will also influence the propagation process. In this paper we improved the candidate similarity calculation method in existing method. Existing label propagation based algorithm used Jaccard Index to indicate candidate similarity, which counts the number of shared Chinese characters of candidates, however, this method is rough because it only considers the morphological features. This method can easily lead to propagation error and affect the confidence sorting thus reducing the final extraction effect. On the basis of existing research, this paper proposed a Tongyici Cilin and morphological characteristic based candidate similarity calculation method for Chinese microblog, this method will integrate morphological and semantic features to calculate similarity of candidates.

Keywords: Hashtag Segmentation; Opinion Target Extraction; Label Propagation

目 录

第一章 绪论

情感信息抽取是情感分析的主要子任务之一，在产品评论挖掘、舆情分析等多个方面都具有广泛的应用。情感信息抽取的目的是从文本数据中自动抽取出与情感相关的信息。近年来，该研究领域吸引了诸多研究者的研究和讨论。本章的内容为对文本情感分析和情感信息抽取的相关研究进行阐述，并对评论对象抽取任务的相关研究进行展示。

1.1 研究背景与意义

1.1.1 研究背景

进入互联网时代，各种智能联网设备已广泛普及，互联网已经成为人们获取、共享、交流各种信息的主要途径之一。今年年初，中国互联网络信息中心发布了《第37次中国互联网络发展状况统计报告》，报告显示，截至去年年底，中国互联网使用者规模已达6.88亿，其中去年新增网民3951万，增长了6.1%，相对于2014年提升了1.1%，互联网普及率达到50.3%[1]。互联网对人们的生活面貌产生了巨大的影响，除网络购物、在线旅游预定、在线支付等活动以外，人们还通过微博、贴吧、论坛等各种在线交流平台发表自己的意见与看法，进行信息发布共享以及实时资讯交流。互联网极大的改变了人们的生活方式，便捷了人们的生活。

随着以微博为首的自媒体信息平台的快速发展，广大网民逐渐从信息的被动接收者向主动发布共享者转变，互联网的共享信息时代来临。微博作为社交网络平台的代表，早已成为广大网络用户发布信息、获取信息、进行意见交换的主要平台之一。微博具有文本短小精悍、可实时互动等一系列优秀的内在特性，这使其成为涵盖多个层次网络用户的信息发布和交流的重要平台，包括个人、组织、机构、团体等等。由于用户层次多、背景复杂，微博已成为各类网络信息的集散地。内容方面，微博的内容所涉及的话题十分广泛，内容充实，上至军国大事下至百姓的日常小事，从科学前沿报道至日常生活小常识，无所不有。此外，分析微博用户的构成发展，可以发现其用户分布正逐渐向下渗透，早已不是仅仅集中于一二线城市，而是向三四线城市甚至更低级别的乡村地区渗透。从应用价值角度看，微博用户数量大、使用率高，导致数据积累量巨大，这使得微博在舆情管理、网络营销等方向具有极大的应用价值和广阔的应用空间。

情感分析，又称为意见挖掘[2]，是数据挖掘与计算语言学的研究分支。其目的为帮助人们精准地获取文本中包含评论内容的信息，提取出表达了一定情感的文本，然后对其进行进一步的分析。情感分析由多个子任务构成，如主观文本识别、情感极性分类、情感信息抽取等。根据待处理的本的粒度级别不同，可以将情感分析分为多个不同的研究层次：词语级、语句级、文档级以及多文档级等[3][4]。

1.1.2 研究意义

作为新兴互动交流平台，微博已经成为人们宣泄感情、发表观点的重要社交平台之一，微博内容覆盖面极广，往往包含用户对某个人或事物的看法，或对某部电影的评论，或者对某商品的评价，这些信息通常具有强烈的感情色彩，表达着用户对相关事物的观点。

无论是在理论研究还是实际应用方面，情感分析研究工作都具有重大的意义。在理论方面，情感分析技术与文本摘要、问答系统等众多其它研究领域密切相关。以文本摘要为例，在抽取摘要时，通常是直接将文中权重较高的语句进行拼接，或者是将权重较高的段落直接抽取出来作为摘要，这种方法通常会导致信息冗余，甚至是抽取出大量无关信息。然而而情感分析，则可以使得组建摘要时将更多的信息进行有效挖掘和整合，从而使得抽取的摘要更接近文本所表达中心意图。在传统的问答系统中，系统返回的答案常常让人感觉生硬而且语义模糊。而借助情感分析，可以对初步生成的答案作进一步的处理，让答案显得更加简练且准确。

在实际应用方面可以概括为以下几点：

（1）决策支持

在网上买东西的时候，用户往往习惯参考商品评论信息来决定是否购买。然而往往在一个商品下就有成千上万条评论，使得人们没有时间和精力去阅读每一条评论内容。借助文本情感分析技术可以解决此问题，既可以对所有评论内容进行极性分类以展示褒贬比例，也可以针对不同商品属性来对评论进行分类展示。借助情感分析技术，可以更方便用户快速获取需要的信息以作出购买决策。

（2）舆情分析

及时了解和掌控民众舆论动向，是政府部门维护社会稳定的基础。面对突发性事件，政府部门需要通过网络言论了解民众间的主流观点与情绪，获取评论所指向的具体问题和事物以采取相应的对策妥善处理相关问题。借助情感分析技术，可以自动从大规模网络文本中获取民众对于某热点事件的各层次粒度的意见和情感走向，帮助相关部门及时而正确的作出应对策略。

（3）信息预测

互联网逐渐深入人们的日常生活，网络舆论常常会影响更多人的思想和行为，对于好的影响需要倡导，而坏的影响则需要及时制止。情感分析技术可以在现有的评论信息基础上，对舆论的发展趋势进行很好地预测，因此可以为政府部门或其他相关人员提供指导以采取正确的把控措施。

1.1.3 国内外研究现状

情感分析领域早期的研究主要集中在文本情感分类方面，核心任务为情感倾向判别，以篇章级的文本内容作为研究对象。经过多年发展，研究体系逐渐完善，研究方向则则更系统化、深入化。随着时代变化，研究任务和研究对象也随之发生较大变化，任务划分更为细致，研究对象更加多元化，除行文较为规范的新闻稿、博客外，还包含进微博等写作形式更为自由的网络文本。

情感分析主要由三个子任务组成：主客观句分类、情感极性识别以及情感要素抽取。本文仅对情感要素抽取任务的研究现状作出介绍。

情感信息主要包括三个内容：观点持有者、评价词、评价对象。

（1）观点持有者抽取

观点持有者抽取是指识别出某主观句中的观点持有主体，通常是人或者机构。对于该任务，目前的较普遍的方法是非监督的基于启发式规则的方法。具体方法主要分为 a)基于命名实体识别的抽取方法[5] [6] ；b)基于语义角色标注的抽取方法：Bethard 等[7] 用语义分析识别观点命题和持有者。Choi 和 Cardie[8]以及 Kim 和 Hovy[9]将 FrameNet 的语义角色映射到的观点持有者上面，在 MPQA 语料(Wilson 和 Wiebe[10])上进行了观点持有者的识别。

（2）评价词抽取

评价词是指带有感情色彩的词语，是情感分析任务的关键之一。评价词抽取方法主要有两种：基于语料库以及基于情感词典的方法[12][13][14][15]。Rao D. and D. Ravichandran[11]利用语料库的统计规律来抽取评价词，通过观察语料抽取语料库中的评价词语并判断其极性。

基于语料库的方法易实现但依赖语料，语料规模不足会影响抽取效果。代表方法为 Hatzivassiloglou 和 McKeown[16]，该方法通过文本中的关联词来挖掘词语之间的同义或反义关系，例如通过 and、or 来连接的词语通常是具有相同倾向的词语。如句子 This car is beautiful and spacious 中，如果知道 beautiful 是正面的，那可以推断 spacious 也是正面的。然而该方法不能处理由连接词连接的词语以外的形容词。

Wiebe 等[13]通过计算词语的分布相似性，然后依此进行聚类，识别出形容词。然而，他们并未对识别出的形容词作出情感极性识别。Turney 和 Littman[12]通过计算与种子词语的 PMI 来获取语义相关性。

基于词典的方法基本思路是通过词典中不同词语的词义联系来抽取评价词，所用到的词典通常是 WordNet 或 HowNet 等[17][18][19]。其中比较具有代表性的工作如下：Kamps[20]通过 WordNet 构建同义网络，词语的情感倾向由它到种子词的最短路径计算得到。Esuli 和 Sebastiani[21]则是利用词汇表或字典中的注释来完成抽取任务。Takamura 等[22]也使用了字典中注释信息构建词汇网，通过权重来决定两个词是否具有相同的情感极性。Hu 和 Liu[23]和 Kim 和 Hovy[24]仅利用词典中人工定义的同反义词来计算词的情感倾向。

1.2 评论对象抽取研究概述

1.2.1 评论对象概念介绍

评论对象是指在主观性文本中表达情感或观点的的形容词所修饰的对象。以产品评论为例，一般是产品本身或者某一部分或属性，如汽车的外观、价格等。在新闻事件评论中，则是人们表达的观点的对象。因此，无论是对于需要掌握产品市场反馈的商家，还是需要掌握新闻舆论的政府机构，评论对象抽取工作具有极大的现实意义。

1.2.2 评论对象抽取相关研究

目前的评论对象抽取算法主要包括两种：基于非监督学习和监督学习的方法。

（1）基于非监督学习的方法

Hu 和 Liu[23]最早提出这个问题，他们在观察了大量文本之后，发现评论对象通常是在文中出现频率较高的名词，并提出使用关联规则来进行评论对象抽取。Popescu 和 Etzioni[25]提出了一种基于网络搜索的方法来实现评论对象识别，利用网页搜索计算待分短语和指定鉴别器的 PMI 值来决定一个名词或名词短语是否为一个特征。Scaffidi 等[26]在假定产品特征在产品评论中被提到的次数较普通英文文本中多的情况下，提出了一种建立语言模型的方法识别产品特征。Stoyanov 和 Cardie[27] 利用话题同指消解来进行建模，也比较有效。该方法将具有相同对象的观点聚集到一起，然后利用设计好的分类器来进行判断，以确定两个观点是否具有一致的评论对象。Li 和 Zhou[28]基于情感词典和主题词典抽取<情感词,评论对象>二

元组，一定程度上获取了上下文信息。Popescu 和 Nguyen[29]也是通过计算点互信息来抽取产品的 aspect，利用 OPINE 系统抽取显性特征，对情感词进行聚类并标明类别标签，进而抽取隐性特征。Ma 和 Wan[30]提出基于中心理论的抽取方法，总结出某个语句的 center 通常是讨论的焦点，在新闻评论文本中倘若某句子包含中心，则会被被抽取为评论对象。

还有人尝试利用话题模型来处理情感分析的任务[31][32]。Mei 等[33]利用多粒度的话题模型抽取产品领域中的评论对象，有效提高了召回率。

Zhou等[39]提出利用标签传播算法来进行评论对象抽取，并取得了良好的效果。本文就是受到他们工作的启发，在其基础上提出了基于聚类的多图平行标签传播算法来进一步优化算法，提升抽取效果。

（2）基于监督学习的方法

这种方法起步晚于非监督学习抽取方法。Zhuang 等[34] 针对意见描述-评论对象序偶的抽取提出，可以从已标注的数据中获取潜在节点以及与序偶相关的依存和词类路径的结合信息。Kessler 和 Nicolov[35]提出了基于机器学习来实现评论描述和评论对象的识别，识别结果远比非监督学习方法要好。Jakob 等[36]将该任务建模成序列标注（sequence labeling）问题，然后通过 CRF 来学习，效果比 Zhuang[34]更好。Liu[45]等提出一种基于图的方法对产品评论中的评论对象和评论词进行协同抽取，与以往算法仅利用词与词之间的评论关系进行抽取不同，他们对语义相关性和评论相关性两种关系来对词语的关系进行挖掘，然后利用基于图的协同游走算法来对候选词进行提纯以识别评论对象和评论词。

1.3 研究内容

本文面向中文微博文本，将评论对象抽取任务划分为候选词抽取和标准对象抽取两个步骤，分别对两个步骤的现有方法进行了改进和优化。本文提出了一种改进的基于话题标签分词和扩展规则的评论对象候选词抽取方法，同时提出了一种改进的基于聚类的多图平行标签传播的候选词提纯方法。具体内容主要有：

（1）基于中文微博话题标签分词以及融合第三方细胞词库来构建面向中文微博的用户分词词典，将其应用于微博文本分词，然后优化和拓展评论对象候选词抽取规则。通过实验研究新词典与优化后的规则对评论对象抽取的作用。

（2）改进现有的基于 LPA 的评论对象抽取模型。对同一个话题下的微博通过聚类进行划分，分别构建无向图，然而实行多图平行标签传播算法。

（3）改进现有工作的计算微博语句相似度的算法，将微博上下文和语句浅层词汇特征相结合来优化语句相似度计算方法。

（4）改进现有的评论对象候选词相似度计算方法，融合词形与词义特征，提出基于同义词词林和词形特征相结合的候选短语相似度计算方法。

1.4 组织结构

本文的组织结构如下：

第一章 绪论

本章主要介绍了文本情感分析的相关知识，然后详细介绍了评论对象抽取技术的相关研究。最后描述了本文的具体研究内容以及组织结构。

第二章 相关知识介绍

本章主要介绍了本文所涉及到的相关技术，比如中文分词技术、标签传播算法等、AP 聚类算法等相关理论。

第三章 评论对象候选词抽取

本章详细介绍了中文微博的内在特性，并针对其所特有的缺乏语言规范、句子结构不清晰等特点，提出一种改进的基于话题标签分词和第三方细胞词库相结合的微博文本分词方法，还对现有的候选词抽取规则进行了优化和扩展。

第四章 评论对象对象抽取

本章的主要内容为首先通过对传统的面向新闻语料、商品评论语料等领域文本的评论对象抽取算法进行总结，然后提出一种改进的基于聚类的多图平行标签传播算法来对评论对象进行抽取。本章中，还提出一种改进的基于微博上下文和浅层词汇特征的面向微博的语句节点相似度计算方法以及一种新的基于同义词词林和词形特征的候选词相似度计算方法。

第五章 总结与展望

本章对本文的研究进行了大致描述，并提出了下一步可以跟进的研究计划。

第二章 相关知识介绍

2.1 数据表示模型

向量空间模型(Vector Space Model)[43]是较为常用的文档表示模型，其将文档内容根据一定的规则进行切分，然后用向量的形式进行表示，每个维度表示一个切分单元的统计量，从而使文档语义得以在向量空间上表示，并且可以通过公式进行文档的相关计算。

在将文本转化为向量空间模型进行表示的过程中，我们进行以下假定，一共有 N 篇文档的文档集 $\{D_i, i=1,2,\cdots,N\}$，每一篇文档 D_i 包括 M_i 个无序的特征 $T_i=\{t_j, j=1,2,\cdots,M_i\}$（这些特征一般是指词，所构建出的特征集一般被称为词袋），则将文档集 $\{D_i, i=1,2,\cdots,N\}$ 转化为向量空间表示的过程如下：

1. 将文档集中的每篇文档所包括的特征集进行合并，得到整体的具有 M 维的特征集 T：

$$T=\bigcup_{k=1}^{i} T_k=\{t'_q, q=1,2,\cdots,M\} \quad \text{(公式 2.1)}$$

2. 统计每篇文档 $D_i(i=1,2,\cdots,N)$ 在特征集 T 上进行统计量计算，得到 M 维的向量表示 Q_i：

$$Q_i=[v_{ij}], j=1,2,\cdots M \quad \text{(公式 2.2)}$$

3. 将每篇文档的向量表示作为一个行向量，把 N 篇文档的向量表示放在一起，即组成了文档集的向量空间矩阵：

$$\begin{bmatrix} Q_1 \\ \vdots \\ Q_N \end{bmatrix}=\begin{bmatrix} v_{11} & \cdots & v_{1M} \\ \vdots & \ddots & \vdots \\ v_{N1} & \cdots & v_{NM} \end{bmatrix} \quad \text{(公式 2.3)}$$

利用 VSM 进行建模时，每个向量都是由若干统计量的权重组成。统计量有多种，比较常用的有词频 TF、倒排文档频率 IDF、$TF-IDF$ 等。

IDF 为总文档数 N 与在文档 D_i 中出现的频次的比值 n_{ij} 的对数，作为文档 D_i 的表

示向量第 j 维的值 w_{ij}：

$$w_{ij}=\log\frac{N}{n_{ij}},i=1,2,\cdots N \quad j=1,2,\cdots,M \tag{公式 2.4}$$

$TF-IDF$ 是结合绝对词频和倒排文档频率的权重计算方法：

$$w_{ij}=tf_{ij}\times\log\frac{N}{n_{ij}} \tag{公式 2.5}$$

对文本长度进行归一化处理后的 $TF-IDF$ 计算方法：

$$w_{ij}=\frac{tf_{ij}\times\log(N/n_{ij})}{\sqrt{\sum_{t_j\in D_i}\left[tf_{ij}\times\log(N/n_{ij})\right]^2}} \tag{公式 2.6}$$

当前普遍使用的的 $TF-IDF$ 权重计算公式：

$$w_{ij}=\frac{tf_{ij}\times\log(N/n_{ij}+0.01)}{\sqrt{\sum_{t_j\in D_i}\left[tf_{ij}\times\log(N/n_{ij}+0.01)\right]^2}} \tag{公式 2.7}$$

将文档用同维向量进行表示后，在文本分类中，一般使用余弦距离进行相似度衡量。假设文档 D_i 为 $\underline{Q}_i=[v_{ij}],j=1,2,\cdots M$，文档 D_k 为 $\underline{Q}_k=[v_{kj}],j=1,2,\cdots M$，则余弦相似度为：

$$Sim(D_i,D_k)=\frac{\sum_{j=1}^{M}v_{ij}\times v_{kj}}{\sqrt{\left(\sum_{j=1}^{M}v_{ij}^2\right)\times\left(\sum_{j=1}^{M}v_{kj}^2\right)}} \tag{公式 2.8}$$

2.2 汉语自动分词

中文分词指将汉字序列分成多个单独的词语的过程。这个问题看似很简单，实际处理却并不容易，主要难点为以下三点：分词规范、歧义切分和未登录词识别。

(1) 分词规范

“词”的概念向来是不易定义又不可避免的内容，但是迄今尚无权威的标准，

从计算的严格意义上说，自动分词是一个没有明确定义的问题[37]。

（2）歧义切分

歧义问题则比较常见，严重影响着分词精准度，是一个难点，也是重点。歧义类型主要包含两类：a）交集型切分歧义 b）组合型切分歧义。

（3）未登录词

未登录词指已有的词表中没有收录或已有的语料中没有的词。未登录词可粗略划分为如下几种类型：一是新出现的普通词汇，如房姐、起来嗨、给力等，尤其是在网络用语中这种新生词汇层出不穷；二是专有名词，主要指人名、地名等实体名词；三是专业名词和研究领域名称，如苏丹红、禽流感等等；四是其它专有名词，如新出的电影、书籍等文艺作品的名称。对于微博语料这种网络文本而言，未登录词的识别显得尤为重要，严重影响着分词精度。

现有的分词算法可分为三类：基于字符串匹配的方法、基于理解的方法和基于统计的分词方法。基于字符串匹配的算法是指按照特定策略将汉子字串串与词典的词条进行配。理解法是通过让计算机模拟人对句子的理解，达到识别词的效果。统计法是指利用汉字之间的互现信息来进行分词。

还有一类是基于统计机器学习的方法。利用已标注语料训练模型来进一步切分其它文本。

2.3 标签传播算法

标签传播算法（LPA）[42]，是一种基于图的半监督学习算法，该算法利用样本间的关系建立关系完全图模型，然后用已知节点的信息去预测未知节点的信息。在图中，数据被建模为节点，数据之间的相似度被建模为边，节点之间可以根据相似度进行标签传递，传播概率与相似度的值成正比。

在本算法中，已知数据的标签可以通过传播网络将标签信息传向其它数据。通过设定迭代次数来控制迭代过程，或者一直迭代直到标签分布趋向收敛停止。

具体算法如下：

$(x_1,y_1)\cdots(x_l,y_l)$为已标注数据，$Y_L=\{y_1\cdots y_l\}\in\{1\cdots C\}$ 为类别标签，类别数 C 已提前设定。$(x_{l+1},y_{l+1})\cdots(x_{l+u},y_{l+u})$为未标注数据，$Y_U=\{y_{l+1}\cdots y_{l+u}\}$不可观测，$l\ll u$ ，数据集$X=\{x_1\cdots x_{l+u}\}\in R^D$。问题转换为：从 X 中通过对 Y_L 进行学习，为 Y_U 中的点分配合适的标签。

然后构建完全连接图，图中节点为所有的数据，边权重计算为：

$$w_{ij} = \exp\left(-\frac{d_{ij}^2}{\sigma^2}\right) = \exp\left(-\frac{\sum_{d=1}^{D}(x_i^d - x_j^d)^2}{\sigma^2}\right) \quad \text{（公式 2.9）}$$

式中d_{ij}为节点的欧氏距离，权重w_{ij}受到参数σ的控制。

定义$(l+u) * (l+u)$的概率传播矩阵 T：

$$T_{ij} = P(j \to i) = \frac{w_{ij}}{\sum_{k=1}^{l+u} w_{kj}} \quad \text{（公式 2.10）}$$

其中，T_{ij}是节点j到节点i的传播概率。

定义$(l+u) \times C$的标注矩阵 Y，使得$Y_{ic} = \delta(y_i, c)$，i 行表示节点y_i的标注概率，c 列表示类别，如果$Y_{ic} = 1$则表明 y_i 为 c 类，如果不是则为 0。算法描述如下：

输入：未知数据，已标注数据以及类别。

输出：未知数据的标注。

a) 通过公式 2.9 计算边权重矩阵 w_{ij} 。

b) 通过公式 2.10 计算传播概率。

c) 定义一个$(l+u) \times C$维的标注矩阵 Y。

d) 每个节点按传播概率把它周围节点传播的标注值按权重相加，并更新自己的概率分布：

$$F_{ij} = \sum_{k=1}^{l+u} T_{ij} \qquad 1 \le i \le l+u; 1 \le j \le C \quad \text{（公式 2.11）}$$

e) 恢复已标注数据的概率分布为初始值。重复步骤 d)，直到收敛。

$$F_{ij} = Y_{ij} \qquad 1 \le i \le l; 1 \le j \le C \quad \text{（公式 2.12）}$$

由于 LPA 较强的实用性和不需要大量人工标注等系列优点，目前该算法已经在诸多领域都取得了十分良好的成果，如文本检索与分类、多媒体信息标注、检索与分类等。

2.4 ICTCLAS

ICTCLAS 系统由中科院计算所的张华平博士、刘群博士等所开发，该系统依靠如图 2.1 所示的五层模型，利用多层马尔可夫模型进行分词。

该系统对多种编码格式提供良好的支持，最重要的是支持用户自己制定具有特定需求的词典，这使得该系统在处理不同文本时具有更好的可扩展性。 ICTCLAS 分词速度单机 500KB/s，分词精度 98.45%，API 不超过 100kb，各种词典数据压缩

后不到 3M，且可以由用户自己制定具有特定需求的用户词典来辅助进行分词，该系统堪称目前最优的汉语分析工具之一。

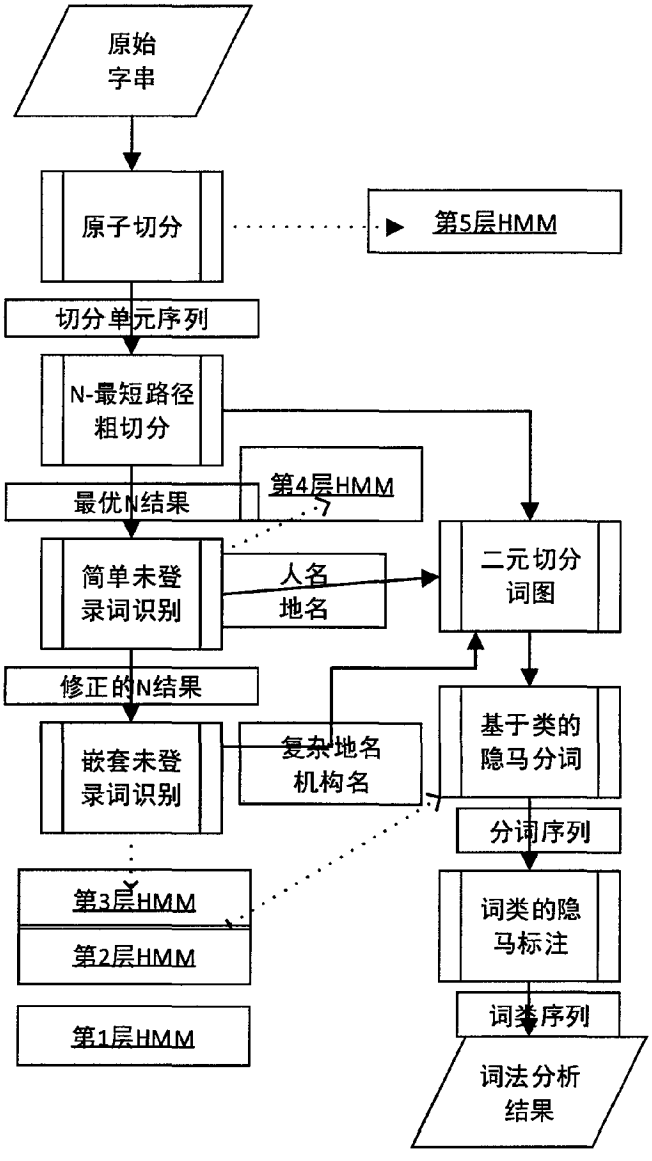

图 2.1 基于 CHMM 的汉语词法分析框架

2.5 SCP 理论

1999 年，Silva 和 Lopes 等提出 SCP(Symmetrical Conditional Probability)理论用以衡量字符串中字符之间的关系强度[44]。

为了衡量考虑二元组 x 和 y 之间的相关性，该文章提出一种新的标准，通过将两个分别给定对方的情况下的条件概率相成得到，我们称之为“对称条件概率”如下：

$$SCP(x,y) = p(x|y)\cdot p(y|x) = \frac{p(x,y)}{p(y)}\cdot\frac{p(x,y)}{p(x)} = \frac{p(x,y)^2}{p(x)\cdot p(y)} \qquad \text{（公式 2.13）}$$

推广至 n 元模型时，进行一般化，则有：

$$SCP(((w_1\cdots w_{n-1}),w_n)) = \frac{p(w_1\cdots w_n)^2}{p(w_1\cdots w_{n-1})\cdot p(w_n)} \qquad \text{（公式 2.14）}$$

其中 $p(x|y)$、 $p(y|x)$为条件概率。$p(w_1\cdots w_n)$ 为 $w_1...w_n$ 在语料中出现的概率。

2.6 AP 聚类算法

近邻传播聚类（Affinity Propogation Cluster）算法[40]是一种基于近邻信息传播的聚类算法，利用的是点与点之间的相似度。首先求得的相似度矩阵 s，该矩阵的维数应该是 n*n 的，n 为数据点个数，然后寻找类代表点集，每个类代表点对应一个数据点（exemplar），应满足的条件是所有的数据点与最近的类代表点的相似度之和最大。

AP 聚类算法不需要人工指定聚类数目，甚至所有的数据点都有可能是 exemplar。观察相似度矩阵，如果对角线上的数值 s(k, k)越大，表示 k 点越有可能成为聚类中心，这个值又叫做参考度 p (preference)，最终聚类结果的数量会受到 p 的影响。

Similarity：该算法中相似度计算一般使用欧氏距离，如 $-\sqrt{(x_i-x_j)^2+(y_i-y_j)^2}$。

Sreference： p(i)表示点 i 作为聚类中心的参考度。一般取 s 相似度值的中值。

Responsibility： r(i,k)用来表示点 k 作为数据点 i 的聚类中心的合适度。

Availability：a(i,k)用来表示点 i 选择点 k 作为其聚类中心的合适度。

两者的关系如下图：

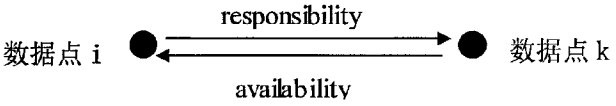

图 2.2 数据点关系图

下面是 r 与 a 的计算公式：

$$r(i,k)=s(i,k)-\max_{k'\neq k}\{a(i,k')+s(i,k')\} \quad \text{（公式 2.15）}$$

$$a(i,k)=\begin{cases}\min\left\{0,r(k,k)+\sum_{i'\in\{i,k\}}\max\{0,r(i',k)\}\right\}, & i\neq k\\ \sum_{i'\neq k}\max\{0,r(i',k)\}, & i=k\end{cases} \quad \text{（公式 2.16）}$$

公式 2.15 和 2.16 表示，当 s(k, k)较大导致 r(k, k)较大时，a(i, k)也会较大，进而表示 k 称为最终的聚类中心的可能性也较大。因此可以看出，只要改变 s(k, k)的值就可以改变最终的聚类数量。

Damping factor(阻尼系数)：$\lambda(\lambda\in[0.5,1))$，主要是起收敛作用的。

$$r_{new}(i,k)=\lambda*r_{old}(i,k)+(1-\lambda)*r(i,k) \quad \text{（公式 2.17）}$$

$$a_{new}(i,k)=\lambda*a_{old}(i,k)+(1-\lambda)*a(i,k) \quad \text{（公式 2.18）}$$

AP 算法的具体工作过程如下：

首先通过计算 n 个数据点间的相互相似度构建相似度矩阵 s，然后设置 p 值、确定一个最大的迭代次数 maxits，迭代启动。

算法运行时主要是更新两个矩阵：代表(Responsibility)矩阵$r=[r(i,k)]^{n\times n}$和适选(Availabilities)矩阵$a=[a(i,k)]^{n\times n}$。 这两个矩阵初始化为 0，n 是所有样本的数目。迭代更新公式如公式 2.15 和 2.16。迭代过程结束的条件是所有的样本的聚类结果逐渐收敛或者迭代次数达到 maxits。

为了避免算法在迭代过程产生震荡的情况，通常迭代公式都会被修改为 2.17 和公式 2.18。在 AP 算法中，聚数数目主要受 p 值（负值）的影响。表 2.1 为以 200 个数据点的规模样本为例，展示不同 Prefrence 值对聚类数目的影响。

由表 2.1，我们可以看到，preference 取值与最终聚类数目成正相关。

λ取值不同时，迭代次数和迭代过程中数据的摆动也会不同， λ值越小，迭代次数越少，不过 net Similarity 值波动会越大，倘若数据点比较大，就会很难收敛。

λ值越大，迭代次数就会越大，然而 net Similarity 比较稳定。

表 2.1 preference 对聚类数目影响

Preference 值	聚类数目
$\frac{\text{median(s)}}{2}$	16
median(s)	11
2×median(s)	8

从公式 2.17 和公式 2.18 可以看出，每一轮迭代的$r(i,k)$和$a(i,k)$受到λ的影响。当λ取较小的值时，$r_{new}(i,k)$和$a_{new}(i,k)$相比上一次迭代的$r_{old}(i,k)$和$a_{old}(i,k)$会发生较大的变化，也是由于这个原因导致 net Similarity 值摆动比较大；当λ取较大值时，$r_{new}(i,k)$和$a_{new}(i,k)$和上一次迭代的$r_{old}(i,k)$和$a_{old}(i,k)$比较接近，这也是导致迭代次数比较多的原因。

2.7 性能评测指标

本文使用的评价指标包括准确率、召回率、F 值。由于在语料标注任务中短语边界精确界定十分困难，因此本文采用严格评价指标和宽松评价指标两个评价指标来对抽取效果进行评价。

（1）严格评价指标

严格评价指标要求抽取出来的结果必须与标注语料中的标准标注完全吻合才算抽取正确。

（2）宽松评价指标

宽松评价指标最早由 Johansson and Moschitti 提出，在 2013 年中国计算机学会组办的第二届自然语言处理与中文计算会议 (NLP&CC 2013)评测中被采用为衡量抽取效果的标准之一。在此标准中，对于抽取出的结果与预料中标注的标准结果对指间定义一个跨度覆盖率 c：

$$c(s,s') = \frac{|s \cap s'|}{|s'|} \quad \text{（公式 2.19）}$$

在公式 2.7 中$|\cdot|$表示中文字符数的计数函数，$\cap$表示 s 与s'共有字符的集合。根据跨度覆盖率定义集合覆盖率$C(S,S')$:

$$C(S,S') = \sum_{s \in S} \sum_{s' \in S'} c(s,s') \quad \text{（公式 2.20）}$$

对于抽取出的词项集合 S 与语料标注的标准词项集合S'，我们为其定义宽松准确率 P 以及召回率 R ：

$$P = \frac{C(S,S)}{|S|} \qquad R = \frac{C(S,S)}{|S|} \quad \text{（公式 2.21）}$$

F 值由宽松准确率和宽松召回率计算调和平均数得到。

$$F = \frac{2P \cdot R}{P+R} \quad \text{（公式 2.22）}$$

第三章 评论对象候选词抽取

3.1 中文微博分词特性

目前大多数的微博情感分析研究主要集中于Twitter等英文微博语料。但是，对于中文微博的分析与英文微博的研究存在着以下几点不同：

（1）对于中文文本情感分析来讲，中文分词是一个必不可少的步骤，但是纵观目前多种中文分词工具，对于微博文本的处理是不尽如人意的，甚至可以说是表现很差，主要是由于微博文本与一般文本存在很大的区别，比如字数限制、表达方式自由化、缺少语言规范、句子结构不清晰等；

（2）英文微博中的标签（hashtag）经常被用于突出博文所表达的感情倾向，比如“#love”，“#sucks”，或者可以作为用户自标注的、较为粗糙的话题，比如“#news”，“#sports”。但是在中文微博中，绝大多数的标签（hashtag）经常会指示一个细粒度的话题名称，比如“#NBA 全明星总决赛#”。除此之外，在英文微博中，话题标签（hashtag）多是出现在某个句子之中，比如“I love the song #California Hotel！”然而在中文微博中，微博本身和话题标签是隔离的，一般标签由两个“#”包围表示，如“#加州旅馆#我很喜欢这首歌！”

在中文微博网站上，一个很常见的现象是同一个话题标签所聚集起来的微博会形成一个天然的话题，这一点对于中文微博的处理和分析是很重要的。国内的中文微博平台（如新浪微博、腾讯微博）经常提供一个单独的页面来列举显示当前热门的话题并且邀请人们参与进来并进行讨论。这些话题通常包含有数万条微博，拥有相同的话题标签（hashtag）。对这些话题中的评论对象进行一番研究可以帮助人们更深入的去了解人们对出现在这些热门话题和事件中的实体所表现出来的公众态度。

正如前面所说，中文微博中的#标签#常常能指示出一些细粒度的话题（或主题）。并且通常情况下微博话题标签中会包含有一部分评论对象，如“周星驰美人鱼”。所以针对话题标签找到合适的分词方式，然后将分词结果添加进已有中文分词工具的用户词典中，将显著提升整体分词效果。

再举一个例子，将更好地帮助理解。在话题“#00后谈恋爱#”中，“00后”是一个十分重要的词汇因为它是很多句子的评论对象。但是现有的中文分词工具一般都会将这个词汇拆分为两个词：“00”和“后”。然后在词性标注过程中，“00”被识别为数字，“后”被识别为位置指示词。由于我们只抽取名词性短语作为评论对象候

选词，所以这两个分开的词并不会被抽取出来。因此，错误的分词方法会导致无法抽取出正确的评论对象。这种错误会在包含有“00 后”以及对“00 后”表达了评论的句子中出现多次，从而会影响后续的抽取结果。

表 3.1 传统分词工具错分举例

待分词语句	传统分词效果	正确分词效果
一脸懵逼	一脸 / 懵 / 逼	一脸 / 懵逼
你这个魂淡	你 / 这个 / 魂 / 淡	你 / 这个 / 魂淡
我勒个去	我 / 勒 / 个 /去	我勒个去（不用切分）

在我们的方法中，话题下的微博文本将被用来识别类似“00 后”这样的未登录词，主要是基于它们在话题中出现的频率。例如“00 后”在不同微博中的高频率出现强烈表明这应该是一个词，不应该被分词程序分开。

在将话题标签“#00 后谈恋爱#”正确切分为“00 后/谈恋爱”之后，我们可以将标签分词加入分词工具用户词典来进一步对话题下的所有微博文本进行分词。

3.2 中文微博话题标签分词算法

3.2.1 基于粘度值的话题标签分词算法

现有的中文分词工具在微博语料上效果并不尽如人意。并且分词错误尤其是在评论对象上的分词错误将直接影响后续的词性标注和候选词抽取结果。

本方法的基本思想就是将在一个话题中出现频率很高的字符串作为一个词语看待。给定一个标签 h，h 中包含 n 个中文字符 $c_1c_2\ldots c_n$，我们想将其分割为若干个词语 $w_1w_2\ldots w_m$，其中每个词语包含一个或多个字符。

首先，我们基于 SCP（Symmetrical Conditional Probability,Silva and Lopes,1999）理论为中文字符串 $c_1c_2\ldots c_n$ 定义“粘度值”(Strickiness Score)这个概念。

$$SCP(c_1c_2\cdots c_n)=\frac{p(c_1c_2\cdots c_n)^2}{\frac{1}{n-1}\sum_{i=1}^{n}p(c_1\ldots c_i)p(c_{i+1}\cdots c_n)} \qquad \text{（公式 3.1）}$$

其中当 n=1 时，也就是当字符串只有一个字符时，$SCP(c_1)=p(c_1)^2$。$p(c_1c_2\ldots c_n)$是该字符串在该话题中出现的频率。

本文拟采用 Li 等[38]文章中的方法，通过进行对数运算来对 SCP 值进行平滑处理。同时将字符串的长度 n 考虑在内，从而有：

$$SCP'(c_1c_2\cdots c_n)=n\times logSCP(c_1c_2\cdots c_n) \qquad \text{（公式 3.2）}$$

其中，n 为字符串长度。

然后，将平滑结果进行 Sigmoid 函数变换，得到：

$$Stickiness(c_1c_2 \dots c_n) = \frac{2}{1+e^{-SCP'(c_1c_2\cdots c_n)}} \quad \text{（公式 3.3）}$$

对于话题标签$h = c_1c_2 \dots c_n$，我们拟将其分割为 m 个词语$w_1w_2 \dots w_m$并使得公式 3.4 取得最大值

$$max \sum_{i=1}^{m} Stickness(w_i) \quad \text{（公式 3.4）}$$

求解上式的最优解的方法可以用动态规划的方法，递归地将字符串分割为两个字串来完成求解完成。本文中的标签分词算法并不需要额外的训练语料，只需要利用话题本身的内容就可以完成。

3.2.2 第三方细胞词库辅助分词

众所周知，中文输入法工具在现如今已经成为广大电脑使用者不可缺少的系统辅助工具，借助智能的中文输入法，用户将更方便的进行文字录入，极大的提高了电脑的使用效率，而反过来，伴随着数以亿计的使用者的“调教”，中文输入法也已经积累了大量的第一手的用户输入习惯相关资料，见证着中国广大网民近年来的语言使用风格的变迁。

细胞词库的概念由搜狗输入法首创，相比系统的默认词库，其主要突出开放共享的特征，可以更好地满足用户个性化的需求，并且支持在线升级。细胞词库是一种细分化词库，外在表现就是一个.scel 文件（搜狗输入法词库格式，其它各输入法的细胞词库格式不同）每一个词库都是一种具体的细分类别的集合，如动物词汇大全、植物词汇大全、网络流行新词、日剧动漫词汇大全等等。2007 年 7 月，搜狗在 3.0Beta2 版本中提出了第一个细胞词库的概念并推出了第一个细胞词库。截止到 2016 年 3 月份，搜狗输入法已共有两万余名网友创建了 27695 个词库，共计 48482247 个词条。国内如搜狗、百度、QQ 等市场占有率最高的几款输入法软件经过多年积累，已经形成了大量的经过海量用户“调教”出来的词库，并且随着网络生活多元化的不断发展，包含更多新词的词库文件也正在形成。这些词库往往包含着近阶段中国网民使用频率较高的一些流行语或者词汇，其中不乏一些旧词新解，甚至是新造词汇短语，这是中国网民语言现状最直接的呈现方式，也是第一手的资料。在对微博、网络论坛等具有活跃的语言使用氛围的网络平台的文本进行分析时，套用传统的分词或解析手法往往带来与真实语境不同甚至相反的结果，有效的利用这些词库文件，用以佐助分词过程，将能很好地解决解析错误的问题。

本文中，作为对微博语料话题标签分词结果的补充，我们将同时加入搜狗、百度、qq 输入法等主流输入法的细胞词库，这些细胞词库包含许多用户已经形成习惯

的输入词汇，以及新出现但使用频次比较高的流行语，包含相当多的非书面的口语用法文本，一些比较新奇的词汇，如“十动然拒”甫一出现时由于并不满足一般组词规则且不属于常用短语或成语，并不能为传统的分词工具所识别，但是对于中文输入法工具来讲，类似的用语自出现并广为人所用之后，便已将这种词汇记录在细胞词库中。中文输入法更像是一个训练工具，利用数以亿计的用户来为词汇识别进行大规模的训练。以微博句“这个魂淡然后走了出去”为例，句中“魂淡”相对于传统分词工具来讲为未登录词，因为这是在网络语言环境下网民新造的词语，意义为“混蛋”，传统分词工具在对该句进行切分之时因无法识别“魂淡”这个词语会将淡然切分为一个词，使得分词结果为“这个/魂/淡然/后/走/了/出去 ”，然而这样分词是错误的，正确切分方式为“这个/魂淡/然后/走/了/出去”。倘若导入细胞词库，分词工具将能很好地识别出类似于“魂淡”这种网络新词，这对提高微博文本的分词准确率将大有帮助。

本文中，为了进一步降低网络新词对微博文本分词任务的影响，提高分词准确率，以及最大可能提高分词结果中评论对象的包含率，我们收集了包括搜狗输入法、QQ 输入法、百度输入法等多个输入法的 7 个细胞词库文件，共计 35482 个词汇，去除不同词库中的重复项后共计 35105 个词汇。这些词汇将与标签分词结果一道全部加入分词工具 ICTCLAS 的用户词典中用来为微博语料进行分词。

表 3.2 细胞词库规模表

词库来源	词库名称	词条数量
搜狗输入法	网络流行新词	30021
	最新汉语新词语选目	169
	潮词潮语	86
	推荐实用流行新词	11
百度输入法	网络用语词库	4918
QQ 输入法	常用聊天短语词库	80
	网络用语词库	197
总计	35482	
去重后	35105	

3.2.3 分词冲突解决

由于基于粘度值理论将微博话题标签进行分词后会与第三方细胞词库一起加入分词工具的用户词典中进行进一步的微博文本分词，因此很可能会出现标签分词结果与第三方词库词条冲突的问题。当此类冲突情况出现时，将以细胞词库词条作为标准。

3.3 评论对象候选词抽取

在完成分词工作后，所有的标签分词结果被加入 ICTCLAS 的用户词典中。接下来，更进一步的将该话题下所有的微博文本进行分词，同时，该工具也可以将分词结果的词性进行标注。

在进行评论对象抽取之前，要做的工作是先进行候选词抽取，因为在一个句子中，并不是所有的词素都有可能是评论对象。一般情况下，只有一个句子中的名词或者充当名词成分的短语才可能是评论对象。基于微博文本之于一般性文本在基本文法、字数、表达方式、成分完整度方面具有所区别，例如表达方式自由化、部分成分缺失（无主语、宾语等），本文将抛弃以往研究中利用句子成分解析的方法来抽取微博中的名词性短语，因为传统的语句解析工具应用于微博时往往会出现较大的噪声，存在较高的解析错误。

在完成词性标注后，我们将抽取出句子中的以下几种成分组合作为候选词：

表 3.3 候选词抽取模版

标号	形式	例句
1	名词 n	美人鱼真是太好看了！
2	名词+的+名词 n+DE+n	周星驰的电影果然不错。
3	名词+的+动词 n + de+v	美人鱼的播出简直是星爷粉的福音。
4	形容词+名词 adj+n	美丽海岸不再美丽。
5	形容词+的+名词 adj+de+n	美丽的人鱼
6	名词+名词	美人鱼小姐
7	名词+位置指示词	中国东部
8	位置指示词+名词	西太平洋
9	名词+动词	中国制造走向世界

在汉语中，能够作为名词性成分的并不仅限于名词和被形容词或名词修饰的名词这两种情况，如上所述，与英语中的动名词（x-ing 形式）相似，汉语中名词与动词搭配也可能是名词效果。

综上，我们按照如上 9 种规则编写正则表达式来抽取出句子中的名词性成分作为评论对象候选词。我们设定抽取出的名词性短语长度控制在 1 到 7 个字的长度，以往有研究者要么不设定长度范围要么规定最少 2 个，本文认为在某些特定的表达场景下，评论对象可以是一个字的名词，如人称代词“你我他”、指示代词“这”“那”、其他名词“盐”“水”等。

在句子中，凡是与前述九种正则表达式相匹配且满足长度限制的名词性短语将被抽取出来作为显式评论对象候选词。除此之外，在某些微博中，并没有出现名词性短语，不包含显式评论对象，如“真是太难看了”。类似的这些微博所包含的评论对象通常是在前文中出现过或者在话题标签中出现，我们称之为隐式评论对象。因此，若在之前相似的句子中已经抽取出显式的评论对象，可将这些显式评论对象作为这些句子的饮食评论对象。在后续的实验部分，我们我们将本文的方法和目前表现最优的基于解析的方法进行了对比。

3.4 实验与结果

3.4.1 实验设计及评价方法

本次实验所用语料来自 2013 年 CCF 中文信息技术专委会学术年会(NLP&CC 2013)微博观点句中的评论对象与极性识别任务所用语料。语料统计规模如表 3.4。

表 3.4 实验语料统计

话题数目	10
句子数目	206564
观点句数目	1762
标准评论对象数目	3133

该语料来自于新浪微博，由组委会组织相关人员进行采集，可作为标准的训练及测试数据使用。实验所用语料为多个 xml 格式文档，每个文档对应一个话题，包含数百条微博，数千条语句。每条微博已被切分为多条语句，并已标注是否为主观句。文档格式如图 3.1 所示。

```
<?xml version='1.0' encoding='UTF-16'?>
<topic title="bu_dong_chan_deng_ji_tiao_li">
<weibo id="11437">
<sentence id="1">#保利-业内观点# 【不动产登记将
<sentence id="2">这种统一的登记条例，应当指全国
<sentence id="3">政策意义有三，一摸清房产投资状
<sentence id="4">利益攸关，推进的难度可想而知。<
<hashtag id="1">不动产登记条例</hashtag>
</weibo>
<weibo id="11438">
<sentence id="1">分享了一篇文章：《牛刀：谈谈不
<hashtag id="1">不动产登记条例</hashtag>
</weibo>
<weibo id="11439">
<sentence id="1">[转载]牛刀：谈谈不动产登记条例
<hashtag id="1">不动产登记条例</hashtag>
</weibo>
<weibo id="11440">
<sentence id="1">发表了一篇转载博文 《[转载]牛刀
<hashtag id="1">不动产登记条例</hashtag>
</weibo>
<weibo id="11441">
<sentence id="1" opinionated="N">分享自牛刀 《
<sentence id="2" opinionated="Y">最腐败的就是北
<sentence id="3" opinionated="N">http://t.cn/zT
```

图 3.1 语料文档格式

本次评论对象候选词抽取实验，我们选取了两种方法与本文的方法进行对比，其一是目前最优的句法分析工具 BerkeleyParser 工具包，该工具对文本进行处理分析后可直接抽取出名词性短语；其二是利用 ICTCLAS 完成简单分词并进行词性标注后直接采用基于规则的方法来抽取评论对象候选词的方法。

本实验采用抽取正确率 AC 与命中率 HIT 为实验评价指标。准确率为取出候选词中为标准评论对象的数目与抽取候选词总数目的比值，命中率是指抽取出候选词中为标准评论对象的数目与语料中标准评论对象总数目的比值。

$$AC = \frac{\text{抽取的标准评论对象数}}{\text{评论对象候选词总数}} \times 100\% \qquad HIT = \frac{\text{抽取的标准评论对象数}}{\text{标准评论对象总数}} \quad \text{（公式 3.5）}$$

评论对象候选词总数：实验抽取出的评论对象候选词数量；

标准评论对象数：抽取出的候选词中为标准评论对象数量。

3.4.2 实验结果与分析

为各方法候选词抽取效果对比。表中第二列为各方法所抽取评论对象候选词总数，第三列为所抽取的候选词中包含的标准评论对象数量。

表 3.5 各方法下评论对象候选词抽取效果

方法	抽取总数	正确数量	准确率	命中率
Berkeley Parser	3728	981	26.3%	31.3%
ICTCLAS+规则	3361	1023	30.4%	32.7%
Zhou	3342	1102	33.0%	35.2%
标签分词+扩展规则	3353	1165	34.7%	37.2%

根据实验结果，我们可以得出的结论有：

（1）传统的文本处理工具在面对微博文本时，无论是基本的分词还是评论对象抽取任务，效果均不理想，而基于规则的方法尽管在各项指标方面较前者效果均有所提升，然而却并不明显，这表明对于表达形式自由、缺少基本的语言使用规范的微博文本进行自动化分析是十分困难的。

（2）本文方法相对于其他的方法，效果要明显更好一些。从表中数据可以看出，基于规则的三种方法效果均明显优于 Berkeley Parser，其中 ICTCLAS 与规则相结合的方法在抽取标准评论对象数量方面较前者提升了 4.3%，命中率提升了 15.6%；基于标签分词与粗糙规则的算法较 Berkeley Parser 在抽取标准评论对象数量方面提升了 12.3%，命中率提升了 25.5%；而本文基于标签分词与扩展规则的方法较 Berkeley Parser 在两个指标上分别提升了 18.8%和 31.9%。而本文的基于话题标签分词与扩展规则的方法也优于 Zhou 的简单的基于规则的方法，所抽取的正确评论对象数量提升了 1.02%，命中率提升 5.2%。由此我们可以得出结论，即针对中文微博中的话题标签进行初步解析，将其分词结果加入分词工具的用户词典并配合以输入法的细胞词库，将有效地提升中文微博的分词效果，从而进一步提升评论对象候选词的抽取效果。这也印证了之前所论证的利用话题标签信息以及细胞词库提升微博分词效果的有效性。

3.5 本章小结

本章主要描述了一个改进的从中文微博文本中抽取评论对象候选词的方法。这个模型可以有效地利用中文微博话题标签中的信息来进一步指导话题下所有微博文本的分词工作，加入第三方细胞词库后，模型将能更好的处理中文微博这种表达形式自由化、缺少基本标准文法的文本，有效地提升分词等工作的效果。利用精心设计的扩展的规则集，模型将能更有效地从中文微博文本中提取具有名词同等效力的、有可能充当评论对象的词汇或者短语，从而有效提升微博评论对象抽取的效果。本章的实验部分也证明了上述方法的可行性与有效性。在实验中，我们首先利用基于粘度值函数的算法对中文微博标签进行初步的分词，然后将分词结果与第三方中文输入法的细胞词库进行有机融合，组成新的用户词典加入现有的中文分词工具 ICTCLAS 的用户词典中对话题下的所有微博文本进行进一步的分词。与文中提到的其它方法不同，本文选择性的加入了网络流行语、网络新词等实时第三方细胞词库，有效地提升了对不规则微博文本的处理效果。实验证明，无论是抽选词抽取数量还是标准评论对象的命中率，本文的方法相较于其他方法的抽取效果均有明显的提升。

第四章 评论对象抽取

4.1 微博语料介绍

微博文本相较于传统的中文文本，如博客文本、新闻语料文本及商品评论文本而言，无论是在词语的使用搭配还是语言的表达方式上都具有明显的不同之处。传统的中文文本如博客等大多用词比较正常和规范，据法结构完整且严谨，语句结构清晰、明确，句中各种语言成分位置和使用方式均比较规范，因此在应用比较成熟的词法分析、句法分析算法和工具以及句子成分依存解析时，均具有不错的分析效果和稳定的分析效率。相反，微博作为一种新兴的网络媒体和文本载体，具有其自身所特有的若干特性。一是由于微博数以亿计的使用用户，涵盖了各行各业的使用者，传统媒体如博客、新闻等作者集中于文字工作者职业，多经过严格的文字训练，用语相对严谨和规范，相反，微博用户的极端分散性导致其在书写微博时并未刻意追求用语规范，而形成一种自由不羁的书写和表达习惯，语句成分缺失、简写、标点错用等等现象屡见不鲜，这种现象也造成微博文本数量增长过于迅速等情况；二是微博文本具有字数限制，一般不多于 140 字，这使得微博文本的表达更简洁化，也从另一方面促进了微博文本用语不规范、简写、句法成分缺失等现象的发生；三是近年来随着网络媒体平台的快速发展，尤其是微博、贴吧等自由媒体的兴起，部分年轻网民在追求特立独行、与众不同的个性时，创造出诸多网络用语并迅速席卷网络，客观上形成一种新的语言形态，如“我去年买了个表”、“魂淡”、“我也是醉了”等，这些网络用语具有与传统解释意义不同甚至相反的意义，如果按照传统的语言解析方式，将不能得到其真正的含义，更适合将其作为一个完整的不可割裂的意涵对待，如“我去年买了个表”并非真的买了个手表，而是表达一种极端反感的意思，“我也是醉了”并非真的喝酒醉了，而是表达一种不认同又无可奈何的意思。

以上所述微博文本与传统媒体文本的不同之处，极大地影响了基于微博的文本情感分析工作。微博用语不规范、语句成分缺失等现象使得传统的基于句子依存解析的文本情感分析无法适用。语句过短使得情感表达不甚明显，加大了分析难度。微博的随意转发现象使得文本中产生了大量与微博内容本身并无关联的冗余信息，导致文本清洗等预处理工作变得更加繁琐和复杂。以上现象综合导致对于微博文本的相关研究比对于传统文本的情感分析研究要落后得多，并且分析效果远不如传统文本的分析效果好。

本文使用的微博语料来源于NLP&CC 2013评测任务相关语料，该次评测共包含五个独立的评测任务，其中之一便是中文微博观点要素抽取，评论对象抽取是该任务的子任务之一。该语料具有以下几个特点：

（1） 单条微博极其简短，通常只包含2-3个句子，单个句子或者只有几个字的情况也很常见。此外，文本表达极其自由随意，多口语化用词，且富有网络色彩，包含大量的网络流行语。

具有该特点的具体句子如：

#锤子ROM# 锤子ROM看起来确实不错，在人人上看介绍的时候看到这段话笑尿了。。

#锤子ROM# 有思想的公司，有情怀的公司。

#锤子ROM# 可惜丑就一个字。

#中国方言式英语# 虽然赞比亚这谁口音也很奇怪，但这中式英语简直屌爆了，无法直闻。

#中国方言式英语# 雷的外焦里嫩啊！

#笑傲江湖# 我勒个去的，神剧呀，有木有！

#厨子戏子痞子# 看过《厨子戏子痞子》：前20分钟：这是扯的什么淡啊！

这些句子当中的“笑尿了”、“屌爆了”、“雷”、 “有木有”等等词汇均带有十分浓厚的网络用语色彩以及口语表达色彩，传统的分词工具将无法很好的处理这类词语，因此这些句子的分词难度会大大提升。

（2） 缺失显性情感词，情感表达不明

由于微博文本用语极度不规范，存在大量的句子成分缺失以及口语化表达现象，因此常常存在某个句子是观点句却又不带有明显的情感词，而是通过对比、比喻等来表达相关情感。

具有该特点的句子实例如下：

#厨子戏子痞子# 值得票价！

通过票价来间接表达电影不错的意思。

#科比# 我喜欢科比，科比打球比詹姆斯帅多了。

通过对比来表达科比打篮球很帅的意思。

#不动产登记条例# 我宁愿做牢。

本句并不含有情感词，但却明显表达出对该条例的不满。

（3） 用语表达缺乏规范

首先可能存在因输入错误或者语文常识缺失导致的错字、漏字现象；其次由于

大量转载的现象时有发生，因此微博文本中通常含有大量重复、无意义的文字和符号；然后，部分用户在书写比较特别的微博时倾向于附上额外的网页信息作为进一步的说明和补充，因此微博文本中包含网址的情况屡见不鲜，然而这些内容对我们分析微博文本情感并无帮助；最后，由于微博字数限制以及网络用语习惯的影响，微博文本中通常包含大量的缩写现象。

具体实例：

1）非中文噪声

#锤子 ROM#【老罗锤子 ROM 发布会-人性化篇之自拍优化】http://t.cn/zT7fKc8

2）转发、@某人产生无意义字符

#曼联 VS 皇马# @KimmiKemina 被裁判黑出翔的感觉有木有

3）标点错用

#曼联 VS 皇马# 我鹿鍋哭了沒 T^T~~~

4）评论对象缺失

#墨镜仙踪# 迪士尼威武，还是很好看的。

本句中实际上应该包含两个评论对象，其一是“迪士尼”，其二是隐性的“墨镜仙踪”。人们在就某一个话题或者主题进行讨论和评价的时候，往往为了叙述简洁而缺失评论对象。因此在评论对象抽取任务中，如果没有处理好这类状况，会导致分析结果正确率和召回率下降。

4.2 评论对象抽取系统框架

本文的研究目的就是探究如何对具有相同的话题标签（如在同一个话题下）的多条微博中的评论对象进行协同抽取。在一个句子中，评论对象一般分为两类，一类是显式评论对象，直接出现在句子中，如“我很喜欢《加州旅馆》这首歌”；还有一类叫做隐式评论对象，出现于句子之外，如“真的很好听！”

在目前的大多数研究中，隐式评论对象一般都未被考虑在内，并且，相对于显式评论对象抽取来说也更加困难。但是我们认为，如果将上下文信息利用好将很好的帮助我们同时抽取出这两类评论对象，因为在相同话题下的相似的句子有很大可能具有相同的评论对象。比如，在“# 周星驰美人鱼#”这个话题下，有将近 78.9 万条微博参与了相关讨论，通过观察，我们发现，在其中绝大多数带有主观倾向性的微博文本评论对象集中于《美人鱼》电影本身、周星驰（导演）以及林允（女主角）等几个实体。这种情况并不少见，而这也正为我们对相同话题下的多条微博的

评论对象进行协同抽取提供了可能性。

下表中用两个话题的四个句子进行举例说明。

表 4.1 动机举例

话题	句子
#周星驰美人鱼#	确实好看！
	钱花得值，电影确实好看！
#sunshine 组合#	这几个姑娘颜值简直“逆天”啊！
	还有哪个组合比这几个姑娘颜值还要“逆天”的吗？

上表中每个话题分别有两条微博句子，我们可以认为这两条微博是相似的，因为它们具有相同的一部分子序列。

在话题#周星驰美人鱼#中，第一个句子省略了评论对象，第二个句子包含有一个显式的评论对象“电影”。如果我们准确地为第二个句子抽取出评论对象，我们就可以推断出句子 1 也很有可能拥有和句子 2 相似的隐式评论对象。在第二个话题中 ，每个句子都包含一个名词性词语“姑娘”。这两个句子之间存在的相似性预示着这两个句子均对“姑娘”产生了评论。

基于以上观测，我们可以作出如下假设：同一个话题下相似的句子可能包含相同的评论对象。该假设将有效的帮助我们同时抽取显式的和隐式的评论对象。

本文方法的流程框架如图 4.1 所示，首先对微博语料进行数据清洗，然后进行响亮空间模型的建模，将每条微博句子向量化，然后利用本文提出的微博语句相似度计算方法进行 AP 聚类，形成多个细分的讨论主题的微博文本聚簇，然后针对每个聚类进行无向图构建，并在该图上进行标签传播算法，抽取出每个句子中得分最高的候选词作为该句的评论对象。

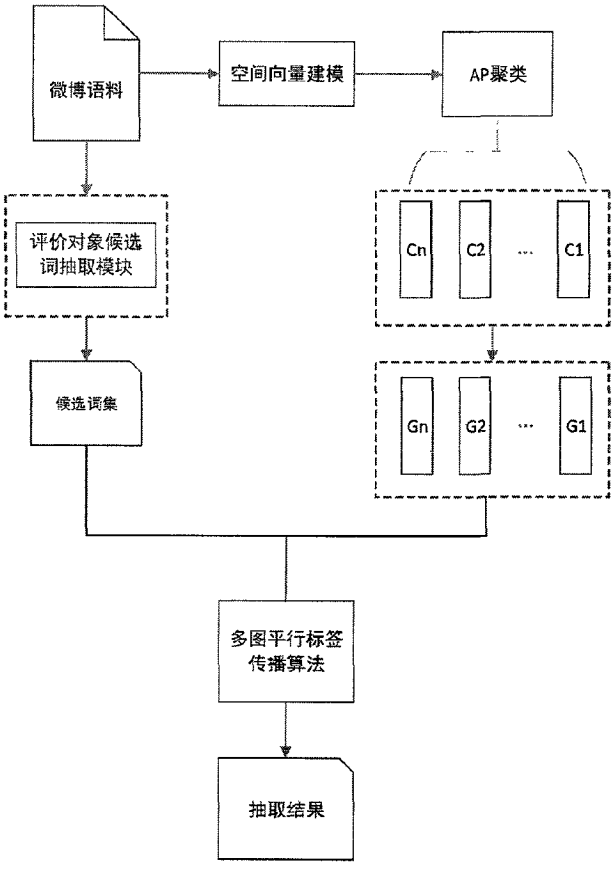

图 4.1 系统流程图

4.3 微博语句相似度计算与候选词相似度计算方法

4.3.1 基于微博上下文的语句相似度计算方法

在 Zhou[39]的文章中，计算微博语句相似度使用的是标准的向量空间模型的向量余弦值计算方法，该方法实现较为简单，效果也还可以，然而对于微博这种自由文本时，向量由于过于稀疏，导致在计算相似度过程中会流失很多信息。此外，由于微博所具有的表达自由化、缺乏语言规范等内在特性，一条语句自身携带的信息量极其有限，其语句本身的词汇特征往往并不能很好地反应语句本身所携带的信息量，若要对某条语句进行理解分析，不仅需要对语句本身的词汇特性等内在特性进

行考虑，还应该结合该条语句的上下文来进行辅助理解，通常情况下，语句的上下文信息对于语句的理解是不可或缺的。此外，在对微博文本进行观察之后，我们可以发现，不同微博中两条语句的相似度不仅与它们自身有关，还与它们所处的两条微博之间的关系有较大的关系——假设两条微博相似度较高，那么两条语句相似度也很高的可能性会很大，因为这表明两条微博所要表达的意义是相似的，从而两条语句之间的关联可能是很密切的。

例如 T_1、T_2 两条微博语句可能由于内在特性（比如不具有相同中文字符）看似没有关联而相似度为 0，但是由于两条微博语句所处的微博上下文具有很高的相似度，我们便可以认为这两条语句也是具有一定相似性的。

表 4.2 中，句子 T_{12} 与 T_{22} 因为没有共有的词向量，因此在用向量空间模型表示的向量求相似度时，余弦值将会是 0，表示两条语句相似度为 0，然而如果我们分析一下这两条微博，将会发现这两个句子实际上评价的是同一个评论对象，即电影本身。因此两条语句并不是完全没有相似度，在语义上，两条语句的理解是一致的。

表 4.2 微博语句相关性示例

M_1	M_2
T_{11} 昨天看了《厨子戏子痞子》	T_{21} 刚刚看了《厨子戏子痞子》
T_{12} 真的很不错	T_{22} 还是值得一看的
T_{13} 影帝的表演果然不同反响	T_{23} 几个影帝的表演还是很到位的

此外，我们还可以发现，倘若两条语句中所包含的共有字符数数量越多，两条语句的相似度通常会越高。相反，当两条语句字符数量差距越大，两条语句相度通常会越小。因此，我们也将把以上两点因素考虑在内。

综上所述，仅仅依靠浅层的词汇特征进行语句相似度计算具有很大的不足，通常会因为向量的稀疏导致相似度计算错误产生，这种错误将会随着后续的标签传播过程不断积累，影响到最后的评论对象抽取结果。

基于以上考虑，本文设计了一种基于微博上下文与语句自身特性相结合的语句相似度计算方法。在本方法中，我们用余弦函数来建模两条语句在词汇等内在特性上的关联程度，用所在微博的余弦函数来建模两条微博的关联程度。将上述因素考虑在内，我们设计了如下的相似度计算公式

$$Sim_{ab} = \log \frac{(\cos(Ma,Mb)+1)\times(cos(T_a,T_b)+1)\times(A(T_a,T_b)+1)}{|cont(Ma)-cont(Mb)+1|\times|cont(Ta)-cont(Tb)+1|} + 1 \quad \text{（公式 4.1）}$$

公式 4.1 中，$Ta \in R_+^{1\times n}$和$Tb \in R_+^{1\times n}$为待计算相似度的两条微博语句，$Ma \in R_+^{1\times n}$和 $Mb \in R_+^{1\times n}$分别为两条语句所在的两条微博，T_a、T_b、M_a、M_b 均为标准向量

空间模型表示，由词频表示权重。Cos(·)为求两个向量余弦值操作，cont(·)表示计算微博或者句子中中文字符个数，A(Ta,Tb)为求两条语句中相同字符数，|·|表示求绝对值。考虑到两条语句（或者微博）可能具有相同的中文字符数量，我们将用加 1 法来对公式分母两个分式进行平滑。并且，由于可能存在两条微博或者语句共有字符数为 0 的情况，我们也用加 1 法来对公式中相应部分进行平滑。最后，避免最后所得结果数值过小，我们对其进行取对数处理。

在得到相似度矩阵 W 后，我们将 W 中每行标准化得到$\widehat{W}$，使得

$$\sum_b \widehat{W}_{ab} = 1 \qquad \text{（公式 4.2）}$$

4.3.2 基于同义词词林与浅层词汇特征的候选词相似度计算方法

在 Zhou[39]等文章中，计算词汇相似度采用的是杰卡德距离指数(Jaccard Index)：

$$\text{JacIndex}(c_i, c_j) = \left|\frac{A(CT_i) \cap A(CT_j)}{A(CT_i) \cup A(CT_j)}\right| \qquad 1 \le I \ne J \le M \qquad \text{（公式 4.3）}$$

其中 $A(Cti)$和 $A(CTj)$表示 CT 中第 i 个候选词 c_i 和第 j 个候选词 c_j 的中文字符串。该方法只考虑词语中出现的共有字符一个因素，充其量只能对词汇相似性进行极浅层的建模，尽管简单易实现却十分粗糙，在实际的微博文本处理过程中，将会出现较为严重的错误，比如“手机”和“飞机”、“电影”和“电话”等词语对，尽管双方共享部分字符，但是在微博语句中并没有明显的关联，相反，在进行评论对象抽取的时候，还会对抽取工作带来错误传播。考虑到这种情况，我们拟采用田久乐[41]提出的基于词汇语义的词汇相关性计算算法与浅层词汇特征相结合的词汇相关性计算算法来计算候选词之间的相关性。

《同义词词林》是梅家驹等人编纂而成，其考虑的是词语的广义相关性，除同义词外，同类词也包含在内。哈工大扩展板词林收录词语近 7 万条，为五层分类体系，词语之间形成良好的层次关系。

图 4.2 为词语树形结构。

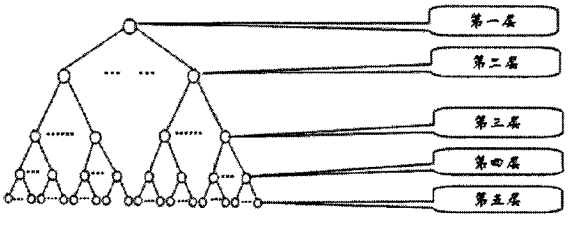

图 4.2 词语树形结构

词语的编码规则如下：

表 4.3 词语编码表

编码位	1	2	3	4	5	6	7	8
符号举例	A	b	1	5	B	0	2	=\#\@
符号性质	大类	中类	小类		词群	原子词群		
级 别	第 1 级	第 2 级	第 3 级		第 4 级	第 5 级		

在对不同的候选词进行相似度计算时，我们采用如下计算方法，两个词汇的相似度用 Sim 表示

若两个义项不在同一棵树上，Sim(A,B)=f

若两个义项在同一棵树上：

若在第 2 层分支，系数为 a Sim(A,B)=1*a*cos*(n*π/180)((n-k+1)/n)

若在第 3 层分支，系数为 b Sim(A,B)=1*1*b*cos*(n*π/180)((n-k+1)/n)

若在第 4 层分支，系数为 c Sim(A,B)=1*1*1*c×cos*(n*π/180)((n-k+1)/n)

若在第 5 层分支，系数为 d Sim(A,B)=1*1*1*1*d*cos*(n*π/180)((n-k+1)/n)

如果编码相同，而只有末尾是"#"，那么我们就可以认为两者相似度为 e。

公式中 n 为分支层的节点分支总数，k 为两个分支的距离。

该文献中给出的参数值为 a=0.65，b=0.8，c=0.9，d=0.96，e=0.5，f=0.1。

我们用基于同义词词林的算法来衡量词汇间的语义相关，利用杰卡德距离指数（Jaccard Index）来建模浅层词汇特性，则候选词$c_1 = (a_1, a_2, \dots, a_n)$和$c_2 = (b_1, b_2, \dots, b_m)$之间的相似度

$$\text{Similar}(c_1, c_2) = \gamma_1 \times \sum_{j=1}^{m} \sum_{i=1}^{n} Sim(a_i, b_j) + \gamma_2 \times JacIndex(c_1, c_2)\lambda_1 \quad \text{（公式 4.4）}$$

公式 4.4 中，γ_1和γ_2为调和系数且$\gamma_1 + \gamma_2 = 1$。

4.4 基于聚类的多图平行标签传播评论对象抽取算法

在对本文实验所用微博语料进行统计分析后，我们发现绝大部分情况下一个句子中包含有不超过 2 个评论对象，且平均每个句子包含 1.77 个评论对象。因此在本文中，我们假设一个句子存在 2 个评论对象。也就是说，在句子中，我们可以将抽取出的候选词中可信度最高的两个作为该句子的评论对象。本文采用的方法是无监督的基于聚类的多图平行标签传播算法来对某个话题下的所有句子的候选词进行

可信度排序。

标签传播算法（Label propagation）[42][43]是一种半监督的算法，它将标签分布从图中已被标注的某一小部分种子节点向整个图进行传播扩散。其基本思想就是图中相邻的节点之间所携带的信息具有相关性，因此可以利用这种相关性来进行全图的标签分布标注。基于标签传播的评论对象抽取由 Zhou[39]提出，他们将同一个话题下的所有句子构建成一个无向图（Graph），其中句子建模为图中节点（node），句子与句子之间的相关性建模边（edge）。然后在整个大图上使用标签传播算法。该算法忽略了微博文本中的表达跳跃性问题，并且将同一个话题下所有句子构建成一个图的思路过于粗糙，忽略了微博文本的上下文依赖性，因此算法的效果受到影响。本文基于对微博文本的特点，对 Zhou 等方法加以改进，将同一个话题下的所有句子通过 Affinity Propagation 聚类算法分割成多个相关性密集群，并为每个群构建一个图，然后在每个图上平行地运行标签传播算法。此外，原算法中衡量句子相似性时使用的是向量空间模型化后的余弦值，这就遗失了大量的前后文相关性信息。因此本文中对此加以改进，提出一种基于上下文与浅层词汇特征相结合的节点相似度算法。具体算法流程如下：

输入：

某话题下所有微博句子集合 Sen=<sen_1,sen_2,…,sen_n>

候选词相似性矩阵：　$S \in R_+^{M \times M}$

原始标注信息：　$Y_v \in R_+^{1 \times M}$　其中 $v \in V$

过滤矩阵：　$F_v \in R_+^{M \times M}$　其中　$v \in V$

概率参数：　p^i, p^c

输出：

标签向量：　$\hat{Y}_v \in R_+^{1 \times M}$

算法过程：

1.对 Sen 进行空间向量模型建模

2.计算 Sen 中任意两个句子的相似度，构建相似度矩阵

3.根据微博语句相似度进行 AP 聚类，得到 cluster 集合 Cls=<c_1,c_2,…,c_n>

4.对所有 cluster，分别构建无向连接图，得到图集合 Graph=<g_1,g_2,…,g_n>

其中$g = < V, E, \tilde{w} >$

0. 对所有 g in Graph

对所有 v ∈ V

Do

$\hat{Y}_v \leftarrow Y_v$

END

循环：

……

对所有 v ∈ V

DO

$D_v \leftarrow \sum_{u \in V, u \neq v} \widetilde{W}_{uv}(\hat{Y}_u \times S) \times F_v$

$\hat{Y}_v \leftarrow p^i Y_v + p^c D_v$

END

……

结果收敛时停止循环

本算法中，微博中的句子作为图中的节点（node），每个句子的候选词作为节点的标签（label）。节点的初始标签在前述候选词抽取步骤的结果基础上生成，这也意味着不需要进行人工标注等复杂耗时的工作 。

每轮迭代中，节点的标签向量向邻近的节点进行传播。本算法中，候选词相似性与句子相似性均考虑在内。最后我们选择候选词中得分最高的两个候选词作为句子的评论对象。

算法的具体形式逻辑阐述如下：

首先，为微博语料中的语句进行空间向量模型建模，然后利用本文 4.3 节中提出的方法计算任意两条语句的相似度并构建相似度矩阵 W；然后，根据语句间相似度进行语句聚类，本文使用的是 AP 聚类算法，利用语句相似度矩阵，将所有微博语句聚类得到 n 个聚簇 $c_1, c_2, \ldots, c_n$。

接下来，对每个聚簇 c 构建无向图$G = < V, E, \widetilde{W} >$,图中每个节点（node）表示一个句子，节点之间的边$e = (a, b) \in E$表示两个节点的标注（label）之间是相似的。$\widetilde{W}$是一个标准化的权重矩阵，指示节点间标注的相似性强度。节点标签之间的相似度W_{ab}为由 4.3 节中语句相似度计算方法求得。

对于每一个句子（节点）v，前面的步骤已经抽取出了候选词集合 C_v，则整个话题的候选词集合是所有 C_v 的并集$CT = \cup C_v$。整个话题的候选词数量 $N = |CT|$。

接下来计算候选词的相似度矩阵$S \in R_+^{M \times M}$，

在我们的模型中，候选词是被作为标签。我们设定整个话题的标签为$L =$

$\{1 \ldots M\}$，L 中的每个标签与对应 CT 中候选词一一对应。对于 V 集合中每一个 v，将标签向量$Y_v \in R_+^{1\times M}$进行初始化：

$$(Y_v)_k = \begin{cases} w & L_k \in C_v \\ 0 & L_k \notin C_v \end{cases} \qquad 1 \le k \le M \qquad \text{（公式 4.5）}$$

其中 w 是候选词的初始权重。并且显式评论候选词的权重与隐式对象候选词的权重也不同，如果是显示评论对象，$w=w_e$，如果是隐式评论对象，$w=w_i$。

如果 L_k 不是本句子的候选词，那么句子的标签向量中对象的位置为 0。在这些值中，如果初始时为 0，那么在后续的传播过程中也应该一直为 0。为了在传播过程中将这些位置重置为 0，我们为所有节点 v 构造一个对角矩阵$F_v \in R_+^{M\times M}$：

$$(F_v)_{kk} = \begin{cases} 1 & (Y_v)_k > 0 \\ 0 & (Y_v)_k = 0 \end{cases} \qquad 1 \le k \le M \qquad \text{（公式 4.6）}$$

式中的脚标 kk 表示 F_v 中的第 k 个位置。将 Y_v 右乘 F_v，清除无效候选词。

将传播过程形式化表示，分为两个动作：注射和继续，分别具有预先定义的概率值 p^i 和 p^c，两个概率值的和为 1。

在每一轮迭代中每个节点都将被其周围的节点影响。对于每个节点 v，传播效应公式：

$$D_v = \sum_{u\in V, u\neq v} \hat{W}_{uv}(\hat{Y}_u \times S) \times F_v \qquad \text{（公式 4.7）}$$

其中$\hat{Y}_u$是节点的当前迭代轮的标签向量。而之所以要进行$\hat{Y}_u \times S$的操作，是因为我们不仅是为了将节点 u 的第 i 个候选词的得分传递到节点 v 的第 i 个候选词，还要将其传递给其他所有候选词。W_{uv} 表征这种传播的强度。如之前所述，F_v 是为了清理传播过程中的无效候选词。随后，当所有标签向量中最大的值在十轮迭代内不再变化，表示算法已经收敛。

4.5 实验与性能分析

4.5.1 实验设计及评价方法

本章内容的实验所用语料同第三章实验部分语料一致，语料统计规模如表 4.4。

表 4.4 实验语料统计

话题数目	10
句子数目	206564
观点句数目	1762
标准评论对象数目	3133

如本章 4.3 节所描述，本算法中主要包含表 4.4 所示的多个参数，在实验中，本文依据 Zhou[39]等论文中的相关研究，对各项参数进行如表 4.5 设置。

表 4.5 实验语料统计

参数	值
p^i	0.5
p^c	0.5
w_e	1
w_i	0.5
γ_1	0.5
γ_2	0.5

4.5.2 Baseline 介绍

在实验分析部分，我们选择了四种目前效果较好的方法作为 Baseline 进行效果对比，它们分别是：

（1）2013 年 CCF 测评参赛队伍提交的算法。我们选取了该次评测所有队伍中效果最好的三支队伍的算法进行比较，分别为 T-1、T-2、T-3 。其中效果最好的一支队伍的方法的核心为一个被称为本体表的话题依赖的评论对象词典，只要句子中出现了本体表中的词项，则该词项被抽取出来作为评论对象。该本体表由人手工工制定，因此该方法并不能有效适应于新的话题。此外，我们还将计算出该次评测所有队伍方法的平均值 T-a，并作为一组对比项。

（2）基于关联挖掘的评论对象抽取算法。该算法是由 Hu and Liu[23]提出的一种无监督的算法。该算法是依赖关联挖掘和情感词典来抽取高频的和低频的产品特征。

（3）基于 CRF 的评论对象抽取算法。该算法包含单领域模型和跨领域模型两种模型，单领域模型（CRF-SIN）为每个话题训练一个不同的模型，而跨领域模型（CRF-CRO）由多个话题数据进行训练，由其余话题数据进行测试。

（4）基于单图标签传播的评论对象抽取算法。该算法由 Zhou[39] 等提出，将同

一话题下的所有句子构建一个单图，然后在该图上进行标签传播算法。本文与之不同的是，本文考虑到同一个话题下的微博按照评论对象不同，在表达方式、遣词用句、文法词法等各方面均会有所不同，倘若将所有句子构建成一个图来实行标签传播算法，则将会出现较多的噪音以及较大的冗余计算量。此外该算法在计算节点之间相似度的时候，使用的是简单的空间向量余弦值的方法。本文在对微博文本进行观察和研究之后认为该相似度计算法方法无法有效对微博语句指间的相似度进行度量，因此本文还提出一种新的基于上下文和浅层词汇特征相结合的语句相似度计算方法。本文的实验部分对该方法进行实现，并与本文提出的方法进行了对比。

4.5.3 实验结果与分析

4.5.3.1 与评测队伍各方法对比实验结果分析

由于该次评测要求各参赛队伍首先对微博语料的主客观性及情感倾向都要进行识别，评论对象抽取只针对主观性语句进行，因此为了公平起见，本文延续 Zhou[39] 的思路，主客观识别和倾向识别直接使用该次评测中效果最好的队伍的结果。然后在该结果之上继续进行本算法的实验。

表 4.6 评测队伍方法抽取结果

方法	宽松标准			严格标准		
	准确率	召回率	F 值	准确率	召回率	F 值
T-1	0.39	0.36	0.37	0.30	0.27	0.29
T-2	0.40	0.22	0.29	0.31	0.18	0.23
T-3	0.40	0.25	0.31	0.26	0.16	0.20
T-a	0.29	0.15	0.20	0.17	0.09	0.12
ULP	0.48	0.37	0.42	0.37	0.27	0.32
MGULP	**0.51**	**0.39**	**0.44**	**0.40**	**0.28**	**0.33**

图 4.3、4.4 为各项标准的折线图表示，可以清晰的看出各方法的实验结果对比。

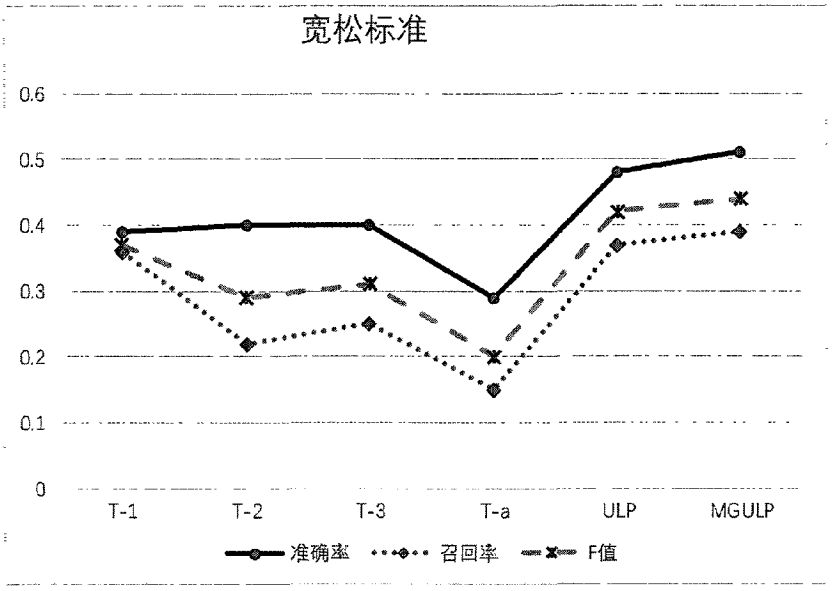

图 4.3 宽松标准下抽取效果对比

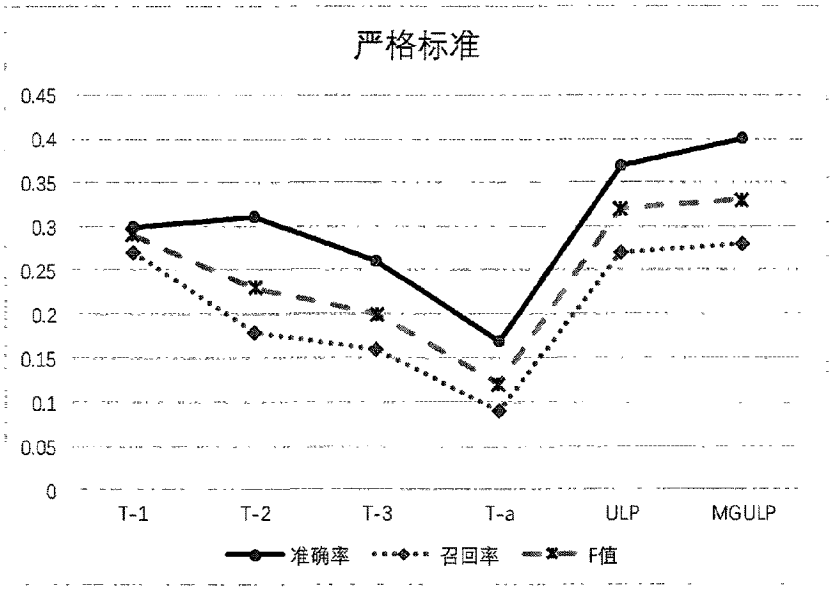

图 4.4 严格标准下抽取效果对比

由上面表格和图示可以看出，各参赛队伍抽取结果的平均 F 值为 0.20 和 0.12，其中严格标准的 F 值为 0.12 宽松标准的 F 值为 0.20，由此可见面向微博文本的评论对象抽取是非常困难的。除此之外，本文的方法相较于各参赛队伍而言，在各个标准上均有不小的效果提升，这也说明了本算法的有效性和高效性。

4.5.3.2 与其它算法对比实验

与上节实验不同，在与其它算法进行对比实验时，使用得是标准主观文本语料，并且对评论对象的情感倾向并不进行识别处理。表 4.7 为各算法结果对比。

表 4.7 各算法抽取结果展示

方法	宽松标准			严格标准		
	准确率	召回率	F 值	准确率	召回率	F 值
HandL	0.47	0.43	0.45	0.22	0.20	0.21
CRF-SIN	0.73	0.31	0.41	0.61	0.27	0.35
CRF-CRO	0.70	0.18	0.28	0.59	0.15	0.24
ULP	0.61	0.55	0.58	0.43	0.39	0.41
MGULP	0.65	0.56	0.60	0.44	0.41	0.43

图 4.5、4.6 为各项标准的折线图表示，可以清晰的看出各方法的实验结果对比

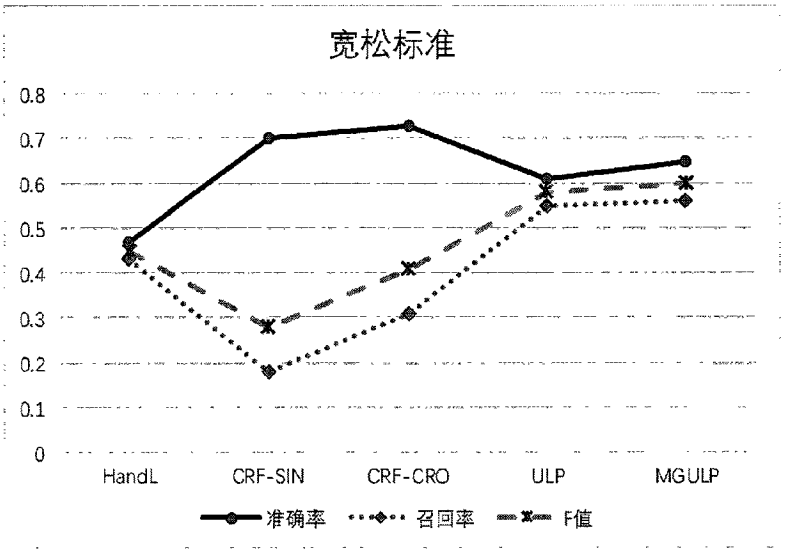

图 4.5 宽松标准下抽取效果对比

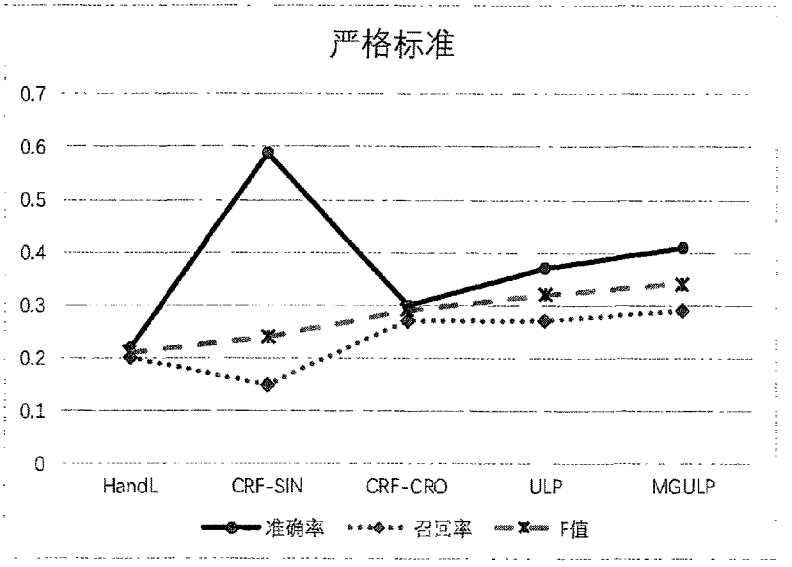

图 4.6 严格标准下抽取效果对比

从上表的数据可以看出，本方法取得了较为客观的效果。Hu and Liu 等提出的基于关联挖掘的算法在严格标准下表现最差，但在宽松标准下表现要比基于 CRF 的方法好一些。基于 CRF 的方法在准确率方面遥遥领先于其它算法，但在召回率方面表现过于糟糕。此外，单领域 CRF 方法相对于跨领域 CRF 方法效果要稍好一些，或许是由于为每一个领域（话题）构建一个特定的模型能够更好地为单个模型的数据进行建模，具有更好的适应性，因此分析效果很好，然而在具体实用中，不大可能为每一个待分析的领域（话题）构建单独的模型，因为对于新话题来讲并无已标注的实例来进行模型训练，否则工作量过大。本方法为无监督的方法，不需要提前标注数据作为训练数据，并且从算法效果来看，相对于各对比算法，均有不小提升。

4.5.3.3 参数敏感性实验

在本文的算法中，主要参数包括 p^i 、p^c 、w_e、w_i、γ_1、γ_2。根据现有文献，我们可以确定 p^i 、p^c 、w_e、w_i 等四个参数的取值方法，在现有的基于标签传播的方法中，在以相似语料进行实验时，这四个参数中 w_e 和 w_i 分别取 1 和 0.5 时实验效果最佳，而 p^i 、p^c 两个参数则除了在 0 和 1 两个极端值处效果骤降以外，在 0 到 1 这个区间中其它位置表现比较稳定，在本文中，我们将这四个参数分别设置为 1、0.5、0.5、0.5，重点考察 γ_1、γ_2 两个参数取值对实验结果的影响。图 4.7 为 γ_1 不

同取值时评论对象抽取结果（由于两者和为 1，所以我们只需设置其中一个参数，另外一个随即可以确定）。

由图 4.7 可以看出，除了在两个端点附近取值时抽取效果有所下降以外，整体趋势还是比较稳定的，表明本文的方法对参数设置并不是特别敏感，还是具有一定鲁棒性的。此外，在取值为 0.5 到 0.6 之间时算法效果表现最好，这也证明将词汇的词义相关性与词形相关性相结合的候选词相似度计算方法比仅仅使用词形相似度效果好好一些，从而证明了本文提出方法的有效性。

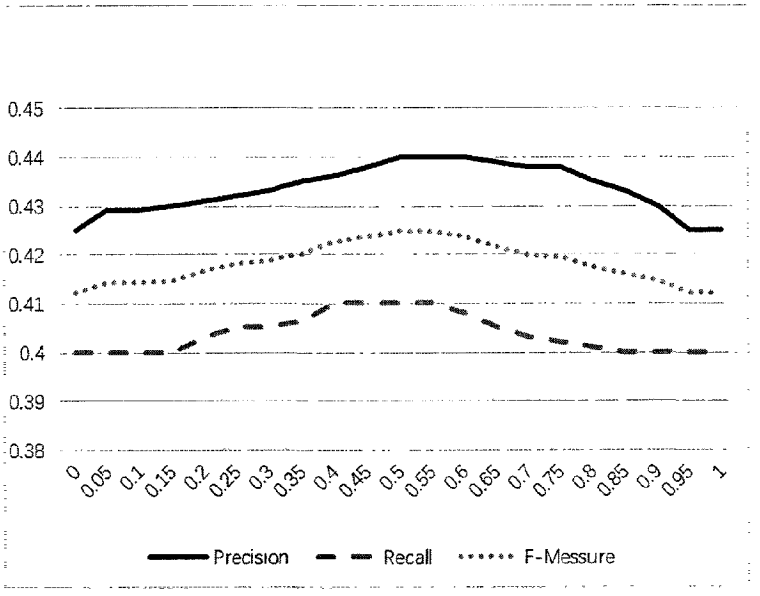

图 4.7 γ_1 参数设置对抽取效果影响

4.6 本章小结

本章主要描述一个改进的基于标签传播的评论对象抽取算法，与以往文献将微博语料整个话题的语句无差别构建为一个无向图的方法不同，本文提出基于聚类的多图平行标签传播算法。首先，本文提出了一种新的适用于中文微博的基于微博上下文的语句相似度计算方法，区别于以往仅仅利用空间向量这种利用浅层词汇特征进行建模的方法，本文将语句所处微博的上下文信息以及语句浅层词汇信息进行有机整合，以克服仅利用浅层词汇信息建模所具有的信息缺失、错误传播等缺陷，从

而提高语句相似度计算的精度与准确率。其次，本章提出了一种新的候选词相似度计算方法。在以往的评论对象抽取相关文献中，词汇相似度通过简单的杰卡德距离指数（Jaccard Index）来计算，然而这种方法仅仅考虑到共有字符这一浅层的词汇特征，在评论对象抽取过程中，这种粗放的计算方法将导致较大的计算错误，并且这种错误会随着在无向图的传播过程中不断积累，从而影响最终的评论对象候选词的可信度排名，进而影响到最终的抽取结果的正确率。本章提出的方法除共有字符这一词形相似性外，还将词义相似性也考虑在内，借鉴现有的基于同义词词林的词语相似度计算方法，将词形和词义特征进行有机整合来规范候选词相似度计算过程。最终实验结果也证明，两种方法对提高评论对象抽取的准确率和召回率是有效的。

第五章 总结与展望

5.1 本文总结

微博作为社交网络平台的代表，已成为广大网络用户获取和发布信息以及进行意见交换的重要工具之一。由于微博用户数量大、使用率高、数据积累量大，如何利用机器对海量微博文本进行自动分析并抽取出有价值的信息，已成为数据挖掘领域中十分重要的一个研究任务。本文以面向微博文本的情感分析领域中的评论对象抽取问题为主要研究内容进行了研究和探索，通过对相关研究地学习，本文选取了基于标签传播的算法来完成评论对象抽取的任务。本文主要工作包括以下几点：

首先，本文以中文微博文本为研究对象，详细阐述了微博文本与一般文本在表达方式、文法词法等方面的异同，对中文微博文本的内在特性进行了细致地分析与归纳。由于微博文本具有文本长度短、表达方式自由、缺乏正常的语言规范、句子结构不规则等一系列的内在特性，使得传统的针对一般性文本的分析算法如语句依存分析在迁移到微博文本时效果会大大降低。本文通过基于规则的方法来抽取出文本中评论对象候选词，使用本文提出的改进的基于标签传播算法的方法来对候选词进行可信度排序，从而实现以话题为单位同时对大量微博进行评论对象抽取工作。

其次，本文针对现有的基于标签传播的评论对象抽取算法进行了分析和改进。现有的相关方法是将一个话题下的所有微博无差别的构建为一个无向图，然后在图上进行标签传播，然而在对微博语料进行了观察和分析后，我们发现这种方法忽略了即使是在同一个话题下的微博也可能是讨论不同方面的主题这一现象，并且不同讨论方面的微博在表达方式、遣词用句等方面也有较大差别，无差别的构图方法将会在标签传播过程中产生错误的传播路径和效果，并且这种错误会随着传播的进行而不断积累。基于以上分析，本文提出了一种改进的方法，即基于聚类的多图平行标签传播算法，本文的实验部分的相关结果展示也证明该方法相较之前的算法，在准确率方面具有明显的提升。

然后，本文在对微博文本进行了较为仔细的分析后，提出一种基于上下文信息与浅层词汇特性相结合的的微博语句相似度计算方法。在基于标签传播的抽取算法中，语句作为图的节点，其相似度计算对于整个标签传播算法是非常重要的，相似度计算的是否准确将直接影响传播过程从而影响到最后的抽取结果。现有的基于标签传播算法的评论对象抽取方法中的语句相似度是计算语句的标准向量空间表示

的余弦值，这种方法虽然简单却忽略了语句所处的上下文信息，然而对于微博这种短文本来讲，上下文信息对于某条语句的理解是非常重要的，因此，本文在计算语句的相似度时，不仅考虑了语句本身的词汇特征，还将所在微博的上下文考虑在内，设计了更为合理的相似度计算方法。

最后，本文在计算候选词的相似度方面做出了改进，现有的方法是利用杰卡德距离指数，通过共有字符数来衡量相似程度，然而这种仅仅考虑词形特征的方法是粗放的，很容易造成误传播，影响候选词可信度的排序结果从而影响到最后的抽取结果。本文在现有研究基础上，提出一种面向微博的基于同义词词林与词形特征相结合的候选词相似度计算方法，将词汇的词形与词义特征相结合来计算候选词的相似度。

5.2 展望

本文针对现有的基于标签传播的评论对象抽取算法提出了三个改进方法，尽管最后的实验结果表明抽取效果有所提升，但是依然有比较大的改进和提升空间。在后续的研究中，我们将从以下几方面着手做进一步的分析和改进：

（1）本文选择了基于标签传播的方法来完成评论对象抽取任务，尽管效果较其它方法稍好，但是 CRF 等方法在这个任务上的表现也很好，因此，在下一步的工作中，我们将探索思路，尝试以层次化的方法来整合两种方法，甚至是直接将两种方法进行结合，以更大可能的挖掘和利用微博文本的内在特性信息以提高抽取效果。

（2）本文选取的实验语料为 NLP&CC 会议评测语料，只包含了直接发表的微博，转发、评论等信息都没有包含在内，我们认为如果将转发和评论的微博也建模到本文的系统中来，将会有效地提升对评论对象抽取效果。

（3）接下来，我们将设计程序自动下载更多微博数据并完成相关信息标注以扩展实验所用语料，在更大的数据集上验证本文提出的方法的有效性。

参考文献

[1](中国互联网络信息中心.第 37 次《中国互联网络发展状况统计报告》http://cnnic.cn/gywm/xwzx/rdxw/2015/201601/t20160122_53283.htm)

[2] 赵妍妍，秦兵，刘挺．文本情感分析[J]．软件学报，2010，21(8)：1834-1848.

[3] 黄萱菁，赵军．中文文本情感分析[C]．中国计算机学会通讯.2008，4（2）.

[4] 赵军，许洪波，黄萱菁，等．中文倾向性分析评测技术报告[C]．第一届中文倾向性分析评测(COAE2008)，2008.

[5] Kim S. and E. Hovy. 2006. Extracting Opinions, Opinion Holders, and Topics Expressed in Online News Media Text. In Proceedings of the ACL Workshop on Sentiment and Subjectivity in Text.1-8

[6] Ruppenhofer J., S. Somasundaran and J. Wiebe. 2008. Finding the Sources and Targets ofSubjective Expressions. In Proceedings of LREC-08.

[7] Bethard S., H. Yu and A. Thornton. 2004. Automatic Extraction of Opinion Propositions and Their Holders. In Proceedings of the AAAI Spring Symposium on Exploring Attitude and Affect in Text.22-24.

[8] Choi Y., C. Cardie and E. Riloff. 2005. Identifying Sources of Opinions with Conditional Random Fields and Extraction Patterns. In Proceedings of HLT/EMNLP-05. 355-362.

[9] Kim S. and E. Hovy. 2005. Identifying Opinion Holders for Question Answering in Opinion Texts, In Proceedings of AAAI-05 Workshop on Question Answering in Restricted Domains. Pittsburgh PA.

[10] Wilson T. and J. Wiebe. 2003. Annotating opinions in the world press. In Proceedings of the 4thACL SIGdial Workshop on Discourse and Dialogue (SIGdial-03)

[11] Rao D. and D. Ravichandran. 2009. Semi-supervised Polarity Lexicon Induction. In Proceedings of EACL-2009. 675-682.。

[12] Turney P. and M. Littman. 2003. Measuring Praise and Criticism: Inference of Semantic Orientation from Association . ACM Transactions on Information Systems (TOIS), 21(4). 315-346

[13] Wiebe J., T. Wilson, R. Bruce, M. Bell and M. Martin. 2004. Learning Subjective Language.Computational Linguistics, 30(3). 277-308.

[14] Kanayama H. and T. Nasukawa. 2006. Fully Automatic Lexicon Expansion for

Domain-oriented Sentiment Analysis. In Proceedings of EMNLP-06. 355-363.

[15] Wiebe J. 2000. Learning Subjective Adjectives from Corpora. In Proceedings of AAAI. 735-740.

[16] Hatzivassiloglou V. and K. McKeown. 1997. Predicting the Semantic Orientation of Adjectives. In Proceedings of ACL-97, Stroudsburg, PA. 174-181.

[17] Kim S. and E. Hovy. 2006. Extracting Opinions, Opinion Holders, and Topics Expressed in OnlineNews Media Text. In Proceedings of the ACL Workshop on Sentiment and Subjectivity in Text.1-8.

[18] Breck E., Y. Choi and C. Cardie. 2007. Identifying Expressions of Opinion in Context. In Proceedings of IJCAI. 2683-2688.

[19] Takamura H., T. Inui and M. Okumura. 2007. Extracting Semantic Orientations of Phrases from Dictionary. In Proceedings of NAACL HLT-07. 292-299.

[20] Kamps J., M. Marx, R. Mokken and M. Rijke. 2004. Using WordNet to Measure Semantic Orientation of Adjectives. In Proceedings of LREC-04. 1115-1118.

[21] Esuli A. and F. Sebastiani. 2005. Determining the Semantic Orientation of Terms through Gloss Classification. In Proceedings of CIKM'05. 617-624.

[22] Takamura H., T. Inui and M. Okumura. 2005. Extracting Semantic Orientations of Words using Spin Model. In Proceedings of ACL-05. 133-140.

[23] Hu M. and B. Liu. 2004. Mining and Summarizing Customer Reviews. In Proceedings of SIGKDD-04, 168–177.

[24] Kim S. and E. Hovy. 2004. Determining the Sentiment of Opinions. In Proceedings of COLING-04.1367-1373.

[25] Popescu A. and O. Etzioni. 2005. Extracting Product Features and Opinions from Reviews. InProceedings of EMNLP-05. 339-346.

[26] Scaffidi C., K. Bierhoff, E. Chang, M. Felker, H. Ng and C. Jin. 2007. Red Opal: Productfeature Scoring from Reviews. In Proceedings of EC'07. 182–191.

[27] Stoyanov V. and C. Cardie. 2008. Topic Identification for Fine-Grained Opinion Analysis. In Proceedings of Coling-2008. 817-824.

[28] Li B., L. Zhou, S. Feng and K. Wong. 2010. A Unified Graph Model for Sentence-based Opinion Retrieval. In Proceedings of ACL. 1367-1375.

[29] Popescu A., B. Nguyen and O. Etzioni. 2005. OPINE: Extracting Product Features and Opinions from Reviews. In Proceedings of HLT/EMNLP. 32-33.

[30] Ma T. and X. Wan. 2010. Opinion Target Extraction in Chinese News Comments. In Proceedings of COLING. 782-790.

[31] Blei D., A. Ng and M. Jordan. 2006. Correlated Topic Models. In Proceedings of Advances in NIPS. 147-154.

[32] Titov I. and R. McDonald. 2008. Modeling Online Reviews with Multi-grain Topic Models[C]. In Proceedings of WWW-08. 111-120.

[33] Mei Q., X. Ling, M. Wondra, H. Su and C. Zhai. 2007. Topic Sentiment Mixture: Modeling Facets and Opinions in Weblogs. In Proceedings of WWW-07. 171-180.

[34] Zhuang L., F. Jing and X. Zhu. 2006. Movie Review Mining and Summarization. In Proceedings of CIKM-06, November. 43-50.

[35] Kessler J. and N. Nicolov. 2009. Targeting Sentiment Expressions through Supervised Ranking of Linguistic Configurations. In Proceedings of the Third International AAAI Conference on Weblogs and Social Media, San Jose, California, USA, May. 90-97.

[36] Jakob N. and I. Gurevych. 2010. Extracting Opinion Targets in a Single and Cross-Domain Setting with Conditional Random Fields. In Proceedings of EMNLP-10. 1035-1045.

[37]黄昌宁，高剑锋，李沐.2003.对自动分词的反思.见：语言计算与基于内容的文本处理.全国第七届计算语言学联合学术会议论文集.北京：清华大学出版社，26-38

[38] Chenliang Li, Jianshu Weng, Qi He, Yuxia Yao,Anwitaman Datta, Aixin Sun and Bu-Sung Lee.2012b. Twiner: Named entity recognition in targeted twitter stream. In Proceedings of the 35th international ACM SIGIR conference on Research and development in information retrieval, pp. 721-730.ACM.

[39]Xinjie Zhou, Xiaojun Wan and Jianguo Xiao.Collective Opinion Target Extraction in Chinese Microblogs. Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing, pages 1840–1850

[40] Frey B J, Dueck D. Clustering by passing messages between data points[J]. science, 2007, 315(5814): 972-976.

[41] 田久乐，赵蔚. 基于同义词词林的词语相似度计算方法[J]. 吉林大学学报：信息科学版, 2010 (6): 602-608.

[42] X. Zhu and Z. Ghahramani. 2002. Learning from labeled and unlabeled data with label propagation.Technical report, CMU CALD tech report.

[43] G. Salton, A. Wong and C. S. Yang. 1975. A Vector Space Model for Automatic

Indexing, Communications of the ACM, vol. 18, nr. 11, pages 613–620.

[44] J. F. da Silva and G. P. Lopes. 1999. A local maxima method and a fair dispersion normalization for extracting multi-word units from corpora. In Proc. Of the 6th Meeting on Mathematics of Language

[45] Liu K, Xu L, Zhao J. Extracting Opinion Targets and Opinion Words from Online Reviews with Graph Co-ranking[C]//ACL (1). 2014: 314-324.

.

攻读硕士学位期间参加的科研项目与发表的论文

[1] 国家语委“十二五”科研规划重点项目：国家语言资源监测语料库建设及相关技术研究(ZDI125-1).2012-2014

[2] 国家“十二五”科技支撑计划课题子课题：基于互联网的地铁施工全过程风险识别、可视化预警与现场安全检测关键技术研究(2012BAK24B01).

[3] 首届中国“互联网+”大学生创新创业大赛国家级银奖：《指间童年—智能化的幼教管理及家园共育平台》2015.10

致 谢

转眼间三年已过，在众多老师、同学等陪伴和帮助下，我度过了人生中最快乐的三年美好时光。于我而言，华师就像一个大家庭一般，正所谓爱在华师，各位老师和同学宛如家人一般，无时不刻给予我陪伴和关怀，让我的学习生活变得更加多姿多彩。三年中，随着知识的不断积累，对于自然语言处理这个曾经对我而言十分陌生的领域如今已经了解得越来越全面和深刻。三年中，除努力学习基本的理论知识外，我还积极参与到实验室的项目中，努力锻炼自己的实践能力。充实的三年学习生活让我的知识水平得到提升，也让我变得越来越从容和自信。在三年研究生生活即将结束的时刻，我对给予过我关心和帮助却从未曾要求回报的各位老师、同学、亲朋好友以及我的家人致以最衷心的感谢！

首先我要感谢我的导师胡小华教授。他为人随和谦逊，在我任何需要帮助、有疑问需要解答的时候，他都会及时的回复并耐心的给予我指导和帮助。胡老师经常会组织专家学者为我们开展学科领域前沿报告，帮助我们了解最前沿的学科领域动态，这为我们了解前沿知识、开拓国际视野提供了莫大的帮助。我由衷感谢胡老师在我的研究生生涯中提供给我的指导和帮助。

然后，我要感谢我的导师何婷婷教授。除学科导师外，何教授于我而言更是一位人生导师。我不仅仅要感谢何老师在专业学习方面给予我的指导和帮助，更要感谢何老师在人生规划、为人处世等方面给予我的影响和教诲。多年来，何老师的学生无不佩服她缜密的心思和严谨的治学态度，在研究中遇到问题时总是能够为我们及时解惑，帮助我们解决各种问题。在生活中，何老师像一位家长一样，在我遇到难题时给予我帮助，在我遇到困惑时给予我指引，在我取得成绩时给予我祝贺。

我还要感谢张勇和胡珀两位老师，感谢两位老师极为负责的态度和无比的耐心，在我需要帮助的任何时刻，及时、耐心地给予我指导和帮助。

我也要感谢我身边一直陪伴着我和关心着我的同学和朋友。感谢郭博、潘博、马博、李博等几位博士师兄在学术上和生活上给予我关心和帮助；感谢呼呼、莫总、杰仔、小龙、兰姐、高明、罗星、易阳等同学在生活上给予我关心和帮助；感谢易丽、刘洋等众位师弟师妹的活泼可爱，让我的研究生生活更加丰富多彩。

感谢我的女朋友冰婵，是她的不离不弃给了我不断奋斗的动力和勇气。

感谢我的家人，感谢他们一直以来给予我坚定的支持以及毫无保留的爱。