意见目标网络与意见目标抽取研究

郝博一

摘 要

意见目标抽取是意见挖掘领域的重要子任务，同时由于意见目标抽取的对象是客观性信息，又使得它与信息抽取技术有着密切的关系。先前的意见目标抽取研究，存在四个主要弊病：1）对意见目标定义含糊。2）对意见目标管理低效。3）意见目标扩展抽取时使用的种子颗粒度偏大。4）过分依赖统计方法，句法分析不足。针对以上问题，本文在首先明晰了意见目标定义的基础上，提出了一种能结构化表示意见目标的高效管理体系——意见目标网络，以及一套基于泛化与繁殖的自举式意见目标抽取算法。

意见目标网络是一个双层有向图，它以原子意见目标（广义实体和属性）同义词集为结点，通过意见目标模式实现了对复合意见目标的表示。意见目标网络的构建过程恰恰是未知意见目标抽取过程，配合基于泛化和繁殖的多轮自举处理，显著提高了意见目标抽取覆盖率。本文在中文评价文本上进行了实验，结果表明：意见目标网络对发现未知意见目标具有很好的性能。

关键词：意见目标抽取　　意见挖掘　　信息抽取　　术语抽取　　意见目标网络

Abstract

Opinion Target Extraction (OTE) is an important subtask of Opinion Mining (OM). Meanwhile, as opinion targets carry factual information, OTE task has a close relationship with Information Extraction (IE). There are four disadvantages in previous research: 1) Having no clear definition of opinion target. 2) Inefficient management of opinion targets. 3) Manually compiled opinion targets are too large to be sound seeds. 4) Depending too much upon statistical methods, lack of parsing. To deal with this, a definition of opinion target is proposed first in this paper, followed by a structual management model of opinion target with high efficiency and a new method for opinion target extraction based on generalization, propagation and bootstrapping.

The opinion target network (OTN) is proposed in this paper to organize atom opinion targets (AOT) of generalized entity and attribute in a two-layer directed graph. OTN use nodes to show synsets of AOT and paths to show compound opinion targets (COT). With multiple cycles of OTN construction, a higher coverage of opinion target extraction is achieved via generalization and propagation. Experiments on Chinese opinion target extraction show the OTN is promising in handling the unknown opinion targets.

Keywords: opinion target extraction opinion mining information extraction term extraction opinion target network

目 录

第1章　引言

1.1 事实与意见

人们日常面对的文本中普遍包含两类信息——事实（Fact）与意见（Opinion）。它们分别对应了人类两种不同的认识世界的方式。正如古希腊哲学家柏拉图在他的著作《理想国》中所指出的[1]，人类具有两种认知世界的方式，一种是事实认知（也称为真理认知），一种是意见认知。这两种认知方式的主体都是人类本身，客体都是客观世界，但是这两种认知方式有着明显的不同。所谓事实认知，是不以个人的意志为转移的，是人类共同意志的体现。它具有固定性、肯定性和公共性。比如，“奥巴马当选了美国总统”就是一个事实认知。而意见认知则完全取决于个体，随着个体的不同而产生差异。它具有变化性、流动性和个体性。比如，有人支持奥巴马当选美国总统，有人反对。这就是意见认知。对于人类来说，由这两种认知方式所产生的两类信息——事实和意见，都有着重要的意义。

人类认识世界与改造世界的过程中也相互传递着这两种类型的信息，从早期的口耳相传，到后来的印刷品，再到互联网。互联网的产生，极大的改变了人类传递信息的速度和范围，影响到社会生活的方方面面，开创了一场信息革命。互联网，以其广泛性、快捷性、便利性，成为了现今世界人类转递信息的主要渠道。它可以轻松连接两个毫无关系的人类个体，为他们提供通信平台。由于互联网信息的公开性与易获取性，它也为大规模的信息处理提供了可能。近些年来的数据挖掘技术，信息检索技术等都是伴随互联网的快速发展而产生的。一个显而易见的事实是，互联网上绝大多数信息是以文本的形式存在的。

对应于文本中包含的两类信息，人们对文本的处理也可以分为两类——对事实的处理与对意见的处理。由此产生了两个研究领域：信息抽取（Information Extraction，IE）和意见挖掘（Opinion Mining，OM）。信息抽取（严格应称为事实信息抽取，鉴于历史沿革，仍称为信息抽取），

是指对文本里包含的事实信息进行结构化处理。它又包含命名实体识别（Named Entity Recognition，NER）、术语抽取（Term Extraction，TE）等技术。意见挖掘，是指从文本中抽取非事实的主观性信息，也就是个人、群体、组织等主体在主观性文本中表达的意见、情感和态度。意见挖掘包含有意见目标抽取（Aspect Extraction，AE）、持有者识别（Holder Identification，HI）、情感分析（Sentiment Analysis，SA）等子任务。可以看出，其中的意见目标抽取任务既是意见挖掘的子任务，同时其抽取对象——意见目标本身是事实信息，这使得意见目标抽取任务又与信息抽取紧密相关。意见目标抽取是在文本中寻找意见的表达对象，并将其抽取出来。意见本身包含情感，是非客观信息，但是意见目标往往是客观信息。所以说，意见目标抽取是跨越信息抽取与意见挖掘两个领域的一项综合任务。一方面可以使用信息抽取的基础技术，另一方面又从意见挖掘技术中得到更多的提升。

1.2 信息抽取

信息抽取，其主要功能是从文本中抽取出特定的事实信息（factual information）。主要是实体（比如时间、地点、人物名、组织名等）以及实体之间的关系（比如“位于”关系、“任职”关系等）信息[2]。图 1.1 表示了一个信息抽取的实例。

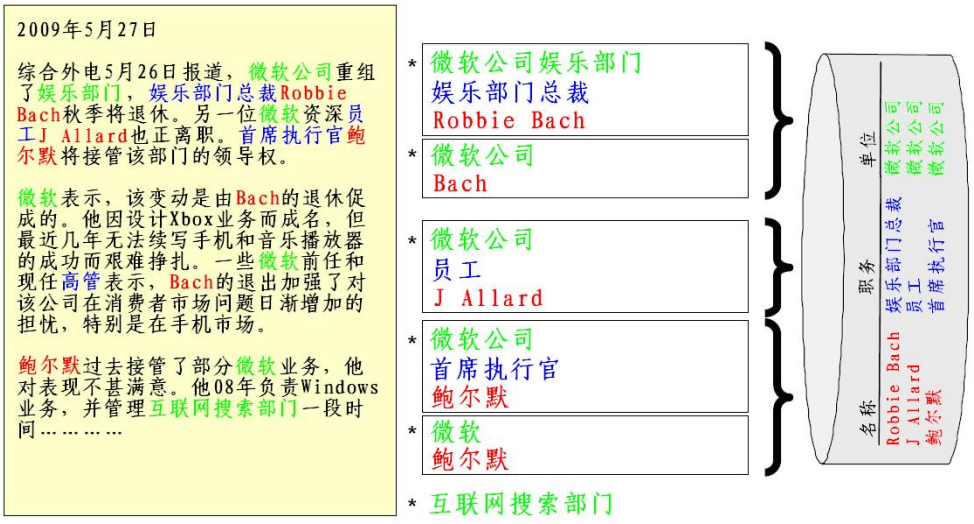

图1.1 信息抽取技术实例

在上图中，原始文本是一篇包含各种信息的介绍性文字。通过命名实体识别技术，可以在文本中确定出组织名、职务名和人名等实体信息。再通过实体关系识别技术，可以将这些名称配对，从而提取出组织名-职务名-人名三元组，构成任职关系搭配，将文本中的事实信息结构化。完成了信息抽取流程。得到的结构化信息，可以提供给用户检索，也可以作为其他信息处理技术的输入，进行更深入的信息挖掘。

信息抽取研究开始于 20 世纪 60 年代，而其蓬勃发展得益于 80 年代末开始的消息理解系列会议（Message Understanding Conference，MUC）的召开。当 1998 年最后一届 MUC 会议结束时，信息抽取已经发展成为自然语言处理领域的一个重要分支。其后，美国国家标准技术研究所（NIST）组织的自动内容抽取（Automatic Content Extraction，ACE）评测会议成为推动信息抽取技术发展的动力。它研究的主要内容是自动抽取新闻语料中出现的实体、关系和事件等内容，也就是对新闻语料中实体、关系和事件的识别与描述。现在，信息抽取已经发展出不少成熟的技术与方法，并且投入实际应用。带来了良好的社会效益。

从图 1.1 中的例子我们可以看出，信息抽取技术也需要对文本有一定程度的理解。但是它与文本理解（Text Understanding，TU）技术是迥然不同的。在信息抽取中，用户只关心有限的一些令其感兴趣的事实信息，而不关心诸如作者的意图等深层理解问题，远远达不到文本理解的水平。从这个角度说，信息抽取技术只是浅层的文本理解技术。从后文我们可以看到，意见挖掘比信息抽取的文本理解程度要高，处理难度会更大。

1.3 意见挖掘

意见挖掘，处理的对象是主观性文本，比如评论（Comments）或者断言（Allegations），其功能是自动获取关于意见的信息和知识[3]。主观性文本是相对于客观性文本而言的，它主要描述了文本作者对人物、事件等的想法和观点。再以上节中的文本为例，展示意见挖掘技术的一个实例，表示在图 1.2 中：

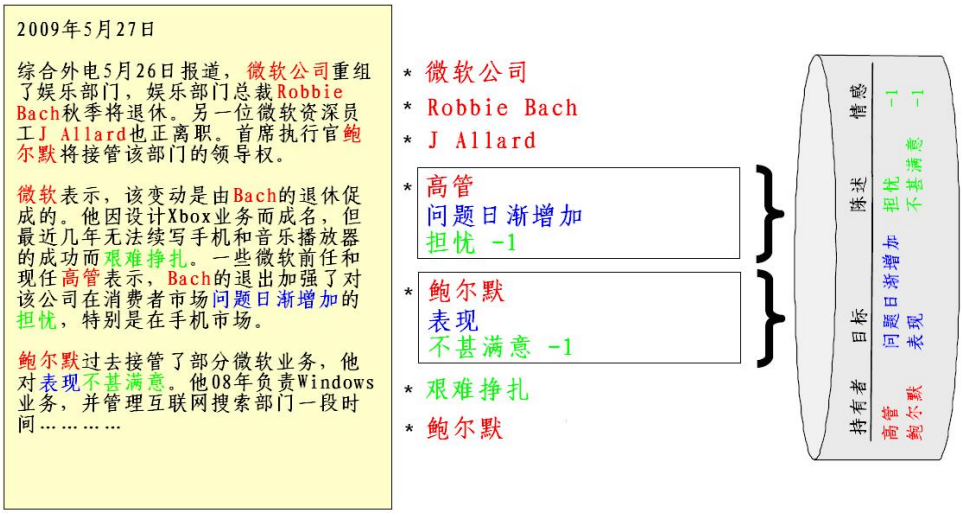

图1.2 意见挖掘技术实例

从图中我们可以看出，意见需要有持有者，也就是发表意见的个人或者组织。除此之外，意见还需要有对象，可以是具体事物，也可以是事件或者现象，它是承担意见的客体。再有就是意见本身，它包括文本中出现的陈述，还包括陈述中隐含的意见持有者的态度、情感等信息。通过意见挖掘技术，将上述一些信息以及其他的有用信息抽取出来，并按照一定的格式存储，这就是意见挖掘的过程。简单来说，意见挖掘所做的事情即是如此，但实际上所面临的问题并非如此简单。在后文中还将有所介绍。

一个完整的意见，是由多个要素组成的。对意见的挖掘，就是对意见要素的挖掘。至于表达一个完整意见所需要的要素个数，目前学术界还没有共识。文献中可以看到有三要素说[4]，四要素说[5]，七要素说[6]等说法。意见的四要素说由 Kim 和 Hovy 提出，他们认为：一个完整的意见包括四个要素，即主题（Topic）、持有者（Holder）、陈述（Claim）和情感（Sentiment）。也就是，意见的持有者针对主题发表了具有情感的意见陈述。需要指出的是，针对不同的意见，有时候主题又称为对象。这是根据意见的颗粒度决定的，对于粗颗粒度的意见，针对的事件和人物比较概括，此时称为意见的主题；对于细颗粒度的意见，针对的事件、人物、目标比较具体，此时称为意见的对象（Feature 或 Aspect）。意见的七要素说由 Kobayashi 提出，他认为：一个完整的意见需要用七个要

素描述，即持有者（Holder）、主题（Subject）、主题部件（Part）、主题属性（Attribute）、评价（Evaluation）、前提条件（Condition）、支持条件（Support）。相对于四要素说，七要素说将主题（Topic）分为了主题（Subject）、主题组分（Part）和主题属性（Attribute）；合并陈述（Claim）和情感（Sentiment）为评价（Evaluation）；并新添了要素前提条件（Condition）、支持条件（Support）。图 1.3 给出了使用意见七要素说的一个处理实例。而意见四要素说，体现在图 1.2 所表达的实例中。

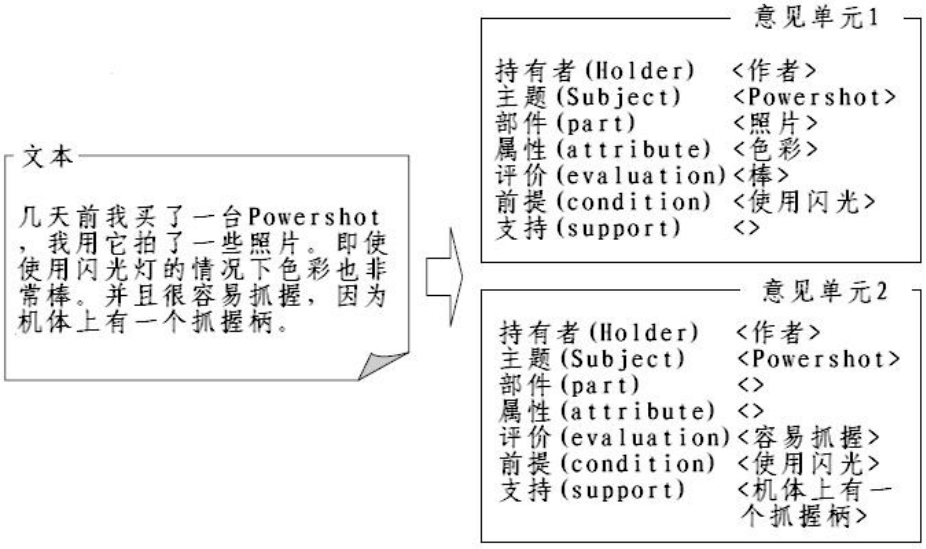

图1.3 意见七要素说处理实例

很难说这两种说法孰优孰劣，因为意见挖掘面临不同颗粒度处理的问题。所谓颗粒度，是指意见的概括程度，比如针对如下一段文字：

“尼康 D90 是一款优秀的单反相机。对于刚刚痴迷于单反相机的初级发烧友来说，拥有一台 D90 是梦寐以求的事。抛开它相对诱人的价格不谈（虽然对于普通数码相机来说是高昂的），它更让人津津乐道的是强大的功能、出色的操控性和优秀的光学性能。另外，由于配备了机身马达，所以它可以支持便宜的镜头。唯一美中不足的是它的防抖能力不尽如人意。”

我们可以说整段文字包含了一个意见——“尼康 D90 是一款优秀的单反相机”。也可以说它包含了七个意见：

（1）尼康 D90 优秀

（2）尼康 D90 有相对诱人的价格

（3）尼康 D90 有强大的功能

（4）尼康 D90 有出色的操控性

（5）尼康 D90 有优秀的光学性能

（6）尼康 D90 支持便宜的镜头

（7）尼康 D90 防抖能力不尽如人意

说这段文字包含一个意见，是从粗颗粒度角度考虑。因为整段文字所要表达的就是“尼康 D90 是一款优秀的单反相机”这个观点。虽然它遗失了许多详细的信息，但是对于特定用户——例如只关心产品名誉的用户来说，这就足够了。然而，对于那些更关注产品细节，比如相机防抖能力的用户来说，这个意见显然是不够的。必须将其细化，才能发现用户感兴趣的具体意见信息。也就是说，选择何种颗粒度，是根据不同的需求来确定的。这就使意见颗粒度的选择具有很大的灵活性和不确定性。一般来说，有粗颗粒度、细颗粒度和特定颗粒度等不同层次的选择[3]。

对于不同意见颗粒度的选择，影响了研究者们对于意见要素的定义。四要素定义法偏向于粗颗粒度的意见选择，而七要素定义法偏向于细颗粒度的意见选择。我们需要注意，从细颗粒度的意见中，可以复原粗颗粒度的意见，因为采用细颗粒度模式记录的意见包含了更多原文本中的信息。然而它付出的代价是处理流程的繁杂，处理速度的下降，以及错误率的上升。

颗粒度选择问题只是意见挖掘技术所面临的众多难题中的一个，除此之外，还有挖掘方法的精度和鲁棒性问题、隐式主题（不显式存在于文本中的主题）的识别问题、对应情感关系识别（针对多对象多情感的意见）问题等诸多问题的存在。这些问题都制约着意见挖掘技术的发展。

从意见挖掘技术诞生以来，十几年间取得了迅速的发展，新技术新方法新应用不断涌现。成为自然语言处理领域的研究热点。但由于起步晚，面临的难题众多，距离信息抽取技术那样的广泛应用，还有一定的距离。

1.4 意见目标抽取

意见目标抽取的任务是要抽取出意见所针对的目标，也就是意见表达的对象。一方面，意见目标是依赖于意见而存在的，如果不存在意见也就无所谓意见目标；另一方面，意见目标往往是实体，比如人物、事物、事件或者现象等等，所以它与信息抽取密切相关。作为意见挖掘与事实挖掘的交叉任务，意见目标抽取具有如下的意义：首先，确定意见目标，对于明晰意见本身是很重要的。有些时候，意见挖掘任务具有预设的主题，此时意见目标可能并不重要，因为所有获得的意见情感可以都归入预设的主题之下。但很多情况下，意见挖掘任务是没有预设主题的，或者虽然有预设主题，但是对次级主题（或者称详细目标）也非常关注。此时，抽取正确的意见目标可以使意见本身更加清晰准确，对挖掘结果具有重要意义。比如，同在“美国总统奥巴马的施政纲领”这个主题下，有人更关注教育，有人更关注外交，有人更关注经济，这时次级主题就显得非常重要；其次，抽取出意见目标之后，有利于联合抽取意见陈述。意见目标与意见陈述往往是成对出现的，那么，意见目标的出现表明上下文中存在意见的可能性增大。利用这个启发信息，可以提高意见陈述抽取的效率；再次，抽取准确的意见目标，有利于完成意见统计或者意见摘要。意见的汇总和归并，很多情况下需要按照意见的目标分类进行，这也是最自然的一种处理方式。如果没有准确的抽取出意见目标，就会给这个过程带来很大的困扰，导致意见统计结果出现偏差。所以说，意见目标抽取，是意见挖掘过程中的一项重要任务。

对于“意见目标”这个术语的使用，学界也并没有达成统一。本文在此处将对这一术语进行说明，一方面厘清概念，另一方面为本文中使用的表述提供依据。前文提到，对于意见所表达的那个人物或者事物、事件、现象，一般文献中会使用“意见主题（Topic）”、“意见特征（Feature）、意见属性（Aspect）、意见焦点（Focus）”等术语。而实际上，“意见主题”这个概念颗粒度偏大，适用于对大段文字的描述，而其余三个概念则颗粒度偏小，适用于描述某一个事物的具体特征。在统称这两类概念时，使用任何一个都难免偏颇。所以本文采用“意见目标”（也可称为“意见对象”）来统称以上所有提到的术语。

意见目标抽取过程中还面临一个困难的问题，那就是隐含意见目标

问题。也就是说，在文本中没有具体出现意见目标本身，而是通过上下文指代，或者习惯性省略等方式传递给阅读者。这时候需要对隐含的意见目标进行恢复。此外，对于有些意见目标，虽然表达形式不同，但它们所指的是同一个概念，属于同一概念下的不同意见目标具体表达。如何将这些不同的意见表达归纳在统一的概念下，也是一个很有挑战性的问题。进而，如何对一个特定领域进行基于本体的目标抽取，或者通过意见目标抽取技术达到本体的自动构建，都是非常有趣且意义重大的问题。

后文将按照如下结构组织：第二章介绍相关领域的研究现状，介绍一些经典算法；第三章分析面临的实际问题，明确研究难点以及现有研究方法的优缺点；第四章将介绍一种融合了浅层句法分析和统计规律的算法；第五章介绍意见目标网络，以及利用意见目标网络进行的基于泛化和繁殖的意见目标抽取算法。第六章总结前文，并展望未来工作。

第2章　相关技术综述

意见目标抽取，是信息抽取和意见挖掘的交叉任务。在本节中，将从信息抽取领域的术语抽取技术、意见挖掘领域的意见目标抽取技术两个方面来介绍前人的工作。

2.1 术语抽取

术语，是特定领域中表达特定概念的词或短语[7]。也就是说，术语可以是词，也可以是词组。术语抽取就是从大规模语料中抽取出特定领域术语的过程。

术语抽取技术最开始依赖于利用规则进行模板匹配的方法。基于这种方法，准确率高而且计算量小。但缺点是覆盖率低，并且依赖于大量的花费在制定规则上的人力资源。后来，研究者开始引入统计理论处理术语抽取问题。产生了互信息（Mutual Information）方法、Log-likelihood 方法等典型统计方法，取得了很好的抽取效果。使术语抽取技术上了一个新的台阶。同时，一些学者将统计方法与规则方法结合起来进行术语抽取，也取得了很好的效果。本节中将对统计方法以及统计与规则结合的方法进行介绍。

2.1.1　基于统计的术语抽取

基于统计的术语抽取技术大致可以分为两个部分[8]：一部分的主要作用是判断一个完整的语言单位，称之为单元度计算；另一部分的主要作用是判断这个完整的语言单位是否是术语，称之为领域度计算。

单元度计算的常用方法有互信息（MI）方法[9]，Log-likelihood 方法[10][11]，左右熵方法[12]等。其中，MI 方法和 Log-likelihood 方法是从字符串内部各部分之间的关系考虑单元度，通过考察内部各部分之间的结合强度给出单元度结果；左右熵方法则是从字符串与外部串之间的关系考虑单元度。

MI 方法是对信息论的应用，它定义概率为 P(x)和 P(y)的两个点 x 和 y 之间的互信息 MI(x,y)为：

$$MI(x,y)=\log_2\frac{P(x,y)}{P(x)P(y)} \qquad (2\text{-}1)$$

其中 P(x,y)是 x 和 y 同时出现的联合概率，P(x)和 P(y)分别为 x 和 y 独立出现的概率。由概率论我们可知，如果 x 和 y 相互独立，则 P(x,y)= P(x)P(y)，即 MI(x,y)=0；否则，如果 x 和 y 存在某种关系，则 P(x,y)> P(x)P(y)，即 MI(x,y)>>0。这样，通过计算 MI(x,y)就可以判断 x 和 y 之间是否存在联系。在术语抽取中，假设 x 和 y 是两个字串，它们在语料中出现的概率是 P(x)和 P(y)，P(xy)代表词串 xy 在语料中出现的概率。通过计算 x 和 y 的互信息，可以判断 x 和 y 的结合紧密度。互信息越高，则说明 x 和 y 更倾向于同时出现，xy 更有可能是一个语言单位。反之，则 x 和 y 的关联性越低。MI 方法给出了一个简单的计算词汇关联度的途径，但是它对概率值（字串频率）非常敏感，不适合数据稀疏情况下的处理，所以当术语的出现频率过低时，不能被有效的抽取[13]。

Log-likelihood 方法是另一种衡量两个事件是否共现的方法，对于相邻的两个字串 u 和 v，它们的 Log-likelihood 可以定义为：

$$\begin{aligned}-\log uv = {} & N\log N + a\log a + b\log b + c\log c + d\log d \\ & -(a+c)\log(a+c) - (a+b)\log(a+b) \\ & -(c+d)\log(c+d) - (d+b)\log(d+b)\end{aligned} \qquad (2\text{-}2)$$

其中 a 为词串 uv 在语料中出现的频率，b 为 uy 在语料中出现的频率（y 为除 v 以外的其他词汇），c 为 xv 在语料中出现的频率（x 为除 u 以外的其他词汇），d 为不包含 u 或 v 的候选术语的个数，N 为语料中的总词数。这种方法在抽取低频率术语时具有很好的效果。

MI 方法和 Log-likelihood 方法的本质都是用一种统计量表征两个字串 x 和 y 的同现关系，进而确定词串 xy 作为一个语言单位的概率。那么类似的，还可以尝试更多的统计量，比如 Frequency、Selectional Association、Symmetric Conditional Probability、Dice Formula、Chi-squared、Z-score 和 Student’s t-score 等。罗盛芬等人考察了以上全部九种统计量在术语抽取中的表现，得出结论：MI 方法的抽取能力最强，

各种方法组合后的最优效果也只比 MI 方法在 F 分数上高 0.7%，改进效果不明显[14]。

左右熵方法给出了另一种计算单元度的思路。它通过考察词串在边界上的特征来进行单元度计算[12]。首先计算词串边界的熵值，熵值越大，说明边界越活跃，本词串越可能是一个完整的语言单位。反之，边界熵越小，说明边界越稳定，本词串与外部词串的关系密切，不能作为完整语言单位。词串 s 左侧边界熵的计算方法是：

$$e_{left}(s)=\sum_{u,us\in C}h(\frac{|us|}{|s|}) \tag{2-3}$$

其中，u 是词串 s 左侧边界上出现的词，|s|为词串 s 在语料中出现的频率。$h(s)=s\log s$。同理，词串 s 右侧的边界熵为：

$$e_{right}(s)=\sum_{u,su\in C}h(\frac{|su|}{|s|}) \tag{2-4}$$

其中，u 为词串 s 右侧边界上出现的词。综合左右两侧的情况，词串 s 的平均边界熵为：

$$e(s)=\frac{e_{left}(s)+e_{right}(s)}{2} \tag{2-5}$$

使用边界熵的方法，不需要考虑词串 s 的内部组成，所以便于处理词长较长的词串，尤其在提取多词术语时，效果良好。

关于领域度的研究不如单元度这样受关注。目前主要是基于 TfIDf 方法以及对它的改进。TfIDf 方法在信息检索领域应用十分广泛，它基于如下的一些前提假设：第一，术语应该在特定领域中出现；第二，术语不能是领域中的平常词；第三，术语不能频繁出现在其他领域中[15]。TfIDf 值的计算公式为：

$$TfIDf(CT)=\frac{Tf(CT)}{Df(CT)} \tag{2-6}$$

其中，CT 为候选术语词串，Tf(CT)为 CT 在领域文档中出现的频率，Df(CT)为 CT 出现的领域文档数目。由式 2-6 可知，CT 在领域文档中出现的次数越多，出现的领域文档数越少，TfIDf 值越大，它越可能是领

域术语。可以发现，使用 TfIDf 方法来计算领域度，利用的是术语与所在领域文档的关系。另外还有一些研究者在 TfIDf 方法的基础上进行了改进[15] [16]，也取得了不错的效果。基于 TfIDf 思想的算法优缺点比较明显。优点是一般来说相对简单，并且除了领域语料不需要任何特定的领域信息。缺点是过分依赖于术语出现的频率，不能剔除频繁出现但是无意义的词语。

2.1.2　统计与规则相结合的术语抽取

Keh-Yih Su 等人提出一种通过提取术语组合的方式来进行术语抽取的方法[17]。此方法基于这样的假设：新术语可以由更基本的现有的术语组合构造而成。本质上，此方法将术语抽取问题看作一个二值分类问题。其中使用互信息、相关频率信息和词性模板匹配等作为术语组合抽取时使用的特征，并利用似然比（likelihood ratio）λ 作为决策的依据，其计算方法是：

$$\lambda = \frac{P(\vec{x} \mid M_c) \times P(M_c)}{P(\vec{x} \mid M_{nc}) \times P(M_{nc})} \qquad (2\text{-}7)$$

其中，Mc 代表词串 c 是由多个单元组成的，Mnc 代表词串 c 不是由多个单元组成的，$\vec{x}$ 是术语互信息、相关频率信息和词性模板的一个综合观测值。如果 λ >1，则词串 c 可以被认为是术语，否则就是非术语。研究表明这种方法在由两到三个单元（2-gram 或 3-gram）组成的术语组合抽取中是非常有效的。然而，由于对新术语的词性标记相对困难，这种方法也有其局限性。

Luning Ji 等人提出了一种基于窗口上下文的中文术语提取方法[18]。这种方法主要利用一个来自于很小窗口的上下文的句法和语义信息来进行单个领域的术语提取。其基本假设是如果候选术语一定数量的邻居词是属于特定领域时，认为候选术语也是特定领域的术语。

总的说来，这些方法综合考虑了术语内部成分的结合信息、术语与所处领域的关系、术语的词性信息等，相对于单纯使用统计的方法，增加了更多的决策依据，为我们继续研究术语抽取问题开拓了思路。

2.2 意见目标抽取

意见目标抽取与术语抽取的不同在于它可以利用与意见相关的信息，从而获得帮助，但同时抽取到的目标必须依赖于意见的存在，这也使问题的难度上升。意见目标抽取的这些特点，使它产生了许多不同于术语抽取的方法。在本节中将对现有意见目标抽取方法做简要介绍。意见目标抽取，并不像情感分析那样受到关注，甚至一些意见挖掘任务中不存在意见目标抽取的问题。但是在产品意见挖掘任务中，意见目标抽取是一个很重要的课题。

2.2.1　基于规则的意见目标抽取

Yi 等人根据名词短语的组成和位置特点，采用相似性测试（Likelihood test）方法来确定意见目标[19]。他们认为，意见目标通常表现为名词或者名词短语。这样一来，就大大减小了候选目标的规模。此外，还通过基于词性标注（Part-Of-Speech，POS）模型的算法对名词和名词短语进行二次筛选。方法中提到了三条启发性词性标注模型规则，分别是：（1）基础名词短语，（2）定指的基础名词短语，（3）句首的定指基础名词短语。通过以上两次筛选，保证了意见目标的语法完整性。其后，利用领域相关性，挑选领域局限性高的候选词作为本领域意见目标。领域相关性，使用相似性测试（Likelihood test）来计算。此方法在数码相机领域本文的实验中，取得 82%的平均正确率；在音乐领域文本的实验中，平均正确率为 96%。

Hu 和 Liu 等人利用标记序列规则（Label sequential rules，LSR）实现意见目标抽取[20]。基于标记序列规则（LSR）的方法是有监督的，通过标注好的实例，训练 LSR 模型，再利用模型获得意见目标。对于训练数据比如，

“Included memory is stingy”

第一步将其变成包含词性标注的序列：

<{included, VB}{memory, NN}{is, VB}{stingy, JJ}>

第二步标记出意见目标，形成一条规则。

<{included, VB}{$feature, NN}{is, VB}{stingy, JJ}>

识别过程与训练过程相反，首先根据规则去匹配模型，从而在相应

的$feature 位置上获得意见目标。

基于规则的方法可以达到很高的抽取正确率，但是不能有效的解决意见目标的覆盖性问题。

2.2.2　基于同现的意见目标抽取

Hu 和 Liu 根据意见目标和一些指示词的同现特征来识别常现（Frequent）和非常现（Infrequent）意见目标[21]。他们同样先抽取名词和名词短语作为候选的意见目标，同时他们认为，常现的名词和名词短语更可能是意见目标。于是，第一步，通过词性标注技术选取高频的名词和名词短语。第二步要对这些名词和名词短语进行剪枝，剔除不需要的部分。这里面主要包含两类错误的候选词，（1）多词短语，但是词间的顺序信息不符合语法规则。（2）独立词，但是它出现在更长的意见目标中。也就是，它只是其他意见目标的一部分。通过这两步之后，可以获得正确率相对较高的意见目标集合。下面开始抽取非常现的意见目标。这里用到意见目标与意见情感词汇的同现信息。借助已有的常现意见目标集合，训练意见情感词汇集合，之后借助意见情感词汇集合，启发性的寻找非常现意见目标。就这样，此方法首先使用规则得到高正确率的常现意见目标，之后借助意见情感词汇和同现概率模型，由常现意见目标向非常现意见目标前进。分两步走的达到意见目标抽取的目的。

2.2.3　基于关系的意见目标抽取

Popescu 和 Etzioni 对问题有不同的看法，他们不再把意见目标看作是孤立的，而想到利用意见目标之间的关系[22]。例如，对于扫描仪，尺寸是描述它的一种属性，而翻盖是它的一个组成部分。利用意见目标与主题词汇之间的关系，可以帮助寻找意见目标。这中间的桥梁就是关系识别符。对于扫描仪，关系识别符是这样的短语：of scanner，scanner's，scanner has 等。通过计算候选词汇与关系识别符之间的点互信息（Point-wise Mutual Information，PMI）来获取意见目标。计算公式如下：

$$PMI(f,d)=\frac{Hit(f,d)}{Hit(f)Hit(d)} \tag{2-8}$$

其中，f 是候选词汇，d 是关系识别符。点互信息越高，说明候选词

与关系识别符关联越密切，也就说明它更可能是主题词汇的一个属性或者组成部分。这样，相比于 Hu 和 Liu 的结果，此方法以牺牲 0.03 的召回率为代价，换来了准确率 0.22 的提升。

从公式 2-8 可以看出，点互信息不单单可以在有限训练数据中获得，也可以从整个互联网获得。通过搜索引擎技术，可以在互联网上寻找相关信息。

2.3 其他

Liu 等人对于给定主题的任务，利用搜索引擎在互联网上抽取结构化信息，以获得意见目标[23]。在现实中往往面临这样的问题：有些任务具有预设的意见主题，但是没有细化的具体意见目标。此方法正是利用意见主题作为关键词，使用信息检索技术得到命中文档，从这些页面的结构化组织结构中，发现副标题、分栏目等条目作为意见目标。它很大程度上依赖于网页的结构化程度，对于高度结构化的网页，可以获得出色的效果。它不依赖于意见语料，更接近信息抽取技术。

在意见抽取过程中，很多研究者也关注了同义词识别的问题。所谓同义词识别，是指在意见抽取完成后，对表述不同而意义相同的意见目标进行同义关联。这个环节对于许多后续处理过程是非常有意义的。一个简单的同义词识别方法是使用现有的同义词词典和语言学资源，比如 WordNet 和 HowNet。借助这些语言资源中所包含同义词信息，可以将意见目标进行同义词关联[24]。这样做的优点是正确率高（接近 100%）。而缺点是召回率偏低。因为许多意见目标是领域词组，没有被相关的语言资源收录。要解决这个问题，Carenini 等人提出了一种基于相似度计算的同义词识别方法[25]。对两个意见目标之间的相似度计算依赖于一个相似度矩阵，这个矩阵表征了意见目标的组成词汇之间的形似度距离，距离度量的依据是语言学资源（WordNet）。作者使用 DVD 领域的语料进行实验，取得了不错的效果。

第3章　问题分析

3.1 任务目标

意见目标抽取任务，作为意见挖掘的重要组成部分，有着广泛而重要的应用。随着电子商务的逐步发展，越来越多的人开始在互联网上购物，而各种针对产品的评价与交流网站也蓬勃发展，受到各方的关注，无论是商家还是消费者。意见挖据技术可以帮助人们更加快速的收集评价信息，为合理决策提供依据。这种应用一出现，就受到了一致欢迎。我们可以看到多款针对特定产品的意见挖掘系统[19][24] [26][27] [28]。商品领域的意见挖掘，不同于新闻意见挖掘。往往在一款产品下，讨论的焦点众多，品评对象纷繁复杂。不同的用户对商品不同的侧面感兴趣。这样就带来了意见目标数目的倍增。如何准确的抽取意见目标，在此时成为一个必须解决的重大问题。它成为影响系统性能的重要因素。

在众多产品意见挖掘系统中，多数都会按照意见目标对意见进行归类统计。例如，图 3.1 展示了一款针对汽车的意见挖掘系统，由 Gamon 等开发[27]。其中就按照“price”、“feel”等意见目标进行了归类，统计文本中对此目标的综合评价，对于不同的评价程度，给定不同的颜色，这样的方法直观、方便。

基于这种应用，就需要完善意见目标抽取算法，尽可能准确且尽可能全面地抽取文本中包含的意见目标，并能将这些意见目标正确归类。而另一方面，技术层面上我们可以发现，意见目标的准确抽取，可以帮助情感的挖掘与意见消歧，从而促进意见挖掘性能的提升。许多意见挖掘研究都关注于“属性-评价（aspect-evaluation）”的联合挖掘[4][6]。

由于意见目标抽取并不等同于意见挖掘，具有某种程度的独立性，所以既可以与意见挖掘过程联合进行，也可以独立于意见挖掘过程单独进行。也就是说，意见目标抽取任务使用的语料不局限于意见语料（产品评价）。

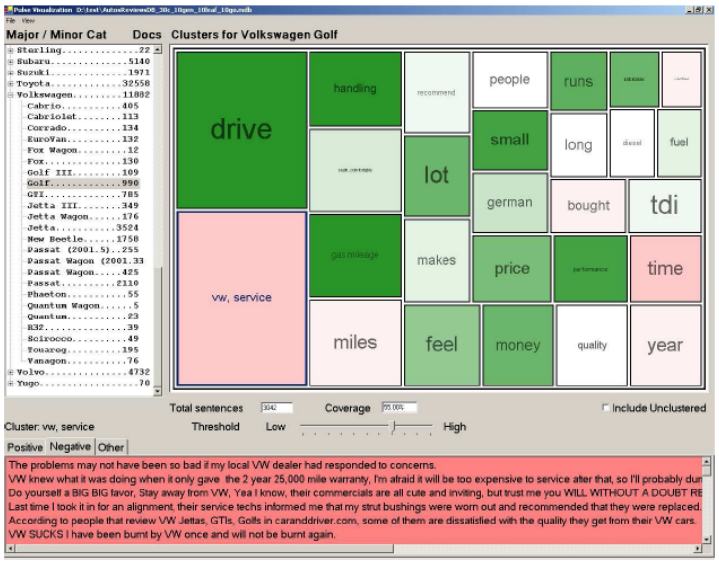

图3.1 汽车领域意见挖掘系统实例

3.2 难点分析

现有的意见目标抽取技术，都不能很好的解决问题。这主要是因为意见目标抽取面临着许多难点。

首先，意见目标数量大，种类多，造成意见目标抽取很难形成单一的有效的方法。一方面，不同领域的意见目标有不同的特点，没有统一的规律，这使得在某领域中有效的意见目标抽取算法移植到另一个领域时未必奏效。如果对每个领域设计不同的意见抽取算法，一来会造成算法的复杂度上升，二来也需要花费极大的人力劳动，对于变化迅速的现代社会，未知的新领域也在以很高的速度被产生出来，造成无力承担对算法的维护成本。另一方面，就是在某特定领域内，由于人们可以对本领域任何实体、事件、现象发表意见，这也造成了意见目标抽取的难度。同时，有些意见目标的出现频率并不高，这使得利用统计学方法很难将它们抽取出来。另外，意见抽取面临的一个困扰是新生词汇众多，结果是现有的语言资源未必能够带来帮助。现代科学技术的飞速发展，使新事物新现象不断涌现，直接后果就是随时都有从未见过的新术语词汇产生，给意见目标抽取带来压力。对于基于统计模型的意见目标抽取算法，如果不能及时更新统计模型，即使算法没有改变，也会随时间流逝而丧

失抽取性能。

第二，隐式意见目标的存在，是造成意见目标抽取困难的另一个因素。分析意见文本之后我们可以发现，有些意见目标是显式包含在文本中的。而另一些意见目标并不在文本中出现，通过上下文承接、逻辑推理等方式隐含在文本中。比如如下这个例子：

“它的性能很出众，可惜贵了点。”

这段评价文本中含有两个意见，分别是：

意见 1：<性能，出众>

意见 2：<价格，贵>

对于意见 1，它的意见目标是出现在原文本中的，它称为显式意见目标。而对于意见 2，它的意见目标“价格”没有出现在文本中，它称为隐式意见目标。隐式意见目标可以通过一定的方法来找回。在意见 2 中，由于出现意见情感词汇“贵”，通过逻辑关系，我们可以判断出它的评价对象是“价格”，从而实现了隐式意见目标的恢复。在意见目标抽取任务中，对隐式意见目标的恢复并不容易，它必须借助意见挖掘技术，借助意见目标与意见情感词汇的搭配关系。而建立这种搭配关系模型，需要训练语料，并且希望训练语料的覆盖程度尽可能高。

第三，意见目标的表达形式多样，归类难。网络文本依据风格样式可以分为两类，一类是正规文本，比如新闻、专著、评估报告等；另一类是自由文本，比如博客、点评、论坛留言等。两类文本具有不同的特点，正规文本的特点是，文本中语法和词汇错误少，用词准确规范，文意清晰通顺；自由文本则不然，里面会有很多不合语法的表达方式，会有更多的口语词汇，甚至自造词汇，文章结构也不完整，充斥着省略和意义跳跃。在产品评论文本中，大多数是自由文本。这就带来了意见目标抽取难度的增加。对于同一个意见目标，可能有不止一种表达方式，甚至出现十多种不同词汇表达同一个概念的情况。比如，产品“价格”可以被表达为：“价格”、“价值”、“售价”、“价钱”、“价”、“卖价”、“要价”、“成交价”、“门市价”等等。对于这样的情况，意见目标抽取任务不仅要将这些具体意见目标抽取出来，还要能将这些目标归于同一概念主题之下，无疑又增加了难度。

第四，修饰限制条件复杂。有些时候，单一的意见目标并不能完全

表达文意，作者可能对意见目标进行了修饰和限制。如果不考虑这些修饰限制条件，就会造成文意理解的偏差。比如下面这两段意见评论文本：

“跟那些高端货色比，它的价格当然是出色的了。”

“只有在晴朗的天气里，它的感光性能才能令人满意。”

如果我们简单的抽取出意见：

意见 1：<价格，出色>

意见 2：<感光性能，令人满意>

这显然是有悖于原文内容的。所以需要加上修饰限制条件：

意见 1：<（跟那些高端货色比）价格，出色>

意见 2：<（在晴朗的天气里）感光性能，令人满意>

目前，学界对此问题的关注甚少，Kobayashi 在他的博士论文中提出了这个问题，并进行了一定的分析研究[6]。

3.3 解决思路

前面几节中介绍了意见目标抽取任务的作用和意义、研究现状、存在难点等内容，本节中，将给出本文作者对意见目标的定义，并在分析现有抽取方法弊病的基础上，给出意见目标抽取任务的解决方案。

3.3.1　意见目标

现今，对意见目标抽取的研究很多，但是对意见目标本身的研究却并不多。如果没有认真分析意见目标的内涵和外延，将无法切中意见目标抽取任务的要害。Kobayashi 给出意见目标的较为详细的定义[6]：意见目标可以是如下一些概念——主题（Subject）、主题的部件（Part）、主题的属性或者部件的属性（Attribute）；意见目标可以包含限定条件。他在论文中将意见目标称为 Aspect。类似的，Liu 也在细致分析意见目标组成的基础上，给出了意见目标的定义：意见目标可以是——物体（Object）、物体的组分（Component）、物体或其组分的属性（Attribute）。他将意见目标称为 Feature。这两种类似的定义方法看上去很细致，但是，它们已经将意见目标定义清楚了么？

还没有。从上面的定义，我们看出，意见目标表面上可以分为两类，一类是实体，另一类是实体的属性。哲学告诉我们，世界是物质的，物

质是组成世界的基础。所以实体具有第一性。所谓实体，可以是一个活生生的客观存在，比如某个人、某匹马、某台照相机，也可以是对这些客观存在的抽象概念，比如人类、马、照相机。由于实体是第一性的，是其它一切东西的基础和主体，这样，实体就是一切属性的承担者，属性依附于实体而存在。离开实体，属性是没有意义的。另一方面，由于实体是客观存在，所以不能成为意见的直接目标。意见的直接目标只能是属性。即使对实体本身表述意见时，也是在指它的存在性。这样一来，意见的直接目标是属性，而属性依附于实体存在。那么，意见目标可以定义为：

意见目标 = 实体 + 属性

对于意见目标来说，实体和属性都是不可或缺的。前面的两种定义都显得冗余，意见目标就是实体的属性。

然而，实际的意见目标并不是一定包含实体和属性两部分，在多数情况下，会产生省略。实体和属性任何一方都可以被省略。比如，下面的两个例子：

“相机镜头很快。”

“这是一款好相机，性能优异，清晰度高，而且便宜。”

第一句评论中，意见目标只有实体而没有属性，但是通过形容词“快”，我们可以知道完整的意见目标是“相机镜头速度”。在第二句评论中，对于“性能优异，清晰度高”，意见目标只有属性没有实体；对于“而且便宜”，意见目标既没有实体又没有属性。其中，“性能”是承接前文“相机”所说的，属于承前省略；“清晰度”的实体只能是“图像”，属于无歧义性省略；“便宜”则是这两种省略的综合，一方面承接前文的实体，另一方面由于所修饰的属性只能是“价格”，所以也省略掉了。而这种既无实体又无属性的情况，就是前面提到的“隐式意见目标”。

进一步分析实际意见目标会发现，意见目标的主体并不局限于实体。还包含另外一些客观实在——实体的运动。于是提出广义实体的概念：广义实体包括实体（Entity）、功能（Function）、现象（Phenomenon）。以数码相机为例，“相机”是实体，作为相机的一个组成部分——“镜头”也是实体；“对焦”是相机的一项功能；“红眼”是照相中产生的一种现象。它们都可以是意见的评价目标，都属于广义实体的范畴。这样，我

们将意见目标的定义扩展为：

1. 意见目标 = 广义实体 + 属性

2. 广义实体 ={实体，功能，现象}

3. 在实际中，广义实体和属性都可以被省略

3.3.2　现有方法的弊病

通过前面的一些分析，我们可以总结出，现有的意见目标抽取方法有四大弊病：

第一，对意见目标定义不准确，导致抽取任务无法切中要害。从 3.3.1 节中我们得知，意见目标抽取任务不仅仅是简单的字符串抽取，它的核心任务是要挖掘出广义实体和属性两部分，缺少任何一部分，都不能算是完整的意见目标抽取。对于传统的意见目标抽取任务来说，他们对如下文本的处理是：

“相机非常漂亮” → 意见目标抽取算法 →“相机”

“相机很贵” → 意见目标抽取算法 → “相机”

“相机轻便” → 意见目标抽取算法 → “相机”

虽然抽取出的意见目标是一致的，但是用户可以轻易发现其中的不同，对于用户来说，他们需要的意见目标抽取算法应该是：

“相机非常漂亮” → 意见目标抽取算法 →“相机+外观”

“相机很贵” → 意见目标抽取算法 → “相机+价格”

“相机轻便” → 意见目标抽取算法 → “相机+重量”

这就是由于对意见目标的定义不清晰，造成了抽取结果不能代表真实的意见目标。

第二，过分倾向于统计方法，对语法分析工具的利用不足。我们知道，意见目标抽取的主要瓶颈在于长度大的组合词汇以及频率底的词汇。那些长度不长、词频高的意见目标，是容易抽取的。图 3.2 是一张意见目标抽取准确率/召回率/F1 分数随词长变化的曲线。

在图 3.2 中，随着词长的增长，意见目标的抽取准确率是逐步上升的，也就是说，对于长词汇，只要字符串匹配成功，则它是意见目标的概率就很大；随着词长的增长，意见目标的抽取召回率是逐步下降的，这表示对长词汇意见目标的挖掘是困难的。往往找不到它，或者抽取不

完整；F1 分数作为抽取准确率和召回率的折中，它随着词长的增加，出现极大值。既然抽取的一个主要困难是对长词汇意见目标的抽取性能低，那么分析长词汇意见目标的特点，可以发现，长词汇意见目标往往是由多个短词汇组合而成的词组。由此可以推知，引入浅层语法分析会对长词汇意见目标的抽取带来帮助。现有方法忽视语法分析技术的使用，是一个损失。

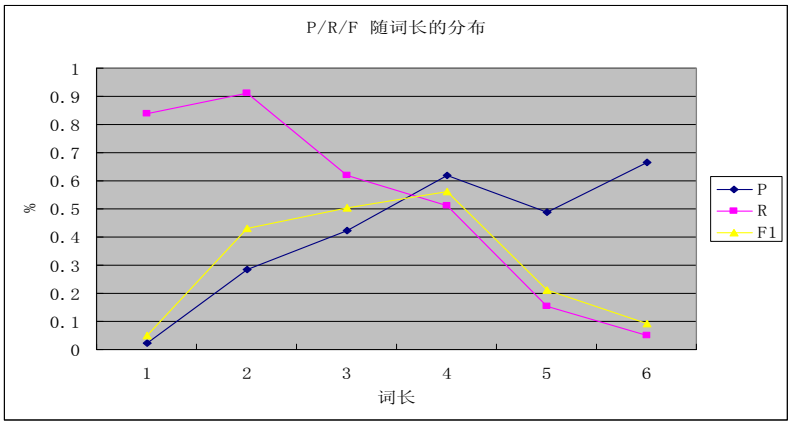

图3.2 意见目标抽取准确率/召回率/F1 分数随词长变化的曲线

第三，对意见目标的管理体制不高效，导致无法利用意见目标之间的启发信息。现有方法对意见目标的管理主要有两种模式——列表型和树型。所谓列表型，就是把已知的意见目标全部按照无序表来存储，每个意见目标之间没有任何关系，只有列表本身包含一些统计信息。有些时候，还会给列表中的每个意见目标指派一个概念。图 3.3 给出了一个列表型管理模式的示意图例：

列表型的优点是简单，管理成本低。但随之带来的还有管理效率的低下和信息的浪费。所谓树型管理模式，是以领域主题为根结点，以各次级概念为子结点构建的一棵概念树。将已知的意见目标分别归类于相应的结点下。图 3.4 给出了一个树型管理模式的示意图例：

镜头性能	镜头
对焦速度	对焦
红眼问题	红眼
LCD亮度	LCD
原配电池的寿命	电池
单机售价	售价
国产镜头价格	镜头
相机的质量	质量
快门速度	快门
电池卡仓的开关锁按钮	按钮
背光时的图像质量	图像
……	……

图3.3 列表型管理模式示意图

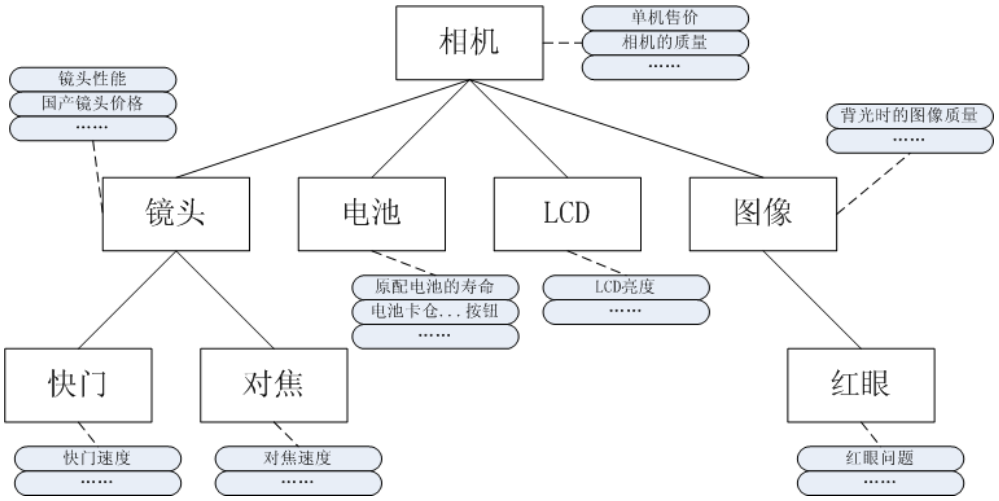

图3.4 树型管理模式示意图

树型比列表型有进步，它把意见目标升级为二维管理。在一定程度上归并了同概念意见目标，并且对意见目标的概念按亲子关系进行组织。将一些有用信息表现在组织结构中，一方面为抽取算法提供了支持，另

一方面更加贴近用户，给用户一个更加结构化的视觉效果。

但是这两种管理模式都还远远不能满足意见目标抽取算法的要求和用户的需求。 需要探索一种更加有效的管理机制，让意见目标之间的各种联系和启发信息得到充分的表现，使意见目标以方便用户使用的前提进行组织。

第四，基于扩展抽取的意见目标抽取方法，所使用的种子颗粒度偏大，不能发挥良好的效果。现有意见抽取方法中，扩展抽取是一个比较常见的方法，它基于已知的意见目标，通过对语料中已知意见目标进行字符串扩展来发现新的意见目标。但是这种方法一般并不对用来做扩展的种子意见目标进行筛选，导致有些种子的颗粒度偏大，扩展能力降低，进而导致了抽取性能的损失。

3.3.3　解决方案

针对以上问题，本文提出的解决方案如下：

一、在统计方法的基础上，加入浅层句法分析算法。着力挖掘意见目标内部的句法信息，通过句法扩展，获得新的意见目标。首先，使用浅层句法分析工具，如分词-词性标注器、依存分析器等，对处理文本进行句法分析处理。之后在文本中提取句法特征和其他一些统计特征，来表征一个候选意见目标。通过特征向量的统计特性，产生候选意见目标排序。设置一定的置信度，则可以实现对候选意见目标的二值分类，从而实现意见目标的抽取。

二、提出意见目标网络，按照网络的形式管理意见目标，摒弃了以往的列表型和树型结构。通过网络的形式，更大程度上保存了意见目标结点之间的关系信息，更加有助于通过结点之间的相互关系来实习启发式挖掘。另外，进一步解决种子意见目标颗粒度过大的问题。通过将已有意见目标进行泛化，得到更一般、颗粒度更小的优质种子，再将种子进行扩展繁殖，获取更多意见目标。

3.4 解决思路后文结构

在后文中，第四章将介绍一种融合了浅层句法分析和统计规律的算法，第五章介绍意见目标网络，以及利用意见目标网络进行的基于泛化

和繁殖的意见目标抽取算法。这两种方法都能有效的从评价语料库中抽取出意见目标。

第4章　统计与句法分析相结合的意见目标抽取方法

4.1 介绍

本章算法基于一个小规模基本意见目标集，从浅层语言分析结果中提取与已有意见目标相关的统计特征，作为未登录意见目标判别的依据。同时利用这些统计特征对未登录意见目标进行排队，将可信度较高的未登录意见目标置于队列前部。实验结果证明，在排队后的前 200 个意见目标中能达到 87.5%的抽取正确率。

基于本章意见目标抽取算法的系统称为 OPINAX 意见目标抽取系统。它以产品领域的评论文本为语料，通过规则与统计相结合的方式抽取意见目标。本系统使用 OPINMIE 系统[28]的意见语料库，以及 ICTCLAS 语法分析器[29]和 D-parser 依存分析器[30]等浅层句法分析工具。OPINMINE 系统由香港中文大学开发，在第六届 NTCIR 意见挖掘国际评测中取得优异成绩。它所使用的语料库是数码相机产品领域的网络评论文本，全部来源于互联网，属于非规范文本。语料中标注了意见目标、意见表达、意见极性等诸多要素，可以支持意见目标抽取、意见抽取、意见情感分析等任务。ICTCLAS 语法分析器由中科院计算所开发，是目前为止性能最好、使用最广泛的中文分词和词性标注系统。DParser 依存分析器由哈尔滨工业大学设计开发，它能对汉语文本进行依存分析。依存分析是一种浅层的句法分析，它不同于句法分析，不能得出严格的句法结构，只能得出句子中不同词汇间的依存关系。但同时，由于依存分析对语言的处理不再局限于表层匹配，而是深入语言的内部结构。其分析结果为机器翻译、信息检索、信息抽取等应用领域提供有力的支持。

本章将首先介绍算法的基本框架、具体处理流程，并详细讲解关键算法，在之后的实验部分，设计多种实验验证算法的可行性和效率。并对实验结果进行分析，探讨算法的优缺点。最后得出结论，说明本算法是有效的意见抽取算法。

4.2 算法架构

4.2.1　算法结构及流程

OPINAX 系统的目标是从评价文本中抽取产品意见目标。图 4.1 描述了 OPINAX 系统的整体架构。

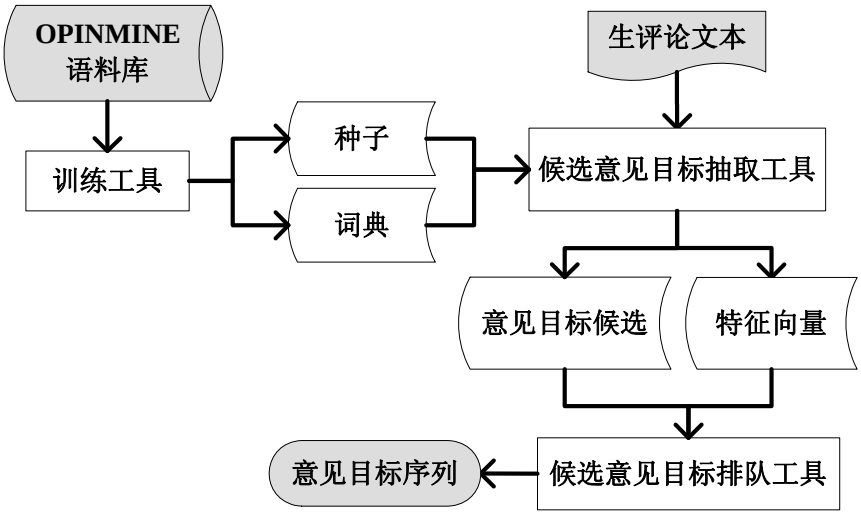

图4.1 OPINAX 系统的整体架构图

OPINAX 系统可以分为：语料训练阶段、候选意见目标抽取阶段、候选意见目标排队阶段等三个阶段。

首先，在语料训练阶段，OPINAX 系统借助训练工具从 OPINMINE 语料库建立意见目标种子集和三个词典：意见情感词词典、意见目标-意见情感词搭配词典以及意见指示词词典（这三个词典将在后面章节中详述）。其中，意见目标种子集作为意见目标扩展抽取的基础；而词典则提供了启发信息和统计信息，为之后的意见目标抽取提供资源。

其次，在候选意见目标抽取阶段，对输入的一批生评价文本，OPINAX 系统首先调用候选意见目标抽取（CE）工具，根据意见目标种子在输入文本中扩展性的寻找候选意见目标。为了获得尽可能多、且尽可能优质的候选意见目标，这里引入了两种浅层语言分析工具——ICTCLAS 语法分析器和 DParser 依存分析器。它们负责指导意见目标种子的扩展行为，使得这些扩展具有更强的句法依据。扩展出的候选意见目标集合作为 CE 工具的一项输出。而 CE 工具的另一项输出是每一个

候选意见目标所对应的特征向量。每个特征向量包含多个维度，其中包含有依存关系类型、候选词长度、候选词词频、拉丁字母出现与否、意见指示词、意见情感词等特征的统计信息（细节请参见 4.2.3 节）。这些信息将作为候选意见目标可信度的依据。将 CE 工具的这两项输出作为输入，进入候选意见目标排队阶段。

第三，在候选意见目标排队阶段，候选意见目标排队（CR）工具为每个候选意见目标计算可信度，从而实现对候选意见目标的排队。为了实现这个目标，我们基于候选意见目标的特征向量设计可信度的计算公式（细节请参见 4.2.4 节），根据可信度公式对每个候选意见目标的打分来决定它的可信性。

最后，经过三个阶段的处理，得到候选意见目标序列，更有可能成为真实意见目标的候选项将排在队列的前端。通过这种方式实现了意见目标抽取。

在接下来的几节中，将介绍本系统的核心算法，4.2.2 介绍如何抽取候选意见目标，4.2.3 介绍特征向量的生成过程，4.2.4 介绍如何对候选意见目标进行可信度排队。

4.2.2　候选意见目标抽取

候选意见目标抽取算法是本系统的核心算法之一。它实现了从意见目标种子扩展为候选意见目标的过程。它使用意见目标种子和依存关系来定位候选意见目标。意见目标种子来自于训练语料，是网络用户实际使用的评价对象。依存分析将句子从线性序列转化为一棵结构化的依存分析树，利用依存弧来反映句子中各词汇之间的依存关系。图 4.2 和图 4.3 给出了依存分析结果的一般形式。

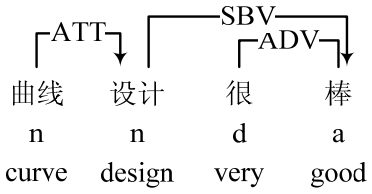

图4.2 依存分析实例 1

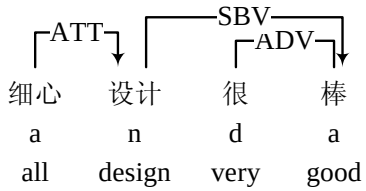

图4.3 依存分析实例 2

实例图中，汉语句子被 ICTCLAS 语法分析器分成基本词汇，其下方第一行是词性标注结果。n 表示名词词性，d 表示副词，a 表示形容词。其上方是 DParser 依存分析器给出的依存分析结果，用线段连接的两个词汇之间产生依存关系，线段上标记有不同的依存关系类型：ATT 表示修饰关系，SBV 表示主谓关系，ADV 表示状语中心语关系。在每一组依存关系中存在一个中心词（也称为核心词），它是这个关系的核心成分，支配另一个词。在本例中，箭头指向的词为中心词。同样的，一个依存关系也可以表示为三元组的形式。例如，三元组（曲线，设计，ATT）表示了一种修饰依存关系：即“曲线”修饰“设计”。其中，“设计”是中心词。

在依存关系中，若一方为意见目标种子，则可以进行依存扩展，从而抽取出候选意见目标。比如在实例 1 中，若“设计”是意见目标种子，则可以通过合并这两个相互依存的词汇，作为意见目标种子“设计”的一个扩展，将其抽取出来作为一个新的候选意见目标，即“曲线设计”。

但并不是所有的依存关系都可以用来做扩展。比如，实例 1 中“设计”和“棒”之间是 SBV 主谓关系，如果将意见目标种子“设计”扩展为“设计棒”，从语法的角度来说，有理由认为这种组合方式不大可能会产生新的意见目标。而实际上，“设计棒”也确实不是意见目标。这个例子说明，必须对扩展所使用的依存关系进行约束和限定。只有特定的几种依存关系可以用来对意见目标种子进行扩展。

上面的两个例子体现了 CE 工具的核心思想：若某意见目标种子出现在依存分析的结果中，而且另一个与意见目标种子邻接的词与之形成特定的依存关系（比如 ATT），那么这个词就可以与属性种子合并，生成一个新的候选属性。D-parser 分析器可以产生 24 种依存关系，经过人

为的分析研究，OPINAX 系统只针对三种依存关系进行扩展处理，它们是：ATT（修饰关系），COO（并列关系），VOB（动宾关系）。

另一方面，词性对依存分析的结果影响很大。比如，ATT 关系的中心词通常会是一个名词，但也存在其他可能词性。这就造成了依存关系中词性的搭配多种多样。对于候选意见目标来说，不同的词性搭配可信度不同。这一点我们可以从实例 1 与实例 2 的比较中得到验证。在实例 1 中，通过 ATT 关系扩展得到候选意见目标“曲线设计”；而在实例 2 中，通过 ATT 关系扩展得到候选意见目标“细心设计”。两者的差别仅在于扩展部分的词性不同。在本例中，扩展部分的词汇是对中心词“设计”的修饰。由于修饰部分的词性不同，导致了“曲线设计”是一个常见的产品属性，而“细心设计”却鲜被作为产品属性对待。基于以上的观察，OPINAX 将词性搭配信息也纳入到候选意见目标抽取的方法中。

更一般的，对产品评论更多的观察表明，有些产品意见目标包含两层甚至更多层的依存关系。图 4.4 中所示的实例 3 提供了两个邻接的依存关系，即（外形，曲线，ATT）和（曲线，设计，ATT）。它们可以被合并成一个更长的意见目标，即“外形曲线设计”。受此启发，OPINAX 引入多层依存关系以抽取更多的候选意见目标。同时，考虑到引入过多层的依存关系一方面会导致算法复杂度上升，另一方面会导致噪声增多，在 OPINAX 系统中限制用于扩展的依存关系层数不得超过两层。

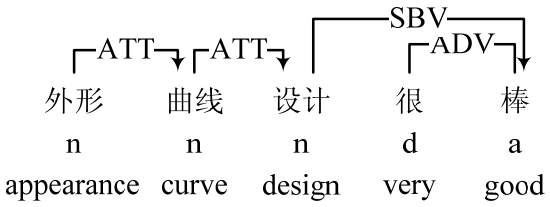

图4.4 依存分析实例 3

CE 工具通过对意见目标种子进行依存关系扩展得方式，可以在生评价文本中获得大量的候选意见目标。但是，这些候选意见目标是否都是正确的意见目标呢？答案是否定的。如何辨别这些候选意见目标的质量，是随之而来的问题。OPINAX 系统通过计算候选意见目标来判断它作为意见目标的质量。而这种可信性是利用候选意见目标的特征向量来

表征的。

4.2.3　特征向量生成

特征向量在 OPINAX 系统中起重要作用，它由候选意见目标抽取工具（CE）产生，用于候选意见目标排队工具（CR）以计算候选意见目标的可信度。为了叙述方便，此处先给出特征向量的描述。

特征向量包含 6 个维度的特征，用以下六元组定义：

$$FV =< DR, INDO, TF, TL, LTO, SENO > \quad (4\text{-}1)$$

对公式（4.1）中的符号元素解释如下：

4.2.3.1 DR：依存关系类型

依存关系类型在候选意见目标抽取和候选意见目标排队过程中都有重要意义。本系统只考虑三种类型的依存关系：ATT，COO 和 VOB。由于 OPINAX 系统支持两层依存关系扩展，特征向量中也相应地包含了两层依存关系：直接关联到意见目标种子上的依存关系被称作主依存关系（DRPri），其余的称作次依存关系（DRSec）。另外，我们已经知道词性对于依存关系扩展的重要性，有必要把扩展词汇的词性作为一个特征引入特征向量，用 POS 表示。这样，特征向量被扩展为如公式（4.2）的形式。

$$FV =< DRPri + POS, DRSec + POS, INDO, TF, TL, LTO, SENO > \quad (4\text{-}2)$$

4.2.3.2 INDO：指示词出现与否

观察产品评价文本可以发现，产品意见目标的周边词汇具有统计规律。也就是说，产品意见目标经常出现在某些特定词汇的前面或者后面，这些特定的词汇在 OPINAX 系统中统称为意见指示词。

比如，“比较”总是经常出现在意见目标的后面，而“但”这个词则常常出现在意见目标的前面。对比两种情况，将“比较”称作后指示词，而将“但”称作前指示词。意见指示词词典可以通过对 OPINMINE 语料库的训练过程得到。OPINAX 统计语料库中直接出现在意见目标前面或者意见目标后面的词汇的词频，将词频大于 2 的词汇认定为意见指示词。

通过这种方法，最终从语料库中学习到 201 个意见目标指示词。

需要指出的是，INDO 特征属于布尔型变量。若在该候选意见目标的前后发现指示词，就将此维特征赋值为 1，否则为 0。

4.2.3.3 TF：词频

TF 特征反映了该候选意见目标在评价文本中出现的次数。一般情况下，高频率的候选意见目标比低频率的候选意见目标更有可能是一个真实的意见目标。所以词频可以作为一维特征。

4.2.3.4 TL：词长

TL 特征反映了候选意见目标的字符串长度。对产品评价文本的观察显示，字符串长度越长的候选意见目标，越可能成为真实的意见目标。

4.2.3.5 LTO：拉丁字符出现与否

LTO 特征指示在候选意见目标中是否出现了拉丁字符。可以假设，如果在中文文本中出现了含有拉丁字符的候选意见目标，则它更有可能是真实意见目标。这是因为拉丁字符通常被用作缩写或者英文原词，而这些词汇更可能作为一个实体，代表一个具体的意见目标。

LTO 是布尔型特征。含有拉丁字符的项取值为 1，否则为 0。

4.2.3.6 SENO：情感关键词出现与否

产品意见目标和意见情感词之间的搭配关系可以帮助辨别真实的意见目标。如果一个候选意见目标与一个情感词在文句中存在搭配关系，这个候选意见目标就更有可能是一个真实意见目标。这是因为，意见目标与意见情感词的搭配出现，意味着意见的产生。其中，意见情感词词典和意见目标-意见情感词搭配词典在训练阶段从语料库中获得。

SENO 是布尔型特征。若候选意见目标与情感词搭配出现，并且满足 SBV 关系（主谓关系）或者 ATT 关系（修饰关系），则 SENO= 1，否则为 0。

在候选意见目标抽取的过程中，同时为每个候选意见目标建立如下

的特征向量：计算它的字符串长度并且检测其是否包含拉丁字符；在全部的生文本上计算它的词频；搜索上下文，检查是否存在意见指示词或者搭配出现的情感关键词。在得到这些信息之后，建立特征向量。比如，候选意见目标“XD 卡价格”的特征向量如下：

$$fv =< ATT\ n-n, ATT\ n-ws, 0, 12, 8, 1, 1 > \quad (4\text{-}3)$$

在这个例子中，“ATT n-n”代表了主依存关系的特征值，而 ATT n-ws 代表了次依存关系的特征值。其他的特征取值表明，此候选意见目标的前后没有发现意见指示词（INDO=0），它在文本中共出现了 12 次（TF=12），它的字符串长度是 8（TL=8），它包含有拉丁字符（LTO=1），且在它附近发现了搭配使用的情感关键词（SENO=1）。

4.2.4　候选意见目标排队

候选意见目标排队算法是本系统的另一个核心算法。它实现了把候选意见目标根据可信度分数排队的过程。在候选意见目标被抽取出来之后，无法确定它是否就是真实的意见目标。可以使用分类器对候选意见目标进行二值分类，确定某候选意见目标是不是正确。但是现有信息不足以达到一个优秀的分类结果，且对于此任务而言，只是需要推荐一些新的可靠的意见目标，所以选择了候选意见目标排队的算法，将更有可能是意见目标的候选词排在队列的前端，这样在使用时就可以从表头取出所需的意见目标。排队的依据，是由候选意见目标排队工具对每一个候选意见目标计算的一个可信度值，依据可信度将高分的候选意见目标排在队列的顶端。现在的核心问题是，六维特征对候选意见目标的排队都有影响，但它们的影响程度又各不相同。如何衡量它们的影响程度呢？OPINAX 使用可信度计算公式来解决这个问题。文中尝试了两种方法来构造可信度计算公式。

第一种采取决策树方法。也就是不依赖于训练语料，由人工对每一维特征设定一个确定的优先级，例如 F_1 的优先级是 P_1，F_2 的优先级是 P_2，…，F_n 的优先级是 P_n ($P_1>P_2>\ldots>P_n$)。F_1 到 F_n 都是取值为{0，1}的布尔型变量。对于非布尔型特征，通过添加二级特征的方法转化为布尔型特征。例如，特征 DRPri+POS=（ATT n-n），可以转化为 DRPri+POS

（ATT n-n）=1。二级 DRPri+POS（ATT n-n）的优先级，继承原特征 DRPri+POS 的优先级。于是每个候选意见目标的可信度分数为：

$$Rd(fv) = P_1 \cdot F_1 + P_2 \cdot F_2 + \ldots + P_n \cdot F_n \quad (4\text{-}4)$$

优先级 P_1 到 P_n 由对本领域的数据有一定了解的研究人员手工设定。不同优先级之间需要有比较大的阶差。

第二种权值自动学习的方法。该方法从数据中学习得到不同特征对排队的不同贡献。处理流程如下：将生评价文本按照特定比例（例如 1：9）分成两个部分，1/9 数据作为开发集，剩余作为测试集。在开发集上运行候选意见目标抽取工具，得到候选意见目标和对应的特征向量。接下来，对每个候选意见目标人工标注“正确”和“错误”的标签。之后，我们根据下面的方法来确定每个特征的影响因子：对于特征 F_i，用 nf_i 表示 F_i=1 的候选词数目，用 np_i 表示 F_i=1 并且 tag=“正确”的候选词数目。计算特征 F_i 的准确率 PF_i，$PF_i=np_i/nf_i$。使用这个准确率 PF_i 来衡量特征 F_i 在决定候选意见目标排队中的重要性。于是每个候选意见目标的可信度分数可以通过下面的公式计算：

$$Rw(fv) = PF_1 \cdot F_1 + PF_2 \cdot F_2 + \ldots + PF_n \cdot F_n \quad (4\text{-}5)$$

对于非布尔型特征，按照前面所描述的方法转化为二级特征。

4.3 实验

4.3.1 实验数据与评测标准

4.3.1.1 数据

本实验数据分为两部分，一部分是训练数据，一部分是评测数据。训练数据来源于 OPINMINE 系统语料库[28]。是数码相机领域的网络评价文本。以 xml 格式进行了人工标准。

其中，对每条意见都标注有意见目标，并且可以表示意见目标缺省的情况。在全部训练数据中，共计有评论 4001 篇，意见 9950 个。

测试数据为来源于互联网的数码相机领域评价文本，没有进行人工标注。通过爬虫获得之后，进行简单的预处理，比如查重、过滤无关信

息等，存储在文本文件中。文件总大小为 13.2M。因实验需要，又将测试数据分为两部分，一部分 2.0M，作为开发；另一部分 11.2M，作为测试。

4.3.1.2 评测标准

在意见目标抽取领域，常用的评测标准为准确率（P）、召回率（R）和 F1 分数（F1-score）。其中，准确率 P 的计算公式为：设抽取出的全部候选意见目标数为 N_e，其中正确的意见目标数目为 N_r，而文本中实际包含的全部意见目标数为 N。那么，

$$P = \frac{N_r}{N_e} \tag{4-6}$$

召回率 R 的计算公式为：

$$R = \frac{N_r}{N} \tag{4-7}$$

准确率和召回率分别考察了抽取的两方面性能，这两个指标之间有一个平衡。一般的，如果准确率上升，就会带来召回率的下降，而准确率的下降会带来召回率的上升。所以，为了融合两个指标，定义 F1 分数：

$$F1\text{-}score = \frac{2PR}{P+R} \tag{4-8}$$

但是，本实验选择 P@N 参数作为 OPINAX 系统的评测指标。P@N 参数表示在候选意见目标序列前 N 个候选意见目标中的抽取准确率。定义式为：

$$P@N = \frac{n}{N} \tag{4-9}$$

其中，n 表示前 N 个候选意见目标中正确的意见目标的个数。本实验选取 P@N 作为评测指标，是由 OPINAX 系统采用候选意见目标排队策略决定的。这种策略下，没有对候选意见目标进行二值分类，所以也就不存在严格意义上的准确率 P 和召回率 R。而这种策略也是有实际意

义的，用户有时候只是需要更多的意见目标，然后使用这些意见目标，结合意见挖掘算法来挖掘意见。所以，OPINAX 给出的是一个可任意选取的意见目标集合，用户可以根据实际情况，选取某特定准确率条件下的意见目标。于是，P@N 也就成为了一种有效的考察 OPINAX 系统性能的指标。

4.3.2　实验方法

4.3.2.1 实验一：单层依存关系扩展与多层依存关系扩展性能对比实验

依存关系是本文中用来抽取候选意见目标的主要依据。本实验中，我们分别使用一层依存关系和两层依存关系进行候选意见目标抽取。在抽取候选意见目标时，将意见目标种子在依存分析结果中进行匹配，合并符合依存关系种类的词对作为候选意见目标。

在一层依存关系扩展情况下，我们只考虑(W_1_W_2) 类型的依存合并。也就是说，W_1 和 W_2 邻接，它们之间存在依存关系，且满足依存关系类型条件。而对于两层依存关系扩展，我们限制以下扩展模式：

W_0_(W_1_W_2)

(W_1_W_2)_W_3

W_0_(W_1_W_2)_W_3

W_{-1}_W_0_(W_1_W_2)

其中，W_1 和 W_2 是第一层依存关系扩展，它们满足一层依存扩展的条件，第二层依存扩展的词汇（W_0、W_{-1}、W_3）需要和第一层依存中的词汇 W_1 或者 W_2 具有依存关系，且满足依存关系类型条件。

在候选意见目标排队时，本实验采用决策树的方法，暂时只考虑 5 个特征，人工设定优先级为： DRPri+POS > DRSec+POS > LTO > TF > TL。实验结果如图 4.5 所示，其中横坐标是排队序列中的前 N 个候选意见目标的序列编号，纵坐标是 P@N 指标。蓝线所代表的，是只考虑一层依存关系扩展时的结果曲线，红线代表的是考虑两层依存关系扩展时的结果曲线。

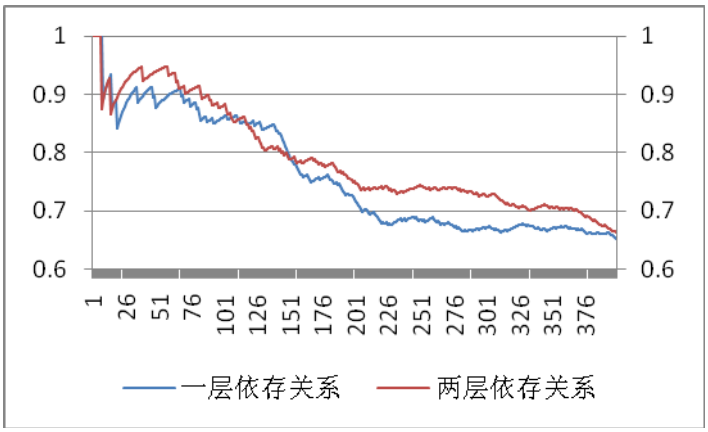

图4.5 使用不同依存关系层数进行扩展的结果比较

4.3.2.2 实验二：统计特征的影响对意见目标排队的影响

本文采用了多个统计学特征，为了验证这些统计特征对于候选意见目标排队的影响，设计了如下实验：以实验一中使用二层依存关系扩展进行候选意见目标抽取的实验作为基准实验。分别在基准实验的基础上依次添加意见指示词特征、情感词特征以及两者的结合。将分别得到的实验结果与基准实验结果进行对比，就可以看出这两个统计学特征对于意见目标排队的影响。

在意见目标排队时均使用决策树的方法，三个对比实验中优先级的设定分别为：

1. DRPri+POS > DRSec+POS > LTO > INDO > TF > TL；
2. DRPri+POS > DRSec+POS > LTO > SENO > TF > TL；
3. DRPri+POS > DRSec+POS > LTO > INDO > SENO > TF > TL。

之所以挑选意见指示词特征和情感词特征来做实验，是因为这两个特征与候选意见目标抽取的过程无关，更能体现出统计学特征对于意见目标排队过程的影响，更有说服力。同时，由于意见指示词特征和情感词特征是领域无关特征，也就是说，这两个特征不止可以用在数码相机产品评价领域，还可以用在其它领域的意见目标抽取中。这两个特征具有普遍意义。

将三次对比实验的结果分别表示在图 4.6、图 4.7 和图 4.8 中：

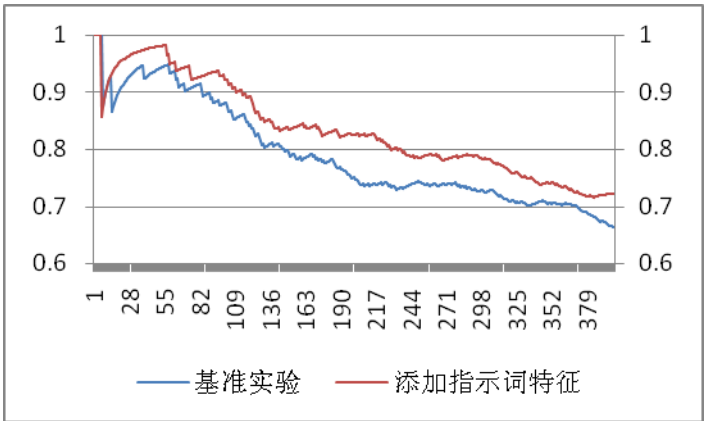

图4.6 加入指示词特征前后的效果对比

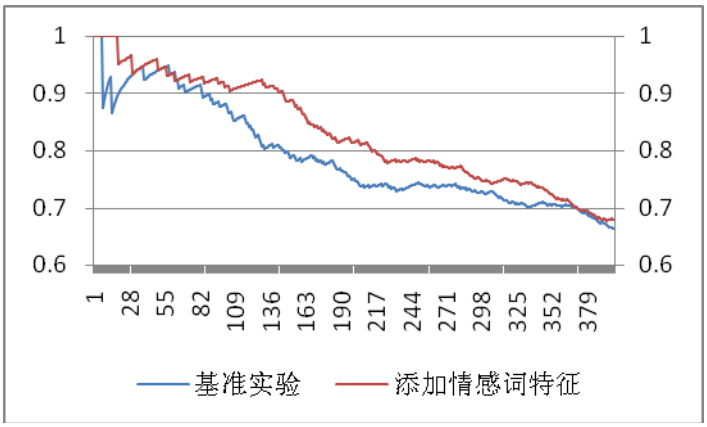

图4.7 加入情感词特征前后的效果对比

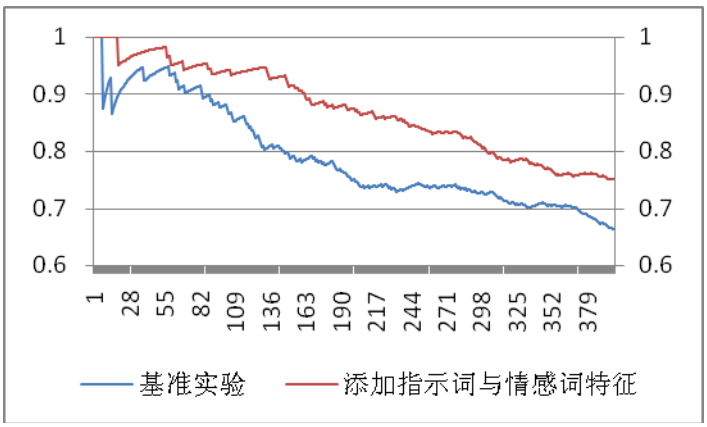

图4.8 结合了指示词特征和情感词特征的实验结果

在以上三图中，横坐标是排队序列中的前 N 个候选意见目标的序列

编号，纵坐标是 P@N 指标。蓝线所代表的，是基准实验的结果曲线。在图 4.6 中，红线代表了加入意见指示词特征后的实验结果。在图 4.7 中，红线代表了加入情感词特征后的实验结果。在图 4.8 中，红线代表了同时加入指示词特征和情感词特征后的实验结果。

4.3.2.3 实验三：决策树公式排队法与权值自学习公式排队法性能比较实验

本实验对前文所述的两种候选意见目标可信度排队方法进行了对比。以采用决策树方法，并添加了全部特征的实验二作为基准实验。对比实验采用权值自学习公式排队法，其设计如下：在开发集上运行候选意见目标抽取工具，得到大约 1500 个不同的候选意见目标。人工给每个候选意见目标做一个标记，标记它是否是正确的意见目标。之后计算每个特征的准确率 PF_i，作为该特征的重要性因子。接下来，在测试集上运行候选意见目标抽取工具，并使用公式（4.5）对每个候选意见目标进行可信度打分，从而实现候选意见目标排队。

实验结果如图 4.9 所示：

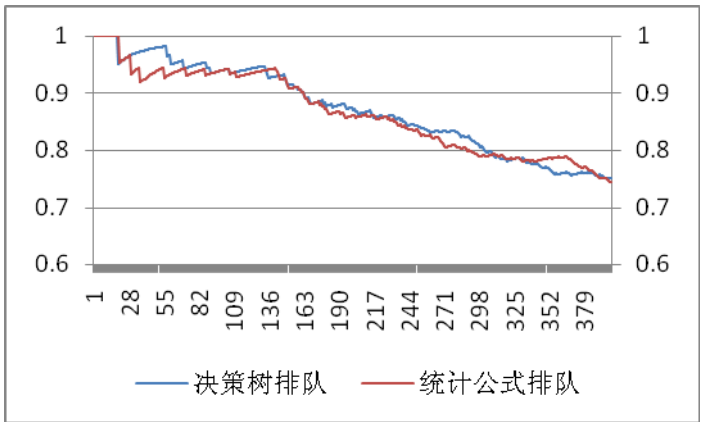

图4.9 两种不同排队方法的结果比较

在上图中，蓝线表示使用决策树方法排队的实验结果，而红线表示使用权值自学习打分排队法的实验结果。

4.3.3　实验结果及分析

实验一的结果表明，在 P@N 指标上，两层依存关系扩展好于一层依存关系扩展。在前 100 个候选意见目标中，能达到 90%以上的准确率。

（1）依存关系层数的扩展在总体上带来了准确率性能的提升。分析这个提升可知，一方面来自于候选意见目标抽取阶段，在这个阶段中，两层依存关系扩展的引入，帮助抽取到更长的意见目标，提高了抽取的召回率，同时减少了不完整短语的出现概率，也有利于准确率的提高。另一方面来自于候选意见目标排队阶段，这个阶段不会产生新的候选意见目标，但是由于两层依存关系扩展的引入带来了一维新的特征，它对排队起到了促进作用，于是也带来了 P@N 指标的提高。

（2）候选意见目标序列前端的性能波动，来源于 P@N 指标本身。由 P@N 的计算公式可知，N 作为除数，从 N=1 开始增加，在前端（N 较小）时比值是不稳定的，看不出统计特性，任意一个错例的出现都会大幅拉低准确率。当 N 逐渐增大后，统计特性开始发挥作用，也就看不到这种准确率的波动了。

（3）有理由相信，随着依存关系层数的增加，在严格控制扩展模式的条件下，候选意见目标的准确率会持续提高。从上面的分析可以看出，依存关系的增加可以帮助抽取更长的意见目标，同时减少了不完整短语的出现。在排队阶段亦可以增加特征，提高排队性能。但是，重要的前提是严格控制扩展模式。如果放松对扩展模式的控制，会导致在候选意见目标抽取阶段大量引入非意见目标，对候选意见目标排队阶段的区分性能造成压力。也就是说，从本质上讲，实验一的结果表明，引入严格的人工规则和人工控制，有助于意见目标的抽取。

实验二的结果表明，寻找有效的特征可以帮助候选意见目标排队性能的提升。可以看到，添加了指示词特征的实验在前 400 个候选意见目标中的准确率几乎处处优于没有指示词特征的实验。这证明了意见指示词特征对判断候选意见目标的可信度非常有效。同样的情况也发生在关于情感词特征的实验中。情感词特征也对候选属性的排队具有积极影响。可以相信，如果能找到更多对排队有帮助的统计特征，系统的性能会进一步得到提升。在现有的七维特征中，只有指示词特征和词频特征是纯

统计特征，其余如依存关系类型和词性标注是句法特征，情感词特征是搭配特征，词长和拉丁字符是字符串特征，它们都基于规则或者语义理解的。也就是说，非统计特征也是非常有效的。在统计方法中融入规则信息，可以提升系统性能。

实验三的结果表明，决策树排队法和权值自学习排队法的性能大致相当，权值自学习排队法的性能要低一些。决策树排队法，人为给定每一维向量的优先级，简单但是人工干预的程度过大，对于高维向量的情况是不实用的。而权值自学习排队法让数据说话，它反映了已知数据的一种倾向。尤其当统计特征向量维数很高的时候，权值自学习排队法的优势会很明显。但它也存在问题，如果用来学习的开发集与测试集数据不匹配，或者开发集的规模过小，都会导致习得的权值不能体现每一维特征的影响。

4.4 结论

本章介绍了一种以依存关系扩展为核心的意见目标抽取系统 OPINAX。它采用了先抽取后排队这样一种两段式的意见目标抽取算法。在每个阶段都着力挖掘了意见目标内部的句法信息，并利用这些信息来指示候选意见目标的可信性。在统计方法占主导的今天，此方法把统计规律和浅层句法分析结合起来，实验证明这种方法是可行的。并且取得了不错的效果。

本系统的任务与传统的意见目标抽取有一定的出入。它更关注抽取的准确率而不是召回率。本系统并没有采用二值分类器对抽取到的意见目标特征进行分类。但是本系统可以很容易的改造为二值分类方法，只要对可信度分数取一阈值，将超过阈值的候选词判定为意见目标即可。

第5章　基于意见目标网络的抽取方法

5.1 介绍

本章将介绍一种新的管理意见目标的结构化模式——意见目标网络，并提出一种基于泛化与繁殖的自举式意见目标抽取方法。意见目标网络打破了以往采用列表型或树型结构管理意见目标的窠臼，将意见目标以网络路径的形式进行管理。这与之前的管理模式有了根本的区别，它将意见目标经过拆分与组合之后再表现出来，更加关注意见目标内部的信息。意见目标网络具有列表型和树型结构所不具备的优势。基于泛化与繁殖的自举式意见目标抽取方法是本章的另一个重点，它秉持着与意见目标网络相同的思想——将意见目标打散从而挖掘更多内部信息。此方法有三个重点：泛化、繁殖、自举式抽取。所谓泛化，就是对意见目标进行拆分，提取出组成意见目标的元素；所谓繁殖，是利用泛化的结果，进行意见目标扩展，发掘更多意见目标；所谓自举式抽取，就是不断重复处理流程，不断提高抽取性能，并使抽取效果趋于收敛。

意见目标网络的构建和基于泛化与繁殖的自举式抽取流程是相辅相成的。在运行基于泛化与繁殖的自举式抽取流程的同时，会不断构建及更新意见目标网络；在构建和更新意见目标网络的同时，也会促进意见目标的抽取。从而使两部分有机结合起来。但是，这两部分之间并没有必然的联系。基于泛化与繁殖的自举式抽取流程也可以用于构建列表型或者树型的意见目标结构；而意见目标网络也可以通过其他的形式来构建。这就使两部分之间具有相对的独立性，从而更好地实现了模块化思想，有利于它们的后续研究与发展。

本章提出意见目标网络和基于泛化与繁殖的自举式意见目标抽取方法，是为了改进意见目标抽取方法研究中所存在的一些问题：（1）人工编撰的意见目标粒度太大，以他们作为种子必然导致覆盖率低的问题。同时，大部分人工编撰的意见目标包含多个词汇，因此无法在同义词集中找到，这导致基于同义词集的方法无法奏效。（2）种子和意见目标往

往都存放在一维列表或者数型机构中，很难有效表示二者关系。（3）基于种子的意见目标扩展方法无法一次取得满意的覆盖率，通常需要多次递增式学习才能取得良好效果。而本文所介绍的方法，经过多次循环，在意见目标网络构建完成的同时，也能最大限度地获得评价文本中的未知意见目标。中文意见目标抽取实验结果表明：本文方法在第八轮循环中比基线方法在 F1 分数上提高了 0.117，在召回率上提高了 0.239。

本章将按照如下方式组织：首先介绍意见目标网络的定义及如何构建意见目标网络等内容，之后详述基于泛化与繁殖的自举算法的基本框架和具体流程。接下来设计实验验证本章方法的有效性，给出实验结果并做分析。最后给出结论。

5.2 意见目标网络

5.2.1 介绍

意见目标网络（opinion target network, OTN）是一个有向图，其中图的节点（node）代表原子意见目标（atom opinion targets, AOT）同义词集，边（edge）揭示出 AOT 之间的关系，路径（path）则有效地表示了由 AOT 有序组合而成的意见目标（compound opinion targets, COT）。其次，OTN 又是一个双层有向图。根据本文第三章的论述，我们可以据此将 AOT 划分为两个部分——广义实体（generalized entity）和属性（attribute）。这样，必须采取双层图分别管理广义实体和属性。

5.2.2 基本思想

本文定义了原子意见目标（AOT）和复合意见目标（COT）。原子意见目标指内聚力强、外在搭配灵活的意见目标。换句话说，从 AOT 内部看，其词汇在统计上彼此依赖很强；而从 AOT 外部看，AOT 能够同很多 AOT 搭配形成真实的意见目标。复合意见目标是指评价文本中以不同模式将 AOT 组合而成的真实意见目标。例如，“图像亮度”是一个复合意见目标，“图像”和“亮度”是原子意见目标。本文工作中，原子意见目标进一步被划分为广义实体和属性两类，划分的依据是前文中对意见目标的定义。在复合意见目标后面，隐藏这意见目标本身。它必须

包含广义实体和属性两部分，无论这两部分是否出现在复合意见目标中。

本文提出的意见目标网络体现了四个设计意图：（1）该网络能够表示 AOT、COT 以及它们之间的关系；（2）该网络能区分广义实体和属性；（3）该网络能通过同义词集有效表示数万个种子；（4）该网络可便于以泛化和繁殖方式自动构建。

5.2.3　定义

“原子意见目标”和“复合意见目标”是本节的两个重要概念。在前文中已经提到过。为便于算法描述，我们先给出这两个术语的定义。并描述它们与意见目标之间的关系。

5.2.3.1 定义 1：原子意见目标

原子意见目标（Atom Opinion Target, AOT）是满足如下条件的意见目标：

1. 凝聚度强，即该意见目标的构成成分（字符或词）之间能形成较为稳定的组合。
2. 灵活度强，即该意见目标可与较多的原子意见目标组合形成意见目标。
3. 有明确的含义，即该意见目标可以独立存在，不能是没有意义的字符串。一般说来，AOT 可以是广义实体或者属性。

例如，“色彩”和“感光度”是满足条件的原子意见目标，而“图像的色彩”凝聚度低，不能被选择为原子意见目标。提出原子意见目标是对意见目标进行泛化的要求。显然，原子意见目标是比一般意见目标更好的“种子”。

5.2.3.2 定义 2：复合意见目标

复合意见目标（Compound Opinion Target, COT）是由原子意见目标复合而成的意见目标。

例如，“镜头对焦能力”由“镜头”、“对焦”和“能力”三个原子意见目标组成。“图像的色彩”由“图像”和“色彩”两个原子意见目标组成。

复合意见目标是抽取任务中实际得到的意见目标，它们是原子意见目标复合的结果。

5.2.3.3 定义 3：意见目标

“意见目标”是一个概念，它指意见的持有者所针对的对象。从逻辑上说，它包括广义实体和属性两部分。现实文本中的意见目标，都是它的一个实例化结果。无论是复合意见目标还是原子意见目标，是“意见目标”这个概念下的特例。

5.2.4　形式化表示

意见目标网络（OTN）是一个双层有向图 G^{OTN}，定义为如下五元组：

$$G^{OTN} = < V^{COM}, E^{COM}; V^{ATT}, E^{ATT}; E^{\Theta} > \quad (5\text{-}1)$$

其中 V^{GET} 和 V^{ATT} 分别表示广义实体节点和属性节点；E^{GET} 和 E^{ATT} 分别表示广义实体边和属性边；E^{θ} 则表示广义实体–属性交叉边。在 OTN 中，路径通常包含了多个边，实际上表示了多个原子意见目标以特定模式有序组合而成的复合意见目标。需要指出：OTN 中的节点都是 AOT 所对应的同义词集，因此一个节点实际上代表了一组 AOT。图 5.1 给出了 OTN 示例。其中，属性层的虚线边代表了属性分类关系，并不实际用于意见目标抽取。

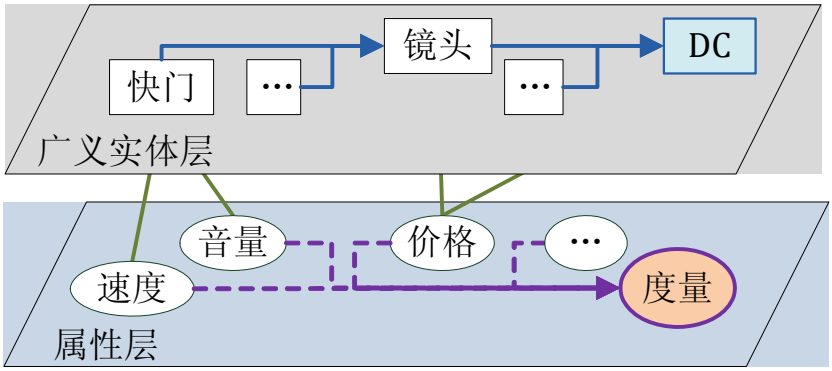

图5.1　一个包含广义实体层和属性层的意见目标网络示例

观察图 5.1 发现：（1）属性和广义实体必须结合在一起才能形成有

效意见目标（尽管在真实评价文本中，有时它们可以缺省）。意见目标的核心是属性，它们显式或隐式地与意见关键词搭配组成意见单元。这一发现让我们了解了意见的形成方法。（2）OTN 中的同义词集和模式能揭示出概念和语义关系，例如整体部分关系。因此我们预测，OTN 可能对自动构建领域本体有帮助。

5.3 基于泛化与繁殖的自举式抽取

本部分介绍的一种自举式的意见目标抽取算法，本方法可用于构建意见目标网络，经过多次循环，在意见目标网络构建完成的同时，也能最大限度地获得评价文本中的未知意见目标。使用中文文本进行意见目标抽取实验，结果表明：本文方法在最后一轮抽取结束时，F1 分数比基线方法提高了 0.117，召回率提高了 0.239。改进效果明显。

5.3.1　算法框架

本文方法将意见目标抽取任务和意见目标网络构建任务通过泛化和繁殖多轮自举的方式自动完成，其工作流程如图 5.2 所示。

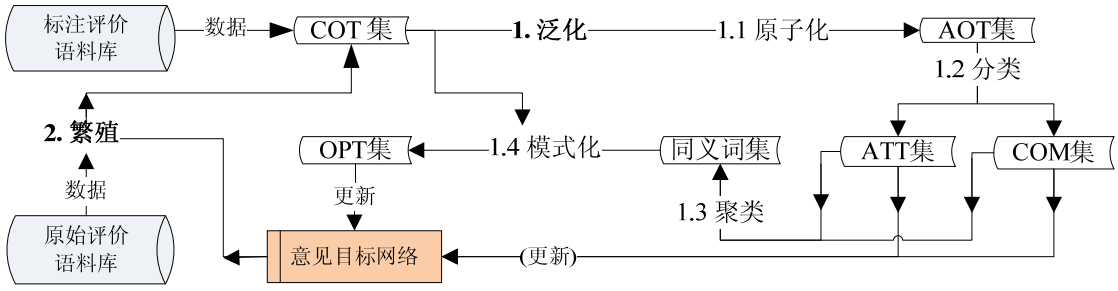

图5.2　工作流程如图

图 5.2 显示：标注评价语料库提供最初的 COT 集合。泛化过程从 COT 集中提取 AOT 集，并将 AOT 划分为广义实体和属性，同时为每个 AOT 赋予特定的同义词集标签，经过意见目标模式化过程形成意见目标模式，最终构建意见目标网络。繁殖过程借助意见目标网络和依存关系发现未知意见目标。自举算法则执行泛化和繁殖过程，通过多轮学习获

得越来越多的意见目标，以及不断完善的意见目标网络，从而实现高覆盖率的未知意见目标的抽取。

标注评价语料库为本文方法提供了初始复合意见目标集，即人工标注的意见目标。本文方法从这些初始复合意见目标集出发，在泛化、繁殖和自举过程中实现了意见目标抽取。

在泛化过程中，我们先对复合意见目标进行词法分析和依存分析，接下来根据凝聚度和灵活度提炼原子意见目标，结合词性和依存关系提炼意见目标模式。泛化处理完成后，新获得的原子意见目标将作为新词更新到语言分析器，以确保这些原子意见目标不会被错误切开。

在繁殖过程中，我们先基于原子意见目标和意见目标模式进行概率推理，再以依存关系进行句法推理，从大规模原始评价语料中抽取有效的意见目标，以充实复合意见目标集。

自举学习过程是串联了泛化和繁殖处理的多轮回递增学习过程。每次泛化处理完成后都会得到新的原子意见目标集合，以便于繁殖过程从原始评价语料中抽取复合意见目标。这样，经过几轮处理后，原子意见目标集和意见目标模式集的规模将趋于稳定，自举过程终止。此时，复合意见目标集也同时建立起来，系统至此完成了意见目标抽取任务。

本文在以下章节介绍上述过程的算法细节。

5.3.2　泛化过程

5.3.2.1 原子化

泛化算法获取原子意见目标的依据是意见目标的凝聚度和灵活度。对应这两个指标，原子意见目标获取过程应在词法分析和依存分析后分为两步完成（注：词法分析后，复合意见目标可能被拆分为若干子串）：

第一、紧密度计算。对构成复合意见目标的两邻近子串进行递归两两组合，计算某更长子串的紧密度，并接纳紧密度较高的长子串为原子意见目标。同时，对现有子串进行检验，如果不满足紧密度条件，则进行拆分，形成长度更短的原子意见目标。两邻近子串紧密度通过计算点式互信息（pointwise mutual information）获得。公式如下：

$$PMI(W_1,W_2)=\ln\frac{P(W_1,W_2)}{P(W_1)P(W_2)} \quad (5\text{-}2)$$

其中 W_1 和 W_2 是相邻的两个子串。P(W)表示字符串 W 在语料库中出现的概率。

第二、灵活度计算。紧密度计算之后，再对复合意见目标的子串进行灵活度计算，并接纳灵活度较高的子串为原子意见目标。灵活度计算公式如下：

$$F(W)=\frac{1}{2}\left(\frac{\sum_{W_i\in N^L}\frac{1}{N^R(W_i)}}{N^L(W)}+\frac{\sum_{W_i\in N^R}\frac{1}{N^L(W_i)}}{N^R(W)}\right) \quad (5\text{-}3)$$

其中，N^L 代表意见目标集合中与子串 W 左邻的子串集合，函数 $N^L(W)$ 表示与子串 W 左邻的子串的种类数；N^R 代表意见目标集合中与子串 W 右邻的子串集合，函数 $N^R(W)$ 表示与子串 X 右邻子串的种类数。公式 5-3 表明，一个子串的灵活度依赖于其左邻子串的右灵活度和右邻子串的左灵活度。

我们设定阈值，选取紧密度和灵活度满足条件的词汇作为原子意见目标。

5.3.2.2 分类

本文设计了一个基于概率的分类器，用于识别广义实体和属性。该分类器考虑如下两类特征：

1）平均编辑距离（d^{AVG}）

平均编辑距离以字符串编辑距离公式度量将其分类为广义实体或者属性的概率：

$$d^{AVG}(t\mid X)=\frac{1}{|X|}\sum_{x_i\in X}d(t,x_i) \quad (5\text{-}4)$$

其中，t 表示被分类 AOT，$X=\{x_i\}$代表已知广义实体 AOT 集合或属性 AOT 集合，|X|表示集合所包含的元素个数，$d(t, x_i)$是 t 和 x_i 编辑距离度量函数。根据公式（3），我们可分别度量 t 被分类为广义实体（C）或

属性（A）的概率，并取概率较大者为预测结果。

2）综合位置倾向值（t^{OVA}）

综合位置倾向值利用位置启发信息度量某 AOT 是广义实体或属性的概率。在某些语言中，位置信息对 AOT 分类具有决定意义。综合位置倾向值计算方法如下：

$$t^{OVA}(t)=\frac{count(t,A)}{count(C,t)} \tag{5-5}$$

其中 count(t, A)是该 AOT t 出现在属性词前的次数，count(C,t)是该 AOT t 出现在属性词后的次数。

需要指出：初始广义实体集合和属性集合从标注评价语料库中获得。为提高覆盖率，我们从 HowNet[31]中抽取更多属性。

5.3.2.3 聚类

为了给新发现的 AOT 赋予同义词集标签，我们采取 K-Means 聚类方法将所有 AOT 聚类到适当数目的类簇中。由于 K-Means 聚类方法可通过调节参数获得不同数目的类簇，因此可通过调节参数获得满足如下两个条件的类簇：（1）该类簇包含 3 个以上 AOT，且这些 AOT 同属一个同义词集。（2）该类簇包含至少一个新 AOT。一旦我们找到了这样的类簇，我们将已知 AOT 的同义词集标签赋予所有新 AOT。我们重复上述聚类和赋值过程，直到所有新 AOT 都被赋予了同义词集标签。聚类处理中我们采取了两类特征：（1）原始评价语料库中 AOT 的临近词；（2）新 AOT 和已知 AOT 的编辑距离。

通过上述聚类处理可能无法给所有新 AOT 赋予同义词集标签。这时，我们将所有未获同义词集标签的新 AOT 进行再聚类处理，从而试图获取新的同义词集。如果形成满足如下条件的类簇，则认为发现了新的同义词集：（1）该类簇包含 3 个以上新 AOT。（2）所有 AOT 在原始评价语料库中出现次数都超过 3 次。

在发现了新的同义词集后，我们需要给新发现的同义词集赋予一个标签。我们采取在原始评价语料库中出现次数最多的 AOT 作为标签。上述处理后，仍然会有一些新 AOT 无法获得同义词集标签。我们将这

些 AOT 暂时搁置，期望在下一轮聚类处理中参与发现新的同义词集。

5.3.2.4 模式化

意见目标模式刻画意见目标的构成规律，包含对原子意见目标、顺序和词性的描述。其形式化定义如下正则表达式：

{A}*{string{B}*}*,

其中，A_c 和 B_c 代表 AOT 的同义词集标签，string 代表模式中的字符串常量。举例来说，“<图像>的<颜色>”通过字符串“的”组合了“图像”和“颜色”两个同义词集标签。由于 COT 除了包含 AOT，还包含一些非 AOT 字符，我们用 string 常量表示他们。

在具体使用时，还会演化出三种具体形式：

{A_c_}*{string_{ B_c}*}*　　## 原子意见目标

{A_v_}*{string_ {B_v}*}*　　## 顺序

{A_{pos}_}*{string_ {B_{pos}}*}*　## 词性

其中 string 仍然代表模式字符串，而

A_c、B_c 表示原子意见目标常量（constant），是具体出现在文本中的字符串，这个层次上的模型，可以统计出字符串之间的统计关系，类似于 N-gram。

A_v、B_v 表示原子意见目标变量（variable），是将文本中的字符串进行抽象，它们可以代表文本中所有的字符串。这个层面上的模型，主要用来统计词汇顺序的统计模型。

A_{pos}、B_{pos} 表示原子意见目标的词性标签（part-of-speech tag），是本位置上词汇的词性特征。使用这个层次上的模型，为了统计原子意见目标在词性顺序上的统计规律。

于是，上面例子的演化模式表示如下：

图像_的_颜色

X_的_Y

n_的_n

这里，“图像的颜色”是语料中出现的具体文本；“X 的 Y”不会出现在语料中，但它表示了一种原子意见目标的结合方式，它是对语料中众多复合意见目标的一种抽象；同理，“n 的 n”也是对语料中众多复合

意见目标的一种抽象，只不过这种抽象是建立在词性的层面上的。

插入由于拥有一定规模的意见目标模式集，我们可基于统计信息计算出原子意见目标在模式中同时出现且满足特定顺序的概率。例如，我们以最大似然估计方法计算出 A_c、B_c 满足模式 p 且 A_c 出现在 B_c 之前的概率：

$$P(A_c > B_c \mid p) = \frac{C(A_c > B_c \mid p)}{C(A_c > B_c \mid p) + C(A_c < B_c \mid p)} \tag{5-6}$$

在我们获得了所有同时包含 A_c、B_c 的模式后，可计算宏观上 A_c 出现在 B_c 之前的概率：

$$P(A_c > B_c) = \sum_{p_i} P(A_c > B_c \mid p_i) P(p_i) \tag{5-7}$$

其中 p_i 是同时包含 A_c、B_c 的模式。$P(p_i)$是意见目标模式概率模型，可通过最大似然估计方法从复合意见目标集合中获得。同理，我们可计算某两个词性 A_{pos} 和 B_{pos} 在特定模式 p 下按某一顺序出现的概率以及宏观上 A_{pos} 出现在 B_{pos} 之前的概率：

$$\begin{aligned} & P(A_{pos} > B_{pos} \mid p) \\ & = \frac{C(A_{pos} > B_{pos} \mid p)}{C(A_{pos} > B_{pos} \mid p) + C(A_{pos} < B_{pos} \mid p)} \end{aligned} \tag{5-8}$$

$$P(A_{pos} > B_{pos}) = \sum_{p_i} P(A_{pos} > B_{pos} \mid p_i) P(p_i) \tag{5-9}$$

意见目标模式是对大量 COT 的提炼。它们用模式的方式表示出，原子意见目标是如何拼接成复合意见目标的。掌握了意见目标模式，就可以根据规则，利用原子意见目标创造出新的复合意见目标，而这正是意见目标繁殖的一个重要途径。

需要指出的是：在利用意见目标模式进行意见目标繁殖时，要求至少一个子串为原子意见目标，这样才能不断发现未知的原子意见目标。

5.3.2.5 意见目标网络形成

我们以 AOT 所对应的同义词集为结点，以意见目标模式画边，就

建立起一个意见目标网络。也就是说，意见目标网络中的结点集合，就是已知的广义实体和属性的集合，而意见目标网络中的路径集合，就是所有已知的复合意见目标的集合。通过意见目标网络可以很容易的重构出列表型和树型存储结构。注意：如果一个新发现的 AOT 被赋予一个已有的同义词集标签，这时不会在意见目标网络上建立新的结点，而是将这些 AOT 加入该同义词集。

5.3.3　繁殖过程

繁殖算法以原子意见目标集和意见目标模式集为输入，运行于原始评价语料库，完成后输出复合意见目标集。繁殖算法主要通过两个途径发现候选意见目标：第一个途径是以原子意见目标集和意见目标模式集进行概率推理；第二个途径是借助依存关系进行句法推理。

5.3.3.1 基于 OTN 的繁殖

基于 OTN 的繁殖，就是基于模式的概率推理。以原子意见目标集和意见目标模式集进行概率推理的基本原理是：若我们以原子意见目标填充所有意见目标模式，可繁殖出大量不出现在现有复合意见目标集中的意见目标。例如可将所有原子意见目标填充到模式“X_的_Y| n_的_n”中，从而繁殖出类似如下形式的候选意见目标：

图像_的_分辨率

图像_的_音质（*）

……

镜头_的_尺寸

镜头_的_价格

….

当然，上述例子中难免包含不合实际的伪意见目标，例如“图像_的_音质”。所谓概率推理，就是为上述繁殖过程赋予特定置信度。具体做法是：我们在意见目标模式概率模型 $P(p_i)$的指导下取最长匹配，并滤除词汇顺序置信度低于阈值的候选意见目标。词汇顺序置信度计算公式如下：

$$SC(c)=\sum_{S_i>S_j}\frac{P(S_i>S_j)}{C^2_{N(c)}}=\sum_{S_i>S_j}\frac{2P(S_i>S_j)}{N(c)(N(c)-1)} \quad (5\text{-}10)$$

$$(N(c)>1) \quad (5\text{-}11)$$

其中，c 是意见目标，N(c)表示 c 中原子意见目标的个数，S_i、S_j 表示 c 中出现的原子意见目标。

按照词汇顺序原则进行候选意见目标过滤的依据是：原子意见目标在构成意见目标时遵循严格顺序。例如“光学取景器”是一个意见目标，而“取景器光学”则不是有效意见目标。

同时，高于词汇顺序置信度阈值的意见目标可继续遵循特定模式进行扩充，直至新获得的意见目标不再满足词汇顺序置信度约束为止。需要指出的是，采取原子意见目标进行模式匹配可能面临数据稀疏问题。此时我们采取回退策略，退而以词性顺序置信度进行校验。

意见目标网络具备利用 AOT 和模式进行意见目标推理的能力。换句话说，在意见目标网络中，如果同义词集 A 和 B 之间存在某两个 AOT 所形成的边，那么 A 中的所有 AOT 都有可能和 B 中所有的 AOT 建立同样关系。基于这个假设，我们可借助意见目标网络推理产生候选意见目标。自动推理会导致错误候选，因此必须设计过滤机制排除错误。我们采取序列可信度过滤方法，从原始评价语料库中估计某 AOT 出现在另一个 AOT 之前的概率。给定一个候选意见目标 X，它包含 N 个 AOT，即 $X=\{A_i\}i=1,\ldots,N$，序列可信度计算公式如下：

$$SC(X)=\frac{1}{C^2_N}\sum_{i<j}count(A_i,A_j) \quad (5\text{-}12)$$

其中 $count(A_i, A_j)$表示语料库中 A_i 出现在 A_j 之前的次数。我们设定经验阈值过滤候选意见目标。

5.3.3.2 基于依存关系的繁殖

基于依存关系的繁殖就是基于依存关系的句法推理。基于模式不能发现新同义词集，制约了意见目标抽取覆盖率。于是添加基于依存关系的句法推理，以发现新的原子意见目标概念。为发现新同义词集，我们考察与已知 AOT 发生依存关系的词汇。以依存关系为依据进行句法推

理的基本原理是：观察与意见目标 O 发生依存关系的非原子意见目标词 W，以如下四条原则判定是否将 W 与 O 合并成意见目标：

1. 依存关系为以下四种之一：ATT（修饰关系）、COO（并列关系）、QUN（数量关系）和 DE（“的”字结构）。
2. 除并列连词和“的”字以外，W 与 O 应至少在一个评价语句中邻接。
3. W 不能位于谓语语法块。
4. W 不能为形容词、代词。

上述原则的确定基于如下考虑：

首先，我们选取 ATT、COO、QUN 和 DE 四类依存关系是因为他们能准确地定位意见目标。修饰关系能反映原子意见目标间的上下位关系和限定关系，因而可根据这种关系发现未知意见目标。COO 代表并列关系，多以并列连词引导，对发现未知意见目标意义显著。“的”字结构与 ATT 的功能类似，QUN 关系则容易发现数字限定的意见目标。

其次，限定 W 与 O 之邻接关系的主要目的是便于二者的合并。若二者之间存在距离，则容易在合并中引入不必要的噪声，大大降低合并后意见单元的正确性。

最后，对语法块的限定也是基于语言学和统计信息的综合考虑。句法块分析器能获得句子的主语块、谓语块、宾语块和状语块。我们对大规模评价文本进行了考察，发现谓语块几乎不包含意见目标。由于某些状语块是包含表示途径的状语，因而不能武断滤除。

为保证繁殖的准确性，我们只考虑一层依存关系的扩展。但由于我们采取自举学习方法，因此能够发现起初与意见目标发生多层依存关系的原子意见目标。

实验表明，依存关系对发现新同义词集很有帮助。

5.3.4　自举算法

自举算法是一种调度泛化算法和繁殖算法的递增学习机制。第一轮处理从标注语料库中提取初始复合意见目标集。经过泛化和繁殖处理后，算法可形成原子意见目标集和意见目标模式集，进而从原始语料库抽取到新的复合意见目标。

第二轮及其之后的各轮自举处理与标注评价语料库脱离关系，而以上一轮构建的复合意见目标集作为泛化起点，产生更加完备的原子意见目标集和意见目标模式集，进而从原始语料库抽取到更多新的复合意见目标。

自举算法的终止条件是复合意见目标的规模趋于稳定，即不再能够发现新的复合意见目标。通常情况下，若参数设置恰当，经过有限次自举过程即可达到这一目标。

5.4 实验

5.4.1 实验设置

我们在实验中采用了两个语料库。一个是 Opinmine 语料库[28]，它包含 8,990 个人工标注的关于数码相机的意见。另一个是原始评价语料库，包含了 6,000 篇来自相同领域的用户评价。为对本文方法进行评测，在实验中我们将 OPINMINE 标注语料库随机平均划分为两部分，一部分作为训练集用于初始复合意见目标获取，另一部分作为测试集用于意见目标抽取算法评测。

本文方法以哈尔滨工业大学提供的 LTP[30]实现中文分词和依存分析，同时我们从中文 HowNet[31]中手工提取初始属性同义词集。

5.4.2 实验指标

为评测最终获取的复合意见目标是否有效，我们用他们对测试集进行自动标注。我们将自动标注结果与人工标注结果进行对比，从而计算出意见目标标注的准确率（p）、召回率（r）和 F1 分数（f）。召回率可按常规进行，定义为自动标注正确个数占人工标注总数的比例。准确率的定义略有不同，这是因为评价文本中某些语句可能只涉及意见目标但不涉及意见，但目前的标注方法尚未分析是否涉及意见，因此会对所有意见目标进行标注。若以常规方法统计，准确率会失去评测意义。为此我们限定：只统计与人工标注意见目标发生字符重叠的自动标注结果。这样准确率定义为自动标注正确个数占自动标注与人工标注发生字符重叠的标注总数的比例。F1 分数计算亦按照常规公式计算。

需要特别指出的是：在单个评判自动标注结果是否正确时，我们采取了宽松评测方法，只要求二者相互包含即可，不要求完全匹配。这是因为人工标注有时只标注最重要的原子意见目标，而我们的方法则总是标注完整复合意见目标，因此不能采取严格评测方法。

5.4.3　实验设计

本实验的目的是对比本文方法和现有方法的性能。为此，我们采取效果较好的文献[32]方法作为基线（baseline）方法。该方法视标注语料库中的人工标注意见目标为“种子”，依据依存关系从原始评价语料中抽取意见目标。

基线方法：直接以人工编撰的意见目标为种子进行未知意见目标的抽取。为了获取未知意见目标，我们在基线方法中使用了依存分析和意见目标模式。本文方法则以原子意见目标为种子，基于意见目标网络进行多轮自举过程实现未知意见目标抽取。设置基线方法的目的，是要同本文方法进行对比，证明意见目标网络在未知意见目标抽取中的贡献。

在本实验中，我们采用了哈尔滨工业大学的依存分析器[30]和清华大学语法块分析器[33]。原子意见目标获取中的凝聚度和灵活度阈值、繁殖算法中词序置信度阈值都从实际观察中获得。

5.4.4　实验结果

5.4.4.1 原子意见目标的扩展

随着自举式算法的进行，基于依存关系的繁殖方式保证了原子意见目标的扩展，而原子意见目标的扩展必然带来复合意见目标的扩展。考察原子意见目标的扩展情况，有助于评价系统的性能。图 5.3 给出了在 N 轮自举式抽取过程中，每一轮过后意见目标的增长情况：

其中，横坐标表示自举式抽取的轮次，纵坐标表示词条个数。蓝色线代表所有原子意见目标（AOT）的数目，而红色线与绿色线分别代表对 AOT 进行分类后所得的广义实体（GET）和属性（ATT）的数目。

可以看出，广义实体数目的增长，远远超过属性数目的增长。这个很符合客观规律，因为世界是不断发展变化的，新的事物不断涌现，所以实体的数目会名目繁多，且不断扩大。而作为人类对世界的认识——

属性的种类并没有那么多。因为人类的区分能力有限，而且不会随着时间的推移发生重大变化。

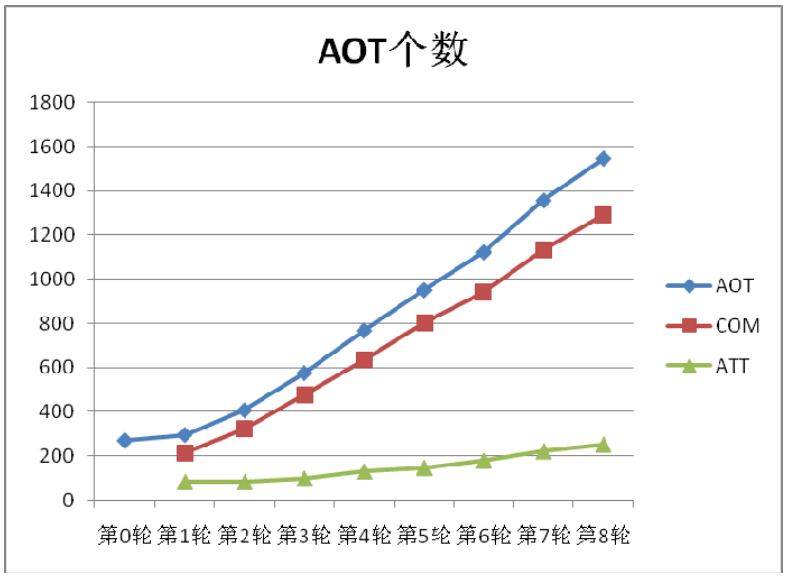

图5.3　AOT 规模增长图

在考查原子意见目标规模增长速度的同时，还对它的正确率做了考察。图 5.4 给出了在 N 轮自举式抽取过程中，每一轮过后意见目标的正确率情况：

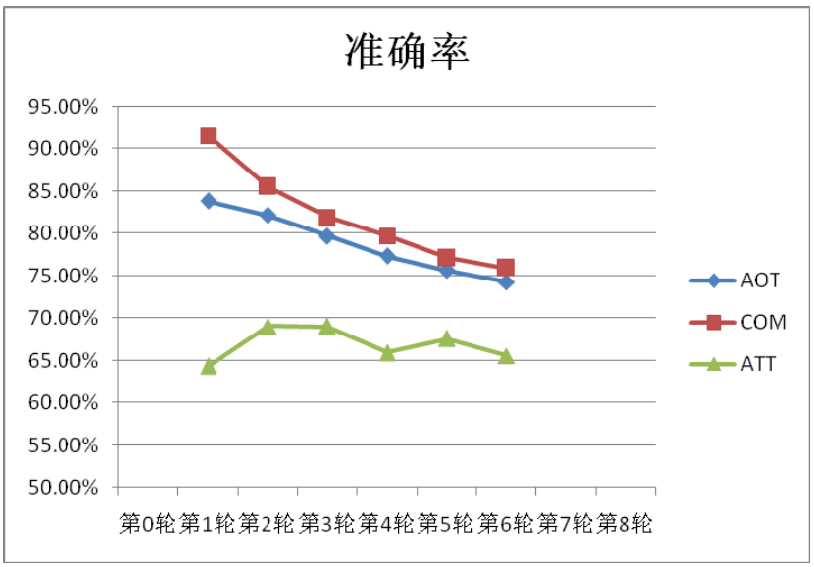

图5.4　AOT 正确率图

在图中，横坐标表示自举式抽取的轮次，纵坐标表示词条的正确率。蓝色线代表所有原子意见目标（AOT），而红色线与绿色线分别代表广义实体（GET）和属性（ATT）。

可以发现，随着自举式抽取的不断进行，系统中的噪声在不断放大。AOT 规模在急剧增长的同时，正确率也在不断下降。但所幸正确率下降的幅度并没有规模增长的那么大。从开始到最后一轮结束，AOT 的规模变为原来的 6 倍，而正确率从 90%滑落到 70%。从图中可以看出，因为 GET 的规模比 ATT 小了很多，所以它对 AOT 的正确率影响很小。而 GET 的整体正确率偏低。这很大程度上来自于分类算法的性能欠佳。

5.4.4.2 复合意见目标的扩展

复合意见目标是文本中真实出现的意见目标字符串，也是在意见挖掘系统中直接会使用到的数据。因此，COT 的抽取性能对本系统来说至关重要。在本节中，主要考察复合意见目标规模的增长情况。图 5.5 给出了在 N 轮自举式抽取过程中，每一轮过后意见目标的增长情况：

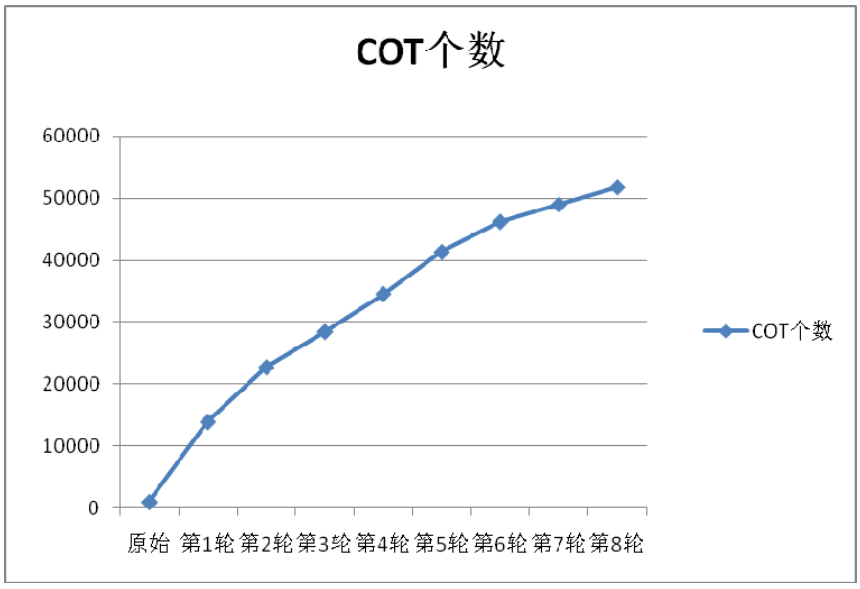

图5.5 COT 规模增长图

在图中，横坐标表示自举式抽取的轮次，纵坐标表示复合意见目标 COT 词条的数目。从图中可以看出，COT 在每轮结束后都会大幅度增加，但是增长幅度却在逐渐减小。从上面 AOT 的增长趋势图我们可以推测

出，COT 一定是不断增长的。因为 COT 由 AOT 组成，如果 AOT 呈现增长的态势，那么 COT 也会增长。而 COT 的增速没有 AOT 的增速快，这证明有大部分 AOT 在 COT 中出现的频率很低。对于 COT 的抽取性能，会在后面论述。

5.4.4.3 意见目标模式的扩展

随着 AOT 数目和 COT 数目的增加，意见目标模式的规模也会逐渐扩大。图 5.6 给出了在 N 轮自举式抽取过程中，每一轮过后意见目标模式的增长情况：

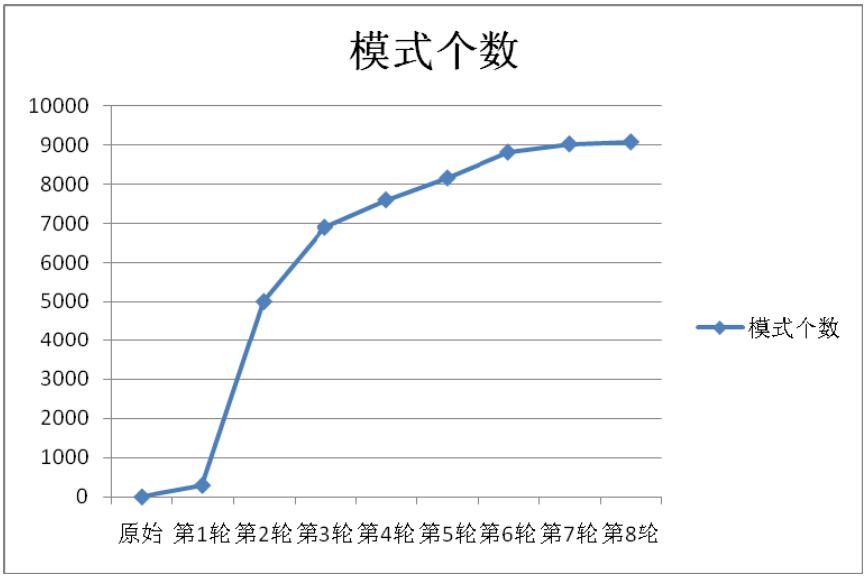

图5.6 意见目标模式增长变化图

在图中，横坐标表示自举式抽取的轮次，纵坐标表示意见目标模式的数目。从图中可以看出，意见目标模式随着自举式抽取轮次的增加而增加，但是呈现先急剧增长，后缓慢增长的态势。这说明，意见目标模式并不是无穷无尽的，它存在一个合理的规模。这个规模足以维持丰富多彩的复合意见目标，而又不至于让它们陷入泛滥。也证明，众多 COT 是分享意见目标模式的，所以随着 COT 的高速增长，意见目标模式并没有高速增长。根据图中的曲线可以推测，意见目标模式会趋向于收敛在某固定值。

5.4.4.4 意见目标抽取性能

由于 COT 的规模庞大，本实验没有直接计算抽取出的 COT 的准确率、召回率。因为这将耗费大量的人工。本文采取了间接的方法测试意见目标抽取的性能。借助于意见挖掘技术，在已标注的意见语料库中匹配意见目标。对每轮结束后的意见目标集合运行此过程，根据这个过程中的匹配性能，间接考察意见目标集合的准确率和召回率。可以分析出，此方法虽然不能全面反映出意见目标集合的完整情况，但是是一个很好的估算方法。图 5.7 给出了在 N 轮自举式抽取过程中，每一轮过后意见目标匹配的情况：

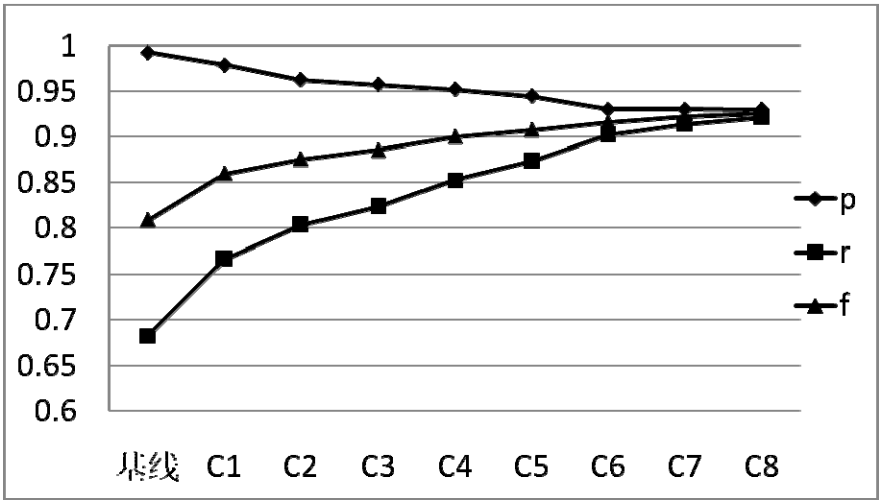

图5.7 意见目标抽取性能图

在图中，横坐标表示自举式抽取的轮次，纵坐标表示意见目标匹配的准确率、召回率和 F1 分数。从图中可以看出，在第一轮处理后，本文方法在 F1 分数上超出基线方法 0.051。这时，召回率提高了 0.085，准确率损失了 0.014。可见，本文方法在第一轮处理中就以较小准确率代价取得召回率的较大提高。这表明：原子意见目标的确是更好的种子，基于意见目标网络的抽取方法具有较大潜力。第八轮处理后，本文方法在 F1 分数上超出基线方法 0.117。这时，召回率提高了 0.239，准确率损失了 0.063。这说明自举过程对意见目标抽取具有的重要贡献。同时我们发现，从第六轮开始，F1 分数开始收敛，在第八轮基本趋于稳定。

这说明自举方法在提升性能上是可收敛。即，总可以通过有限轮自举完成较好性能的意见目标抽取。

但同时也要发现，随着三集合规模不断扩大，系统性能提升的加速度下降，在第四轮只产生了 0.016 的提升。这说明：大量候选复合意见目标给意见目标确认处理带了较大压力。令方法性能接近了极限。因此，进一步研究确认方法，可进一步提升本文方法的性能。

5.5 总结

未知意见目标是影响意见挖掘系统覆盖率的重要因素。现有意见目标抽取方法大多直接将人工标注的意见目标作为种子，通过语法/统计模板从真实评价文本中抽取未知意见目标。存在三个问题：（1）手工标注的意见目标粒度过大，不适合作为种子；（2）以列表作为管理种子的数据结构难以表达种子之间的关系；（3）一次意见目标挖掘往往难以取得满意的效果。

针对现有意见目标抽取方法存在的灵活性弱、扩展性差等问题，本文提出了一种基于泛化、繁殖和自举的意见目标抽取方法。实验结果显示，本文方法在自举过程的第一轮就在 F1 分数上超出基线方法 0.051。自举过程完成后，本文方法在 F1 分数上提高了 0.117。这说明，泛化处理对意见目标充分繁殖和抽取意义重大，自举过程则有助于泛化能力和繁殖能力的充分发挥。

第6章　总结与工作展望

意见目标抽取任务，在意见挖掘领域占据重要地位。尤其在产品意见挖掘等关注意见目标的任务下，意见目标抽取就显得十分关键。对意见目标抽取的研究成为近年来的研究热点之一。但是意见目标抽取领域的一些难题，诸如隐式意见目标抽取、同义词归并等，严重影响意见目标抽取的性能。同时，对意见目标定义的不明确，也阻碍了研究的深入。

本文主要分为两部分内容：一是针对目前对意见目标定义不明确的情况，提出了“意见目标=广义实体+属性”的定义公式，从而将意见目标划分为两个层面，一种是意义层面上的，一种是字符层面上的。当下，意见目标抽取任务的主要对象，还是字符层面的，这也是导致意见目标抽取任务不能满足用户需求的主要原因之一。二是针对现有意见目标抽取方法存在的灵活性弱、扩展性差等问题，本文提出并实现了意见目标网络及一种基于泛化、繁殖和自举的意见目标抽取方法，以提高意见目标抽取的覆盖率。意见目标网络是一个双层有向图，它以原子意见目标（广义实体和属性）同义词集为节点，通过意见目标模式实现了对复合意见目标的表示。意见目标网络的构建过程恰恰是未知意见目标抽取过程，经过泛化和繁殖的多轮自举处理，显著提高了意见目标抽取覆盖率。本文在中文评价文本上进行了实验，结果表明：意见目标网络对发现未知意见目标具有很大潜力。泛化处理对意见目标充分繁殖和抽取意义重大，自举过程则有助于泛化能力和繁殖能力的充分发挥。

我们在意见目标抽取方法研究上的未来工作包括：（1）将意见目标抽取升级为意见目标挖掘，增加更多的语义理解成分。（2）尝试新方法构建限定领域的意见目标网络。（3）在意见目标确认处理等环节显著提高方法的执行效率。（4）用更多领域的语料对本文方法进行评测。

参考文献

[1] 柏拉图. 理想国. 北京：商务印书馆出版社，1986. 56~60

[2] 李保利,陈玉忠,俞士汶,等. 信息抽取研究综述. 计算机工程与应用，2003，10：1~7

[3] 姚天防，程希文，徐飞玉，等. 文本意见挖掘综述. 中文信息学报，2008， 22（3）：71~80

[4] Hu M and Liu B. Mining and summarizing customer reviews. In Proc. of KDD’04. 2004, 168-177.

[5] Kim S M and Hovy E. Determining the Sentiment of Opinions. In Proceeding of COLING-04, the Conference on Computational Linguistics (COLING-2004). 2004, 1367-1373.

[6] Kobayashi. Opinion Mining from Web documents : Extraction and Structurization. Doctor thesis. 2007

[7] 张秦龙. 基于多特征的中文多词术语提取技术研究：[硕士学位论文]. 北京：北京大学计算机系， 2007

[8] Kageura K and Umino B. Methods of automatic term recognition: A review. Terminology. 1996, 3(2): 259-289.

[9] Church K W and Hanks P. Word association norms, mutual information and lexicography. Computational Linguistics. 1990, 16 (1): 22-29.

[10] Dunning T. Accurate methods for the statistics of surprise and coincidence. Computational Linguistics. 1993, 19 (1).

[11] Béatrice D. Study and Implementation of Combined Techniques for Automatic Extraction of Terminology. In The Balancing Act: Combining Symbolic and Statistical Approaches to Language. 1994.

[12] Patry A and Langlais P. Corpus-based terminology extraction. In 7th International Conference on Terminology and Knowledge Engineering. 2005, 313-321.

[13] Manning C and Schütze H. Foundations of Statistical Natural Language Processing. 1999.

[14] 罗盛芬，等. 基于字串内部结合紧密度的汉语自动抽词实验研究. 中文信息学报，2003

[15] Uchimoto K, Sekine S, Murata M, Ozaku H, and Isa-hara H. Term recognition

by using different field corpora. In Proceedings of the First NTCIR Work-shop on Research in Japanese Text Retrieval and Term Recognition. 1999, 443-450..

[16] Chen Y, Lu Q, Li W, Sui Z, and Ji L. A Study on Terminology Extraction Based on Classified Corpora. the 5th International Conference on Language Resources and Evaluation (LERC2006). 2006, 22-28.

[17] Su K Y, Wu M, and Chang J. A corpus-based approach to automatic compound extraction. Proceedings of the 32nd annual meeting on Association for Computational Linguistics. 1994, 242-247.

[18] Ji L, Sum M, Lu Q, Li W, and Chen Y. Chinese Terminology Extraction Using Window-Based Contextual Information. In Proceedings of the Eighth International Conference on Intelligent Text Processing and Computational Linguistics (CICLing 2007).2007, 62-74.

[19] Yi J and Niblack W. Sentiment Mining in WebFountain. In Proc. of ICDE-2005. 2005, 1073-1083.

[20] Hu M and Liu B. Opinion Extraction and Summarization on the Web. In Proceedings of 21st National Conference on Artificial Intellgience (AAAI-2006, Nectar paper). 2006.

[21] Hu M and Liu B. Mining and summarizing customer reviews. In KDD’04. 2004, 168-177.

[22] Popescu A and Etzioni O. Extracting product features and opinions from reviews. HLT-EMNLP’05. 2005, 339-346.

[23] Liu et al. Mining Topic-Specific Concepts and Definitions on the Web. In Proceedings of the 12th international World Wide Web conference (WWW-2003). 2003.

[24] Liu B, Hu M, and Cheng J. Opinion Observer: Analyzing and Comparing Opinions on the Web. In Proceedings of the 14th international World Wide Web conference (WWW-2005). 2005.

[25] Carenini G, Ng R, and Zwart E. Extracting Knowledge from Evaluative Text. K-CAP'05. 2005.

[26] Wilson, Wiebe, and Riloff. OpinionFinder : A System for Subjectivity Analysis. HLT-EMNLP-2005. 2005.

[27] Gamon M, Aue A, Corston S, et al. Pulse: Mining customer opinions from free text. In Proceedings of IDA. 2005, 121-132.

[28] Xu R, Xia Y, and Wong K F. Opinion Annotation in On-line Chinese Product Reviews. In Proc. of LREC-2008. 2008.

[29] Zhang Z, Yu H, Xiong D, and Liu Q. HMM-based Chinese Lexical Analyzer

ICTCLAS. In the 2nd SIGHAN workshop affiliated with ACL'03. 2003, 184-187.

[30] Ma J, Zhang Y, Liu T, and Li S. A statistical dependency parser of Chinese under small training data. IJCNLP-04. 2004.

[31] Dong Z and Dong Q. HowNet and the Computation of Meaning. World Scientific Publishing. 2006

[32] 郝博一，夏云庆，郑方. OPINAX: 一个有效的产品属性挖掘系统. 第四届全国信息检索与内容安全学术会议（NCIRS-2008）. 2008, 281-290.

[33] Zhou Q and Yu H. Integrate Statistical Model and Lexical Knowledge for Chinese Multiword Chunking. In Proc. of NLP-KE-2008. 2008, 408-415.

致　谢

衷心感谢导师郑方教授及辅导老师夏云庆副研究员对本人的精心指导。他们的言传身教将使我终生受益。

在香港中文大学进行两个月的合作研究期间，承蒙黄锦辉教授热心指导与帮助，不胜感激。

感谢信息研究院语音和语言技术中心全体老师和同窗们学的热情帮助和支持！

本课题承蒙国家自然科学基金资助，特此致谢。

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成果。尽我所知，除文中已经注明引用的内容外，本学位论文的研究成果不包含任何他人享有著作权的内容。对本论文所涉及的研究工作做出贡献的其他个人和集体，均已在文中以明确方式标明。

签　名：_____ 日　期：_____

个人简历、在学期间发表的学术论文与研究成果

个人简历

1984 年 06 月 18 日出生于河北省石家庄市。

2003 年 9 月考入清华大学电子工程系电子信息工程专业，2007 年 7 月本科毕业并获得工学学士学位。

2007 年 9 月免试进入清华大学计算机科学与技术系攻读计算机科学与技术硕士至今。

发表的学术论文

[1] 郝博一，夏云庆，邬晓钧，等. 基于泛化和繁殖的自举式意见目标抽取方法. 清华大学学报自然科学版. 2009, S1: 1333-1338.

[2] 郝博一，夏云庆，郑方. OPINAX: 一个有效的产品属性挖掘系统. 第四届全国信息检索与内容安全学术会议（NCIRS-2008）. 2008, 281-290.

[3] Xia Y, Hao B, and Wong K F. Opinion Target Network and Bootstrapping Method for Chinese Opinion Target Extraction. The Fifth Asia Information Retrieval Symposium (AIRS). 2009, 339-350.

[4] Xia Y, Hao B. Term Extraction from Web Reviews with opinion heuristics. International Conference on Machine Learning and Cybernetics (ICMLC). 2009.

[5] 夏云庆，郝博一，徐睿峰. 意见目标网络与意见目标抽取研究. 中国第十届计算语言学学术会议. 2009.