基于句法分析与依存分析的评价对象抽取①

王卫平，孟翠翠

(中国科学技术大学 管理学院，合肥 230026)

摘　要：随着互联网的不断普及，针对各种产品的评论也不断增多，这些评论中所包含的丰富信息，对制造商和消费者都极具分析价值。只有正确分析评价对象，意见挖掘的结果才会准确可信。在总结前人成果的基础上，针对网络上的客户评论，提出了一种新的评价对象抽取方法。该方法是基于 ICTParser 的句法分析与 IR 的依存关系分析的联合，采用似然值检验的方法筛选掉与主题不相关的候选评价对象。实验结果验证了该方法的有效性。

关键词：意见挖掘；句法分析；依存分析；评价对象；似然值检验

Opinion Object Extraction Based on the Syntax Analysis and Dependence Analysis

WANG Wei-Ping, MENG Cui-Cui

(School of Management, University of Science and Technology of China, Hefei 230026, China)

Abstract: As the Internet becomes more widespread, the comments according to various products are also growing, these comments which contain abundant information are extremely analysis value for the manufacturers and consumers. Based on the correct analysis of opinion object, the results of opinion mining can be accurate and reliable. Based on the summarization of predecessors' achievements, according to the customer comments on the network, this paper proposes a new evaluation object extraction method. This method is based on the ICTParser syntactic analysis and IR dependence analysis, adopts the likelihood testing methods to screen irrelative candidate opinion objects. The experimental results preliminarily verified the validity of this method.

Key words: opinion mining; syntactic analysis; dependence analysis; opinion object; likelihood-ratio test

1　引言

随着 Internet 的迅猛发展和电子商务的不断普及，互联网以其独特的优势吸引着各大制造商在网络上出售产品以及消费者在网络上购买产品。与此同时，网络上关于各种产品的评论语句的数量迅速增长。这些产品评价给制造商和消费者带来巨大好处。一方面，制造商可以从中得到关于产品的反馈信息；另一方面，潜在的消费者可以从已有的产品评价中找到客观真实的购物参考。

但是评论语句庞大的数量在一定程度上非常不利于制造商以及潜在的消费者从中提取有用信息。面对这样的现实问题，如何才能对这浩如烟海的评论语句进行快速查询和统计，意见挖掘技术应运而生。

意见挖掘是当前自然语言处理的研究热点，它帮助人们在大量产品评论中快速定位需要寻找的相关产品意见。根据 Kim 和 Hovy 对意见的定义[1]:意见由四个元素组成，即主题(Topic)、持有者(Holder)、陈述(Claim)、情感(Sentiment)。这四个元素之间存在着内在的联系，即意见的持有者针对某主题发表了具有情感的意见陈述。

在这四个元素中，主题的抽取可谓是重中之重。准确又快速地定位网络客户评论的主题（即评价对象），这是正确进行情感分析的基础，这也是意见挖掘系统准确率的保证。只有正确抽取评价对象, 意见挖掘的结果才会准确可信。

评价对象抽取任务当中，最重要的就是主题的识

① 收稿时间:2010-11-14;收到修改稿时间:2011-01-13

别，而主题的识别的基础就是要正确的词性标注(Part-of-Speech Tagging)。词性标注就是对于目标句子进行分词并标注上相应的词性。目前，国内能够实现词性标注的比较权威的分析器主要有 ICTCLAS 词法分析器、ICTParser 句法分析器以及 IR 分析器。

ICTCLAS 词法分析器是中国科学院计算研究所历时 1 年研究出来的汉语词法分析系统，该系统的功能主要有中文分析、词性标注。

ICTParser 句法分析器由中国科学院计算技术研究所自然语言处理研究组研究出来的能够实现分词、词性标注以及识别短语的功能，确定了词与词之间的关系。

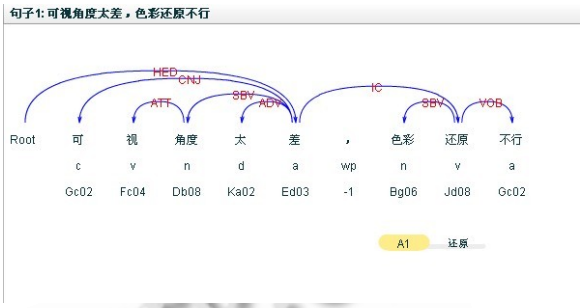
IR 是哈尔滨工业大学信息检索研究中心研究出来的语言技术平台，能够同时实现词性标注、词义消歧、命名实体、句法分析以及语义分析的功能。它使用了依存分析的技术，能够更加明确词与词之间的关系。

例如：利用汉语词法分析系统 ICTCLAS（中国科学院计算技术研究所研制出来的）对“为了省成本，用的 led 面板太差，颜色很冷，可视角度太差，色彩还原不行”进行词性标注的结果为“为了/p 省/n 成本/n ，/wd 用/v 的/ude1 led/x 面板/n 太/d 差/a ，/wd 颜色/n 很/d 冷/a ，/wd 可/v 视/vg 角度/n 太/d 差/a ，/wd 色彩/n 还原/vi 不行/a”。其中，p 表示介词，n 表示名词，wd 表示标点符号，v 表示动词，ude 表示助动词，x 表示字符串，d 表示副词，a 表示形容词。

词法分析的作用是从词典中划分出词，而句法分析的作用是了解这些词之间的关系。

还以上面句子为例，ICTParser 句法分析的结果为“(IP (PP (P 为了)(NP-B (NN 省)(NN 成本)))(PU ，)(IP (IP (IP (NP(CP (IP (VP-B (VV 用)))(DEC 的))(NP-B (NN led)(NN 面板)))(VP (ADVP (AD 太))(VP-B (VA 差)))))(PU ，)(IP (NP-B (NN 颜色))(VP (VP (ADVP (AD 很))(VP-B (VA 冷)))(PU ，)(VP (VV 可)(VP-B (VV 视)(NP-B (NN 角度)))))(VP (ADVP (AD 太))(VP-B (VA 差)))))(PU .)(NP-B (NN 色彩))(VP-B (VCD (VV 还原)(VV 不行))))”。在这个例子当中，会指出名词短语“led 面板”，而不像词法分析的结果那样，把“led”和“面板”分开标注，这样准确率就提高了。其中，NP 表示名词短语，VP 表示动词短语。

与此同时，IR 提供了词与词之间的依存关系，例如：ATT 代表修饰关系，SBV 代表主谓关系，ADV 代表状语中心语关系，COO 代表并列关系，VOB 代表动宾关系。如下图所示：



本文为了更加准确地提取评价对象，本文采用了 ICTParser 句法分析器与 IR 相结合的方法。

为清楚地了解评价对象抽取任务，请参考如下例子：“摄像头显示效果不好，散热超好”。第一个子句的主题是“摄像头显示效果”，第二个子句的主题是“散热”。

本文的任务就是要准确而又快速地抽取出这些评价对象。

2 相关工作

对于评价对象的抽取，研究者作出了很多的研究，也是现在意见挖掘研究的热点之一，2008 年和 2009 年，连续两次被列入中文倾向性分析评测（the Chinese Opinion Analysis Evaluation, COAE）的任务之一。

评价对象抽取最常用的方法就是基于模板的提取，在牺牲召回率的条件下，实现较高的精确率。

Minqing Hu and Bing Liu[2]采用 NLProcessor 2000 先对评论语句进行分词标注(Part-of-Speech Tagging ,POS)，然后再使用关联规则挖掘频繁项作为候选评价对象，并对其进行了剪枝处理。但是该方法没有有效地筛选候选评价对象，提取出来的所有候选评价对象都会被进一步处理，那么最终的处理结果的精确度也是不能保证的。

在 Popescu 和 Etzioni[3]的 OPINE 系统中，他们计算互信息（PMI）值作为参数对候选评价对象进行筛选。相比于 Hu 和 Liu 的结果，OPINE 系统以牺牲 0.03 的召回率为代价，换来了准确率 0.22 的提升。但是该系统需要根据事先设定的模板以及 WordNet 中的上下位关系(WordNet’s IS-A hierarchy)来提取候选评价对象。

OPINAX 系统[4]利用一个小规模标注评价语料库产生候选评价对象的“种子”, 利用哈尔滨工业大学语言技术平台[5]分析出词汇与“种子”之间的依存关系, 对种子进行评价对象扩展。但是利用语言技术平台的句法分析得到的词性标注信息不一定准确。另外，它还需要提前建立一个语料库，降低了系统的效率。

本文利用 ICTParser[6]和 IR 这两种语言技术平台的结合来对评论语句进行句法分析，得到相对来说比较准确的词法分析结果。

CRF(Conditional Random Field)模型[7]由 John Lafferty 和 Andrew McCallum 于 2001 年提出，这是一种序列化标注模型，常用于命名实体识别等任务，并有良好表现。

文献[8]基于 CRF 并结合一些模式匹配的方式，对评价对象进行抽取。但是使用 CRF 进行标注能够得到较高的准确率，但是召回率不理想。

文献[9]是利用外部资源信息来构造相应的词典，但词典的构建本身是一个难题。

综上所述，目前评价对象的抽取还是存在一些问题，不能同时实现精确率高而召回率也高的情况。

本文从一个新的角度来考虑：抽取的评价对象符合抽取规则，但是并不一定跟主题相关；即使跟主题相关，但未必是主观句中的评价对象。本文不需要建立任何词典，也不需要进行主客观句的分析，只需要利用中国科学院计算技术研究所自然语言处理研究组的 ICTParser 的句法分析以及哈尔滨工业大学的语言技术平台 IR 的依存关系对评论语句进行分析，然后根据一定的规则计算候选评价对象的权重，对候选评价对象进行初步筛选，最后利用似然值检验方法计算候选评价对象与主题的相关程度，对候选评价对象进行排序。

3　产品评价对象

根据 Popescu 和 Etzioni[10]于产品评价对象的定义，产品评价对象主要五种类型，包括产品的整体、部件、部件的特征、与产品相关的概念以及与产品相关概念的特征等。以数码相机为例，如表 1 所示。

本文的目标就是正确而又快速的提取出这些评价对象，包含上面五种类型的评价对象。

表 1　五种产品评价对象

产品的整体特征（Properties）	相机的尺寸
产品的某个部件（Parts）	相机的电池
产品的某个部件的特征（Features of Parts）	相机电池的属性
与产品相关的概念（Related Concepts）	生产厂家
与产品相关概念的特征（Related Concepts’ Features）	生产厂家的规模

4　候选评价对象抽取

4.1 候选评价对象抽取框架

候选评价对象抽取框架如图 1 所示。

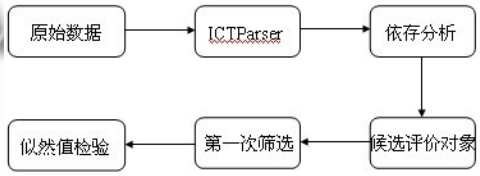


图 1　候选评价对象抽取框架

首先收集网络上的原始数据，即文本；其次利用 ICTParser 和 IR 进行句法分析和依存关系分析，按照一定的规则进行候选评价对象的抽取；最后，利用似然值检验法和权重评分的方法对候选评价对象进行排序。

4.2 预处理

(1) 模糊匹配

在需要处理的原始数据中，由于评论发表者的笔误，而导致错别字的出现。这样就需要模糊匹配成正确的词汇。例如：“显示其”就应该改成“显示器”，“纸纹”改成“指纹”等等。

(2) 标点符号

对于哈尔滨工业大学的 IR 句法分析系统，在实验中发现，它是不识别英文状态下的标点符号的，所以需要把所有在英文状态下的标点符号转换成中文状态下的标点符号。

(3) 由于分析器的不完善，ICTParser 和 IR 有识别不出来的词汇，例如：性价比，显卡，硬盘等等，本文采取词条共现权重的方法，具体如下：

假设 T_{ij}, T_{ik} 为文档 i 中的两个相邻的词，即两个词之间没有任何字符的出现，例如:“散热很好”当中，“散热”和“很好”这两个词就是相邻的词，则词条共现权重

$$W(T_{ij},T_{ik})=\frac{2tf(T_{ij}\bigcap T_{ik})}{tf(T_{ij})+tf(T_{ik})}$$

其中，$tf(T_{ij}\bigcap T_{ik})$代表同时出现的次数。

(4) 对于重复两次或者两次以上出现的评价语句，就被认为是评论垃圾（Review Spam），删除掉重复语句。

4.3 候选评价对象抽取步骤

(1) 基于 ICTParser Demo 平台，进行分词和词性标注。

①对于单个名词，直接提取；

②对于“NN+的+NN”、“NN+NN”、“NN+NN+NN”、“JJ+NN”等形式，直接提取，其中 NN 代表名词，JJ 代表形容词；

③对于“词+的+词”情况，我们认定“的”后面的词汇的词性是名词，不管那个词标注成什么词性。

(2) 在 ICTParser 分析结果的基础上，再基于 IR 平台，进行依存关系分析。

①对于前者分析出来的名词，如果与后者分析出来的结果不一致，以前者的结果为准。

②如果前者的结果当中，存在不能独立表达意思的词语，然后再对照后者的分析结果，提取 ATT、COO、VOB、ADV 等依存关系，进一步精确候选评价对象的提取结果。例如，“可视角度”是一个评价对象，它能够独立表达一个概念，而“角度”就不行，因为也可把“角度”理解成“摄像头的角度”等等。前者只标注出“角度”，所以根据后者的依存关系来完善候选评价对象的抽取。以“屏幕的可视角度太差”为例来分析，如图 2 所示。

ICTParser 的分析结果：

(VP (VV 可) (VP-B (VV 视)(NP-B (NN 角度))))(VP (ADVP (AD 太))(VP-B (VA 差))))))

IR 的分析结果：

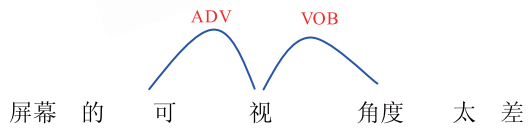


图 2 IR 依存关系分析

于是，得出候选评价对象为“可视角度”。其中，VP 代表动词短语，NP 代表名词短语，NN 代表名词。

③如果前者的结果当中，提取的名词在后者结果中的前面、后面也有名词，或者和这个抽取的名词联合在一起是一个名词，那就提取后者的分析结果。以“显卡温度”为例，前者分析结果为“温度”为名词，而后者分析结果中，“显卡”和“温度”都为名词，所以根据规则，候选评价对象应为“显卡温度”。

4.4 候选评价对象筛选

对于抽取出来的候选评价对象，并不是所有的都符合要求。候选评选对象存在以下两种情况：第一，候选评价对象未必出现在主观句当中；第二，即使是主观句当中的候选评价对象，出现的次数也很多，但是未必与主题相关。这两种情况的候选评价对象都是本文所要处理的对象。

4.4.1 基于规则的候选评价对象筛选

为了确定候选评价对象是否出现在主观句当中，我们采用以下筛选规则：

(1) 候选评价对象前后 5 个字符之内是否有形容词，具有褒贬含义的动词、名词或者其他词性的词汇。在判断动词、名词的词性时，本文利用了情感词汇表，其中包括程度级别词语、负面评价词语、正面评价词语、正面情感词语以及负面情感词语。

(2) 在候选评价对象所在的语句中，出现“没有”、“够”、“尤其”、“如果……”、“不过”、“容易”、“竟然”、“易”、“无”、“需要”、“完全”、“只有”、“要是……”、“值得”等词语的时候，提取出此候选评选对象。

(3) 在对比句当中，“比”前后的候选评价对象都提取出来。

(4) 处在介词短语当中的候选评选对象不需要提取，如“……了”、“在……时候”、“自从……之后”、“……上的”、“在……时”、“……下的”、“对于……”、“当……时”等等。

4.4.2 似然值检验法

上一步骤中，本文把不在主观句当中的候选评价对象进行筛选，那么在剩下的候选评价对象当中，利用似然值检验法进一步对候选评价对象进行排序，以排除与主题不相关的候选评价对象。这个方法是基于 Dunning 提出来的似然率检验（likelihood-ratio test）的方法[11]。

具体方法如下：

选取与评价主题相关的评论语句 D_+ 与评价主题不相关的评论语句 D_-，统计在 D_+ 和 D_- 中出现候选评价对象的评论语句的个数 C_{11}、C_{12}，统计在 D_+ 和

D_- 中没有出现候选评价对象的评论语句的个数 C_{21}、C_{22}，如表 2 所示，其中 bnp 代表候选评价对象出现，$\overline{bnp}$ 代表候选评价对象不出现。

表 2　出现和不出现候选评价对象的语句个数

	D_+	D_-
bnp	C_{11}	C_{12}
$\overline{bnp}$	C_{21}	C_{22}

似然率的计算方法如下：

$$-2\log\lambda=\begin{cases}-2*l_r & if \quad r_2<r_1\\ 0 & if \quad r_2\geq r_1\end{cases}$$

其中，$r_1=\dfrac{C_{11}}{C_{11}+C_{12}}$，$r_2=\dfrac{C_{21}}{C_{21}+C_{22}}$，

$r=\dfrac{C_{11}+C_{21}}{C_{11}+C_{12}+C_{21}+C_{22}}$，

$$\begin{aligned}l_r=&(C_{11}+C_{21})\log r+(C_{12}+C_{22})\log(1-r)\\&-C_{11}\log r_1-C_{12}\log(1-r_1)\\&-C_{21}\log r_2-C_{22}\log(1-r_2)\end{aligned}$$

$-2\log\lambda$ 的值越大，代表候选评价对象 bnp 与主题的相关度越大。根据似然得分值对 bnp 进行排序，根据事先设定的阈值进行筛选。

5　数据实验

5.1 数据来源

本文在中关村在线网站上（http://detail.zol.com.cn/）选取了 HP 4411s 系列笔记本、Canon 500D 数码照相机以及 Nokia 5230 相机这三种商品的网络评论语句作为实验语料进行数据实验，分别选取了 98 篇、80 篇、397 篇评论，每篇评论包含三种类型：优点、缺点和总结。对比评论语句也是选自中关村在线网站上的评论。用人工标注的方法对这些评论中所提到的该商品属性进行识别和标注。以 HP 4411s 为例，人工标注属性的集合如表 3 所示。

表 3　HP 4411s 属性的人工标注结果

商品名称	人工标注属性集合	人工标注属性数量
HP 4411s	散热，led 面板，可视角度，色彩还原，价格，外观，HP LOGO，屏幕，性能，键盘设计，性价比，蓝牙，配置，键盘手感，触摸板手感，控制键，摄像头显示效果，显卡温度，cpu 温度，发热量，商务机，家用机，音响，速度， 硬盘，屏幕显示效果，CPU 主板运行温度，做工，电池续航能力，噪音，品牌，厚度，触控板上的双键，硬盘声音，驱动盘，外壳质量，风扇声音，摄像头质量，无线网卡，3D Drive Guard，读盘速度，噪声，底板质量，面板材料，音箱，面板质量，鼠标，键盘做工，重量，显示器外壳，硬盘容量，屏幕尺寸，驱动，触摸板，左键，电源线，质量，售后服务，硬盘转速，屏幕亮度，画面，系统稳定性，屏幕反应速度，电源适配器，包装，反应速度，售后服务态度，Vista 操作系统运行，音量，内存，显卡，用户界面，指纹识别，电脑锁孔，显示器，智能声音触摸调节，外接口设计，分区，LED 屏幕色彩，BIOS 密码，屏幕解析度，指纹采集器	82

5.2 评测标准

本文的方法分成两个部分，第一部分是怎么样提高候选评价对象抽取的准确率；第二部分就是对候选评价对象进行筛选，验证筛选候选评价对象的方法的有效性。评判筛选候选评价对象的方法的有效性的标准如下：

查全率：筛选出来的正确的评价对象数量/人工标注的评价对象数量，用 R 表示：

$$R=\frac{A}{A+C}$$

查准率：筛选出来的正确的评价对象数量/本文方法提取出来的评价对象数量，用 P 表示：

$$P=\frac{A}{A+B}$$

F 值的计算公式如下：

$$F=\frac{\left(b^2+1\right)*PR}{b^2*P+R}$$

其中，A 代表本文方法提取出来的评价对象中正确的数量；B 代表本文方法提取出来的评价对象中不正确的数量；C 代表本文方法没有提取出来的评价对象中

正确的数量；b 是一个预设值，本文设 b 的值为 1，代表 R 和 P 一样重要。

5.3 实验 1：利用 IR 依存分析修正后的结果对比

候选评价对象抽取的过程中，如果仅仅利用 ICTParser 句法分析器，那么提取出来的候选评价对象有 911 个，而通过 IR 依存关系修正的候选评价对象有 82 个，其中 57 个候选评价对象是最终抽取出来的评价对象。

其中，精确率提高了 4.86%，但是由于需要 ICTParser 和 IR 的联合，处理速度降低了 1%，总体来说，这种方法的处理结果还是可以接受的。

表 4 ICTParser 和 ICTParser&IR 处理结果对比

	ICTParser	ICTParser&IR
候选评价对象	829	911
本文方法提取出来的评价对象	128	186
精确率	15.44%	20.30%

5.4 实验 2：验证筛选候选评价对象的方法的有效性

综合 3 种商品的实验结果（如表 5 所示），平均查全率为 79%，平均查准率 81.38%。对比 Popescu 和 Etzioni 的 OPINE 系统所得出的结果（=79%，=76%），证明本文所提出的方法的有效性。

表 5 筛选候选评价对象的方法的有效性

商品名称	查准率(P)(%)	查全率R (%)	F值(%)
HP4411s	81.70	76.00	78.74
Canon 500D	80.45	80.00	80.22
Nokia 5230	82.00	81.00	81.50
平均值	79	81.38	80.15

6 结语

对于中文评论语句的评价对象的抽取方面，国内外都展开了一系列的研究，对于中文评论语句，最关键的就是句法分析。

本文在候选评价对象抽取方面采用了 ICTParser 和 IR 想结合的方法，弥补了各自句法分析的不足。在对候选评价对象进行筛选时，采用了似然值检验法，筛选掉那些与主题不相关的候选评价对象。实验结果证明，达到了很好的效果。

在评论中评价对象抽取方面，还是存在一些问题的，例如，ICTParser 与 IR 的结合有多种方式，采取的方式不同，得到的结果也不同。

参考文献

1 Kim SM, Hovy E. Determining the Sentiment of Opinions. Proc. of COLING-04, the Conference on Computational Linguistics (COLING-2004). Geneva, Switzerland, 2004. 1367-1373.

2 Hu M, Liu B. Mining Opinion Features in Customer Reviews. Proc. of Nineteeth National Conference on Artificial Intellgience (AAAI-2004). San Jose, USA, 2004.

3 Popescu AM, Etzioni O. Extracting Product Features and Opinions from Reviews. Proceedings of HLT-EMNLP-05, the Human Language Technology Conference/Conference on Empirical Methods in Natural Language Processing. Vancouver, Canada, 2005. 339-346.

4 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论文集(上卷):2008.

5 哈尔滨工业大学语言技术平台 IR: http://ir.hit.edu.cn/demo/ltp/.

6 中国科学院计算技术研究所自然语言处理平台 ICTParser: http://nlp.ict.ac.cn/demo/ictparser/index.php.

7 Lafferty J, McCallum A, Pereira F. Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data. Proc. of 18th International Conference on Machine Learning, 2001:282-289.

8 Choi Y, Cardie C, Riloff E, Patwardhan S. Identifying sources of opinions with conditional random fields and extraction patterns. Proc. of the Human Language Technology Conference and the Conference on Empirical Methods in Natural Language Processing (HLT/EMNLP). 2005: 355-362.

9 赵妍妍,浏鸿宇,秦兵等.HIT_IR_OMS:情感分析系统. Proc. of the COAE2008, Harbin, 2008.81-88.

10 Popescu AM, Etzioni O. Extracting Product Features and Opinions from Reviews. Proc. of HLT-EMNLP-05, the Human Language Technology Conference/ Conference on Empirical Methods in Natural Language Processing. Vancouver, Canada, 2005. 339-346.

11 Dunning TE. Accurate methods for the statistics of surprise and coincidence. Computational Linguistics, 1993,19(1): 61-74.