# 基于语义块的事件倾向性分析研究*

韦向峰，张 全，缪建明，池毓焕
中国科学院 声学研究所，北京 100190
E-mail: wxf@mail.ioa.ac.cn

**摘 要**：事件的倾向性分析对网络舆情分析和事件趋势分析都具有重要意义。本文把影响倾向性分析的词语分为四类：对象词、褒贬词、逻辑词和程度词，建立了语句倾向性分析的二元模型和三元模型，在语句语义块分析的基础上实现对语句和篇章的倾向性获取。实验中首先确定三个事件实例的关键对象和立场，然后根据语句倾向性分析获得文章对于对象的褒贬态度和立场。实验表明语义块的范围限制有助于提高事件倾向性分析的准确性，立场分析则是事件倾向性分析的关键所在。

**关键词**：倾向性；语义块；句类分析；褒贬词；立场分析

# Research on Analyzing Opinion in Events Based on Semantic Chunks

Wei Xiang-feng, Zhang Quan, Miao Jian-ming, Chi Yu-huan
Institute of Acoustics, Chinese Academy of Sciences, Beijing 100190
E-mail: wxf@mail.ioa.ac.cn

**Abstract:** It is very important for analyzing opinion in events in public opinion and hot topics. The words which impact the effect of analyzing opinion are classified into four categories: object, polarity, logic and grade. This paper proposes a bi-gram and tri-gram model of analyzing the opinion of a sentence, and the opinion of article can also be achieved. The key objects and their opinions are confirmed by human, then our analyzing system can obtain the opinion of an article based on the analyzing the opinion of a sentence. Results show that semantic chunks improve the accuracy of opinion analyzing and standpoint analyzing is a key step in analyzing opinion in events.

**Keywords:** opinion; semantic chunks; analyzing sentence categories; polar words; standpoint analyzing.

## 1 引言

由于互联网的开放性，新闻报道、评论、博客、BBS、微博等都可以表达对事件、人或物品的倾向性观点。其中不乏具有爆炸性、争议性的事件，人们对这些事件的倾向和态度有可能影响社会价值取向，甚至造成社会的不稳定。如何自动分析文本的倾向性态度，及时发现和防止不良内容的传播和扩散，成为维护互联网健康发展的重要课题。其中，如何利用自然语言处理的技术获取文本对评价对象的态度，成为研究者们关注的焦点之一。

文本态度倾向性分析的主流方法是通过建立相关的褒贬词汇词典等褒贬评价资源，采用统计处理手段获取文本的态度倾向性；也可以把文本的倾向性类别作为文本分类问题进行处理，得到文本的态度倾向性类别。词语的倾向性是文本倾向性分析的重要基础，也是人判断文本倾向性的重要依据之一。具有强烈的倾向性或极性是某些词语的属性，利用已有的语义词典资源如WordNet[1]或HowNet[2]，根据已知极性的词语和PMI算法可计算得到未知极性词语的倾向性[3]。相对词语倾向性分析而言，句子倾向性分析模型研究较少，有的依据“上下文中相邻句子应该具有相同类别”[4]做分析，有的用CRFs研究句子的倾向性标注序列问题[5]；还有的依据主谓、动宾以及格语法赋予词语不同的权重，作为对句子倾向性分析的补充[6,7,8]。文本倾向性分析一般是

* 本文承中国科学院声学研究所所长择优基金（GS13SJJ04）、中国科学院青年人才领域前沿项目（O754021432）、中国科学院声学研究所创新工程项目（O654091431）和中国科学院知识创新工程重要方向项目（Y02A081431）的资助。

把文本看作是特征词语表达的向量，以特征词语的极性或已标注的训练语料为基础，用文本分类算法（如SVM）[9]或者聚类算法获得文本的倾向性类别。

文本态度倾向性分析的目的，就是获取文本中观点持有者对某一评价对象的态度（或称观点、情感）。从文本的评价对象来看，可分为对人的评价、对物品的评价和对事件的评价。本文主要关注事件文本中对事件的倾向性态度，首先获取文本语句中的对象词、极性词等词语，然后把语句的语义块分析结果转化为二元或三元倾向性分析模型，计算得到语句倾向性，根据语句倾向性的统计结果得到事件文本的倾向性态度。

## 2 倾向性分析模型和方法

从目前的文本倾向性分析的研究来看，涉及的被评价对象绝大多数是物品和人物，关于事件的倾向性分析研究得较少。这主要是因为事件的定义和识别复杂得多，事件包含的要素包括人或机构、时间、地点、事件过程等等。但是，事件也有其自身的特点，重大热点事件往往会形成“特定简称”（例如911事件、马德里爆炸案、汶川地震），一些有争议的事件常出现辩论的正反两方，这两方一般都是人或机构。因此，事件的定位可以通过关键字、关键人名发现相关的文章，而有争议的事件很容易使不同文章具有不同的倾向性。本文着重研究在语句倾向性分析的基础上实现有争议事件文章的倾向性分类，首先人工选定和收集某些事件的相关文章，然后确定出事件的关键词以及关键人物，同时为这些关键人物分配事件立场倾向。以这些关键词和关键人物作为对象词，在已建立的褒贬词语知识库、否定逻辑词集、程度词集的基础上，根据语句倾向性分析的模型和算法，获得语句关于事件的倾向性态度。然后对文章中包含倾向性态度的语句分析结果的权重设置、加权统计后可以得到文章对于事件的倾向性态度。

语句倾向性分析是事件倾向性分析的基础，我们首先把影响语句倾向性态度的主要词语分为四类：褒贬词、对象词、逻辑词和程度词，然后根据语句的语义块分析结果转化为最简单的二元模型或三元模型，通过褒贬词的倾向性计算得到语句对于某个被评价对象的倾向性态度，再通过事件文本中语句倾向性统计得到文本的倾向性态度。

### 2.1 词语分类

褒贬词，是指本身具有某种或多种倾向性态度的词语，例如“好”、“坏”、“赞扬”、“谴责”、“骄傲”、“风骚”等等。为便于计算，具有正面积极评价信息的词称为褒义词（属性值取“+1”），具有负面消极评价信息的词称为贬义词（属性值取“-1”），没有倾向性态度信息的词称为中性词（属性值取“0”），具有多种倾向性态度的词属性值取“2”。

对象词，是表示被评价对象或态度持有者的词语，可以是表示人、物品、事件、属性等的各种各样的词语。对象词可以具有立场属性，属性值“+1”表示事件正方，“-1”表示事件反方，“0”表示中立方。对象间的褒贬、立场具有传递性，并在一定程度上反映了文本述者的褒贬立场。

逻辑词，主要是表示肯定或否定的词语，如“是”、“不是”、“不”、“未必”等。褒贬词经逻辑词修饰后倾向性可能会发生反转，例如“好”加上否定修饰“不”后变成了“不好”，“好”与“不好”二者的倾向性态度完全相反。

程度词，是指一些描述程度的修饰词语，如“最”、“很大”、“较大”、“一定程度”等，按照程度从小到大人工确定属性值（大于0且小于1），如“较大”取0.6，“很大”取0.8，“最”取0.9。程度词在运算中不影响倾向性的极性（褒贬性），但可改变句子倾向性（褒或贬）的强弱程度。

### 2.2 二元模型和三元模型

对象词和褒贬词是构成语句倾向性分析模型的最基本要素，在具有倾向性态度的语句中一般

会同时出现被评价对象词和褒贬词。因此，最简单的二元模型评价形式为“A B”，其中 A 是褒贬词 B 是对象词，反之亦可。例如，一语句经简化处理后为“该方法 好”，那么语句陈述者对于“该方法”的倾向性态度是正面的（“好”），语句倾向性态度的取值为“+1”。

倾向性评价分析的三元模型的基本形式为“C X D”，其中 C 是评价者（对象词），X 是褒贬词，D 是被评价对象（对象词）。如“他 鄙视 这种做法”，则评价者“他”对被评价对象“这种做法”的倾向性态度是反面的。从本质上看，三元模型省略评价者后就是二元模型，此时评价者默认为语句文本的陈述者。当不关心评价者，只关心被评价对象和对其的倾向性态度时，三元模型可以转化为二元模型处理。而四元模型或更多元的模型，如“A 陈述 B X”（其中 B 是对象词，X 是褒贬词），也可以转化为二元或三元模型处理。

根据二元模型或三元模型，语句倾向性分析的计算结果要么为+1，要么为-1，其结果值由褒贬词的属性值确定。当语句中出现逻辑词时，在逻辑词的辖域范围内如果有褒贬词，那么应乘上逻辑词的属性值（否定为-1，肯定为+1）。当语句中出现程度词时，在程度词的辖域范围内如果有褒贬词，那么应乘上程度词的属性值。为了把语句转化为二元或三元模型，我们利用了概念层次网络理论（HNC）的句类分析技术[10]和语义块分析结果，具体如 2.3 所述。

## 2.3 基于语义块的语句倾向性分析

HNC 的句类是语句的语义类型，分为 59 组基本句类。句类表示式由主语义块构成，主语义块可以是词、短语或下一级句子。主语义块又分为特征语义块 EK 和广义对象语义块 GBK。根据句类表示式中主语义块的个数，语句分为两块句、三块句和四块句。其中，四块句的句类表示式的基本格式为 GBK1+EK+GBK2+GBK3，三块句为 GBK1+EK+GBK2，两块句为 GBK+EK 或 GBK1+GBK2。两块句和三块句的表示式可与二元模型和三元模型直接对应。当语句语义块为简单构成（不含句子或句子变形）时，可以利用句类分析的结果表示式直接计算得到语句的倾向性。当语义块为包含句子的复杂构成时，需要根据所包含句子的句类表示式进行逐级深入的计算，直到没有语义块的复杂构成为止。具体的转化方法及语句倾向性分析步骤如下：

步骤 1) 如果语句为两块句，且两个主语义块分别是褒贬词和对象词，那么按二元模型计算语句的倾向性态度，转步骤 11)；

步骤 2) 如果语句为三块句，且 GBK1 为对象词、EK 为褒贬词、GBK2 为对象词，那么按三元模型计算语句的倾向性态度，转步骤 11)；

步骤 3) 如果语句为四块句，且 GBK2 和 GBK3 分别是褒贬词和对象词，那么先按二元模型计算 GBK2 和 GBK3 的倾向性，然后转步骤 2)；

步骤 4) 对语句中的每一个语义块，执行步骤 5)到步骤 10)；

步骤 5) 如果语义块内含一个褒贬词和一个对象词，那么按二元模型计算得到倾向性；

步骤 6) 如果语义块内含一个褒贬词和多个对象词，那么取褒贬词与其右边最近的一个对象词，然后按二元模型计算得到倾向性；

步骤 7) 如果语义块内含多个褒贬词和多个对象词，那么分别取褒贬词与其右边最近的一个对象词，然后按二元模型计算得到倾向性；

步骤 8) 如果语义块内含有逻辑词，那么逻辑词右边最近褒贬词的倾向性应乘上逻辑词的属性值；

步骤 9) 如果语义块内含有程度词，那么程度词右边最近褒贬词的倾向性应乘上程度词的属性值；

步骤 10）如果语义块内含语句，那么把内含语句作为新语句，转步骤 1)；

步骤 11）结束，得到语句的倾向性。

## 3 实验

我们选取唐骏“学历门”事件、肯德基“秒杀门”事件、山西疫苗事件作为实验事件，通过一些网站的专题页面和用关键字在搜索引擎中的搜索结果，下载得到了关于三个事件的网络文章各为76篇、34篇和65篇。

我们建立了一个包含6368个褒贬词的词库，其中褒义词2650个、贬义词3718个，整理出12个表示肯定的逻辑词和38个表示否定的逻辑词，人工设定了14个常见的程度词的属性值。在三个事件中，还事先人工设定了各事件的关键对象词和属性值，以实现含有主观倾向性态度语句的定位。因为只有包含对象词和褒贬词的语句才会进入我们的倾向性模型进行分析，而有争议的事件中必然出现对立的双方如“唐骏与方舟子”、“肯德基与消费者”、“王克勤与山西卫生厅”等等。具体设置如表1所示，属性值“+1”表示立场为事件正方、“-1”表示立场为事件反方。

表1　事件中的对象词和属性值

| 事件 | 对象词 | 属性值 | 事件 | 对象词 | 属性值 | 事件 | 对象词 | 属性值 |
|---|---|---|---|---|---|---|---|---|
| 唐骏学历门 | 学历门 | 0 | 德基秒杀门 | 秒杀门 | 0 | 山西疫苗事件 | 山西疫苗 | 0 |
| | 学位门 | 0 | | 肯德基 | 1 | | 问题疫苗 | -1 |
| | 文凭门 | 0 | | 网友 | -1 | | 山西卫生厅 | 1 |
| | 唐骏 | 1 | | 顾客 | -1 | | 山西疾控中心 | 1 |
| | 方舟子 | -1 | | 消费者 | -1 | | 王克勤 | -1 |
| | 禹晋永 | 1 | | | | | 陈涛安 | -1 |
| | | | | | | | 患儿家长 | -1 |

在文章的倾向性分析中，首先定位对事件具有褒贬倾向的主观性评价语句，把既含有对象词又含有褒贬词的语句作为分析依据。然后根据语句的语义块分析结果，按照本文2.3节所述步骤对语义块中的对象词的褒贬倾向性进行分析。将语句的倾向性按对象权值累计正负得分，实验中每个被评价对象的权值相等，如果正值得分大于负值得分那么文章的倾向性为“褒”（“支持”被评价对象），反之如果负值得分大于正值得分那么文章的倾向性为“贬”（“反对”被评价对象）。通过人工评判每篇文章的“褒”、“贬”倾向性，并与系统分析结果进行比较，文章倾向性分析结果如表2和表3所示。

在表2中，“支持”表示文章与表1中属性值为“1”的对象的立场相同，倾向性为“褒”；“反对”表示文章与表1中属性值为“1”的对象的立场相反，倾向性为“贬”。例如，在唐骏“学历门”事件中，系统判定支持唐骏的文章数为15，反对唐骏的文章数为46。由于属性值为“1”的对象与属性值为“-1”立场对立，因此反对唐骏即支持方舟子，反之亦然。

从表3看，肯德基“秒杀门”事件支持倾向的正确率最低，这可能是因为文章中大量出现了含有“肯德基”的对象词，而忽略了属性值为“-1”的反面对象词。山西疫苗事件中反对倾向获得了较高的准确率，这可能是因为下载的文章中大量是质疑“问题疫苗”的。单从表2中人的判定结果数量来看，文章中处于中立态度的文章数量较多，而持支持态度的文章则较少，这实际上也给系统的判定造成了困难。

表2　事件倾向性分析正确数

| 事件 | 支持 | | | 反对 | | | 中立 | | |
|---|---|---|---|---|---|---|---|---|---|
| | 人 | 机 | 正确 | 人 | 机 | 正确 | 人 | 机 | 正确 |
| 唐骏“学历门” | 11 | 15 | 6 | 26 | 46 | 22 | 39 | 15 | 14 |
| 肯德基“秒杀门” | 5 | 22 | 4 | 14 | 8 | 6 | 15 | 4 | 3 |
| 山西疫苗事件 | 10 | 23 | 7 | 29 | 17 | 13 | 26 | 25 | 21 |

表3　事件倾向性分析的正确率和召回率

| 事件 | 正确率 | | | 召回率 | | |
|---|---|---|---|---|---|---|
| | 支持 | 反对 | 中立 | 支持 | 反对 | 中立 |
| 唐骏“学历门” | 40% | 47.8% | 93.3% | 54.5% | 84.6% | 35.9% |
| 肯德基“秒杀门” | 18.2% | 75% | 75% | 80% | 42.9% | 20% |
| 山西疫苗事件 | 30.4% | 76.5% | 84% | 70% | 44.8% | 80.7% |

错误的原因主要在于句子的语义块分析结果不准确，造成了对象词和褒贬词的错误相关，得到错误的倾向性结果。一旦语义块分析结果正确率提高，将会大幅提高倾向性分析结果的正确率。此外，以下的错误原因也应该引起研究者的重视：（1）分词引起的错误，例如“不才”切成了“不”和“才”，会多出一个否定词，得到相反的倾向性；（2）组合词语表达的倾向性，例如“受到了 伤害”，单个词没有明显的倾向性，但组合起来却具有倾向性；（3）假设句和条件句引起的错误，当作者使用这些句式时，他并不一定赞同所述文字的倾向态度，可能是中立也可能是反对；（4）疑问句式，在反讽等修辞文法中，经常使用反问等方式表达作者的态度，这种文本如果只分析字面得不到真正的倾向性态度；（5）表达中立的叙述说明文本，在新闻报道中有的作者只陈述事实（如正反两方的观点），并不加入自己的观点，有的作者则会在陈述当中加入自己的倾向性态度；（6）与法律庭审相关的文本，这些文本往往包含了控辩双方的态度，但从作者的倾向性态度来看是中立的；（7）立场与褒贬息息相关，如果表达了褒贬倾向也就表达了文本倾向，而立场的表达并不一定要通过褒贬来表达，而且立场往往涉及到多个对象和多种态度，比褒贬只有两个极性要复杂。上述几点都会给倾向性的分析带来困难，造成系统分析的错误。

## 6　结束语

事件的倾向性分析对网络舆情分析和事件趋势分析都具有重要意义。事件倾向性分析比物品和人物倾向性分析的范围要广，但事件内部主要还是人物或机构，以及他们之间形成的语义关系。本文从具有争议性的热点事件入手，在事件中区分出关键对象词，并把关键对象词分为立场对立的两类。然后利用语句的语义块分析技术和倾向性分析模型获得语句的倾向性，在此基础上分析得到文章对事件中主要对象的倾向性态度。总体来看，语义块规范了褒贬词的作用范围和对象词的结合关系，可以有效提高倾向性分析的正确性。

在事件倾向性分析的研究过程中，我们发现立场分析对于事件倾向性分析是至关重要的，同时也是褒贬倾向性分析的一个基础。精确的立场分析应该包括文本对象的立场和作者的立场，一般情况下可以把作者的立场等同于语句中第一陈述对象的立场，但在特殊的句式或文体中却不能等同。进一步的研究应该在文章倾向性的基础上细化出倾向性的相关各方，即“褒”是哪一个评价者对哪一个被评价对象的“褒”，作者立场与文本中对象的立场是否一致，如何根据文体或句式获取作者立场，这些都是本文未来进一步的研究方向。

## 参 考 文 献


[1] [OL]. http://wordnet.princeton.edu/

[2] [OL]. http://www.keenage.com/

[3] Turney, Peter, Littman Michael.Measuring praise and criticism: Inference of semantic orientation from association[J]. ACM Transactions on Information Systems, 2003, 21(4): 315-346.

[4] Bo Pang and Lillian Lee. A sentiment education: Sentiment analysis using subjectivity summarization based on minimum cuts[A]. In : Proceedings of ACL 2004[C]. 2004: 271-278.

[5] 刘康, 赵军. 基于层叠 CRFs 模型的句子褒贬度分析研究[J]. 中文信息学报, 2008, 22(1): 123-128.

[6] 江宝林, 刘永丹, 金峰,葛家翔, 胡运发. 一个基于语义分析的倾向性文档过滤系统[J]. 计算机应用与软件, 2005年1月22(1): 10-11.
[7] 金峰, 刘永丹, 江宝林, 胡运发. TTFS: 一个倾向性文本过滤系统的设计与实现. 计算机工程与应用[J], 2003(30): 137-140.
[8] 刘永丹, 曾海泉, 李荣陆, 胡运发. 基于语义分析的倾向性文本过滤[J]. 通信学报, 2004, 25(7): 78-85.
[9] 徐琳宏, 林鸿飞, 杨志豪. 基于语义理解的文本倾向性识别机制[J]. 中文信息学报, 2007, 21(1): 96-100.
[10] 韦向峰. 基于HNC理论的扩展句类分析平台研究[D]. 北京: 中国科学院声学研究所博士论文, 2005.