*Computer Science and Application* 计算机科学与应用, 2012, 2, 282-286
http://dx.doi.org/10.12677/csa.2012.25050 Published Online December 2012 (http://www.hanspub.org/journal/csa.html)

# Semantic-Based Text Topic Sentiment Orientation Analysis*

**Ping Zhu, Benhua Fei, Shaohui Fan**

International Centre for Bamboo Rattan, Beijing
Email: zhuping@icbr.ac.cn

Received: Oct. 14th, 2012; revised: Oct. 29th, 2012; accepted: Nov. 7th, 2012

**Abstract:** The text emotional tendency research is a subdiscipline of artificial intelligence, involving computational linguistics, data mining, natural language processing, and other disciplines. Semantic-based study and research based on machine learning are two divided directions of emotional tendency analysis. In this paper, we applied the semantic-based method, using emotion word dictionary provided by HowNet on text semantic analysis, give each text phrase or word an emotional value, then use the semantic similarity and semantic unit similarity calculation method to calculate the semantic similarity of words in the text, thereby get their emotional polarity and intensity, and finally obtain a quantify value of text emotional tendency. Experiments show that the accuracy of this method is relatively high.

**Keywords:** Semantic Orientation Analysis; HowNet; Primitive; Word Similarity

# 基于语义的文本话题倾向性分析*

**朱　平，费本华，范少辉**

国际竹藤中心，北京
Email: zhuping@icbr.ac.cn

收稿日期：2012 年 10 月 14 日；修回日期：2012 年 10 月 29 日；录用日期：2012 年 11 月 7 日

**摘　要：**文本情感倾向性研究是人工智能的分支学科，涉及了计算语言学，数据挖掘，自然语言处理等多个学科。基于语义的情感倾向研究和基于机器学习的情感倾向研究是情感倾向性分析的两个方向。本文采用了基于语义的方法，利用 HowNet 提供的情感词词典来进行文本的语义分析，对文本短语或词逐一赋予一个情感值，然后用语义和义元相似度计算的方法，计算文本中词语的语义相似度，得到词语的情感极性和强度，从而对文本的情感倾向给出一个量化的倾向程度值。通过实验表明这种方法文本研判的正确率较高。

**关键词：**文本倾向性分析；HowNet；义原；语义相似度

## 1. 引言

传统的文本倾向性分析(Sentiment Classification)就是用户对某个事物看法或评论进行文本的分析，从而得到该看法或者评论是属于对该事物消极的还是积极意见。关于文本倾向性分析的研究近年来得到了长足的发展，它是数据挖掘中研究的热点，涉及自然语言的处理，信息检索，数据挖掘，计算语言学等领域。基于语义的倾向判别为文本过滤、自动文摘等研究工作提供了新的思路和新的手段。我们可以对语义倾向度量值设定一个合适的阈值，对于倾向值低于或高于阈值，也就是态度倾向过于偏激的文章进行过滤操作，或者可将倾向值赋予一定的权值，作为文本过滤中需要考虑的一个因素。文本的倾向性分析主要由完成以下几个任务：1) 找出文档中能够体现情感的词或短语；2) 判断所找出的词或倾向值的极性和强度；3) 找出所抽取词的或短语与主题的关系。

*资助信息：本文由国家林业局国际竹藤网络中心科研专项项目 1632009006 资助。

目前，国内外对于文本倾向性的分析研究大体上分为两大类：基于机器学习的和基于语义的文本倾向性研究。传统文本分类技术是基于机器学习的方法。如 Pang 等人分别使用朴素贝叶斯、最大熵及支持向量机等方法进行文本倾向性研究。文献[1]选取褒贬倾向性比较强烈的词作为特征项，构造了一个支持向量机(Support Vector Machine, SVM)褒贬两类分类器来进行文本倾向性分析。就上面三种方法来说，SVM 的准确率要略高于其他两种分类技术，所以支持向量机法用的更为普遍。

基于语义的文本倾向性研究方法是国内最为普遍的一种情感倾向研究方法，其主要思想是先对待分析的文本中能够明显体现主观色彩的词进行抽取，其中主要是对形容词和副词[2]进行操作，然后对抽取出来的词通过计算逐一附上一个倾向值，这些词倾向度的平均值作为整个文本的倾向度。基于语义的文本倾向性分析方法又可以大致分为基于词和短语模式的文本倾向性分析方法和基于语义模式库的文本倾向性分析方法。朱嫣岚[3]等人利用 HowNet 提供的语义相似度和语义相关场的定义，计算待估词与预先选好的褒贬基准词对组的相似性和相关性，从而得到该词或短语的倾向度。文献[4]在利用 Wordnet 计算词语的语义相似度时，除了结点间的路径长度外，还考虑到了概念层次树的深度，概念层次树的区域密度等因素。李艺红，蒋秀凤[5]等人就使用一个倾向性词汇表和一个倾向性模式库，对抽取出来的句子和短语进行语义关系分析，进而得到文本倾向性。Wenbin Pan[6]等人尝试将商品的评论文本情感倾向识别为三类，不仅包括正面和负面评论，还有中性评论。

本文采用的是基于语义的方法，尝试构建领域情感字典，通过采用 HowNet 的语义和义元相似度计算的方法，计算文本中词语的语义相似度，得到词语的情感极性和强度，从而对文本的情感倾向给出一个量化的倾向程度值。

## 2. 文本倾向性分析相关技术

### 2.1. 知网(HowNet)

《知网》[7]是一个以汉语和英语的词语所代表的概念为描述对象，以揭示概念与概念之间以及概念所具有的属性之间的关系为基本内容的常识知识库。在知网世界中，有相当重要的两个概念：“概念”与“义原”。“概念”是对词汇语义的一种描述。每一个词可以表达为几个概念。“概念”是用一种“知识表示语言”来描述的，这种“知识表示语言”所用的“词汇”叫做“义原”。义原是描述概念的最基本单位。在《知网》中，一共描述了义原之间的 8 种关系，其中最重要的还是上下位关系。

### 2.2. 词语相似度

本文中词语相似度就是两个词语在不同的上下文中可以互相替换使用而不改变文本的句法语义结构的程度。两个词语，如果在不同的上下文中可以互相替换且不改变文本的句法语义结构的可能性越大，二者的相似度就越高，否则相似度就越低。相似度这个概念，涉及到词语的词法、句法、语义甚至语用等方方面面的特点。其中，对词语相似度影响最大的应该是词的语义。在本文中，相似度被定义为一个 0 到 1 之间的实数。

词语距离和词语相似度是一对词语的相同关系特征的不同表现形式，二者之间可以建立一种简单的对应关系。对于两个词语 $W_1$ 和 $W_2$，记其相似度为 $Sim(W_1,W_2)$，其词语距离为 $Dis(W_1,W_2)$，那么我们可以定义一个满足以上条件的简单转换关系[8]：

$$Sim(W_1,W_2)=\frac{\alpha}{Dis(W_1,W_2)+\alpha}$$

其中 $\alpha$ 是一个可调节的参数，具体实验中选 1.6 较为适宜。$\alpha$ 的含义是：当相似度为 0.5 时的词语距离值。这种转换关系并不是唯一的，我们这里只是给出了其中的一种可能。在很多情况下，直接计算词语的相似度比较困难，通常可以先计算词语的距离，然后再转换成词语的相似度。

### 2.3. 情感字典的构建

情感字典的构建[9]以 HowNet 发布的“情感词语集”为基础，为了构建情感词典现采用人工和计算机双重删选，试图删去那些不常用的或情感倾向不明的词语，使得情感字典的规模下降以便于计算机的检索速度提升。

HowNet 提供了两类情感词集，一类为正面词语集，另一类为负面词语集。我们希望为每个单词赋予一

个语义倾向的度量值。其大小由这个单词与基准词的语义关联的紧密程度有关。“种子词”，在这里指褒贬态度非常明显、强烈，具有代性的词语。一个待估情感词与褒义基准词联系越紧密，则词语的褒义倾向越强烈。与贬义基准词联系越紧密，则词语贬义倾向越明显。种子词我们一般采用形容词、副词、名词和动词。

## 3. 文本情感倾向性分析

计算机如何对有主观情感的文本进行分类，判断其是正面还是负面，是持支持态度还是反对态度，这就引出了自然语言处理领域一个重要的研究方向——倾向性分析。我们先了解一下需要抽取的词应该需要满足的特性，确定必须抽取的词性，然后给出语义相似度计算的方法。

### 3.1. 短语和词汇的抽取

情感词和其它的上下文尤其是修饰副词能够体现文本作者的情感倾向[10]，实验表明，存在一个以上形容词、副词的句子属于含主观倾向性的句子的概率为 55.8%。因此，形容词和副词是必须从文档中抽取的词语。一些动词，名词也是具有情感色彩的和褒贬义的。本文中，我们选择抽取形容词、副词、动词和名词这四类的词性的词作为候选情感词。

### 3.2. 情感词的语义倾向计算

本文通过将种子集作为情感词倾向判断的基准词，情感词越是接近褒义种子集，其带有的褒义倾向越是浓烈；越是接近贬义种子集，其带有的负面倾向越是强烈。于是，一个情感词的情感极性及其强度取决于他与两类种子集合的接近程度，即该词与词语集合相似度。从微观角度来说，种子集由常用情感词组成，词语与集合的相似度可以通过词语与词语的相似度来简单计算，以种子情感词集合中元素的相似度的平均值简单作为该词与情感集合的相似度。

#### 3.2.1. 语义相似度的计算原理

利用《知网》计算语义相似度，一个最简单的方法就是直接使用词语语义表达式中的第一基本义原描述式(即该词对应的第一个概念或者说是用法)，把词语相似度等价于第一概念的相似度。任意给定两个词语，计算它们之间的语义相似度取决于它们义原集合的相似度。集合计算本文采用递归程序来实现集合相似度运算。

#### 3.2.2. 词语相似度计算

对于两个汉语词语 $W_1$ 和 $W_2$，如果能够在情感字典中都检索到并且只考虑两词的第一个概念描述式，假设 $W_1$ 有 $n$ 个义原：$S_{11}, S_{12}, \cdots, S_{1n}$，$W_2$ 有 $m$ 个义原：$S_{21}, S_{22}, \cdots, S_{2m}$，我们规定，$W_1$ 和 $W_2$ 的相似度是各个义原的相似度之最大值，则有：

$$Sim(W_1, W_2) = \max_{i=1\cdots n, j=1\cdots m} Sim(S_{1i}, S_{2j})$$

这样，我们就把两个词语之间的相似度问题归结到了两个义原集合之间的相似度问题，我们称之为基于 HowNet 的计算方法[8]。当然，我们这里考虑的是孤立的两个词语的相似度。如果是在一定上下文之中的两个词语，最好是先进行词义排歧，将词语标注为概念，然后再对概念计算相似度。

### 3.3. 义原相似度计算

由于所有的义原根据上下位关系构成了一个树状的义原层次体系，我们采用简单的通过语义距离计算相似度的办法。假设两个义原在这个层次体系中的路径距离为 $d$，根据公式(1)，我们可以得到这两个义原之间的语义距离[8]：

$$Sim(p_1, p_2) = \frac{\alpha}{d + \alpha}$$

其中 $p_1$ 和 $p_2$ 表示两个义原(primitive)，$\alpha$ 是一个可调节的参数。$d$ 是 $p_1$ 和 $p_2$ 在义原层次体系中的路径长度，是一个正整数。给出以下方案，计算两义元之间的距离 $d$：

1) 将两个未知节点之间的距离 distant 设为正无穷。

2) 分别计算两个词在森林中的数组序号，若该词不是义原应返回 1)。

3) 判断该两个词是否都为义原？不是转步骤 4)，否则转步骤 5)。

4) 返回 distant(正无穷大)。

5) 对两个节点分别求其在各自树中的深度 depth。

6) 对两个节点分别求其在各自树中的从该节点到根节点的路径，其中节点用序号表示，节点和节点之间用“-”做分割。

7) 依据打印好的各自路径，计算两者是否有公共路径，即两字符串是否存在公共子串，并返回子串长度。

8) 计算两个义元之间的距离 $d$。

### 3.4. 义原集合运算

集合的相似度计算比特征结构更为复杂，因为集合的元素是无序而且平等的，因此首要任务是要在两个集合的元素之间建立一一对应关系。两个集合的相似度计算模型，必须满足我们对于集合相似度计算的一些直观要求。这里我们列出以下两条：

1) 一个集合和它本身的相似度为 1；

2) 假设两个集合都有 $n$ 个元素，其中 $m(m < n)$ 个元素相同，又假设两个元素的相似度只能是 0(不同)或 1(相同)，那么这两个集合的相似度应该是 $m/n$。

我们采用以下算法来为两个集合的元素之间建立一一对应关系：

1) 先计算两个集合的所有元素两两之间的相似度；

2) 从所有的相似度值中选择最大的一个，将这个相似度值对应的两个元素对应起来；

3) 从所有的相似度值中删去那些已经建立对应关系的元素的相似度值；

4) 重复第 2)、3)步，直到所有的相似度值都被删除；

5) 没有建立起对应关系的元素与空元素对应。

根据上述算法建立起两个集合元素的一一对应关系后，我们就很容易计算两个集合的相似度了：集合的相似度等于其元素对的相似度的加权平均。又因为集合的元素之间都是平等的，所以我们可以将所有的权值取成相同的，于是：集合的相似度等于其元素对的相似度的算术平均。

### 3.5. 语言倾向性程序的实现

微薄文本是比较典型地反映博主情感的文本，我们网上摘录了两段微博言论，包括正面和负面微博信息，见图 1 和图 2。

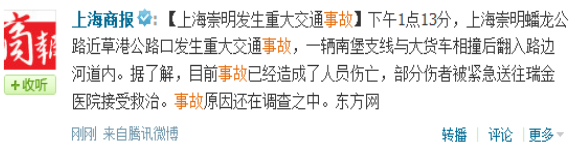


**Figure 1. Microblog text of negative sentiment**
**图 1. 负面微博信息**

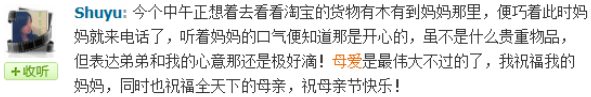


**Figure 2. Microblog text of positive sentiment**
**图 2. 正面微博信息**

#### 3.5.1. 负面信息研判示例(图 3 和 4)


**Figure 3. Word division result of negative microblog text**
**图 3. 负面文本信息分词结果**


**Figure 4. Calculate sentiment value of negative text**
**图 4. 计算负面文本情感值**

#### 3.5.2. 正面信息研判示例(图 5 和 6)

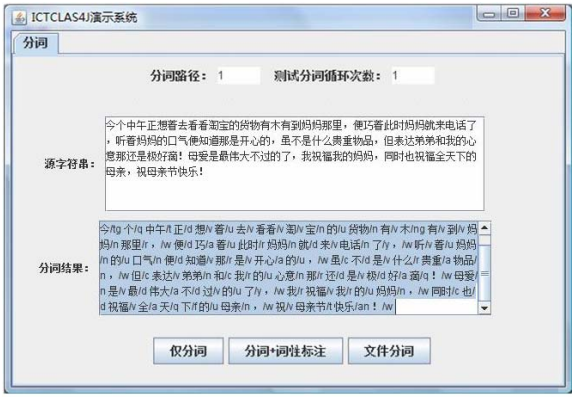


**Figure 5. Word division result of positive microblog text**
**图 5. 正面文本信息分词结果**


**Figure 6. Calculate sentiment value of positive text**
**图 6. 计算正面文本情感值**

## 4. 总结和展望

本文研究文本的情感倾向，通过统计学的原理给出一个精确的度量值，采用客观的外部知识库，以脱离人意识的情况下客观地对文章的情感倾向进行评判并给出一个强度值。研究工作基于 HowNet 提供的世界知识库的构造方法，将注意集中在知网世界的上下位关系，其他各类关系不予考虑，这是一种较为简单的处理方法。文中 HowNet 的提供的语义相似度计算是计算情感倾向的基础，我们定义情感词的相似度取决于褒义倾向和贬义倾向的差值。

在已有工作的基础上可以在以下几个方面进行更深入的研究：

1) 在义原层次体系中我们没有考虑义原树的局部密度，一旦当把它纳入我们考量的范畴时，实验结果可以更为精确。

2) 本文中没有涉及语言所固有的一些结构，如转折和递进等。它们的出现往往会对文本情感判断产生较大的影响。基于统计学的方式将语言中词语的出现位置和词语间的前后联系都忽略了，只以单个情感词为最小的语义单位，这种做法欠妥。以后可以加强这方面的研究。

3) 词程序的结果对文本的倾向的判断也会产生影响，通过优化分词程序能够提取和区分一些非常规结构的句式中的短语和词。

## 参考文献 (References)

[1] 吴琼, 谭松波. 跨领域倾向性分析相关技术研究[J]. 中文信息学报, 2010, 1: 77-83.
[2] 蔺璜, 郭姝慧. 程度副词的特点范围与分类[J]. 山西大学学报: 哲学社会科学版, 2003, 26(2): 71-74.
[3] 朱嫣岚, 闵锦, 周雅倩. 基于 HowNet 的词汇语义倾向计算[J]. 中文信息学报, 2006, 20(1): 14-20.
[4] E. Agirre, G. Rigau. A proposal for word sense disambiguation using conceptual distance. Proceedings of International Conference Recent Advances in Natural Language Processing (RANLP), Tzigov Chark, 1995: 258-264.
[5] 李艺红, 蒋秀凤. 中文句子倾向性分析[J]. 福州大学学报(自然科学版), 2010, 4: 504-508.
[6] W. B. Pan, Y. Q. Zhou. Chinese sentiment orientation analysis. Proceedings of International Conference on Computational Intelligence and Security (CIS), 2010.
[7] http://www.keenage.com/zhiwang/aboutMessage.html
[8] 刘群, 李素建. 基于《知网》的词汇语义相似度的计算[A]. 台北: 第三届汉语词汇语义学研讨会, 2002.
[9] 何凤英. 基于语义理解的中文博文倾向性分析[J]. 计算机应用, 2011, 31(8): 2130-2133.
[10] 林炜, 林世平. 中文倾向性挖掘中情感词修饰极性的研究[J]. 计算机科学, 2008, 35(8): 208-210.