利用词汇分布相似度的中文词汇语义倾向性计算

赵煜，蔡皖东，樊娜，李慧贤
（西北工业大学计算机学院，710072，西安）

摘要：针对现有中文词汇语义倾向性计算方法存在较少考虑深层语义影响因素的问题，提出了一种利用词汇分布相似度的中文语义倾向性计算方法.该方法分2个步骤完成：①利用依存句法分析和统计工具获取词汇在语料库中的分布相似度，并综合知网(HowNet)和汉语连词特征信息优化语料库统计结果，计算中文词汇间的语义相似度；②采用无向带权图划分的聚类方法来实现中文词汇语义倾向推断.由于获取最优聚类结果是一个NP难问题，所以采用贪心算法求解近似最优值.通过在自建的语料库上进行测试，并与利用语料库统计信息、利用HowNet等2个词汇语义倾向性计算系统进行比较，结果是所提方法的准确率达到了80%，表明在提高中文词汇语义倾向性计算的准确性方面是可行、有效的.

关键词：中文信息处理；词汇分布相似度；语义倾向；依存句法分析；知网
中图分类号：TP39 **文献标志码**：A **文章编号**：0253-987X(2009)06-0033-05

Computing Chinese Semantic Orientation Via Distributional Similarity

ZHAO Yu, CAI Wandong, FAN Na, LI Huixian
(College of Computer Science, Northwestern Polytechnical University, Xi′an 710072, China)

Abstract: An algorithm for Chinese semantic orientation calculation that uses distribution similarity is proposed to solve the problem that existing methods take less implied semantic into consideration in semantic orientation inference. The Chinese semantic orientation calculation is carried out in two steps. The first step calculates the distribution similarities using dependency grammar analysis and statistical tools. HowNet and Chinese conjunction features are introduced in semantic similarity calculation to optimize the corpus-based statistical results. The second step adopts an undirected weighted graph clustering algorithm to infer semantic orientation. Because it is an NP-hard problem to obtain the optimal clustering solution, a greedy algorithm is used to get an approximate solution. Experiments on the testing corpus show that the accuracy of the proposed algorithm is 80% and is better than both the corpus-based statistic algorithm and the HowNet-based algorithm. The results demonstrate that the proposed method is feasible and effective to improve the accuracy of Chinese semantic orientation calculation.

Keywords: Chinese information processing; distributional similarity; semantic orientation; dependency grammar analysis; HowNet

舆论分析是自然语言理解领域中一个新的研究热点，主要研究如何获取文本中的主观性信息，可从词汇、句子、篇章和海量信息4个层次展开研究，其中词汇语义倾向性分析是其他3个层次研究的重要基础.

词汇语义倾向性研究的任务在于，分析词汇的褒贬性(或称为极性，通常分为褒义、贬义和中性3种)和褒贬强度.英文词汇语义倾向性分析研究始于

收稿日期：2008-11-03. 作者简介：赵煜(1981—)，男，博士生；蔡皖东(联系人)，男，教授，博士生导师. 基金项目：国家自然科学基金资助项目(60803151).

1997年，Hatzivassiloglou等人针对形容词进行了语义倾向性分析[1]. Kamps从WordNet的同义结构图出发，通过计算词汇与具有强烈倾向意义的种子词汇间的相似度距离，对词汇的褒贬倾向进行赋值，从而量度形容词的语义倾向[2]. 中文词汇语义倾向性分析研究起步较晚，北京大学苏琪博士提出了基于中文概念词典(CCD)同义集合扩展、语言结构规则获取和大规模Web统计信息3种词语倾向性分析方法[3]. 朱嫣岚等人[4]利用知网(HowNet)，提出了基于语义相似度和语义相关场的词汇语义倾向性计算方法. 以上方法都是利用语义倾向待定词汇与基准词汇之间的相似性来衡量词汇语义倾向性的，相似性计算主要利用词汇资源与词汇之间的共现频次，忽视了词汇之间的深层语义信息，影响了词汇语义相似性计算的效果.

针对上述问题，本文提出了一种利用词汇分布相似度的词汇语义倾向性计算方法. 该方法借鉴了文献[5]的英文词汇分布相似度计算方法，结合了词汇间的语义依存关系、HowNet中包含的语义相似信息以及由特殊连接符(连词、标点符号等)构成的词汇结构规则，并利用计算结果构造加权无向带权图，再通过对无向图进行划分聚类，进而计算目标词汇的语义倾向性.

1 词汇语义相似度计算模型

1.1 词汇分布状态描述

词汇分布状态由与该词汇相关的上下文词汇及语义关系构成. 在自然语言处理领域中，词汇的语义相似性被定义为词汇之间在上下文中的可替换程度(词汇来自于相同语义或主题领域)，倾向性是一项重要的语义属性，因此可利用词汇的分布状态来判断词汇语义倾向性. 词汇分布状态为

$$S_{w_i}=\{(w_k,r_{ik},I_{ik})\mid w_k\in W\}$$

式中：W表示特定主题领域语料库中所有词汇的集合；r_{ik}表示词汇w_i与w_k之间的上下文关系；I_{ik}表示词汇w_i在语境(w_k,r_{ik})中所包含的信息量，其中的r_{ik}与r_{ki}是不相同的. 例如，图1中副词(非常/d)是形容词(出色/a)的修饰对象，因此

〈非常，出色〉=状中关系(ADV)

〈出色，相当〉=状中逆关系(ADV_OF)

1.2 语料库预处理

将非结构化的文本信息转换成可进行量化计算的结构化信息，必须对语料库进行预处理，处理过程包括以下4个方面.

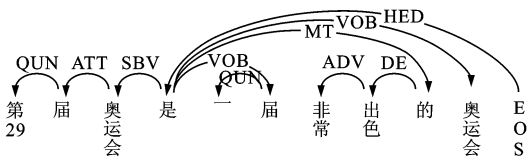

图1 依存句法分析结果举例

(1)对特定主题的语料进行分词、词性标注、命名实体识别、词义消歧等处理.

(2)从语料中提取主观表达. 本文的研究对象是词汇语义的倾向性. 语料中，主观表达是作者情感、意见、观点的主要载体，其中包含了大量具有褒贬含义的词汇，同时词汇的语义在事实叙述语境中与主观表达语境中存在差别. 因此，通过预处理过程，即抽取和分析主观表达，可提高语义倾向性计算的效率和精度.

(3)从语料中获得先验知识. 用户具有一定的领域背景知识，包括：主题领域中的基础词汇的意义；一般的语言结构规则. 前者，用户可以从语料中抽取具有特定属性(主要指词性和语义倾向性)的词汇构造褒贬基准词汇集合，并依据背景知识对语义倾向性计算的结果进行判断，再使用高置信度词汇扩充基准集合；后者，将一般的语言知识，即特殊连接符连接的词汇存在着一定的语义联系，用于词汇语义倾向性计算.

(4)提取词汇的上下文关系. 文献[6]针对英文文本利用依存句法分析，提取了词汇之间的动宾关系、主谓关系、名词修饰关系、形容词修饰关系，并利用其描述词汇的上下文关系，提高词汇相似性的计算精度. 本文采用哈尔滨工业大学的语言技术平台(LTP)完成依存句法分析，针对形容词，采用了其中的8种依存句法关系，并将部分关系进行了双向扩展.

1.3 词汇分布相似度计算方法

文献[5]将2个词汇的分布相似度定义为2个词汇公共信息量与2个词汇信息总量的比值，即

$$D_S(w_1,w_2)=\frac{\sum\limits_{(r,w_k)\in T(w_1)\cap T(w_2)}(I(w_1,w_k,r)+I(w_2,w_k,r))}{\sum\limits_{(r_{1i},w_i)\in T(w_1)}I(w_1,w_i,r_{1i})+\sum\limits_{(r_{2j},w_j)\in T(w_2)}I(w_2,w_j,r_{2j})}$$

$$T(w_1)=\{(r_{1k},w_k)\mid I(w_1,w_k,r_{1k})>0\}$$

式中：$I(w_1,w_k,r_{1k})$表示词汇w_1在语境(w_k,r_{1k})中包含的信息量. 事件A定义为词汇w_1出现在语境

(w_k, r_{1k})中；事件 B 定义为在事件 A 出现次数未知的条件下，事件 A 被随机选中；事件 C 定义为在事件 A 出现次数已知的条件下，事件 A 被随机选中. 那么，$I(w_1, w_k, r_{1k})$可定义为事件 B 与事件 C 包含信息的信息量之差

$$I(w_1, w_k, r_{1k}) = \lg \frac{\| w_1, w_k, r_{1k} \| \ \| *, *, r \|}{\| w_1, *, r_{1*} \| \ \| *, w_k, r_{*k} \|}$$

式中：$*$ 表示可取任意情况；$\| \cdot \|$ 表示事件发生的次数.

1.4　利用先验语言知识优化语义相似性计算

基于大规模语料统计的自然语言处理方法依赖于语料库. 如果词汇语义相似度仅由词汇在语料库中的分布相似度来判定，则分析结果易受资料稀疏和噪声的干扰. 英语词汇研究者的一个重要倾向是，努力寻找合适的先验知识，利用这些先验知识进一步优化语义相似性计算. 词典是语义相似性研究中一种重要的先验资源，它提供了词汇的句法和语义特征，例如 WordNet 中包含了层次信息.

本文将 HowNet 信息和连词结构作为中文先验语言知识，其特点如下.

(1) HowNet 是一种以概念为描述对象、以概念之间以及概念属性之间的关系为基本内容的常识库，它包含了丰富的中文词汇语义知识. 刘群等人提出了基于 HowNet 的词汇语义相似性计算方法[7]，本文利用该算法的分析结果修正了语料库的统计结果.

(2) 连词为词汇倾向计算提供了指示信息. 通过对语料库中 500 个主观句进行依存分析发现：若整句中存在联动结构关系(VV)、独立分句结构关系(IC)、并列结构关系(COO)、同位结构关系(APP)，以这些结构关系为轴心便可将整句划分为 2 个独立分句；如果独立分句中存在连词，则分句可作为词汇语义倾向的计算依据. 如图 2 例句(这种排污方式经济成本低廉，但是社会代价太昂贵)，句中存在 IC 关系，且 2 个分句中存在连词，则可以判断在这个语境中，低廉(/a)与昂贵(/a)是语义倾向相悖的. 针对

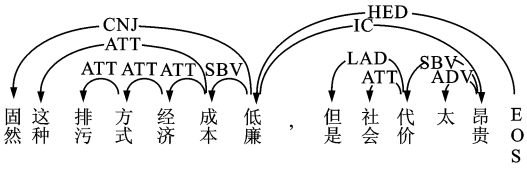

图 2　利用连词关系计算词语倾向性

形容词，研究使用的连词如表 1 所示.

表 1　推测连词表

类型	连词	倾向关系计算
并列连词	和、跟、与、同、而、及、而且、并且	相似
转折连词	却、但是、然而、而、偏偏、只是、不过	相悖
让步连词	虽然、固然、尽管、纵然、即使	相悖

综上，词汇语义相似度定义为

$$S_S(w_1, w_2) = \alpha \frac{F_{ls}(w_1, w_2) + 1}{F_{ld}(w_1, w_2) + 1} D_S(w_1, w_2) + (1 - \alpha) H_S(w_1, w_2)$$

式中：$D_S(w_1, w_2)$是词汇间分布相似度；$H_S(w_1, w_2)$是由 HowNet 计算的词汇间语义相似度；$F_{ls}(w_1, w_2)$是依据连词现象，2 词汇被推断为语义倾向相似的频率；$F_{ld}(w_1, w_2)$为语义倾向相反的频率；α 为混合系数，范围为[0,1]，实验中 $\alpha = 0.65$.

2　基于图划分的词汇语义倾向性计算

将词汇集合描述成无向带权图，词汇集合中的词汇看成是图中的顶点，词汇间的语义相似值看成是连接 2 个顶点的边所对应的权值，这样可将词汇倾向性计算的过程转化成对图划分的过程. 这种方法充分考虑了多个词汇之间的相似关系，当计算一个词汇的倾向性时，不仅简单地考虑了该词汇与其他词汇的相关性，而且还考虑了与该词汇相似的多个词汇之间的相似度，因此有助于减少语境相似但倾向性相异的词汇关系对语义倾向计算的影响.

2.1　目标函数的选择

Newman[8] 研究了无向图中节点之间的连接倾向性，引入了同类连接(assortative mixing)的概念，描述了图中同类点相互连接的程度. 图中同类节点聚合程度的同类系数

$$r = \frac{\sum_i e_{ii} - \sum_i a_i b_i}{1 - \sum_i a_i b_i} = \frac{\mathrm{tr}(\boldsymbol{e}) - \| \boldsymbol{e}^2 \|}{1 - \| \boldsymbol{e}^2 \|}$$

用 e_{ij} 表示 i 类节点与 j 类节点的连接概率，各类节点之间的连接概率构成连接概率矩阵 $\boldsymbol{e}$，并且 $\sum_{ij} e_{ij} = 1$. Newman 将无向图的两端点视为可区别的，则：a_i 表示边的一端属于 i 类节点的概率，

$\sum_j e_{ij} = a_i$；b_i 表示边的另一端属于 i 类的概率，$\sum_j e_{ji} = b_i$；$\mathrm{tr}(\boldsymbol{e})$ 表示图中边的两端点属于同一类型的概率之和；$\boldsymbol{e}^2 = \boldsymbol{e}\boldsymbol{e}^{\mathrm{T}}$，$\| \boldsymbol{e}^2 \|$仅起 到规格化的作用. 真正衡量节点聚合程度的是 $\mathrm{tr}(\boldsymbol{e}) - \| \boldsymbol{e}^2 \|$[9]，将这种衡量方法应用于中文指代消解处理中[10]，其结果证明它对中文信息处理是可行的. 本文定义词汇语义相似图划分的目标函数为

$$Q = \sum_{i=1}^{k} \left(\frac{W(c_i, c_i)}{\sum W(c_m, c_n)} - \left(\frac{W(c_i, c_j)}{\sum W(c_m, c_n)} \right)^2 \right)$$

式中：$W(c_i, c_j) = \sum_{w_l \in c_i, w_k \in c_j}^{lk} S_S(w_l, w_k)$. 当同类节点间的相似度优于异类节点间的相似度时，则该聚类是有效的.

2.2 基于贪心算法的聚类

无向带权图的聚类过程实际上是一个求最优值的过程. 在词汇间语义相似度接近的情况下，若待聚类词汇集合较大，为求得最优值进行穷尽搜索是不实际的，即成为一个 NP 难问题，因此必须使用近似优化算法进行求解. 本文采用贪心算法实现了最优值的近似求解，则词汇语义倾向性计算算法如图 3 所示.

输入：词汇集合W，集合中词汇与目标词汇w语义相似手工构造的基准词汇集合$S, S \subset W$
输出：褒义词汇类c_p和贬义词汇类c_n
过程：
(1) for i=1 to k do $c_i = w_i$
(2) produce adjacency matrix $\boldsymbol{e}$
(3) $C = \{c_i\}$ i=1,2,…,k
(4) repeat
search c_a, c_b in C
where $\arg\max_{c_i, c_j \in C} Q(c_i, c_j) = Q(c_a, c_b)$
if COMPATIBLE(c_a,c_b)
{
$c_{\mathrm{new}} = c_a \cup c_b$
delete c_a, c_b from C
insert c_{new} into C
}
else continue
adjust matrix e
until $|C|$=2
(5) ouput cluster

图 3 词汇语义倾向性计算算法

图中函数 COMPATIBLE(c_a,c_b)的判断条件是

$$|A| \geqslant |c_a| \, |c_b| / 2$$

当条件满足时，函数返回结果为真，反之为假.

$$A = \{ (w_i, w_j) \mid S_S(w_i, w_j) \geqslant \mathrm{Avg}(S_S(w_m, w_n)), w_m \in c_a, w_n \in c_b \}$$

$w_i \in c_a, w_j \in c_b$

Avg()表示簇间语义相似度均值运算.

3 实 验

目前，用于中文文本倾向性分析的语料库比较少，所以人工收集了 2 000 篇文章作为测试语料库. 本文主要是针对领域相关的词汇，因此选取了环保领域的文本进行了测试. 由于研究对象是词汇的语义倾向性，所以语料主要来源于新闻评论、博客以及 BBS.

形容词具有丰富的褒贬色彩，因此本文将其作为实验对象. 根据例句分析，在形容词的语义倾向性判断中采用了 8 种依存关系. 由于 LTP 平台产生的依存句法关系是单向的(见图 1、图 2)，而研究中需要考虑词汇间的相互关系，因此对 LTP 平台产生的结果进行了扩展，见表 2.

表 2 依存句法关系

符号	依存句法关系	扩充方法
ATT	定中关系	ATT_OF
DE	“的”字结构	回溯到“的”字的父节点，利用节点词汇及关系替换“的”字节点
DI	“地”字结构	回溯到“地”字的父节点，利用节点词汇及关系替换“地”字节点
CMP	动补关系	CMP_OF
COO	并列关系	COO
VOB	动宾关系	VOB_OF
ADV	状中关系	ADV_OF
SBV	主谓关系	SBV_OF

实验对语料库中形容词的出现频率进行了统计，选取频率最高的褒贬词汇各 20 个作为基准词汇. 为了保证实验的有效性，手工选取了频率超过 10 次的形容词共 100 个，人工标注了它们的语义倾向性，并作为测试集合.

分别与文献[4]、文献[5]方法进行了比较，以此验证本文方法的有效性. 实验中使用的验证标准——准确率 p、召回率 r 及 F 值定义如下

$$p = \frac{|R_2 \cap R_1|}{|R_2|}; \quad r = \frac{|R_2 \cap R_1|}{|R_1|}; \quad F = \frac{2pr}{p + r}$$

式中：R_1为语料库中手工标注的结果；R_2为系统生成的结果.

实验结果如表 3 所示. 从表 3 可以看出，召回率有所提高，这表明句法依存分析为语义倾向性推断提供了更多、更深层次的语境信息，有利于语境相关

词汇的语义倾向性推断.

表 3　实验结果

方法	p	r	F
文献[4]	0.693	0.187	0.295
文献[5]	0.672	0.103	0.179
本文	0.801	0.233	0.361

4　结　论

针对现有中文词汇语义倾向性计算方法存在的问题,本文提出了一种基于依存句法分析的中文语义倾向性计算方法.它将语料库统计结果与 HowNet 信息、汉语连词特征相结合,提出了中文语义相似度计算方法,并采用无向带权图划分的聚类法实现了中文词汇语义倾向性计算.实验结果表明,语料库的规模和质量对计算结果有较大影响,本文方法较之基准系统法,既考虑了词汇的深层语境特征,又利用先验知识对统计结果进行了修正,由此取得了较好的结果,为中文词汇语义倾向性计算提供了新的思路.

参考文献:

[1] HATZIVASSILOGLOU V, MCKEOWN K R. Predicting the semantic orientation of adjectives [C]// Proceedings of the 35th Annual Meeting of the ACL and the 8th Conference of the European Chapter of the ACL. Stroudsburg, PA, USA: Association for Computational Linguistics, 1997: 174-181.

[2] KAMPS J, MARX M, MOKKEN R J, et al. Using WordNet to measure semantic orientations of adjectives [C]// Proceedings of the 4th International Conference on Language Resources and Evaluation. Paris, France: European Language Resources Association, 2004: 1115-1118.

[3] 苏祺. 面向问答系统的情感倾向分析研究 [D]. 北京:北京大学信息科学技术学院, 2007.

[4] 朱嫣岚, 闵锦, 周雅倩,等. 基于 HowNet 的词汇语义倾向计算 [J]. 中文信息学报, 2006, 20(1): 14-20.
ZHU Yanlan, MIN Jin, ZHOU Yaqian, et al. Semantic orientation computing based on HowNet [J]. Journal of Chinese Information Processing, 2006, 20(1): 14-20.

[5] LIN D. Automatic retrieval and clustering of similar words [C]// Proceedings of COLING/ACL. Stroudsburg, PA, USA: Association for Computational Linguistics, 1998: 768-774.

[6] MCCARTHY D, KOELING R, WEEDS J, et al. Finding predominant word senses in untagged text [C]// Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics. Stroudsburg, PA, USA: Association for Computational Linguistics, 2004:279-286.

[7] 刘群,李素建. 基于《知网》的词汇语义相似度计算 [J]. 中文计算语言学期刊, 2002, 17(2): 59-76.
LIU Qun, LI Sujian. Word similarity computing based on HowNet[J]. Computational Linguistics and Chinese Language Processing, 2002, 17(2): 59-76.

[8] NEWMAN M E J. Mixing patterns in networks [J]. Physical Review: E, 2003, 67(22):26126.

[9] NEWMAN M E J, GIRVAN M. Finding and evaluating community structure in networks [J]. Physical Review: E, 2004, 69(22):26113.

[10] 周俊生,黄书剑,陈家骏,等. 一种基于图划分的无监督汉语指代消解算法[J]. 中文信息学报, 2007,21(2): 77-82.
ZHOU Junsheng, HUANG Shujian, CHEN Jiajun, et al. A new graph clustering algorithm for Chinese noun phrase coreference resolution [J]. Journal of Chinese Information Processing, 2007, 21(2): 77-82.

(编辑　苗凌)