doi:10.11887/j.cn.201403020 http://journal.nudt.edu.cn

应用语义关系自动构建情感词典*

谢松县,刘 博,王 挺
(国防科技大学 计算机学院,湖南 长沙 410073)

摘 要:构建英文情感词典研究相对成熟,形成了丰富可靠的词典资源。而针对中文的研究时间短,中文情感分析词典资源较少。借鉴现有可靠的英文词典资源,提出了基于语义关系的情感词典自动构建算法,算法先从HowNet的概念中进行中文义原和词语抽取及语义分析,再利用HowNet概念中DEF中英文属性值,在英文情感词典SentWordNet中进行义原和词语情感值查询,最后根据词语和义原之间的语义关系进行词语的情感值计算。算法直接利用现有的英文情感词典,无须人工标注,生成的情感词典记录了词语的语义关系、情感极性值等多种信息,弥补了现有词典的不足。评测实验结果表明,根据算法实现的情感词典相比其他词典在准确率接近的情况下,召回率和*F*值最高,取得了较好的评测性能。

关键词:情感分析;情感词典;HowNet;语义关系
中图分类号:TP391 **文献标志码**:A **文章编号**:1001-2486(2014)03-0111-05

Applying semantic relations to construct sentiment lexicon automaticlly

XIE Songxian, LIU Bo, WANG Ting
(College of Computer, National University of Defense Technology, Changsha 410073, China)

Abstract: Researches on constructing English sentiment lexicon is relatively mature, and there are abundant and reliable lexical resources. Whereas for Chinese studies, the research history is short, and there are only a few Chinese sentiment lexicon resources. With reliable English sentiment lexicon as reference, an automatic constructing approach was proposed, based on semantic relationships. Firstly the Chinese sememe and words were extracted from the defination of concepts in HowNet and the semantic analysis was carried out upon them; secondly the sentimental value of each sememe and word was retrieved from the English sentiment lexicon SentiWordNet according to the DEF attributes of concepts in HowNet, and the final sentimental value of each word was calculated on the semantic relations of the sememe and words. The ready English lexicon was used without manual labeling in the method, and diverse information of words was recorded in the final lexicon, including semantic relations and sentimental values, which remedy the lack of other lexicons. The experimental results show that the resulted sentiment lexicon can achieve better performance in the recall and F value measurements under the condition of approaching other lexicons on the precision measurements.

Key words: sentiment analysis; sentiment lexicon; HowNet; semantic relation

随着互联网的发展,尤其是社交网络的发展,以微博为代表的用户发布内容平台中出现了海量含有用户主观情感色彩的文本数据。针对网络文本的信息处理开始由获得关键词[1]、事件[2]、话题[3]等事实信息,开始向情感观点等主观信息深入,情感分析便是近年来迅速发展的信息处理技术[4]。从数据中提炼出用户的主观信息对于商业情报、舆情分析等具有重要意义。情感分析技术就是对带有情感色彩的主观性文本进行自动推理、分析、归纳的过程,涉及自然语言处理、机器学习、认知科学以及社会心理学等方面的研究[5]。语言的情感表达往往使用具有明确情感色彩的词汇,因此构建带有情感色彩的词典资源是进行情感分析研究的基础。情感分析研究在英文上发展迅速,积累了许多情感词典资源,比如:General Inquirer (GI)[6],Opinion Finder (OF)[7],Appraisal Lexicon (AL)[8],SentiWordNet[9]以及Q-WordNet[10]。中文情感分析研究起步较晚,缺乏普遍认可的可靠的中文情感词典[11-13]。目前研究使用主要有HowNet情感词典[14]、NTUSD情感词典[15]以及大连理工大学的情感词汇本体词库[16]。这些词典主要是以手工或半自动方式编辑而成,可靠性和领域适应性受到限制,并且情感词以主要褒贬二值区分,缺少情感强度值的细粒度划

* **收稿日期**:2013-10-25
基金项目:国家自然科学基金资助项目(61170156)
作者简介:谢松县(1977—),男,山东泰安人,博士研究生,E-mail:xsongx@nudt.edu.cn;
王挺(通信作者),男,教授,博士,博士生导师,E-mail:tingewang@nudt.edu.cn

分。能够将资源丰富的英文词典跨语言向资源相对贫乏的语言进行适应性的转化,既可以省去人工标注过程,又可以克服半自动方法的可靠性问题。

1 词典资源简介

1.1 HowNet 语义词典[17]

HowNet 是一个以中英文词语所代表的概念为描述对象,揭示概念与概念之间以及概念的属性与属性之间的关系的知识库。义原是 HowNet 最小语义单元,用于定义和描述概念的属性和概念间的相互关系,义原通过一个树状的层次结构组织构成上下位关系。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。如图 1 所示,HowNet 采用知识词典标记语言(Knowledge Dictionary Mark-up Language, KDML)描述概念。

```
NO.=098818
W_C=医生
G_C=N
E_C=
W_E=doctor
G_E=N
E_E=
DEF=human|人,#occupation|职位,*cure|医治,medical|医
```

图 1 HowNet 中概念的定义方式

Fig. 1 An example of concept definition

其中 W_X 表示词语,G_X 表示词语词性,E_X 表示词语例子,X 为 C 时表示中文,X 为 E 时表示英文。DEF 是对于该概念的定义项,称之为一个语义表达式,其中中英文标注的是义原,"# *"等标示符号来对概念属性之间关系进行描述。

1.2 WordNet 语义词典[18]

WordNet 是由 Princeton 大学的心理学家,语言学家和计算机工程师联合设计的一种基于认知语言学的英文词典。WordNet 是根据词义而不是词形来组织词汇信息。WordNet 使用同义词集合(synset)代表概念,词汇关系在词语之间体现,语义关系在概念之间体现。WordNet 将英语的名词、动词、形容词和副词组织为 Synsets,每一个 Synset 表示一个基本的词汇概念,并在这些概念之间建立了包括同义关系(synonymy)、反义关系(antonymy)等多种语义关系。其中,WordNet 最重要的关系就是词的同义反义关系。

1.3 SentiWordNet 情感词典

SentimentWordNet 是 Baccianella[10] 等在语义词典 WordNet 基础上使用随机游走的图算法得到的情感词典。词典的每条记录都是一个 WordNet 的 Synset,并且每个 Synset 都计算出了褒义、贬义情感强度值,本文就是利用 SentimentWordNet 的情感强度值以及 HowNet 概念的语义关系进行计算得到中文词语的情感值。SentimentWordNet 共有 117 000 多 Synsets,192 493 单词。

2 基于语义关系的情感词典构建方法

HowNet 对义原和概念进行了英汉双语标注,可以作为英文情感词典向中文情感词典转化的"桥梁"。HowNet 中概念的 DEF 是由义原按语义关系进行描述的,可以利用这种语义关系对词语的情感值进行"消歧"。解决方案如图 2 框架所示。构建中文情感词典框架可以分为义原和词语抽取及语义分析、义原和词语情感值查询与计算以及词语的情感值计算三个过程。

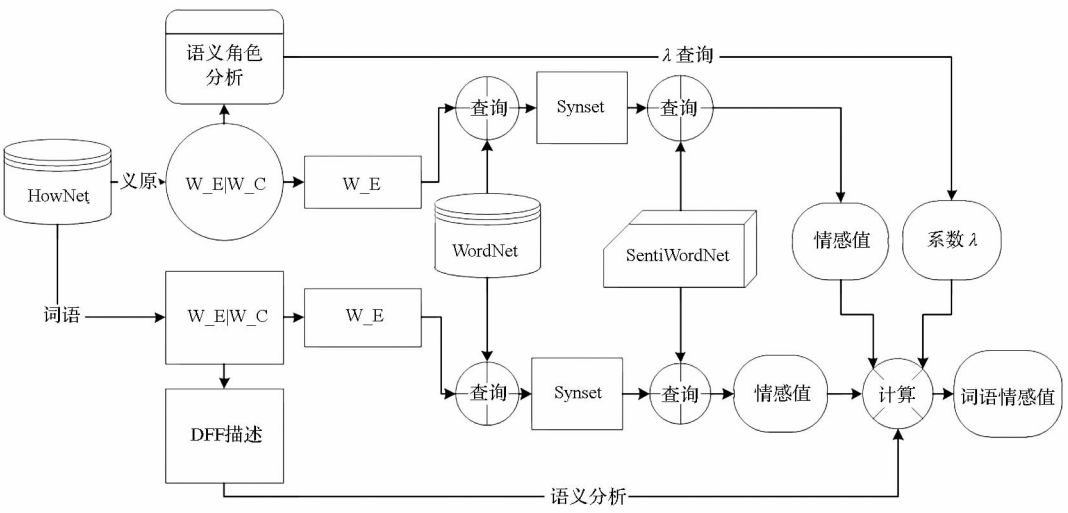


图 2 基于语义关系的情感词典解决方案

Fig. 2 Framework of constructing sentiment lexicon based on semantic relations

2.1　词语抽取和义原抽取及语义分析

词语抽取主要是从 HowNet 词典中抽取词语(W_C)和属性定义(DEF),并对 DEF 进行分析。抽取处理流程如图 3 所示。

在抽取得到的词语记录中,主要关注的内容有词语编号(No.)、中文词语(W_C)、中文词性(G_C)、英文词语(W_E)、英文词性(G_E)、属性(DEF)、第一属性(First_DEF)等。其中第一属性是指位于属性 DEF 第一位置的义原,通过第一属性可以分析出该词语所属的特征类。

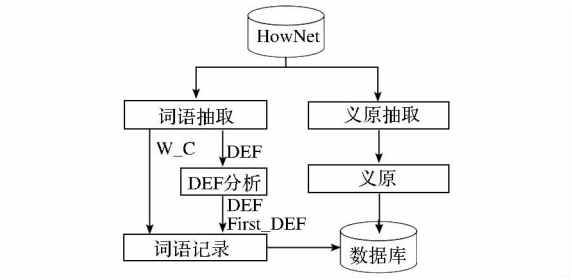


图 3　词语和义原抽取处理流程

Fig. 3　Procedure of extracting words and sememes

在进行词语倾向计算时,需要根据义原进行词语的语义分析和倾向计算。在抽取得到的义原记录中,主要关注的内容有词语编号(No.)、特征类别(Category)、中文词语(W_C)、英文词语(W_E)、属性(DEF)、层次(Layer)、父亲节点编号(Father)等。根据记录中的层次(Layer)和父亲节点编号(Father)可以得到义原之间的语义关系。

2.2　情感值的查询与计算

HowNet 词语是中英双语,因此有的可以直接将抽取到的英文词语(W_E)、英文词性(G_E)直接送入英文情感词典查询其情感值。但是大部分词语英文部分不是一个单词,因此无法直接得到情感值,而且由于词语的多义性,也无法获得唯一的情感值,因此需要进行“消歧”;HowNet 中词语的倾向性值可以通过义原的倾向性值根据语义关系计算获得,一方面可以获得直接查询无法获得情感值的词语,另外一方面也可以利用 DEF 情感值进行修正并消歧。

2.2.1　词语倾向性值查询与计算

WordNet 是以词义(sense)来记录的, sense 以同一词义的词集 Synset 表示。通过查询可以得到词语 W_E 所有的 sense,将每个 sense 映射到 SentiWordNet 就可以得到对应的情感值。

2.2.2　义原倾向性值查询与计算

基于 WordNet 和 SentiWordNet 的义原倾向计算过程如图 4 所示。在 HowNet 中获取义原后将义原对应英文词语(如“good”)映射到 WordNet 中进行查询,得到该词语所有的 Sense(如“good”的 Sense 共有 27 个);将这些 Sense 映射到 SentiWordNet 中,查询得到对应 Sense 情感值;将情感值加权并根据式(1)计算得到义原的情感倾向值(如“good”的倾向值为 PosScore = 0.597, NegScore = 0.004)。

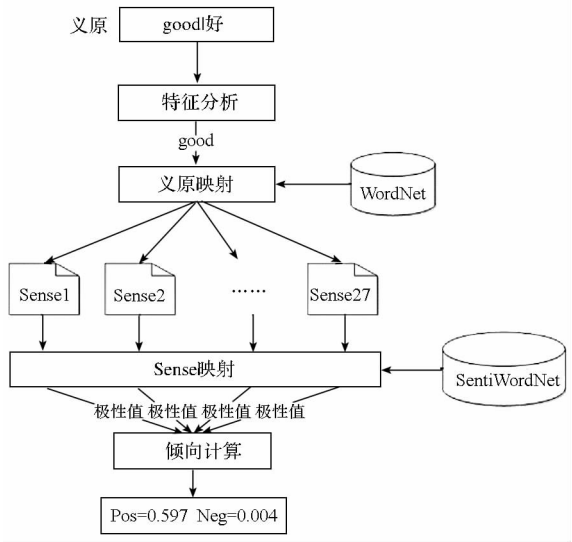


图 4　义原情感值计算过程

Fig. 4　Calculation of sememe sentiment

$$\varphi(s,p) = \frac{\sum_{i=1}^{m}\varphi_i(s,p)}{\sum_{p\in P}\sum_{i=1}^{m}\varphi_i(s,p)} \tag{1}$$

式(1)中 p 表示极性类型(积极、消极、中性,“P、N、O”),m 为与义原相对应的 Sense 的总数,s 表示义原,$\varphi(s,p)$ 表示义原 s 的 p 极性值,$\varphi_i(s,p)$ 表示义原 s 在编号为 i 的 Sense 中的 p 类型极性值。

事件类义原有很多在 DEF 描述中可以引起情感值的变化,比如“DoNot|不做,lose|失去”等会引起情感值符号反转,因此我们标注了 819 个事件类义原的在情感值计算中的语义角色,并用系数 λ 来表示。

2.3　词语情感值计算

通过 2.2.1 部分查询可以获得部分词语的情感倾向值,有些词语由于是多义的,情感值可能有几个,因此需要根据词语 DEF 描述中义原情感值进行计算修正和消歧。对 HowNet 中词语属性描述 DEF 语义关系的不同提出如下定义。

定义 1　情感倾向值取反:词语 s 的 p 极性值 $\varphi(s,p)$ 取反运算是,将 s 的积极倾向值和消极倾向

值互换,过程如式(2)。

$$\overline{\varphi(s,q)} = \varphi(s,p),(p,q) \in P \& p \neq q \quad (2)$$

定义 2 因子乘法运算: λ 因子与词语 s 的 p 极性值 $\varphi(s,p)$ 的乘法运算定义为 λ 乘法运算,过程如式(3)。

$$\lambda \times \varphi(s,p) = \begin{cases} \lambda\varphi(s,p), & \lambda > 0 \\ 0, & \lambda = 0 \\ |\lambda|\varphi(s,p), & \lambda < 0 \end{cases} \quad (3)$$

λ 取值的确定需要根据义原的类别特征、词语 DEF 的组成特征和义原间的语义关系进行确定,这些都已经在抽取部分和义原情感值计算部分记录下来。如词语“好”的 DEF 中每个义原的 λ 可以均取值为 1。词语“扭亏为盈”的 DEF 为“DEF = alter | 改变, StateIni = InDebt | 亏损, StateFin = earn | 赚”,义原“InDebte | 亏损”为初始状态,“earn | 赚”为最终状态,经过分析后,义原“InDebte | 亏损”的 λ 取值为 0,义原“earn | 赚”的 λ 取值为 1。词语倾向计算总结为式(4)。其中 $\varphi(s,p)$ 示词语 s 的 p 极性值,t_i 示词语 DEF 中第 i 个义原,n 为词语 DEF 中义原总数。

$$\varphi(s,p) = \frac{\sum_{i=1}^{n} \lambda_i \times \varphi(t_i,p)}{\sum_{p \in P} \sum_{i=1}^{n} \lambda_i \times \varphi(t_i,p)} \quad (4)$$

其中: $\sum_{p \in P} \varphi_p(s,p) = 1$。

对于已经通过查询得到情感值的词语,可以在多个英文词义 sense 对应的情感值 $\varphi_s(s,p)$ 取最接近 DEF 分析计算得到的情感值的 $\varphi_{\min}(s,p)$,然后加和平均,计算公式为:

$$\psi(s,p) = \frac{\varphi_{\min}(s,p) + \varphi(s,p)}{2} \quad (5)$$

其中:

$$\varphi_{\min}(s,p) = \min\{|\varphi_s(s,p) - \varphi(s,p)|\}$$

3 实验及结果

情感词典的实验评测有两种方法:一是与人工编辑的或者其他可靠性较高的词典进行对比评测;二是将词典应用到情感分析的其他任务上观察性能的提升。本文使用第一种方法。在实验评测时,基准词语由 HowNet 中随机选取了 2000 个词语进行人工判断,人工判断只给出褒贬两种极性。本文生成词典 SentiLex 与 HowNet 情感词典,NTUSD 情感词典以及大连理工大学的情感词汇本体词库 DLLEX 进行对比评价。

3.1 评价指标

评价指标采用准确率、召回率以及 F 值作为评测标准。设 a_1 表示褒义判断正确词数;a_2 表示贬义判断正确词数;b_1 表示判断为褒义的词数;b_2 表示判断为贬义词数;c_1 表示基准词典褒义词数;c_2 表示基准词典贬义词数。准确率计算公式为

$$P = \frac{a_1 + a_2}{b_1 + b_2} \times 100\% \quad (6)$$

召回率计算公式为

$$R = \frac{a_1 + a_2}{c_1 + c_2} \times 100\% \quad (7)$$

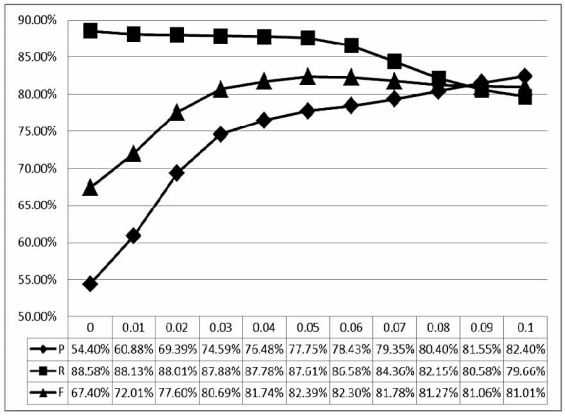
F 值计算公式为

$$F = \frac{2 \times P \times R}{P + R} \times 100\% \quad (8)$$

3.2 性能评测结果

3.2.1 阈值 T 的设定

由于基准词是褒贬二值标注的,因此需要将生成的情感词典连续情感值转换为离散褒贬值。将褒义和贬义情感值相减得到词语的倾向值来判断词语的极性,为了提高判断的准确性,设定阈值 T,高于 T 为褒义,低于 $-T$ 为贬义。图 5 为 T 的不同取值对词典性能指标的影响。在 $T=0$ 时,虽然召回率最高达到 88.58%,但准确率最低仅有 54.40%,F 值仅为 67.40%。当 $T=0.05$ 时,准确率提高到 77.75%,有较大提高,召回率仅下降到 87.61%,下降幅度较小,F 值提高到 82.39%。当 T 提高到 0.05 时性能指标达到最好,因此可以设定 T 为 0.05。



	0	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09	0.1
P	54.40%	60.88%	69.39%	74.59%	76.48%	77.75%	78.43%	79.35%	80.40%	81.55%	82.40%
R	88.58%	88.13%	88.01%	87.88%	87.78%	87.61%	86.58%	84.36%	82.15%	80.58%	79.66%
F	67.40%	72.01%	77.60%	80.69%	81.74%	82.39%	82.30%	81.78%	81.27%	81.06%	81.01%

图 5 不同 T 值时的性能指标

Fig. 5 Performance of different values of T

3.2.2 与其他词典性能对比

在 $T=0.05$ 时,SentiLex 准确率为 77.75%,接近最高的 DLLEX 词典 78.40%,而召回率为 87.61%,F 值为 82.39%,均为 4 个词典中最高。

表 1　$T=0.05$ 时的性能对比

Tab. 1　Performance comparison when $T=0.05$

	准确率(P)	召回率(R)	F 值
HowNet	74.55%	82.35%	78.26%
NTUSD	64.23%	80.27%	71.36%
DLLEX	**78.40%**	85.58%	81.83%
SentiLex	77.75%	**87.61%**	**82.39%**

4　结论

本文对情感词典构建相关研究进行了分析,以英文情感词典为基础,设计了基于语义关系的情感词典自动构建方法。方法以 HowNet、WordNet 语义词典和 SentiWordNet 情感词典为基础,借鉴英文情感词典进行中文情感词典的构建,并且与现有的常用情感词典进行了实验对比。实验结果表明,本文设计的方法取得了较好的评测性能。下一步工作中将重点从以下方面进行研究:如何扩展词典进行基于语料的情感词语选择和倾向性计算方法研究;研究如何利用语义词典对扩展的情感词语进行自动语义标注。

参考文献(References)

[1] Yuan S, Wang J, van der Meer M. Adaptive keywords extraction with contextual bandits for advertising on parked domains[J]. Computing Research Repository, 2013, abs/1307.3573.

[2] 张辉,李国辉,贾立,等. 一种基于 TF · IEF 模型的在线新闻事件探测方法[J]. 国防科技大学学报, 2013, 35(3): 55 – 60.
ZHANG Hui, LI Guohui, JIA Li, et al. On-line news event detection based on TF · IEF model[J]. Journal of National University of Defense Technology, 2013,35(3):55 – 60. (in Chinese)

[3] 刘健,李绮,刘宝宏,等. 基于话题模型的专家发现方法[J]. 国防科技大学学报, 2013, 35(2): 127 – 131.
LIU Jian, LI Qi, LIU Baohong, et al. An expert finding method based on topic model[J]. Journal of National University of Defense Technology, 2013,35(2):127 – 131. (in Chinese)

[4] Liu B. Sentiment analysis and opinion mining[J]. Synthesis Lectures on Human Language Technologies, 2012, 5(1): 1 – 167.

[5] 黄萱菁,张奇,吴苑斌. 文本情感倾向分析[J]. 中文信息学报, 2011, 25(6): 118 – 126.
HUANG Xuanjing, ZHANG Qi, WU Yuanbin. A survey on sentiment analysis [J]. Journal of Chinese Information Processing, 2011, 25(6): 118 – 126. (in Chinese)

[6] 赵妍妍,秦兵,刘挺. 文本情感分析[J]. 软件学报. 2010, 21(8): 1834 – 1848.
ZHAO YanYan, QIN Bing, LIU Ting. Sentiment analysis[J]. Journal of Software, 2010, 21(8): 1834 – 1848. (in Chinese)

[7] Stone P J, Dunphy D C, Smith M S. The General Inquirer: a computer approach to content analysis[M]. Cambridge: MIT Press, 1966.

[8] Taboada M, Grieve J. Analyzing appraisal automatically[C]// Proceedings of the AAAI Spring Symposium on Exploring Attitude and Affect in Text: Theories and Applications, 2004: 186 – 194.

[9] Baccianella S, Esuli A, Sebastiani F. SentiWordNet 3.0: An enhanced lexical resource for sentiment analysis and opinion mining[C]//Proceedings of the Seventh Conference on International Language Resources and Evaluation, 2010: 2200 – 2204.

[10] Agerri R, García-Serrano A. Q-WordNet: extracting polarity from WordNet senses[C]//Proceedings of the Seventh Conference on International Language Resources and Evaluation, 2010:1024 – 1029.

[11] 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. (in Chinese)

[12] 朱征宇,孙俊华. 改进的基于《知网》的词汇语义相似度计算[J]. 计算机应用, 2013,33(8): 2276 – 2279.
ZHU Zhengyu, SUN Junhua. Improved similarity calculation of words based on HowNet[J]. Journal of Computer Applications, 2013,33(8): 2276 – 2279. (in Chinese)

[13] 黄硕,周延泉. 基于知网和同义词词林的词汇语义倾向计算[J]. 软件. 2013, 34(2): 73 – 74, 94.
HUANG Shuo, ZHOU Yanquan. Semantic orientation computing based on HowNet&Cilin[J]. Computer Engineering & Software, 2013, 34(2): 73 – 74, 94. (in Chinese)

[14] 知网 HowNet 评价词词典[EB/OL]. http://www.keenage.com/html/c_index.html, 2013.
Appraisal lexicon of HowNet [EB/OL]. http://www.keenage.com/html/c_index.html, 2013. (in Chinses)

[15] Ku L W, Chen H H. Mining opinions from the web: beyond relevance retrieval[J]. Journal of the American Society for Information Science and Technology, 2007, 58(12): 1838 – 1850.

[16] 情感词汇本体库[EB/OL]. http://ir.dlut.edu.cn/EmotionOntologyDownload.aspx,2013.
Sentiment lexicon ontology[EB/OL]. http://ir.dlut.edu.cn/EmotionOntologyDownload.aspx,2013. (in Chinses)

[17] 刘群,李素建. 基于《知网》的词汇语义相似度计算[C]//第三届中文词汇语义学研讨会论文,2002.
LIU Qun, LI Sujian. Word similarity computing based on Hownet[C]//Proceeding of 3rd Chinese Lexical Semantics Workshop,2002. (in Chinses)

[18] Fellbaum C. WordNet: An electronic lexical database[M]. Cambridge: MIT Press, 1998.