基于句子级的领域倾向词表构建[1]

张小琴 蒋秀凤

（福州大学 数学与计算机科学学院，福建 350108）

（s030401422@163.com）

摘要： 领域倾向性词典有助于提高文本倾向性挖掘的精度，是文本倾向性挖掘研究的一个热门话题。本文分析了文本级算法构建词表的局限性，提出了一种基于句子级的领域词表构建算法。该算法引入了拉普拉斯平滑计算相关性，并将文档词频和逆文档频率的概念扩展到句子级，最后采用 IB 算法来对候选词进行聚类。采用该方法对酒店领域的语料进行领域倾向性词表的构建，得到了准确率为 71.55%的结果。

关键词： 倾向性检索；领域倾向词表；信息瓶颈算法

中图分类号：TP311　　文献标识码：A

Domain-oriented Sentiment Lexicon Based on Sentence-level Corpus

Zhang Xiaoqin, Jiang Xiufeng

(College of Mathematics and Computer Science, Fuzhou University,Fujian 350108,China)

（s030401422@163.com）

Abstract: The domain-oriented sentiment lexicon,which helps improve the accuracy of text opinion mining, is a hot top ic in text opinion mining. Th is pap er an alyzes the li mitations of th e text-level algorithm, and d escribes a s entence-level approach for constructing domain-oriented sentiment lexicon. The algorithm in this paper adopts the Laplace smoothing in calculating the relevance of words, and extends the d ocument term frequency and inverse document frequency of concept to the sentence level, then f inally clus ters th e can didate words b y IB a lgorithm. This m ethod is us ed to constru ct domain-oriented sentiment lexicon in the hotel area. And the accuracy we achieved in the testing corpus is 71.63%.

Key words: opinion retrieval; domain-oriented sentiment lexicon; information bottleneck algorithm

1 引言

随着网络技术的快速发展与应用的普及，越来越多的人通过博客或论坛，记录自己的心情、发表对某些事物的看法和见解，从而形成了海量的文本数据。如何自动地从海量数据中抽取出人们对某一事物的观点和看法，是文本倾向性挖掘所要解决的问题[12]。相同的词在不同的领域中有着不同的倾向性，同一种倾向性在不同的领域表达方式也不尽相同。例如：

这台电脑的内存配置有点低了，你最好别选它。（“低”，贬义）

接近赤道是低纬度地区。（“低”，无倾向性）

低声波是震动频率为 16 赫兹以下的震动波，比超声波所传的距离还远，在海洋考察、矿藏勘探、医疗和宇航方面有很大的应用价值，又称次声波。（“低”，无倾向性）

由此可见，领域知识对倾向性表达有着重要的影响。然而，当前的研究主要关注通用倾向词表的的构建，如英文的 SentiWordNet、General Inquirer 及中文的 HowNet 等。因而有必要构建领域倾向性词典。在语料的分析过程中发现，很多文档中存在不同倾向性的句子，并且含有大量

1. 基金项目：福建省科技创新平台计划项目(2009J1007)；福建省教育厅科研基金资助项目（JA04161）；福建省发展改革委员会基金资助项目（SX2004-29）。

作者简介：张小琴（1987-），女，福建宁德人，福州大学数学与计算机科学学院硕士研究生，研究方向：分布式计算，信息检索；蒋秀凤（1963-），女，江西丰城人，副教授，研究方向：分布式计算，网格计算，数据挖掘。

转折或否定的句子。倾向性文档的具有如下两个特点：(1)倾向性文档中含有相反倾向性的句子。(2) 倾向性文档中包含大量倾向性极性相反的词。因而本文提出了基于该特点的句子级聚类算法来构建倾向词表。

本文接下来的章节安排如下：第 2 节介绍了国内外的相关工作，第 3 节是算法模型，第 4 节是实验及相关分析，最后是结束语。

2 相关工作

目前，国内外对于创建词表大体上有两种方法：基于语义的以及基于语料库的方法[1]。

- 基于语义的分类方法，该方法是基于同义词或同义词库来决定句子的倾向性。 A. Esuli[2]等人使用PMI (Pointwise Mutual Information)方法来估计候选词与基准词(如“好”与“坏”)的相似性度。还有一类方法是基于一个现存的知识库，如英文的WordNet及中文的HowNet，来计算候选词与基准词对的语义距离，进而判断候选词的倾向性。J. Kamps[3]等人就是利用WordNet的同义结构图计算候选词与基准词的语义距离来得到其倾向性的。
- 基于语料库的机器学习方法，它是基于词在语料中的同现率来决定词的倾向性[13]。该算法前提是具有同种倾向极性的词更趋于同现。P. Turney[4]等人提出了PMI算法，其思想是先通过人工标注其中部分词作为基准词，相对应的未知倾向性的词称作候选词，通过计算候选词与基准词之间的相关性，再用分类器对候选词进行分类。M. Gamon[5] 在PMI算法中加入了假设，不同语义倾向的词不会出现在同一句子中，从而得到改进算法（PMI+SO)。以上算法只利用了词与词之间的相关性信息，而Weifu Du[6]等人针对上述方法提出了改进，在算法中添加了候选词与领域内外文档的相关性信息，并通过改进信息瓶颈算法聚类得到了较高的准确率。

但是，在语料的分析过程中发现，很多文档中存在不同倾向性的句子，并且含有大量转折或否定的句子，而这些因素会降低文档级算法的准确率。为了消除这些不利因素，本文提出了一种基于句子级的聚类算法来构建领域倾向词表。首先通过互信息计算词与词之间的相关性；然后将文档词频和逆文档频率的概念扩展到句子级来计算词与句子间的相关性；最后通过改进信息瓶颈算法对候选词进行聚类，得到候选词的语义倾向极性。

3 算法模型

本文提出的基于领域的倾向性词表构建算法主要包括两个步骤：(1) 创建三个矩阵，分别表示候选词与基准词、候选词与领域内句子之间的语义相关性；(2) 基于创建的三个矩阵，利用信息瓶颈聚类算法得到每个候选词的语义倾向，积极或消极。

3.1 相关性矩阵创建

3.1.1 候选词与基准词间的语义相关性

设t_c为候选词，t_s为基准词，则候选词集合可表示为$T_c=\left\{t_{c_i}\middle|1\le i\le m\right\}$；基准词集合为$T_s=\left\{t_{s_j}\middle|1\le j\le n\right\}$，计算词$t_{c_i}$与$t_{s_j}$之间的语义相似性可以有两种方法：基于语义的和基于语料库的[9]。本文使用互信息来计算词语之间的语义相似性[10][11]，其计算公式为：

$$I\left(t_{c_i},t_{s_j}\right)=\log\frac{N\times p\left(t_{c_i},t_{s_j}\right)}{p\left(t_{c_i}\right)\times p\left(t_{s_j}\right)} \qquad (3.1)$$

其中N表示文档数目，$p\left(t_{c_i}\right)$表示词t_{c_i}出现的概率；$p\left(t_{s_j}\right)$类似；$p\left(t_{c_i},t_{s_j}\right)$表示词$t_{c_i}$与$t_{s_j}$在大小固定的窗口内同时出现的概率。采用拉普拉斯平滑后可以得到

$$p\left(t_{c_i},t_{s_j}\right)=\frac{n\left(t_{c_i},t_{s_j}\right)+\lambda}{\sum_{t_i\in T_c,t_j\in T_s}\left[n\left(t_i,t_j\right)+\lambda\right]} \qquad (3.2)$$

其中，$n\left(t_{c_i},t_{s_j}\right)$表示在大小固定的窗口内同时出现词$t_{c_i}$与$t_{s_j}$的句子总数。$\lambda$为平滑参数。我们用矩阵$A=\left[A_{ij}\right]_{m\times 2}$来表示候选词与基准词之间的语义相关性，其中$A_{i1}$表示候选词与所有褒义的基准词之间的语义相关性；相应的，A_{i2}表示候选词与所有贬义的基准词之间的语义相关性。

3.1.2 候选词与领域内句子间的相关度

给定候选词集合$T=\left\{t_i\middle|1\le i\le m\right\}$和句子集合$S=\left\{s_j\middle|1\le j\le n\right\}$，我们可以根据文档中词频（Term Frequency）和逆文档频率(Inverse Document Frequency)的定义相应的给出句子中词频和逆句子频率的定义，并以此来计算候选词与句子之间的相关性强度$aff\left(t_i,s_j\right)$。若句子s_j中出现词t_i，则$aff\left(t_i,s_j\right)$用式（3.3）计算，否则$aff\left(t_i,s_j\right)$为零。

$$aff\left(t_i,s_j\right)=\frac{tf_{t_i}\times idf_{t_i}}{\sum_{t\in s_j}tf_t\times idf_t} \qquad (3.4)$$

其中，词频tf_t表示词t在句子中出现的次数；逆句子频率$idf_t=\log\frac{N}{n_t}$，N表示句子总数，n_t表示包含词t的句子数目。

类似地，我们用矩阵$B=\left[B_{ij}\right]_{m\times 2}$来表示候选词与句子之间的语义相关性，其中$B_{i1}$表示候选词与所有褒义句子之间的语义相关性；相应的，B_{i2}表示候选词与所有贬义句子之间的语义相关性。

根据候选词与基准词之间的语义相关性定义，我们可以按公式（3.4）计算候选词与句子之间的相关性强度

$$sim(t_i,s_j)=\sum_{t_s\in s_j\cap t_s\in T_s}sim(t_i,t_s) \qquad (3.5)$$

与矩阵 B 类似，可以得到矩阵$C=\left[C_{ij}\right]_{m\times 2}$来表示候选词与文档之间的语义相关性，其中$C_{i1}$表示候选词与所有褒义文档之间的语义相关性；相应的，C_{i2}表示候选词与所有贬义文档之间的语义相关性。

归一化：将矩阵A、B、C分别归一化为$\tilde{A}$ 、$\tilde{B}$、$\tilde{C}$。，使得矩阵的每一行之和为1。

3.2 倾向性词表的创建

信息瓶颈算法（IB）是由Slonim和Tishby[7]提出的一种聚类算法。本文中采用的是改进的信息瓶颈算法[14]。

输入：归一化矩阵 $\tilde{A}$，$\tilde{B}$，$\tilde{C}$

输出：各个候选词的语义倾向（即输出类别数目为3）

初始化：创建 $\tilde{W}_c=W_c$　　//W_c为候选词集合

$\forall i,j=1,2,\cdots,|W_c|,i<j$，

计算

$$d_{i,j}=\left(p\left(\tilde{w}_{c_i}\right)+p\left(\tilde{w}_{c_j}\right)\right)\left\{JS_{\Pi_2}\left[p\left(w_s|\tilde{w}_{c_i}\right),p\left(w_s|\tilde{w}_{c_j}\right)\right]+\alpha JS_{\Pi_2}\left[p\left(s_s|\tilde{w}_{c_i}\right),p\left(s_s|\tilde{w}_{c_j}\right)\right]+\alpha JS_{\Pi_2}\left[p\left(s_w|\tilde{w}_{c_i}\right),p\left(s_w|\tilde{w}_{c_j}\right)\right]\right\}$$

循环体：

对$k=|W_c|-1,\cdots,3$

- ✓ 找出一对$\{\alpha,\beta\}$对应的候选词为$\left\{\tilde{w}_{c_\alpha},\tilde{w}_{c_\beta}\right\}$，使得

$$d_{\alpha,\beta}=\arg\min_{i,j}\{d_{i,j}\}$$

- ✓ 合并$\left\{\tilde{w}_{c_\alpha},\tilde{w}_{c_\beta}\right\}\Rightarrow\tilde{w}_{c_*}$，

$$\begin{cases}p\left(\tilde{w}_{c_*}\right)=p\left(\tilde{w}_{c_\alpha}\right)+p\left(\tilde{w}_{c_\beta}\right)\\ p\left(y|\tilde{w}_{c_*}\right)=\dfrac{1}{p\left(\tilde{w}_{c_*}\right)}\left(p\left(\tilde{w}_{c_\alpha},y\right)+p\left(\tilde{w}_{c_\beta},y\right)\right)\\ p\left(\tilde{w}_{c_*}|x\right)=1,\text{若}x\in\tilde{w}_{c_\alpha}\cup\tilde{w}_{c_\beta};\text{否则为}0\end{cases}$$

- ✓ 更新$\tilde{W}_c=\left\{\tilde{W}_c-\left\{\tilde{w}_{c_\alpha},\tilde{w}_{c_\beta}\right\}\right\}\cup\left\{\tilde{w}_{c_*}\right\}$
- ✓ 更新所有与α和β有关的$d_{i,j}$，并打印$\tilde{w}_{c_*}$

结束

图1 信息瓶颈算法

在信息瓶颈算法中，采用条件概率$p(y|x)$和$p(y|c)$之间的KL-距离(Kullback-Leibler divergence)及JS距离来计算合并损失。

$$D_{KL}\left[p(y|x)\|p(y|c)\right]=\sum_y p(y|x)\log\frac{p(y|x)}{p(y|c)}\qquad(3.6)$$

$$JS_{\Pi_2}\left[p_i,p_j\right]=\pi_i D_{KL}\left[p_i\|\hat{p}\right]+\pi_j D_{KL}\left[p_j\|\hat{p}\right]\qquad(3.7)$$

其中

$$\begin{cases}\{p_i,p_j\}=\left\{p\left(y|c_i\right),p\left(y|c_j\right)\right\}\\ \{\pi_i,\pi_j\}=\left\{\dfrac{p\left(c_i\right)}{p\left(c_*\right)},\dfrac{p\left(c_j\right)}{p\left(c_*\right)}\right\}\\ \hat{p}=\pi_i p\left(y|c_i\right)+\pi_j p\left(y|c_j\right)\end{cases}\qquad(3.8)$$

本文在在传统的IB算法中增加了更多的领域内信息，使得聚类结果更加的准确，算法如图

1 所示。本文采用改进后的算法可将候选词聚类为三个词集：褒义、贬义和中性词集。

4 实验及相关分析

4.1 实验数据

本文的语料采自网络的酒店评论（www.ctrip.com），包括 3509 个褒义句子和 4215 个贬义句子。分句、分词采用哈尔滨工业大学信息检索研究室提供的 LTP 分句工具和中国科学院计算技术研究所提供的 ICLCAS 分词工具。

我们抽取所有词性为的 a、ad、an、al、vl 的词作为候选倾向性词，简称候选词。并手动选取褒贬各 40 个基准词，具体选取基准词集的步骤如下：

（1）先将语料进行分句、分词，提取候选词集。

（2）统计正、反面倾向语料中各词汇出现的次数。

（3）将候选词集的各词汇的在褒贬义语料中出现的次数按降序排列。

（4）求候选词集与情感词汇表的交集再人工选出成对的词对构成基准词集。

我们在语料库的所有候选词集中人工标识褒义倾向词 337 个，贬义倾向词 228 个，作为基准倾向词表。

4.2 评价指标

这里采用准确率来评价算法性能。设W_c为候选词集，C为词w的实际倾向极性（即基本词表中的极性），F为本文算法聚类得到的词w的倾向极性。准确率定义为：

$$Accuracy(w)=\frac{\{w \mid w \in W_c \wedge C(w)=F(w)\}}{|W_c|} \qquad (3.9)$$

4.3 实验结果

本文采用句子级的算法进行领域词表构建，并在算法中加入了对转折句和否定句的处理，实验结果采用人工进行评判。

其中对对转折句和否定句采用以下两种方式处理：

（1）对于转折句，计算时只提取转折词之后的倾向词；

（2）对于否定句，在标记句子倾向极性时标记为相反的极性。

实验对转折句和否定句进行处理前后的结果进行比较，具体数据如表 1：

表 1　实验结果

酒店语料	候选词总数	正确结果	准确率
处理前	921	573	62.21%
处理后	921	659	71.55%

由表 1 可以看出增加对否定句与转折句的处理能够有效的提高实验准确率。

本文的算法聚类得到褒义倾向词：“方便、便宜、精致、典雅、好吃，洁净、近、别有情趣、宾至如归、好客、热情等”；贬义倾向词：“碍眼、冷淡、肮脏、陈旧、暗、贵、一般等”。由此可以看出，本算法可以经过聚类有效的得到形容词的倾向性。另外，对于以下词：“别扭、迟钝、粗鲁、低级、丢人、肥胖等”由于在酒店语料中并不常见，使得不能准确聚类，可以通过增加领域外的语料进行改善。

5 结束语

本文分析了文本级算法构建词表的局限性，提出了一种基于句子级的领域词表构建算法。该算法引入了拉普拉斯平滑计算相关性，并将文档词频和逆文档频率的概念扩展到句子级，最后采用 IB 算法来对候选词进行聚类。采用该方法对两种不同领域的语料进行领域倾向性词表的构建，得到了准确率为 71.63%的准确率。但由于时间原因，本文研究的领域倾向词表的构建算法还不

够完善。接下来还有许多工作要做：

1. 将算法与 Weifu Du 的算法进行比较，根据实验评估结果修改算法等。
2. 可以再算法中加入领域外的信息，得到领域独立的倾向词表。

6 致谢

感谢中国科学院计算技术研究所为我们这个任务提供 ICTCAS 分词工具和哈尔滨工业大学信息检索研究室为我们本文工作提供 LTP 分句工具。

感谢廖祥文老师，他在本文工作的创作和实验过程中给了我很多建议。

参考文献

[1] 肖伟. 基于语义的 BLOG 社区文本倾向性分析. 上海交通大学. 2007. 12

[2] A. Esuli and F. Sebastiani. Determining the semantic orientation of terms through gloss classification. CIKM 2005

[3] J. Kamps, M. Marx, R. Mokken, etc. Using WordNet to measure semantic orientation of adjectives. LREC 2004

[4] P. Turney and M. Littman. Measuring Praise and Criticism: Inference of Semantic Orientation from Association. In: ACM Transactions on Information Systems. 2003.

[5] M. Gamon and A. Aue. Automatic id entification of sentiment vocabulary exploiting low association w ith known sentiment terms. ACL 2005 .

[6] Weifu Du, Songbo T an: Bu ilding Dom ain-oriented Sent iment Lexicon b y Im proved Inform ation B ottleneck . CIKM 2009: 1749-1752.

[7] N. Slonim and N. Tishby. Agglomerative information bottleneck. NIPS 1999.

[8]N.Tishby,W.Bialek,andF.C.Pereira.The information bottleneck method.Yet unpublished manuscript, NEC Research Institute TR, 1998.

[9] S. Tan, X. Ch eng, Y. Wang, H. Xu. Adapting Naive Bayes to Dom ain Adaptation for Sentiment Analysis. ECIR 2009.

[10] Q. W u, S. Tan, H. Zhai, G. Zhang, M. Duan an d X. Ch eng. SentiRank: Cr oss-Domain Graph Ranking for Sentiment Classification. WI 2009.

[11] S. Tan, G. Wu, H. T ang and X. Cheng. 2007 . A nov el scheme for domain-trans fer problem in th e context o sentiment analysis. In Proceedings of CIKM 2007.

[12] Seung-Hoon Na, Yeha Lee, Sang-Hyob Nam, and Jong-Hyeok Lee. Improving Opinion R etrieval Based on Query-Specific Sentiment Lexicon. ECIR 2009.

[13] H. Kanay ama and T. Nasukawa. Fully Automatic Lexicon Expansion for Domain-oriented Sentiment Ana lysis. EMNLP 2006.

[14] N. Tishby, W. Bialek, and F. C. Pereira. The information bottleneck method: Extracting relevant information from concurrent data. Yet unpublished manuscript, NEC Research Institute TR, 1998.

[15] A. Kenn edy and D. Inkp en. Sentiment Classif ication of Movie Rev iews Using Contextual Valence Shifters . Computational Intelligence. 2006.