文章编号：1003-0077(2011)02-0089-05

基于概率潜在语义分析的词汇情感倾向判别

宋晓雷[1]，王素格[1,2]，李红霞[3]，李德玉[1,2]

(1. 山西大学 计算机与信息技术学院，山西 太原 030006；
2. 山西大学 计算智能与中文信息处理教育部重点实验室，山西 太原 030006；
3. 山西大学 数学科学学院，山西 太原 030006)

摘 要：该文利用概率潜在语义分析，给出了两种用于判别词汇情感倾向的方法。一是使用概率潜在语义分析获得目标词和基准词之间的相似度矩阵，再利用投票法决定其情感倾向；二是利用概率潜在语义分析获取目标词的语义聚类，然后借鉴基于同义词的词汇情感倾向判别方法对目标词的情感倾向做出判别。两种方法的优点是均可在没有外部资源的条件下，实现词汇情感倾向的判别。

关键词：概率潜在语义分析；数据稀疏；语义聚类；情感倾向

中图分类号：TP391　　**文献标识码**：A

Word Sentiment Orientation Discrimination Based on PLSA

SONG Xiaolei[1], WANG Suge[1,2], LI Hongxia[3], LI Deyu[1,2]

(1. School of Computer & Information Technology, Shanxi University, Taiyuan, Shanxi 030006, China;
2. Key Laboratory of Computational Intelligence and Chinese Information Processing of Ministry of Education, Shanxi University, Taiyuan, Shanxi 030006, China;
3. Department of Mathematics Science, Shanxi University, Taiyuan, Shanxi 030006, China)

Abstract: This paper proposes two kinds of methods to determine the sentiment orientation of a word based on Probabilistic Latent Semantic Analysis(PLSA). In the first method, the similarity matrix between target words and paradigm words is obtained by PLSA, and the polarity of each target word is then determined by voting. In the second method, we obtain the semantic cluster of target words by PLSA, and the polarity of a target word is then determined by a synonym-based method. The advantage to both methods lies in that they can work well without any external knowledge resources.

Key words: probabilistic latent semantic analysis; sparse data; semantic clustering; sentiment orientation

1 引言

在网络信息爆炸的今天，利用计算机自动分析大规模文本的情感倾向技术，在市场营销、客户关系管理以及政府舆情分析等诸多领域有着广阔的应用空间和发展前景。然而，词汇作为语言学的一个基本语义单位，其情感倾向的判别对更大语言粒度的情感倾向性研究有着非常重要的作用[1]。因此，对词汇的褒贬倾向判别是篇章情感倾向研究工作的基础。

Turney[2]使用 PMI-IR 方法研究词汇的情感倾向性，利用点互信息表示目标词与基准词之间的关联强度，进而求出目标词的情感倾向；文献[3]利用 WordNet 计算词汇倾向性，先选择基准词，然后判别待定词与基准词在 WordNet 中是否为同义词，

收稿日期：2010-08-30　**定稿日期**：2010-11-30

基金项目：国家自然科学基金资助项目(60875040，60970014)；教育部高等学校博士点基金资助项目(200801080006)；山西省自然科学基金资助项目(2007011042，2010011021-1)；太原市科技局明星专项资助项目(09121001)；山西省科技攻关项目

作者简介：宋晓雷(1980—)，男，硕士生，主要研究方向为信息抽取；王素格(1964—)，女，博士，副教授，主要研究方向为自然语言理解，信息抽取；李红霞(1983—)，女，硕士生，主要研究方向为自然语言处理。

得出词汇的倾向性；徐琳宏等[4]采用 HowNet 作为基准词，通过计算目标词与基准词的关联度，确定目标词汇的语义倾向；文献[1]对基准词的选取进行了研究(采用 Fisher 准则)，并进一步考虑目标词与其同义词的关系，提出了基于同义词的词汇情感倾向判别方法，该方法不仅考虑了目标词与基准词的关联强度，而且也考虑了目标词的同义词与基准词的关联强度，取得了不错的效果。此外，复旦大学[5]、香港城市大学[6]、中国科学院自动化研究所[7]都进行了相关的研究。

在自然语言处理中，数据稀疏一直是困扰人们的一大问题。对语料规模较小或单纯考察一个词与褒贬义基准词集的同现信息时更容易遇到数据稀疏问题，而数据稀疏问题制约着实验性能的提高。文献[3]的研究发现其性能随着语料规模的减小而急剧变差，当测试集为 2 697 词时，其在 20 亿个词的语料规模上准确率为 83.98%，当语料规模减至 1 000 万个词时，其准确率迅速减为 63.40%，由此，揭示了数据稀疏问题能严重影响实验的性能。文献[1]利用同义词信息在一定程度上解决了数据稀疏问题。文献[4]则采用了扩大基准词范围的策略来解决数据稀疏问题，然而上述研究[1,3-4]都需要用到外部资源(同义词词林、WordNet、HowNet)，外部资源的有限性将会限制其推广性。本文在较小规模的语料上(语料规模为 1 006 篇文档，共有 570 506 个词次)，利用概率潜在语义分析，给出了用于词汇情感倾向判别的两种方法，一定程度上解决了数据稀疏问题。

2 概率潜在语义分析对称参数表示模型

(1) 参数表示模型

概率潜在语义分析(PLSA)最初是 Hoffmann[8]在潜在语义分析(LSA)的基础上提出的一种新方法。该方法引入潜在语义空间概念，使用概率模型来衡量“文档—潜在语义—词”三者之间的关系，文档和词都可以通过计算语义空间上的夹角而得以量化，PLSA 采用了迭代算法来实现，其模型为 PLSA 的对称参数模型(如图 1 所示)。和 LSA(潜在语义分析)相比，PLSA 有明确的物理意义，多义词和同义词的现象均可在潜在的语义空间中得到合理的表示。本文在文献[9-10]的基础上，将 PLSA 的对称参数模型进一步泛化，概括如下：

给定两个集合 $A=\{a_i\}_{i=1}^{s}$ 和 $B=\{b_i\}_{i=1}^{t}$ (A, B 可以代表文档集或特定词集等)以及一个 A 和 B 的索引矩阵 $(m(a_i,b_j))_{s\times t}$，其中 s, t 分别表示集合 A 与集合 B 元素的个数；LSA 利用奇异值分解得到语料中词汇间的统计关系，首先构造词—文档矩阵 A，再对 A 进行 SVD 分解，使得 $A=U\sum V^T$，U, V 为列正交矩阵。类似地，对 PLSA 构造初始映射矩阵 $(p(a_i,z_j))_{s\times k}$ 和 $(p(b_i,z_j))_{t\times k}$，保证任意一行之和等于一。

PLSA 假设“A-B”对之间是条件独立的，并且潜在语义在 A 或 B 上分布也是条件独立的。在上面假设的前提下，根据图 1 所示的模型，依据公式(1)计算出的概率产生每一个观测对 (a,b)。

$$P(a,b)=\sum_{z\in Z}P(a\mid z)P(z)P(b\mid z) \quad (1)$$

其中，$P(a|z)$, $P(b|z)$分别为潜在语义在 A 上和 B 上的分布概率。Z 表示 k 维潜在语义空间，k 为一个经验常数。

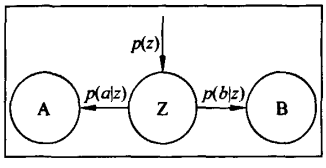

图 1 PLSA 对称参数模型

(2) EM 算法

概率潜在语义分析使用最大期望(Expectation Maximization，EM)算法对潜在语义模型进行拟合，在初始化数据基础上，交替实施 E 步骤和 M 步骤迭代计算。

在 E 步骤中，计算出在每一对 的条件下产生潜在语义块 z 的先验概率；

在 M 步骤中，对模型重新估计；

直到如式(2)所示的似然函数 L 的变化小于某一个给定的阈值，即可认为达到了最优解。

$$L=\sum_{a\in A}\sum_{b\in B}m(a,b)\log P(a,b) \quad (2)$$

其中 $m(a,b)$ 表示 a 和 b 在限定的范围内共现的次数(如果 A, B 分别为词集和文档集，则 $m(a,b)$ 表示词 a 在文档 b 中出现的次数；若 A, B 同为词集，则表示它们在文档的某一限定长度内共现的次数)。

3 基于概率潜在语义分析的词汇情感倾向判别 (6)

将 $A=\{j\_word_i\}_{i=1}^{m}$ 表示基准词集，$B=\{t\_word_i\}_{i=1}^{n}$ 表示目标词集，通过 EM 算法可以获得最优化的 $(z_i)_{k\times 1}$，$(t\_word_i,z_j)_{n\times k}$，$(j\_word_i,z_j)_{m\times k}$ 三个矩阵（m,n,k 分别代表基准词、目标词和语义块的个数）。此时，利用公式(1)可求得 A,B 之间的相似矩阵 $(p(a_i,b_j))_{m\times n}$。进一步利用公式(3)可求得 A,A 之间的相似矩阵 $(p(a_i,a_j))_{m\times m}$。利用相似矩阵 $(p(a_i,b_j))_{m\times n}$ 和 $(p(a_i,a_j))_{m\times m}$，得到两种用于判别词汇情感倾向的方法。

$$(p(a_i,a_j))_{m\times m}=(p(a_i,b_j))_{m\times n}\cdot(p(a_i,b_j))'_{m\times n} \quad (3)$$

3.1 基于词汇相似度的词汇情感倾向判别

目标词和基准词之间相似度利用目标词和基准词之间的相似矩阵 $(p(t\_word_i,j\_word_i))_{n\times m}$ 来度量。

词汇情感倾向类别确定：利用投票法对每个目标词 t_word 的情感倾向 $SO(t\_word)$ 进行判别。其思想为与目标词 t_word 相似度最高的前 k 个基准词中，具有相同倾向类别最多的基准词所在类别为该 t_word 倾向性，其形式化如下：

$$SO(t\_word)=arg\ \underset{i}{most}(SO(j\_word_i)) \quad (4)$$

其中 $j\_word_1,\cdots,j\_word_k$ 为与目标词 t_word 相似度最高的前 k 个基准词。

3.2 基于语义聚类的词汇情感倾向判别

(1) 目标词 t_word 的同义词集合：利用目标词和目标词之间的相似矩阵 $(m(t\_word_i,t\_word_j))_{n\times n}$，自动找到与每个目标词相似度最高的前 k 个目标词集 $\{t\_word_1,\cdots t\_word_k\}$，将其看作目标词 t_word 的相同语义聚类集合。

基于语义聚类的词汇情感倾向强度：

$$FSO\_PMI(t\_word)=\alpha SO\_PMI(t\_word)+(1-\alpha)\sum_{i=1}^{k}SO\_PMI(t\_word_i) \quad (5)$$

其中，$SO\_PMI(t\_word)$ 是利用文献[2]的方法计算每个目标词与基准词集的关联强度。

词汇情感倾向类别确定：每个目标词 t_word 的情感倾向 $SO(t\_word)$ 由判别公式(6)来决定。

$$SO(t\_word)=\begin{cases}\text{褒义} & FSO\_PMI(t\_word)\geqslant\lambda\\ \text{贬义} & FSO\_PMI(t\_word)<\lambda\end{cases}$$

其中 λ 为经验阈值。

(2) 对目标词—基准词索引矩阵 $(m(t\_word_i,j\_word_j))_{n\times m}$ 的不同优化策略。

如果在限定的窗口内目标词和所有基准词均没有同现关系，则目标词—基准词索引矩阵对这部分目标词不能提供任何信息，使这部分目标词的情感倾向无法判别。为此，本文采用两种优化策略。

策略 1："简单策略"，即仅仅扩大同现窗口；

策略 2："融合策略"，即仅对目标词—基准词索引矩阵无法提供信息的目标词进行改进，扩大其窗口，同时对其增加惩罚因子，用于弱化由于扩大窗口所带来的噪声，其他词语处理策略保持不变。其"融合策略"如下：

对目标词—基准词索引矩阵中的全零行，扩大其同现窗口，对求得的新非零矩阵中的元素 $m(t\_word_i,j\_word_j)'$，再利用公式(7)，求得最终 $m(t\_word,j\_word)$ 的值。

$$m(t\_word,j\_word)=floor(\alpha\cdot m(t\_word,j\_word)') \quad (7)$$

其中 $floor(.)$ 为取整算子，为惩罚因子，文中的选取由试验确定。

4 实验结果及其分析

4.1 实验数据与评价标准

实验数据采用文献[1]所提供的语料，语料规模为 1 006 篇文档，570 506 个词次，正面文本 576 篇，反面文本 430 篇，测试数据共有 2 958 个目标词，包括形容词、副词、名词和动词四种类别。

本文对基准词的选取不再做深入研究，参照文献[1]所选用的基准词。评价对象的评价指标：采用精确率、召回率和 F 值；PR、PP、PF、NR、NP、NF、OF 分别表示正面召回率、正面精确率、正面 F 值、反面召回率、反面精确率、反面 F 值、总体 F 值。

4.2 实验结果与分析

为了验证各种情况下词汇情感倾向判别结果，进行了 4 个实验。"PMI-IR"表示文献[2]中的方法；"方法 1"为第 3.1 节中介绍的基于词汇相似度的词汇情感倾向判别；"去零行"表示 PMI-IR 或方

法1中去除无法用相似矩阵计算与基准词相似比较的目标词。“方法2”为第3.2节中介绍的基于语义聚类的词汇情感倾向判别。

实验1：为了验证不同数目的潜在语义块数目 k 对基于目标词—基准词表示模型的词汇情感倾向判别实验性能的影响，利用方法1进行实验，其结果如表1所示。

表1 不同潜在语义块的实验结果

指标 \ K值	K=20	K=40	K=60	K=80
PR/%	64.42	65.91	67.44	64.72
PP/%	75.14	74.47	74.33	74.46
PF/%	69.63	69.93	70.72	69.25
NR/%	54.26	51.49	50.00	52.34
NP/%	41.53	41.30	41.70	40.86
NF/%	47.05	45.83	45.48	45.90
OF/%	61.69	61.33	61.91	60.78

观察表1可知：k 的取值并非越大越好。若 k 太大，则潜在语义块太多，使其粒度过小，失去了采用潜在语义分析的作用。因此，本文取60。

实验2：为了验证不同目标词—基准词相似矩阵的情况下，PMI-IR方法和方法1对目标词极性判断的影响，对其进行实验，实验结果见表2。

表2 两种方法在不同相似矩阵情况下的词汇情感倾向判别实验结果

指标 \ 方法	PMI-IR		方法1	
	未去零行	去零行	未去零行	去零行
PR/%	73.49	60.90	67.44	87.33
PP/%	65.62	76.70	74.33	74.57
PF/%	69.63	67.89	70.72	80.44
NR/%	17.34	57.37	50.00	31.39
NP/%	23.35	38.91	41.70	51.81
NF/%	19.90	46.37	45.48	39.10
OF/%	55.65	59.83	61.90	70.39

由表2可知，PMI-IR直接利用目标词 t_word 与基准词集的相关性来判别其情感倾向，效果不能令人满意，主要原因在于很多目标词与基准词并不在限定的窗口内同现或者仅与极少数基准词同现，使得与这些目标词对应的 $PMI(t\_word, j\_word)$ 值没有意义。利用“方法1”词汇情感倾向判别结果在一定程度上得到了提高，但对于目标词与基准词不同现的情形仍无能为力，当去除这部分目标词后，“方法1”方法的性能将得到了极大的提升，说明了此方法用于那些与目标词同现的词汇的极性判别是有效性。方法1中目标词和基准词的相似性举例如下(按与目标词相似度的降序排列)：哀鸣：撞击、郁闷、缺陷、故障、断裂、失望、倒、降低、担心。

实验3：为了验证采用不同优化策略时方法2的性能，对其进行实验，其实验结果如表3所示。

表3 不同优化策略得到词汇情感倾向判别实验结果

指标 \ 优化策略	为扩大同现窗口	策略1	策略2
PR/%	75.22	85.03	86.92
PP/%	69.95	70.91	72.81
PF/%	72.49	77.33	79.24
NR/%	30.61	25.11	30.32
NP/%	36.50	43.87	51.91
NF/%	33.29	31.94	38.28
OF/%	61.03	66.00	68.93

由表3可知，“方法2”试图对目标词进行语义聚类，然而目标词极性分布的不均衡性(目标词共计2 958个，其中正面2 018个，反面940个)导致了各目标词的语义聚类词集中正面词居多，使其正面的召回率提高。在限定的窗口内，有695个目标词(约占总目标词的23.49%)和所有基准词没有任何同现关系，因此目标词—基准词索引矩阵对这部分目标词不提供任何信息，使得这部分目标词的极性无法有效判。“策略1”扩大了同现窗口，一定程度上解决了数据稀疏问题，但由于窗口的扩大使原来没有同现关系的词汇取得了同现关系，也会带来了噪声，因此，采用“策略2”对目标词进行扩大窗口时，增加了惩罚因子，实验结果表明，其性能达到了提升。

总体上，方法1和方法2相比，方法简单，其索引矩阵的阶数由2 958×2 928降为2 958×80，时间和空间复杂度低，总体性能不低于“方法2”的性能，但对于与所有基准词没有任何同现关系的那些目标词却不能进行情感倾向判别，因此，也有一定的局限性。

5 结束语

本文给出了基于PLSA的词汇倾向判别方法，

该方法在没有任何外部资源的条件下，只需利用少量的基准词，可以解决了语料规模较小时数据稀疏问题。当语料规模为1 000万个词时，文献[3]对于2 697个测试词汇情感倾向判别的准确率迅速减为63.40%，而本文的方法在较小的语料规模上(语料规模不足60万个词)对2 958个测试词汇情感倾向判别的准确率却达到了68.93%。本文所提的方法在性能上还有一定地提升空间。例如，仅假定在一定的上下文范围内，词汇的情感倾向具有连续性，然而，由于转折连接词和否定副词的使用，可对词汇的情感倾向产生影响，下一步可以将这种情形考虑在内，并将中性词也加入到的倾向判别的相关研究中。

参考文献

[1] 王素格，李德玉，魏英杰，等. 基于同义词的词汇情感倾向判别方法[J]. 中文信息学报，2009，23(5)：68-74.

[2] PETER D. Turney: Thumbs up or thumbs down? Semantic orientation applied to unsupervised classification of reviews[C]//Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL) Philadelphia, PA, USA. 2002：417-424 .

[3] Kamps J., M. Marx, R. J. Mokken, and M. D. Rijke. Using WordNet to measure semantic orientation of adjectives [C]//Proceedings of 4th International Conference on Language Resources and Evaluation. Lisbon, 2004：1115-1118.

[4] 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.

[5] 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.

[6] YUEN Raymond W. M., CHAN Terence Y. W., LAI Tom B. Y. et al. Morpheme-based derivation of bipolar semantic orientation of Chinese words [C]//Proceedings of the 20th International Conference on Computational Linguistics. Geneva, Switzerland. 2004：1008-1014.

[7] 王根，赵军. 中文褒贬义词汇倾向性的分析[C]//Proceedings of SWCL2006，沈阳，2006：81-85.

[8] Hofmann T. Probabilistic latent semantic indexing [C]//Proceedings of the 22nd International Conference on Research and Development in Information Retrieval. Berkeley, California, 1999：50-57.

[9] 金千里，赵军，徐波. 弱指导的统计隐含语义分析及其在跨语言信息检索中的应用[C]//Proceedings of CNCCL2003，哈尔滨，2003：527-532.

[10] Hofmann T.. Unsupervised learning by probabilistic latent semantic analysis [J]. Machine Learning, 2001，42：177-196.

基于概率潜在语义分析的词汇情感倾向判别

作者： 宋晓雷， 王素格， 李红霞， 李德玉， SONG Xiaolei， WANG Suge， LI Hongxia， LI Deyu

作者单位： 宋晓雷,SONG Xiaolei(山西大学计算机与信息技术学院,山西太原,030006)， 王素格,李德玉,WANG Suge,LI Deyu(山西大学计算机与信息技术学院,山西太原030006;山西大学计算智能与中文信息处理教育部重点实验室,山西太原030006)， 李红霞,LI Hongxia(山西大学数学科学学院,山西太原,030006)

刊名： 中文信息学报 ISTIC PKU

英文刊名： JOURNAL OF CHINESE INFORMATION PROCESSING

年，卷(期)： 2011,25(2)

参考文献(10条)

1. Hofmann T Unsupervised learning by probabilistic latent semantic analysis 2001

2. 金千里;赵军;徐波 弱指导的统计隐含语义分析及其在跨语言信息检索中的应用 2003

3. Hofmann T Probabilistic latent semantic indexing 1999

4. 王根;赵军 中文褒贬义词汇倾向性的分析 2006

5. YUEN Raymond W.M;CHAN Terence Y.W;LAI Tom B.Y Morpheme-based derivation of bipolar semantic orientation of Chinese words 2004

6. 朱嫣岚;闵锦;周雅倩 基于HowNet的词汇语义倾向计算 2006(01)

7. 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 2007(01)

8. Kamps J;M.Marx;R.J.Mokken;M.D.Rijke Using WordNet to measure semantic orientation of adjectives 2004

9. PETER D Turney:Thumbs up or thumbs down? Semantic orientation applied to unsupervised classification of reviews 2002

10. 王素格;李德玉;魏英杰 基于同义词的词汇情感倾向判别方法 2009(05)