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[1]

山西大学数学科学学院 太原 030006[1],

山西大学计算智能与中文信息处理教育部重点实验室 太原 030006[2]

E-mail: wsg@sxu.edu.cn

摘 要：本文利用概率潜在语义分析，给出了两种用于判别词汇的情感倾向的方法。一种是使用概率潜在语义分析获得每个目标词和基准词之间的相似度矩阵，再利用投票法决定每个目标词的情感倾向；二是利用概率潜在语义分析对目标词进行语义聚类和扩展，自动找到每个目标词的同义词，然后采用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。这两种方法的优点均在没有外部资源的条件下，可以实现情感倾向的判别。

关键词：概率潜在语义分析，数据稀疏，语义聚类，情感倾向

Word Sentiment Orientation Discriminating Based on PLSA

Song Xiaolei[1]， Wang Suge[1,2]， Li Hongxia[1]

Department of Mathematics Science, Shanxi University, Taiyuan 030006[1]

Key Laboratory of Computational Intelligence and Chinese Information Processing of Ministry of Education, Shanxi University, Taiyuan, 030006[2]

E-mail: wsg@sxu.edu.cn

Abstract: This paper proposes two different methods to determine the sentiment orientation of the words, which are both based on Probabilistic Latent Semantic Analysis（PLSA）: 1, By using PLSA, we first obtain the similarity-matrix between the target words and basic words, then determine each target word's polarity through a poll; 2, By making the use of PLSA, we first make a semantic cluster and expansion for the target words, find the synonyms of each target word automatically, then, determine the target word's sentiment orientation by exploiting its synonyms. The advantages of the two methods are that the word sentiment orientation discriminating is realized without employing any external knowledge resources.

Keywords: 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中的情感倾向的技术，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，对比了 PMI-IR 和 LSA 两种方法，实验结果表明：PMI-IR 方法优于 LSA 方法；文献[3]利用 WordNet 计算词汇倾向性，先选择基准词，然后判别

*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(200801080006)；山西省自然科学基金资助项目(2007011042, 2010011021-1)；山西省重点实验室开放基金资助项目(2007031017)；太原市科技局明星专项(09121001)。

待定词与基准词在 WordNet 中是否为同义词，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]、中科院自动化所[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，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 2697 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 1000 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严重地影响了实验的性能。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了其推广性；本文在较小规模的语料上(语料规模为 1006 篇文档，共有 570506 个词次)，利用概率潜在语义分析算法，给出了两种用于词汇情感倾向判别的方法，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

2.1 参数表示模型

概率潜在语义分析(PLSA)最初是 Hoffmann[11]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（如图一所示）。和 LSA 相比，PLSA 有明确的物理意义，多义词和同义词的现象均可以在潜在的语义空间中得到合理的表示。本文在文献[8,9]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。

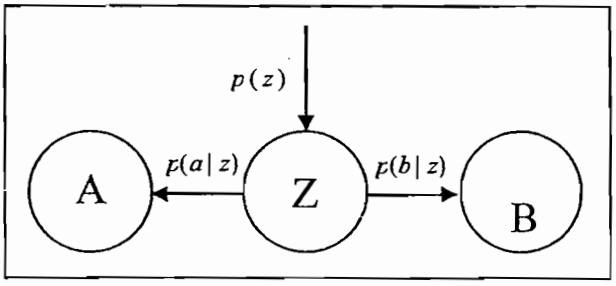

图 1：PLSA 对称参数模型

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，根据图一所示的模型，依据（1）计算出的概率产生每一个观测对（a,b）。

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

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。

2.2 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(a, b)的条件产生潜在语义块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。

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

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；若同为词集，则表示词a和词b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

本文中$A=\{j_word_i\}_{i=1}^{m}$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，再次利用公式（1）可以求得A、B之间的相似度矩阵$(p(a_i,b_j))_{m\times n}$。进一步利用公式（3）可求得A、A之间的相似度矩阵$(p(a_i,a_j))_{m\times m}$。本文利用相似度矩阵$(p(a_i,b_j))_{m\times n}$和$(p(a_i,a_j))_{m\times m}$，分别给出以下两种用于判别词汇情感倾向的方法。

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

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度矩阵$(p(t_word_i,j_word_i))_{n\times m}$来度量。

词汇情感倾向类别确定：对每个目标词t_word的情感倾向$SO(t_word)$利用投票法进行判别，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。模型用公式（4）来表示。

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

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

3.2 基于同义词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度矩阵 $(m(t\_word_i, t\_word_j))_{n\times n}$，自动找到与每个目标词相似度最高的前 k 个目标词集 $\{t\_word_1,\ldots t\_word_k\}$ 做为目标词 t_word 的同义词集合.

基于同义词的词汇情感倾向强度：定义 t_word 的最终词汇情感倾向强度为：

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

其中，$SO\_PMI(t\_word)$ 是利用 PMI-IR[3]方法计算的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。

$$SO(t\_word) = \begin{cases} \text{褒义} & FSO\_PMI(t\_word) \geq \lambda \\ \text{贬义} & FSO\_PMI(t\_word) < \lambda \end{cases} \quad (6)$$

其中 λ 为经验阈值。

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

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，本文对目标词—基准词索引矩阵给用以下两种优化策略。

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，其它词语处理策略保持不变。具体"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），求得最终的值 m(t_word, j_word) 。

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

其中 floor(.)为取整算子，α 为惩罚因子，文中α 的选取由试验确定。

4 实验结果及其分析

4.1 实验数据与评价指标

实验数据采用文献[1]所提供的语料，语料规模为 1006 篇文档，570506 个词次，正面文本 576 篇，反面文本 430 篇，测试数据共有 2958 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 2 节中的基于词汇相似度的词汇情感倾向判别；“去零行 1”与“去零行 2”分别表示 PMI-IR 和方法 1 中去除无法用相似度矩阵计算与基准词相似比较的目标词。“方法 2”为第 2 节中的基于同义词的词汇情感倾向判别。

实验 1：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。利用方法一进行实验，其结果如表 1 所示。

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

表 1：实验 1 的实验结果

评价指标 / K 值	PR%	PP%	PF%	NR%	NP%	NF%	OF%
K=20	64.42	75.14	69.37	54.26	41.53	47.05	61.19
K=40	65.91	74.47	69.93	51.49	41.30	45.83	61.33
K=60	67.44	74.33	70.72	50.00	41.70	45.48	61.90
K=80	64.72	74.46	69.25	52.34	40.86	45.90	60.78

实验 2：验证不同的目标词—基准词相似度矩阵，对目标词极性判断的影响；

实验 3：为了和实验 2 中的“方法 1”进行比较，本验证“方法 2”性能。

实验 4：验证采用不同的优化策略，得到的词汇情感倾向判别结果。

实验 2、3、4 结果如表 2 所示。

实验结果分析：

1）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”虽然结果得到了一定程度的提高，但其效果也不尽人意；原因在于其虽在一定程度上减少了数据的稀疏，但对于目标词与基准词不同现的情形也无能为力；当去除这部分目标词后，“方法 1”方法的性能获得了极

大的提升，说明了此方法用于对那些与目标词同现的词汇的极性判别是有效性。

表 2：多种方法的情感词汇倾向判别实验结果

实验方法 \ 评价指标		PR%	PP%	PF%	NR%	NP%	NF%	OF%
实验 2	PMI-IR	73.49	65.62	69.33	17.34	23.35	19.90	55.65
	去零行 1	60.90	76.70	67.89	57.37	38.91	46.37	59.83
	方法 1	67.44	74.33	70.72	50.00	41.70	45.48	61.90
	去零行 2	87.33	74.57	80.44	31.39	51.81	39.10	70.39
实验 3	方法 2	75.22	69.95	72.49	30.61	36.50	33.29	61.03
实验 4	策略 1	85.03	70.91	77.33	25.11	43.87	31.94	66.00
	策略 2	86.92	72.81	79.24	30.32	51.91	38.28	68.93

2）利用“方法 2”的初衷是对同义词词林进行扩充（目标词共计 2958 个，其中不在同义词词林中的词有 1797 个），为每个目标词都找到一组同义词，然而仅仅依赖于语料所获取的同义词信息并不令人满意;例如，我们依赖语料可以得到以下同义词信息：卑鄙: 质朴、动机、细密、技艺、好处、千秋扩展、挑剔。我们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 2958 个，其中正面 2018 个，反面 940 个）导致了各个目标词的同义词中正面的居多，进而导致了正面的召回率提高。而直接利用“方法 1”判别目标词的情感倾向，方法简单，时间和空间复杂度更低，极大地降低了索引矩阵的大小,其阶数由 2958×2928 降为 2958×80，并且其性能不低于“方法 2”的性能；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；从上述例子中也可直观地看出方法一的有效性。

3）由于有 695 个目标词（约占总目标词的 23.49%）在限定的窗口内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别；“策略 1”仅仅是扩大了同现窗口，虽然一定程度上解决了数据稀疏问题，但由于窗口的扩大使得原来没有同现关系的词汇取得了同现关系，因此带来了噪声；“策略 2”虽对上述的 695 个目标词刻画时扩大了窗口，但增加了惩罚因子，使得性能得到了极大的提高。

5 结语

本文给出的基于 PLSA 的词汇倾向判别方法只需利用少量的基准词,因此，比较容易实现且不受任何外部资源的限制，解决了语料规模较小时数据稀疏问题。当语料规模为 1000 万个词时，文献[3]对于 2697 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 2958 个测试词汇情感倾向判别的准确率达到了 68.93%，验证了方法的有效性。本文所提的方法在性能上还有一定地提升空间。例如:本文假定在一定范围内，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，对词汇的情感倾向产生了影响[10]，下一步可以将这种情形考虑在内，并对中性词的倾向判别进行相关的研究。

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 and MICHAEL L. Littman. Measuring praise and criticism: inference of semantic orientation from association[J]. ACM Transactions on Information Systems, 2003,21(4): 315~346.

[3] Kamps J . , M. Marx , R. J . Mokken , and M. D.Rijke. Using WordNet to measure semantic orientation of

adjectives [A]. In : Proceedings of LREC204 , 4th International Conference on Language Resources and Evaluation[C]. 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,2006,21(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 [A]. In Proc. Of the 20th International Conference on Computational Linguistics (COLING-2004)[C]. Geneva, Switzerland. 2004: 1008~1014.

[7] 王根,赵军.中文褒贬义词汇倾向性的分析,第三届学生计算语言学研讨会论文集[C]. 沈阳. 2006: 81~85.

[8] 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,全国第七届计算语言学联合学术会议论文集[C]. 哈尔滨.2003:527~532.

[9] Hofmann' T..Unsupervised Learning by Probabilistic Latent Semantic Analysis[J]. Machine Learning, 2001,42:177~196.

[10] Hatzivassiloglou, V., & McKeown, K.R.. Predicting the semantic orientation of adjectives[A] . In : Proceedings of ACL297 , 35th Annual Meeting of the Association for Computational Linguistics[C] . Madrid , ES , 1997 : 174~181.

[11] Hofmann T. Probabilistic Latent Semantic Indexing[C]. In: Proceedings of the 22nd Internati- onal Conference on Research and Development in Information Retrieval. Berkeley, Californi- a: [s. n.], 1999: 50~57.