北京大学学报(自然科学版)
Acta Scientiarum Naturalium Universitatis Pekinensis
doi: 10.13209/j.0479-8023.2014.003

基于社会关系网络的半监督情感分类

薛云霞　李寿山†　王中卿

苏州大学自然语言处理实验室，苏州　215006; † 通信作者, E-mail: shoushan.li@gmail.com

摘要　基于样本的社会关系，提出一种新的半监督学习方法，创建一种基于文档—词及社会关系的二部图模型，并根据标签传播算法将未标注样本加入到分类器的构建中。实验结果表明，加入社会关系网络的半监督情感分类方法明显优于传统的仅利用评论文本信息的半监督情感分类方法。

关键词　自然语言处理；情感分类；半监督；社会关系网络；标签传播

中图分类号　TP18

Semi-supervised Sentiment Classification with Social Network

XUE Yunxia, LI Shoushan†, WANG Zhongqing

Natural Language Processing Lab, School of Computer Science and Technology Soochow University, Suzhou 215006;
† Corresponding author, E-mail: shoushan.li@gmail.com

Abstract　Baed on the social connection between the comments in the social network, the authors propose a new approach of semi-supervised sentiment classification, and provide a document-word and social connection bipartite graph structure and apply to label propagation algorithm. Evaluation across three different domains shows that the proposed approach performs better than that which only considers comment textual information.

Key words　natural language processing; sentiment classification; semi-supervised; social network; label propagation

情感文本分类是自然语言处理中一项越来越受关注的研究课题[1−3]。该任务旨在对用户发表的主观性文本进行分析和挖掘，判断其表达的情感色彩是贬义(Negative)还是褒义(Positive)。传统的情感分类任务主要针对评论的文本信息，例如：产品评论、电影评论等。然而，随着互联网迅速发展，众多社交网站(例如新浪微博、Twitter 等)、购物网站(例如亚马逊等)的出现，使得网络用户激增，同时也促使了社会关系网络的形成，面向社会关系网络的情感分类也应运而生。社会关系网络主要体现在人与人之间的关系，例如：某些用户通过在社交网站上的在线活动(如，关注、转帖、赞等)与他人建立联系；或者某些用户之间有相同的兴趣爱好、相同的观点、买过相同的产品等。网络用户的激增也使得社会关系网络更加错综复杂且包含大量信息，因而利用好社会关系网络中人与人的联系也变得尤为重要。

在以往的情感分类研究中，人们一般认为每个文本之间是独立的[4−6]。然而，在某些网络文本中，这些文本之间通过用户有着一定的联系。例如，在购物网站上，两个用户均评论过某个产品；或者在某社交网站上，两个用户相互关注或者发表过类似的文本信息等。鉴于“人以类聚，物以群分”的观点，我们可以认为存在一定社会联系的用户更有可能持有类似观点。同时，同一用户发表的不同评论也更有可能具有同样极性的情感表达。因此，本文提出一种基于社会关系网络的半监督情感分类方法，即将评论间的社会关系与评论的文本信息相结合。具体来讲，我们将评论的文本信息与评论间的社会关系网络相结合来计算文档间的转移概率，并根据标签传播算法(Label Propagation Algorithm)[7]，将未标

国家自然科学基金(61375073, 61273320)资助
收稿日期: 2013−06−17；修回日期: 2013−09−23；网络出版时间: 2013-11-11 10:25

注样本加入到半监督分类器的训练中。实验结果表明，基于社会关系网络的半监督情感分类方法明显优于传统的仅利用文本信息的半监督情感分类方法，进而表明社会关系网络对半监督情感分类确有帮助。

1 相关工作

1.1 半监督情感分类的研究现状

基于半监督学习的情感分类方法是通过结合少量标注样本和大量未标注样本来构建情感分类模型。Dasgupta 等[8]将多种机器学习方法(例如聚类方法、集成学习等)融入基于半监督学习的情感分类中。面对情感分类中中文标注语料匮乏的问题，Wan[9]将两种不同语言(英语和汉语)作为两个不同的视图，采用协同训练方法进行半监督情感分类；Li 等[10]把评价语句分为个人视图(personal view)和非个人视图(impersonal view)并采用协同训练方法进行半监督情感分类。Zhou 等[11]提出基于贝叶斯置信网络的主动学习和半监督学习方法用于情感分类任务。苏艳等[12]对协同训练方法进行改进，提出基于动态随机特征子空间的协同训练算法，并实验验证当特征子空间数目为 4 左右的时候，该半监督分类方法能够取得最佳性能。

1.2 面向社会关系网络的情感分类

面向社会关系网络的情感分析是指使用社会关系网络信息对文本、人物及人物之间的社会关系进行情感倾向性分析。相对于传统面向文本的情感分析，面向社会关系网络的情感分析的研究才刚刚起步，相关的研究还比较缺乏。

面向社会关系网络的情感分析的研究对象同传统的情感分析任务一样，都是对某一文本进行倾向性分析。不同的是，这些文本来源于社交网络。在此情况下，除了文本信息外，大量的文本与文本关系及用户社会关系信息给情感分析提供了更多参考的特征。Jiang 等[13] 和 Speriosu 等[14]都是以 Twitter 为社交网络，利用多条 Tweet 之间的各种关系(例如跟帖、转帖)提升 Twitter 文本的情感分类效果。Tan 等[15]利用类似的社会关系(例如“follow”和“@”关系)，提出一种改进的半监督情感分类框架，使用非标注样本和社会关系共同提升情感分类性能。

本文同 Tan 等[15]的研究目标基本一致，都是使用社会关系去提升半监督情感分类性能。不同的是，我们的研究所使用的社会关系主要是通过评论的作者联系构建，而非评论本身之间的联系。相对而言，评论的作者之间的联系更加普遍。因为我们关注的是产品评论的情感分类任务，在众多的产品评论网站中，并不存在“跟帖”、“转帖”这些连接关系，所以上述方法并不能直接应用到我们的问题中。

2 数据统计

本章旨在通过一些数据分析，直观上说明社会关系网络对提高半监督情感分类存在的可能性。我们利用 Blizter 等[16]①的语料，主要包括 3 个领域的文本：Kitchen，Electronic 和 DVD。每条评论语料包含以下属性：评论内容、评论对象、评论者、评论打分以及评论者所在地等信息。在社会关系网络中，每条评论被看作是一个实体，每个实体具有特定的属性。例如评价对象、评论者及评论者所在地等。我们统计了 3 个领域评论语料中评论总数、评论者总数以及社会联系的总数，结果如表 1 所示。本文只考虑评论通过评论者所建立的社会联系，即如果两条评论有共同评论者，则认为他们之间存在社会联系。

直观上讲，存在社会联系的评论会更有可能情感极性相同。例如，某些评论者在给产品作评论时习惯给产品好评，而某些评论者则会习惯给差评。因此我们认为：1) 有社会联系的实体比无社会联系的实体情感标签一致的可能性大；2) 情感极性标签一致的评论更有可能存在社会联系。本文主要研究评论者相同的评论间的社会联系对半监督情感分类的影响。本文统计以下 4 种情况的概率，统计结果如图 1 和 2 所示，其中符号意义如下。

$p(\text{Co\_label}\,|\,\text{Random})$：随机情况下，评论来自同一评论者的概率；

$p(\text{Co\_reviewer}\,|\,\text{Random})$：随机情况下，评论的情感极性标签一致的概率；

$p(\text{Co\_label}\,|\,\text{Co\_reviewer})$：在评论者相同的条件下，评论的情感极性标签一致的概率；

表 1 语料统计结果

Table 1 Statistics of corpus

领域	评论总数	评论者总数	社会联系总数
Kitchen	5000	4367	21144
Electronic	5000	4609	2438
DVD	5000	3855	2586

① http://www.cs.jhu.edu/~mdredze/datasets/sentiment/index2.html

p(Co_reviewer|Co_label): 在评论极性相同的条件下, 评论来自同一评论者的概率。

从图中可以看出: 1) 随机挑选样本时, 评论的情感极性标签一致的概率平均为 50%, 而在评论者相同的条件下, 评论的情感极性标签一致的概率平均为 72.8%, 比随机情况下高 22.8%; 2) 随机挑选样本时, 评论来自同一评论者的概率平均为 0.693%, 而在评论极性相同的条件下, 评论来自同一评论者的概率为 0.78%, 比随机情况高 0.87%。

从上面的数据可以看出, 样本的社会联系同情感标签存在一种联系, 具有社会联系的样本更倾向于具有一直的情感标签。这一特性可以帮助半监督情感分类, 在传播情感标签的时候, 让那些具有社会联系的样本尽可能获得相同的情感标签。

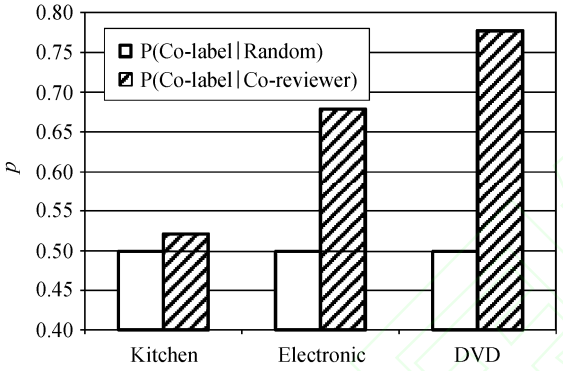

图 1　在随机和相同评价者条件下两个样本标签一致的概率

Fig. 1　Probabilities of common label under random and common reviewer conditions

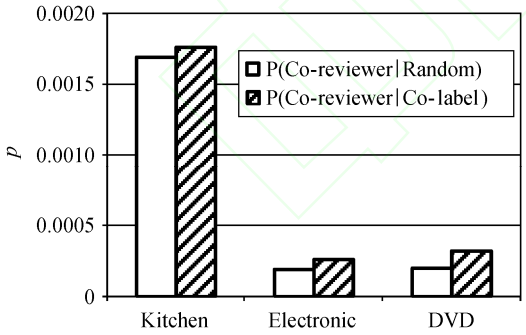

图 2　在随机和相同标签的条件下两个样本来自同一评论者的概率

Fig. 2　Probabilities of common reviewer under random and common label conditions

3　基于社会关系网络的半监督情感分类方法

标签传播算法(Label Propagation)是基于图的半监督学习方法, 基本思路是用已标记的节点的标签信息去预测未标注节点的标签信息。常用的基于二部图的 LP 算法, 利用词——文档的二维模型, 建立文档间的联系。在二部图中, 节点包括已标注和未标注数据, 边代表两个节点的相似度。节点的标签按相似度传给其他节点, 已标注的数据节点是源头, 可以对无标签数据进行标注, 节点的相似度越大, 标签就越容易传播。

3.1　基于社会关系网络的情感文本模型

情感分类中, 文档通常用词袋(bag-of-words)模型化并用向量形式描述, 缺点是文档与单词间的关联不清晰。本文采用的二部图是图论中的一种特殊模型, 其顶点集 V 可分割为两个互不相交的子集, 并且图中每条边依附的两个顶点都分属于这两个互不相交的子集。由于社会关系网络对半监督情感分类有帮助, 因此, 本文将文档的文本信息与社会关系网络相结合, 提出一种基于文档—词及社会关系的二部图, 结构如图 3。其中, 文档用 $d_1,d_2,...,d_V$ 表示, 文档中包含的单词用 $w_1,w_2,...,w_n$ 表示, 文档的评论者用 $r_1,r_2,...,r_m$ 表示。文档—词及社会关系的二部图结构包含文档到词的连接关系及文档间的社会联系。一篇文档包含多个单词, 一个单词会在多个文档中出现; 一篇文档来自一个评论者, 一个评论者会发表多篇评论。显然, 通过构建文档—词及社会关系的二部图可以很清晰地表述文档与文档间的文本联系及社会联系。

文档—词及社会关系的二部图连接关系由文档和词及评论者的连接矩阵表示, 即 $V\times(n+m)$ 矩阵 X; V 为文档数目, n 为词的总数, m 为评论者总数。如果文档 d_i 包含词 w_q, 则将其权重 u_{iq} 赋值为 1, 如果文档 d_i 的评论者是 r_k, 则将其权重 v_{ik} 赋值为 λ。文档到词及评论者的转移概率计算参考文献[17], 具体计算方法如下。

如果文档 d_i 包含词 w_q, 权重为 u_{iq}, 则文档 d_i 到词 w_q 的转移概率为 $\frac{u_{iq}}{\sum_q u_{iq}}$; 同理, 词 w_q 到文档 d_j 的转移概率为 $\frac{u_{jq}}{\sum_j u_{jq}}$。如果文档 d_i 的评论者为 r_k, 权重为 v_{ik}, 则文档 d_i 到评论者 r_k 的转移概率为 $\frac{v_{ik}}{\sum_k v_{ik}}$; 同理, 评论者 r_k 文档 d_j 的转移概率为 $\frac{v_{jk}}{\sum_j v_{jk}}$。则文档 d_i 到文档 d_j 的转移概率 t_{ij} 是由文档

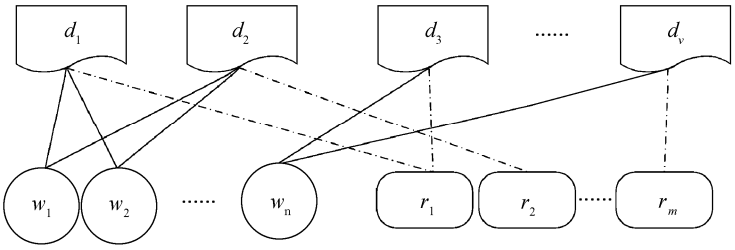

图 3　基于文档—词及社会关系的二部图结构

Fig. 3　Bipartite graph based on document-word and social network structure

d_i 通过该文档里面的所有词到达文档 d_j 的概率之和与该文档的评论者到文档 d_j 的概率之和相加，即

$$t_{ij}=\sum_q \frac{u_{iq}}{\sum_q u_{iq}}\cdot\frac{u_{jq}}{\sum_j u_{jq}}+\sum_k \frac{v_{ik}}{\sum_k v_{ik}}\cdot\frac{v_{jk}}{\sum_j v_{jk}},$$

得到文档间的转移概率后，可以通过标签传播算法计算未标注样本的标签。

3.2　基于社会关系网络的标签传播算法

标签传播算法是一种基于图结构的半监督学习方法。基于社会关系网络的 LP 算法是通过建立文档—词及社会关系的二部图结构计算文档间的转移概率，以此将标注样本的标签传播为未标注样本。具体算法流程如下所示。

输入：已标注样本集合 L，包含 n+个正类样本和 n-个负类样本；未标注样本集合 U;

输出：更新后的标注样本集合 L;

1) 初始化:

$\boldsymbol{P}$：$n\times r$ 标注矩阵，P_{ij} 表示文档 $i(i=0...n)$ 属于类别 $j(j=0...r)$ 的概率;

P_L：P^0 的后 $n-m$ 行对应的 $n-m$ 个未标注实例 U;

$\overline{\boldsymbol{T}}$：$n\times n$ 矩阵，每一项 $\overline{t_{ij}}$ 表示根据上一节计算获得的基于社会关系网络的，从文档 d_i 到文档 d_j 的转移概率;

① 设置迭代标记 $t=0$，根据标注样本设定 P_L^0 的值;

② 初始化 P_U^0;

2) 循环迭代 N 次直到收敛;

① 传播实例的标注信息到相邻的实例，依据公式 $P^{t+1}=\overline{T}P^t$;

② 还原标注实例的标注信息，即用 P_L^0 替代 P^{t+1};

3) 对于每个未标注的实例，根据 $\arg\max_j P_{ij}(j=0...r)$ 得到其正负标签，并添加到 L 中，从 U 中删除。

4　实验

4.1　实验设置

本实验使用的数据包括 3 个领域的产品评论，3 个领域分别为: Kitchen, Electronic 和 DVD。每个领域包含 1300 篇正面评论和 1300 篇负面评论。语料中每篇评论包括以下项目：评论内容、评论者、评论对象、评论打分和评论者所在地。实验中的分类算法是最大熵方法，并使用 MALLET[①]机器学习工具包中的最大熵分类器，分类算法的所有参数均设为默认值。分类选取词作为特征。我们选取正负各 100 个样本作为标注样本，正负各 1000 个样本作为非标注样本，剩下正负各 200 的样本作为测试样本。实验中使用准确率作为结果好坏的评价标准。

4.2　实验结果与分析

我们实现以下几种情感分类方法的比较研究。

1) ME: 不使用未标注样本，只利用标注样本训练最大熵分类器;

2) LP: 仅考虑评论的文本信息，即利用文档—词二部图结构计算文档间的转移概率，然后根据标签传播算法将未标注样本加入到分类器模型的构建中;

3) LP-Social: 将社会关系网络与评论的文本信息相结合，即利用文档—词及社会关系的二部图结构计算文档间的转移概率，然后根据标签传播算法将未标注样本加入到分类器模型的构建中。具体实现时，我们设置社会关系权重为 5，即 $\lambda=5$。

① http://mallet.cs.umass.edu/

表 2 给出了各种情感分类方法的分类结果。

表 2　各种情感分类方法结果比较

Table 2　Performance comparison of different approaches on sentiment classification

领域	ME	LP	LP–Social
Kitchen	0.733	0.755	0.777
Electronic	0.726	0.735	0.754
DVD	0.712	0.724	0.750
平均	*0.723*	*0.738*	*0.760*

从表 2 的结果可以看出, 基于 LP 的半监督情感分类方法比较稳定地优于不使用非标注样本的 ME 分类方法, 平均提高 1~2 个百分点。该结果说明使用非标注样本能够帮助提高情感分类的性能。此外, 我们发现 LP-Social 方法表现明显好于 LP 方法, 平均提高 2~3 个百分点。该结果表明社关系网络对半监督情感分类具有积极的作用。

社会关系权重, 即 λ, 是 LP-Social 方法中的一个重要参数。下面内容给出 LP-Social 方法中该参数的敏感度。我们测试 λ 值从 1 变化到 9 时分类准确率的分类结果, 如图 4 所示。

由图 4 可以看出:

1) 当 $4 \leqslant \lambda \leqslant 9$ 时, LP-Social 方法的分类准确率均高于传统的 LP 方法的准确率, 因此, 本文方法对参数 λ 并不算敏感, 能够在一定范围内都确定较好的分类结果; 2) 当 $\lambda = 5$ 时, 本文方法在 3 个领域中表现最好, 表 2 为当 $\lambda = 5$ 时, Kitchen, Electronic 和 DVD 3 个领域分别使用 ME、传统 LP 和基于社会关系网络的标签传播算法的分类结果; 3) 对于 DVD 领域, 在 λ 小于 5 的情况下, 分类准确率不太稳定, 但是在 λ 高于 5 以后, 分类准确率明显稳定地高于 ME 和普通的 LP 算法。

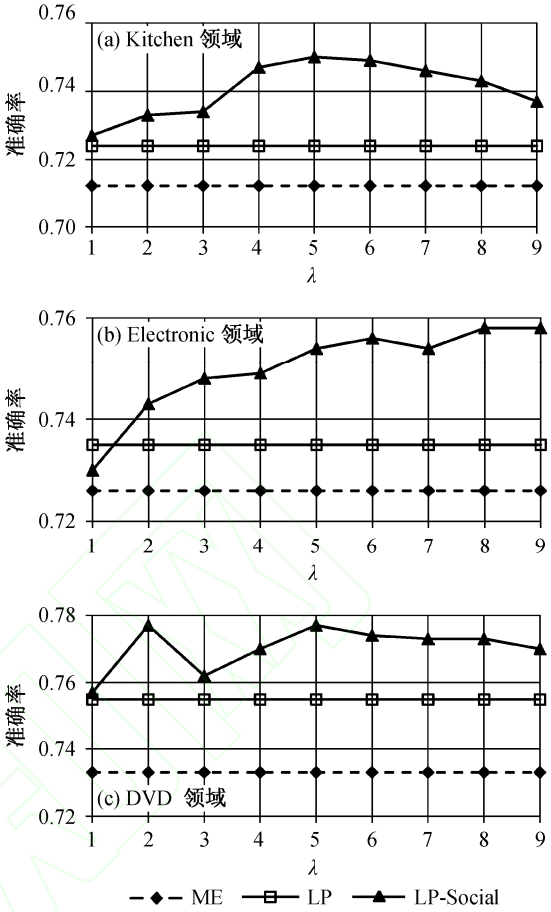

图 4　LP-Social 情感分类方法随λ的变化曲线

Fig.4　Performance comparison on different values of λ

5　结论和下一步工作

本文提出了一种基于文档—词及社会关系网络的二部图结构用于标签传播算法, 用以实现半监督情感分类。该方法不仅可以通过词构建文档与文档之间的联系, 还可以充分利用文档与文档之间的社会关系网络。实验结果表明加入社会关系网络的半监督情感分类方法明显优于传统的只利用评论的文本信息的半监督情感分类方法, 进而表明社会关系网络对半监督情感分类具有的一定的积极作用。

本文实验中, 只加入了评论者这一社会联系。我们计划在下一步工作将尝试评论文本与评论的其他社会联系。例如, 评论对象及评论者所在地等信息, 进一步提高半监督情感分类性能。

参考文献

[1] 黄萱菁, 赵军. 中文文本情感分析. 中国计算机学会通讯, 2008, 4(2):

[2] 赵军, 许洪波, 黄萱菁, 等. 中文倾向性分析评测技术报告//第一届中文倾向性分析评测会议, 2008

[3] 刘鸿宇, 赵妍妍, 秦兵, 等. 评价对象抽取及其倾向性分析. 中文信息学报, 2010, 24(1): 84−88

[4] 唐慧丰,谭松波,程学旗,等. 基于监督学习的中文情感分类技术比较研究. 中文信息学报, 2007, 6(2): 88−94

[5] Pang B, Lee L, Vaithyanathan S. Thumbs up? sentiment classification using machine learning techniques//Proceedings of the Conference on Empirical Methods in Natural Language Processing. Philadelphia, 2002: 79−86

[6] Zagibalov T, Carroll J. Automatic seed word selection for unsupervised sentiment classification of Chinese test//Proceedings of the 22nd International Conference on Computational Linguistics. Manchester, 2008: 1073−1080

[7] Zhu X, Ghahramani Z. Learning from labeled and unlabeled data with label propagation // CMU CALD Technical Report. Pittsburgh, 2002: 107

[8] Dasgupta S, Ng V. Mine the easy, classify the hard: a semi-supervised approach to automatic sentiment classification//Proceedings of Annual Meeting on Association for Computational Linguistics. Singapore, 2009: 701−709

[9] Wan X. Co-training for cross-lingual sentiment classification//Proceedings of Annual Meeting on Association for Computational Linguistics. Singapore, 2009: 235−243

[10] Li Shoushan, Huang C, Zhou Guodong, et al. Employing personal/impersonal views in supervised and semi-supervised sentiment classification// Proceedings of Annual Meeting on Association for Computational Linguistics. Uppsala, 2010: 414−423

[11] Zhou Shusen, Chen Qingcai, Wang Xiaolong. Active deep networks for semi-supervised sentiment classification//Proceedings of the 23rd International Conference on Computational Linguistics. Beijing, 2010: 1515−1523

[12] 苏艳, 王中卿, 居胜峰, 等. 基于随机特征子空间的半监督情感分类方法研究. 中文信息学报, 2012, 26(4): 85−92

[13] Jiang Long, Yu Mo, Zhou Ming, et al. Target-dependent Twitter sentiment classification// Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies. Portland, 2011:151−160

[14] Speriosu M, Sudan N, Upadhyay S, et al. Twitter polarity classification with label propagation over lexical links and the follower graph//Proceedings of the First Workshop on Unsupervised Learning in NLP. Edinburgh, 2011: 53−63

[15] Tan C, Lee L, Tang J, et al. User-level sentiment analysis incorporating social networks//Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. San Diego, 2011: 1397−1405

[16] Blitzer J, Dredze M, Pereira F. Biographies, bollywood, boom-boxes and blenders: domain adaptation for sentiment classification//Proceedings of Annual Meeting on Association for Computational Linguistics. Prague, 2007: 440−447

[17] Sindhwani V, Melville P. Document-word co-regularization for semi-supervised sentiment analysis//Proceedings of The IEEE International Conference on Data Mining. Pisa, 2008: 1025−1030