基于图排序模型的跨领域倾向性分析算法*

吴琼[1,2] 谭松波[1] 张刚[1] 段洣毅[1] 程学旗[1]

1. 中国科学院计算技术研究所 北京 100080；2. 中国科学院研究生院 北京 100080

E-mail: wuqiong@software.ict.ac.cn, tansongbo@software.ict.ac.cn

摘要： 倾向性分析因其重要性而受到广泛关注。通常，监督分类方法对倾向性分析很有效。但是，当训练域与测试域不在同一个领域时，这些算法的性能明显下降。本文提出一个算法，将文本的情感倾向性与图排序算法结合起来进行跨领域倾向性分析。本算法在图排序算法基础上，利用训练域文本的准确标签与测试域文本的伪标签来进行倾向性分析。实验结果表明，本文提出的算法能大幅度提高跨领域倾向性分析的精度。

关键词： 图排序，跨领域，倾向性分析

Cross-Domain Opinion Analysis Based on Graph-Ranking

Wu Qiong[1,2] Tan Songbo[1] Zhang Gang[1] Duan Miyi[1] Cheng Xueqi[1]

1. Institute of Computing Technology, Chinese Academy of Sciences, Beijing 100080

2. 2.Graduate School of CAS, Beijing 100080

E-mail: wuqiong@software.ict.ac.cn, tansongbo@software.ict.ac.cn

Abstract: Sentiment classification is attracting more and more attention because of its great benefits to social and human life. Usually supervised classification approaches perform well in sentiment classification. However, the performance decreases sharply when transferred from one domain to another domain. In this paper, we propose an algorithm which integrates the sentiment orientations of the documents into the graph-ranking algorithm for cross-domain sentiment classification. We apply the graph-ranking algorithm using the accurate labels of old-domain documents as well as the "pseudo" labels of new-domain documents. The experiment results indicate that the proposed algorithm could improve the performance of cross-domain sentiment classification dramatically.

Key words: Graph Ranking, Cross Domain, Opinion Mining

1 引言

随着互联网的快速发展以及论坛、博客等网络交流平台的不断涌现，人们越来越习惯于在网上表达自己对于日常事件、产品、政策等的观点和看法，这使得网上存在大量带有情感倾向性的文本。如何对浩如烟海的网络文本进行快速的倾向性分析，就成为越来越引起广泛关注的研究问题（如[1-7]）。

作为传统文本分类[14-15]的一个特殊分枝，典型的监督分类方法都适用于文本倾向性分析。然而，当训练数据与测试数据不属于同一个领域的时候（例如，已知酒店评论数据集的倾向性，需判断电子评论数据集的倾向性），典型的分类方法的效果就变得很差。这是由于训练域里有强烈倾向性的词在测试域里不再有强烈倾向性，反之亦然。例如，"便携的"在电子评论里就是一个具有正面倾向性的词，而在酒店评论里就不具有强烈的倾向性。这就产生了跨领域倾向性分析问题（也称为跨领域情感分类问题）[8-13]。随着信息量的急速增加、新领域的不断涌现，人们

本文承中科院专项资助基金(0704021000)、国家自然科学基金 (60803085)以及国家重点基础研究计划(2007CB311100)的资助。

需要在越来越多的新领域里进行倾向性分析，而在新领域里重新进行人工标注是个费时费力的事情。因此要尽量基于已经标注好的数据对新领域进行分析，这使得跨领域的倾向性分析具有重要意义。

跨领域倾向性分析是一个全新的研究领域，目前，专门的研究工作还比较少。现有的一些技术主要分为两类：第一类需要在测试域标注少量数据来辅助训练，如[8-9]等；第二类在测试域不需要任何标注好的数据，如[10-13]等。其中本文主要针对应用更为广泛的第二类情况。

图排序算法（如 PageRank[16]）的思想是：图中与其它重要结点紧密相联的结点也很重要。该算法已成功应用于很多领域。本文将文本的倾向性与图排序算法结合起来，提出一种基于图排序的跨领域倾向性分析算法。该算法为测试集中的每一个文本分配一个情感分，来表示该文本“支持”或“反对”的程度，然后利用源领域的准确标签和新领域的伪标签来迭代计算该情感分，算法收敛时得到最终情感分，并据此判别新领域测试数据的倾向性。

2 基于图排序模型的跨领域倾向性分析算法

2.1 算法描述

我们定义跨领域倾向性分析问题如下：

测试集 $D^U=\{d_1,...,d_n\}$和训练集 $D^L=\{d_{n+1},...,d_{n+m}\}$，其中$d_i$表示第$i$个文本的向量，每一个文本应该有一个来自类别集$C$ = {支持, 反对}中的标签。每一个测试文本 $d_i\in D^U(i = 1,...,n)$ 没有被标注，每一个训练文本 $d_j\in D^L(j = n+1,...,n+m)$ 已经被标注了一个类别 C中的标签。假设测试数据集D^U和训练数据集D^L来自相关但不相同的领域。本算法的目标是利用另一个领域的训练数据集D^L来对测试数据集中的每一个文本$d_i\in D^U$ $(i = 1,...,n)$分配一个C中的标签，使得准确率最高。

本算法基于以下前提：

（1）用W^L表示旧领域的词空间，W^U表示新领域的词空间，则$W^L\cap W^U\neq\Phi$。

（2）如果一个文本既存在于训练集中，又存在于测试集中，则标签一致。

基于图排序思想，我们认为如果一个文本与一些具有支持（反对）态度的文本紧密联系，则它也很可能持支持（反对）态度，这也是邻域学习思想。

因此，我们将训练集和测试集看作一个图，里面的每一个文本为图中的一个结点。给每一个结点一个表示其情感类别的分数，称其为情感分。本文提出的算法将文本情感类别间的关系与 graph-ranking 算法结合起来。对于每一个待标注文本，算法通过其在训练域和测试域的邻域来计算它的情感分，并用一个统一的公式进行迭代计算，当算法收敛时，得到待标注文本的最终情感分。如果一个结点的情感分在-1 到 0 之间，表示这个结点所代表的文本是持反对态度，情感分越接近于-1，此文本越倾向于反对态度；如果一个结点的情感分在 0 到 1 之间，表示这个结点所代表的文本是持支持态度，情感分越接近于 1，此文本越倾向于支持态度。

2.2 基于图排序模型的跨领域倾向性分析算法

2.2.1 算法初始化

第一步，本算法需要为训练集与测试集中每一个文本的情感分赋初始值，得到初始情感分向量$S^0=\{s_1^{(0)},...,s_n^{(0)},s_{n+1}^{(0)},...,s_{n+m}^{(0)}\}$。对于训练集中的文本，它们已经有正确标签。对于测试

集中的文本，使用典型的文本分类算法中的任一种分类器，用训练集训练，对测试集分类得到一个伪标签（此时的准确度通常很低）。对于每一个文本，如果它分配到的标签是“反对”，则将它的情感分赋为-1；如果它分配到的标签是“支持”，则将它的情感分赋为 1。

第二步，为保证最终程序的收敛性，将测试集对应的情感分初始值$s_i^{(0)}$ $(i=1,\ldots,n)$归一化，使得正的情感分的和为 1，负的情感分的和为-1：

$$s_i^{(0)}=\begin{cases} s_i^{(0)} \Big/ \sum_{j\in D_{neg}^U}(-s_j^{(0)}), & if\ \ s_i^{(0)}<0 \\ s_i^{(0)} \Big/ \sum_{j\in D_{pos}^U} s_j^{(0)}, & if\ \ s_i^{(0)}>0 \end{cases} \quad i=1,\ldots,n \tag{1}$$

其中D_{neg}^U和D_{pos}^U分别表示D^U中的“反对”、“支持”文本集。同(1)一样，将训练集对应的情感分初始值$s_j^{(0)}$ $(j=n+1,\ldots,n+m)$归一化。

2.2.2 情感分计算策略

得到初始情感分向量S^0后，即可利用训练域的准确情感分和测试域的伪情感分来迭代计算测试集的最终情感分。

（1） 利用训练集的准确情感分来计算测试集的情感分

建立一个图模型，结点表示D^L和D^U中的文本，边表示文本间的内容相似度。如果两个文本间内容相似度为 0，则图中两点间无边，如果不为 0，则图中两点间有边，且边的权重即为此内容相似度。内容相似度有很多方法求出，此处用余弦相似度来计算。我们使用一个联接矩阵U来表示D^U和D^L间的相似矩阵。$U=\left[U_{ij}\right]_{n\times m}$定义如下：

$$U_{ij}=\frac{d_i \bullet d_j}{\|d_i\|\times\|d_j\|},\ i=1,\ldots,n,\ j=n+1,\ldots,n+m \tag{2}$$

其中特征t 的权重用$tf_i idf_i$来计算。为保证算法收敛，将联接矩阵U归一化为矩阵$\hat{U}$，使得$\hat{U}$中每一行的和为 1：

$$\hat{U}_{ij}=\begin{cases} U_{ij} \Big/ \sum_{j=1}^{m} U_{ij}, & if \sum_{j=1}^{m} U_{ij}\neq 0 \\ 0, & otherwise \end{cases} \tag{3}$$

为了找出与一个文本最相似的文本集（此处设此文本集大小为K），我们对$\hat{U}$的每一行进行降序排列得到$\tilde{U}$，也就是：$\tilde{U}_{ij}\geq\tilde{U}_{ik}\ (i=1,\ldots n;\ j,k=1,\ldots m;\ k\geqslant j)$。因此对于$d_i\in D^U$ $(i=1,\ldots,n)$，$\tilde{U}_{ij}\ (j=1,\ldots,K)$就表示它在训练域中的$K$个邻居。我们使用矩阵$N=\left[N_{ij}\right]_{n\times K}$来表示$D^U$在训练域中的$K$个邻居，其中$N_{ij}$对应于$d_i$的第$j$个邻居。

最后，用d_i邻居们的分数来计算它的情感分，公式如下所示：

$$s_i^{(k)} = \sum_{j \in N_{i\bullet}} (\hat{U}_{ij} \times s_j^{(k-1)}), \quad i = 1,...,n \tag{4}$$

其中，$i\bullet$表示矩阵的第i行，$s_i^{(k)}$表示第k次迭代时的情感分s_i。

（2） 利用测试集的“伪”情感分来计算测试集的情感分

类似于上一节所述，建立一个图模型，结点表示D^U中的文本，边的权重由它所连接的两个文本的余弦相似度来计算。我们使用一个联接矩阵V来表示测试集之间的相似矩阵，即$V = \left[V_{ij} \right]_{n \times n}$。同样，我们将$V$归一化为$\hat{V}$，然后将$\hat{V}$的每一行进行降序排列得到$\tilde{V}$，因此得到一个$D^U$在测试域中的邻居矩阵$M = \left[M_{ij} \right]_{n \times K}$。最后，利用测试域的伪情感分来计算测试集的情感分如公式(5)所示：

$$s_i^{(k)} = \sum_{j \in M_{i\bullet}} (\hat{V}_{ij} \times s_j^{(k-1)}), \quad i = 1,...,n \tag{5}$$

2.2.3 算法迭代过程

本算法要同时利用训练域和测试域的信息来对测试域的文本进行标注，因此综合公式（4）（5），得到迭代计算测试数据集的情感分的公式如下所示：

$$s_i^{(k)} = \alpha \sum_{j \in N_{i\bullet}} (\hat{U}_{ij} \times s_j^{(k-1)}) + \beta \sum_{h \in M_{i\bullet}} (\hat{V}_{ih} \times s_h^{(k-1)}), \quad i = 1,...,n \tag{6}$$

矩阵形式为：

$$S^{(k)} = \alpha \hat{U} S^{(k-1)} + \beta \hat{V} S^{(k-1)} \tag{7}$$

其中$\alpha + \beta = 1$，α和β分别表示训练域和测试域对最终情感分的贡献大小。为保证算法收敛，算法每迭代一次都需要将S归一化（如公式(1)），使得正的情感分之和为1，负的情感分之和为-1。迭代计算情感分S并归一化，直到算法收敛为止。

3 实验与分析

3.1 实验数据

我们从互联网上的评论中整理出三个领域的中文数据集，分别是：电子评论（来源于：http://detail.zol.com.cn/），财经评论（来源于：http://blog.sohu.com/stock/）以及酒店评论（来源于：http://www.ctrip.com/）。然后由专家将这些数据集标注为“支持”或“反对”。数据集的具体组成如表1所示（其中“词典长度”表示数据集中不同词的数量）：

表1 数据集构成

数据集	反对评论数	支持评论数	评论平均长度	词典长度
电子	554	1,054	121	6,200
财经	683	364	460	13,012
酒店	2,000	2,000	181	11,336

我们对上述数据集进行以下预处理：首先，我们使用中文分词工具ICTCLAS(http://ictclas.org/)来对这些中文评论进行分词；然后，用向量空间模型来表示文本。在该模型中，每个文本转化为词空间中的词袋表示，词的权重用该词在文本中出现的频率来计算。

3.2 实验评价

本文用支持向量机作为 baseline 算法，它是一种效果很好的监督学习算法，在我们的实验中，我们使用 LibSVM[17]，用它的线性核，并将所有参数设为缺省值。另外，我们将本文算法与结构对应学习算法（记作 SCL）[9]进行比较分析。SCL 算法是一种很新的跨领域倾向性分析算法。该算法思想为：找出在不同领域中频繁出现的情感特征作为枢纽特征，然后通过建模来获得非枢纽特征与枢纽特征之间的关联。本实验中，我们使用 100 个枢纽特征。本文使用精度(accuracy)作为倾向性分析系统的评价标准。

3.3 总体性能

我们提出的算法中有两个参数：K和α （β可以由 1-α 计算得出）。将参数K设为 150，表示算法中为每一个文本求出 150 个邻居；将参数α 设为 0.7，表示训练域对情感分的贡献比测试域略大。同时，我们认为对于每一个测试集中的文本$d_i \in D^U (i = 1,\ldots n)$，如果连续两步计算得到的情感分$s_i$的变化量低于一个给定的阀值，则该算法收敛，本文设定此阀值为 0.00001。

表 2 显示了当进行跨领域倾向性分析时，LibSVM、SCL 以及本文提出的算法的精度，其中我们的算法用 LibSVM 分类器进行初始化。

表 2　跨领域倾向性分析时不同算法性能比较

	LibSVM	SCL	本文提出的算法
电子->财经	0.6478	**0. 7507**	0.7304
电子->酒店	0.7522	**0. 7750**	0.7543
财经->酒店	0.6957	**0. 7683**	0.7457
财经->电子	0.6696	0. 8340	**0.8435**
酒店->财经	0.5978	0. 6571	**0.7848**
酒店->电子	0.6413	0. 7270	**0.8609**
平均	0.6674	0. 7520	**0.7866**

由表 2 可以看出，本文提出的算法大幅度地提高了跨领域倾向性分析的精度。其中第 2 列是 LibSVM 的精度，第 4 列为用 LibSVM 初始化后本算法的精度，对比可见，我们算法的精度均高于 LibSVM 的精度，平均精度提高了 11.9%。精度上如此大幅度的提高表明我们的算法对于跨领域倾向性分析问题非常有效。

表 2 中第 3 列为 SCL 算法的精度，总体上说，我们对于 SCL 算法的实验结果与文章[9]中结果基本一致。SCL 算法的平均精度比 LibSVM 高，这证明 SCL 算法对于跨领域倾向性分析问题很有效。然而从表中可以看出，我们提出的算法的精度优于 SCL 算法。我们算法的平均精度比 SCL 算法高约 3.5%，当从酒店向电子领域移植时，我们算法的精度比 SCL 算法提高地最多，为 13.4%。分析其原因，是因为以下两点。第一，SCL 算法本质上是基于词的共现（窗口大小为整篇文本），因此它很容易被低频词及数据集大小所影响。第二，SCL 算法的枢纽特征是完全由领域专家选定的，因此枢纽特征选择的质量会影响 SCL 算法的性能。

4 结论

本文提出一种跨领域倾向性分析算法，它将文本的情感倾向性与图排序方法结合起来进行跨领域的倾向性分析。该算法首先为每一个待分类文本赋一个情感分，然后利用训练数据的准确标签和测试数据的伪标签迭代计算该情感分，最后根据此情感分将测试文本标注为“反对”或“支持”。我们针对三个领域相关的情感数据集检验本算法。实验结果表明，我们的算法可以大幅度地提高跨领域情感分类的精度。

参 考 文 献

[1] 胡熠,陆汝占,李学宁,段建勇, 陈玉泉. 基于语言建模的文本情感分类研究.计算机研究与发展. 2007, 44(9): 1469-1475.

[2] 姚天昉,娄德成.汉语语句主题语义倾向分析方法的研究. 中文信息学报. 2007, 21(5): 73-79

[3] 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. 中文信息学报. 2007,21(1): 96-100.

[4] 唐慧丰,谭松波,程学旗. 监督学习方法在语气挖掘中的应用研究. 中文信息学报. 2007,21(6): 88-94.

[5] 赵军, 许洪波, 黄萱菁, 谭松波, 刘康, 张奇. 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 (The First Chinese Opinion Analysis Evaluation). COAE 2008.

[6] Weifu Du, Songbo Tan. An Iterative Reinforcement Approach for Fine-Grained Opinion Mining. NAACL-HLT 2009.

[7] Huifeng Tang, Songbo Tan and Xueqi Cheng. A Survey on Sentiment Detection of Reviews. Expert Systems With Applications. Elsevier. 2009, 36(7): 10760-10773.

[8] Aue, A. and Gamon, M. Customizing Sentiment Classifiers to New Domains: a Case Study. RANLP 2005.

[9] John Blitzer, Mark Dredze, Fernando Pereira. Biographies, Bollywood, Boom-boxes and Blenders: Domain Adaptation for Sentiment Classification. ACL 2007.

[10] Songbo Tan, Xueqi Cheng, Yuefen Wang and Hongbo Xu. Adapting Naive Bayes to Domain Adaptation for Sentiment Analysis. ECIR 2009.

[11] Songbo Tan, Xueqi Cheng. Improving SCL Model for Sentiment-Transfer Learning. NAACL-HLT 2009

[12] Songbo Tan, Yuefen Wang, Gaowei Wu, Xueqi Cheng. Using unlabeled data to handle domain-transfer problem of semantic detection. ACM SAC 2008.

[13] Songbo Tan, Gaowei Wu, Huifeng Tang and Xueqi Cheng. A Novel Scheme for Domain-transfer Problem in the context of Sentiment Analysis. ACM CIKM 2007

[14] Songbo Tan, Xueqi Cheng, Moustafa M. Ghanem, Bin Wang, Hongbo Xu. A Novel Refinement Approach for Text Categorization. ACM CIKM 2005

[15] Songbo Tan. An Effective Refinement Strategy for KNN Text Classifier. Expert Systems With Applications. Elsevier. 2006, 30(2): 290-298.

[16] S. Brin, L. Page, R. Motwami, and T. Winograd, The PageRank Citation Ranking: Bringing Order to the Web, Technical Report 1999-0120, Computer Science Department, Stanford University, Stanford, CA, 1999.

[17] Chang CC, Lin CJ. LIBSVM: a library for support vector machines. 2001.