基于权重标准化 SimRank 与半监督学习的产品属性归类*

杨 源，马云龙，林鸿飞

大连理工大学 信息检索研究室，大连 116024

E-mail: yangyuan@mail.dlut.edu.cn; kevinma@mail.dlut.edu.cn; hflin@dlut.edu.cn

摘 要：本文主要把产品评论中属性的不同描述进行归类。在产品评论中，同类的属性会有不同的描述，例如手机的“外形”和“设计”指的是同类属性。同类属性虽然有不同的描述，但是在句中却和相同的情感词搭配使用。本文首先抽取评论句中属性和情感词的搭配关系，形成一个二部图，然后用权重标准化 SimRank 计算不同属性之间的相似度，并把所得的结果与半监督学习中的贝叶斯分类器进行融合，得到了更好的分类结果。通过实验证明了此方法的有效性。

关键词：属性；归类；SimRank；半监督学习

Grouping Product Features Using Weight Normalized SimRank and Semi-supervised Learning

Yang Yuan, Ma Yunlong, Lin Hongfei

Information Retrieval Laboratory of Dalian University of Technology, Dalian 116024

E-mail: yangyuan@mail.dlut.edu.cn; kevinma@mail.dlut.edu.cn; hflin@dlut.edu.cn

Abstract: This paper mainly groups different feature expressions in product reviews into suitable groups. In product reviews, the same feature may have several feature expressions, e.g. “appearance” and “design” of a mobile phone can indicate the same feature. Although the same feature may have several feature expressions, these expressions are always used with same sentimental words in a sentence. Product feature expressions and sentimental words are first extracted in pairs to build a bipartite graph, and then, Weight Normalized SimRank is used to compute similarity between different feature expressions in the bipartite graph, and the similarity is used to optimize the Bayesian classifier in Semi-Supervised Learning. Experimental results show that the proposed method is useful.

Keywords: product features; group; SimRank; semi-supervised learning

1 引言

随着网上购物的发展，网上产品评论的数量也急剧增多，产品评论的情感分析一直是一个热点研究问题。在分析产品情感时，除了分析产品的整体情感外，也从更细的粒度上来分析，如产品的各个属性。由于写评论的人习惯各异，对产品的同一种属性会用不同的描述，例如，手机的“外观”和“外形”指的是同一种属性。把这些属性进行归类，才能更好的进行情感分析。

手工进行属性归类虽然会很准确，但是这是一项耗费时间的工作，需要寻求一种自动的方法来解决这个问题。属性的归类也可以通过现有的同义词资源，把同义的属性归为一类，但是这只能解决其中的一部分，还有很多同类的属性不是同义词，例如，手机的“外形”和“设计”。需要其他的方法更全面的解决这个问题。

Carenini 等人[1]利用 WordNet 得到了几个相似性矩阵，把一些属性描述映射到一个特定领域的属性分类上。Carenini 等人是根据存在的资源来计算词的相似性的，这种方法的缺点是没有考虑词之间的分布相似性，而这是很有用的信息。分布相似性主要考虑了属性周围的一些词，衡量相似性的方法有 Cosine、Jaccard、Dice 等[2]。

* 基金项目：国家自然科学基金资助项目（编号：60673039，60973068）、国家 863 高科技计划资助项目（编号：2006AA01Z151）和教育部博士点基金（编号：20090041110002）。

Guo 等人[3]提出了 mLSA 算法，进行属性归类，mLSA 将 LDA 运行了两次，是一种无监督算法。属性归类问题也与限制性聚类相关，限制性聚类中使用了两种限制，一种是某些节点肯定在一类中，另一种是某些节点不可能在一类中。Andrzejewski 等人[4]把这两种限制引入到 LDA 中，提出了 DF-LDA 算法。

Zhai 等人[5]针对属性归类，提出了一种半监督的 SC-EM 算法，并把以上提到的几种算法作为对比试验，通过实验证明了 SC-EM 算法对属性归类的结果要优于以上几种算法。SC-EM 算法是对 EM 算法的改进，首先从语料中获取每个属性周围的一些词，作为属性对应的文档，然后把其中的一部分属性进行标注，选用朴素贝叶斯模型作为分类器。SC-EM 算法利用了两条自然语言知识，一条是含有相同词的两个属性有可能是同类属性，另一条是同义词的两个属性有可能是同类属性。SC-EM 算法利用这两条知识，得到了更好的初始化效果。

SC-EM 算法既用到了存在的资源，又用到了分布相似性，但是有一个缺点，产品评论中含有丰富的情感信息，对于同类属性，评论中往往会有一些相同的情感词来修饰，例如，手机的“外观”和“外形”常常和“小巧”搭配，而“声音”是不与“小巧”搭配的，显然这些情感词对属性的归类是很有意义的。本文充分考虑了产品评论中的情感因素，从语料中抽取出属性和情感词的搭配对，利用这些搭配对形成二部图，然后用权重标准化 SimRank 算法[6]来计算各个属性之间的相似度，并把所得的结果与 SC-EM 算法中的贝叶斯分类器进行融合，得到了更好的分类结果。

本文的结构安排如下：第 2 节介绍相关术语，第 3 节介绍产品属性和情感词的搭配，第 4 节介绍权重标准化 SimRank 与 SC-EM 算法的结合，第 5 节介绍实验结果和相关分析。

2 相关术语

属性：实验语料选取手机评论，用属性表示手机的一些具体特征，如屏幕、按键等，手机型号或品牌也属于属性的范围。

属性描述：同一类的属性可能会有不同的词或短语来描述，如手机的“外观”和“外形”，这样的词或短语称为属性描述。

同类属性：同类属性是指意思相同的属性，如手机的“外观”和“外形”。

3 属性与情感词的搭配

3.1 产品属性的获取

要进行产品属性归类，首先要获取产品的属性， Hu 等人[7]和 Liu 等人[8]对属性的抽取已经做了很多工作。本文所要解决的主要问题是产品属性的归类而不是抽取，所以预先准备了一个手机属性表，用于属性归类。

3.2 情感词的获取

情感词的识别主要依据大连理工大学信息检索实验室的情感词汇本体[9]。情感词汇本体已经能够识别绝大部分的情感词，但是可能会出现例外的情况，例如，产品评论中有些情感词没有在情感词汇本体中登录，有些产品属性周围找不到情感词。为了更好的解决情感词缺失的问题，把形容词作为了情感词的补充，很多形容词都有情感，而且根据日常用语的习惯，形容词也和产品属性之间存在搭配关系。

3.3 属性与情感词的搭配

产品评论是评论者对产品的评价，有很多都是主观性文本，包含了丰富的情感。根据语言习

惯，对不同的产品属性进行评价时，会用不同的情感词，例如，评价手机的“外观”时，经常用“小巧”，而评价手机的“价格”时，经常用“低廉”。就像修饰名词时要用相应的形容词一样，评价产品属性时也会用相应的情感词，属性和情感词之间自然地产生了一种搭配关系，如表 1 所示

表 1　属性与情感词搭配举例

例句	产品属性	情感词
这款手机的外观比较小巧。	外观	小巧
三星的外形特别小巧。	外形	小巧
索爱的售价比较低廉。	售价	低廉
这款 N97 的价格挺低廉了。	价格	低廉

实验中把离产品属性最近的情感词与产品属性进行搭配，形成候选搭配对，然后通过公式（1）计算情感词与产品属性的互信息。

$$PMI(PF,OW)=\frac{P(PF,OW)}{P(PF)P(OW)} \tag{1}$$

公式（1）中 $PMI(PF,OW)$ 是属性与情感词的互信息，$P(PF,OW)$ 是属性与情感词共现的概率，$P(PF)$ 是属性出现的概率，$P(OW)$ 是情感词出现的概率。

在有限的语料中，情感词和产品属性的共现可能是不均匀的，致使互信息偏低或偏高，为了更好的适应这种不均匀性，先假设含有相同词的两个属性是同类属性，同义词的两个属性是同类属性，计算互信息时，不是计算情感词与单个属性的互信息，而是计算情感词与这类属性的互信息。当然，所有属性中，只有一部分是同义词或含有相同的词，其余的仍计算情感词与单个属性的互信息。实验中主要通过《同义词词林》[10]判断两个词是否是同义词，只是进行简单的判断，并没有利用《同义词词林》的层次信息，避免为初始化引入更多噪音。

计算出所得搭配对的互信息之后，去掉一些互信息过低的搭配对，这些可能是抽取过程中引入的噪音。剩余的搭配对形成了一个二部图的形式。

4　权重标准化 SimRank 与 SC-EM 算法的结合

4.1　SimRank

属性与情感词形成了二部图，可以用 Glen 等人[11]提出的 SimRank 算法来计算属性之间的相似度，SimRank 算法的基本思想是：与相似节点相连的节点相似。这样应用到本文中，被同一个情感词修饰的属性是相似的，修饰同一个属性的情感词是相似的。SimRank 计算公式如下：

$$Sim(v_a,v_b)=\frac{C}{|I(v_a)||I(v_b)|}\sum_{i}^{|I(v_a)|}\sum_{j}^{|I(v_b)|}Sim\left(I_i(v_a),I_j(v_b)\right) \tag{2}$$

在图 $G_T=\{V_T,E_T\}$ 中，$I(v_t)$ 表示节点 v_t 的入边源节点集合，$I_i(v_t)$ 表示节点中第 i 个入边源节点，$|I(v_t)|$ 是节点 v_t 的入度，$Sim(v_a,v_b)$ 表示节点 v_a 和 v_b 的相似度，常数 C 是取值从 0 到 1 的实数，表示相似度在沿有向边传递过程中的衰减系数。

4.2　权重标准化 SimRank

属性与情感词形成的二部图中，不同属性与情感词的相关程度是不一样，它们的互信息也是不一样的，经常共现的属性与情感词应该有较高的互信息，所以连接属性和情感词的边是有权重的，实验中把属性和情感词的互信息作为边的权重，因为它能反映出属性和情感词的相关程度。

直接用 SimRank 计算二部图中属性的相似度会有一个缺点，SimRank 在计算节点间相似度的

时候仅利用了有向图的结构信息，而没有考虑有向边的权重。也就是说直接用 SimRank 进行计算的话，会把相关程度较低的搭配对和相关程度较高的搭配对统一对待，而抽取搭配对的时候难免会引入一些噪音，如果把这些噪音搭配按照正常的搭配进行计算的话，会影响实验结果。为了更好的降低这些噪音的影响，实验采用了马等人提出的权重标准化 SimRank 算法（Weight Normalized SimRank，简称 WNS）。

WNS 算法首先对图中每个节点的入边权重进行标准化，使之对于任意节点 v_i 均满足：

$$\sum_{i}^{|I(v_i)|} \omega\left(I_i(v_i) \to v_i\right) = 1 \tag{3}$$

WNS 的计算公式如下：

$$WNS(v_a, v_b) = \begin{cases} C \sum_{i}^{|I(v_a)|} \sum_{j}^{|I(v_b)|} WNST\left(I_i(v_a), I_j(v_b)\right) & v_a \neq v_b \\ 1 & v_a = v_b \end{cases} \tag{4}$$

$$WNST\left(I_i(v_a), I_j(v_b)\right) = \omega\left(I_i(v_a) \to v_b\right)\omega\left(I_j(v_b) \to v_b\right) WNS\left(I_i(v_a), I_j(v_b)\right) \tag{5}$$

其中，$\omega(v_a \to v_b)$ 表示由 v_a 节点到 v_b 节点的有向边上的权重，其他符号与基础 SimRank 公式中同义。

用 WNS 算法计算二部图中属性的相似性时，既利用了有向图的结构信息，又利用了有向边的权重。假设抽取的搭配对中，有两个属性与同一个情感词搭配，其中有一个是错误的搭配，如果用 SimRank 算法计算，这两个属性会有很高的相似度，而用 WNS 算法计算的话，错误搭配的有向边的权重比较小，从而会降低这两个属性的相似度。

4.3 权重标准化 SimRank 与贝叶斯的结合

如引言所述，Zhai 等人使用半监督的 SC-EM 算法，进行产品属性的归类，本文使用权重标准化 SimRank 对 SC-EM 算法进行了改进。SC-EM 算法首先从语料中获取属性周围左右各 3 个词作为该属性对应的文档，然后对其中的一部分属性标注类别，进行半监督学习，选取贝叶斯分类器，用以下三个公式计算每个属性属于某个类别的概率。

$$P(w_t \mid c_j) = \frac{1 + \sum_{i=1}^{|D|} N_{ti} P(c_j \mid d_i)}{|V| + \sum_{m=1}^{|V|} \sum_{i=1}^{|D|} N_{mi} P(c_j \mid d_i)} \tag{6}$$

$$P(c_j) = \frac{1 + \sum_{i=1}^{|D|} P(c_j \mid d_i)}{|C| + |D|} \tag{7}$$

$$P_1(c_j \mid d_i) = \frac{P(c_j) \prod_{k=1}^{|d_i|} P(w_{d_i,k} \mid c_j)}{\sum_{r=1}^{|C|} P(c_r) \prod_{k=1}^{|d_i|} P(w_{d_i,k} \mid c_r)} \tag{8}$$

其中，D 是训练文档集，d_i 是 D 中的一篇文档，$w_{d_i,k}$ 是文档 d_i 中的第 k 个词，训练集中的总词表是 $V = \{w_1, w_2, ..., w_{|V|}\}$，$C = \{c_1, c_2, ..., c_{|C|}\}$ 是属性的类别集合，N_{ti} 是词 w_t 在文档 d_i 中出现的次数。

通过 WNS 算法可以算出测试集中每个属性与训练集中每个属性的相似度，对应测试集中的每个属性，训练集每个类别的属性中都会有一个属性，使得测试集中的属性与该属性在对应类别中的相似度最大。本文把测试集中属性与该属性的相似度作为测试集中属性与这个类别的相似度，这样可以得到测试集中的属性与各个类别的相似度。本文定义了公式（9），从属性与情感词搭配的角度，计算属性属于某个类别的概率。

$$P_2(c_j \mid f_i) = \frac{WNS(f_i, c_j)}{\sum_{j=1}^{|C|} WNS(f_i, c_j)} \tag{9}$$

其中，f_i 是对应文档 d_i 的属性，$WNS(f_i, c_j)$ 是属性 f_i 与类 c_j 的相似度。

公式（8）是从属性与周围的词共现的角度，公式（9）是从属性与情感词搭配的角度，计算属性的类别，把公式（8）和公式（9）相乘，所得的结果作为属性类别判断的依据更为充分，得到

$$P(c_j \mid d_i) = P_1(c_j \mid d_i) P_2(c_j \mid f_i) \tag{10}$$

通过公式（10）得到测试集中每个属性的类别，然后把训练集和测试集作为训练集进行迭代，完成半监督学习。

5 实验

实验语料是来自 IT168 网站[12]的 31624 句手机评论，其中包含 159 个属性描述，共分为 21 类。选取 SC-EM 算法作为对比试验，采用 SC_EM 中的三个评价标准，实验采用的是半监督学习，可以使用分类准确率作为一个评价标准，另外两个采用聚类的评价标准熵和纯粹度，数据集 DS 上的类别集合是 $C = \{c_1, ..., c_j, ..., c_k\}$，$DS$ 被分为 k 个子集 $DS_1, ...DS_i, ..., DS_k$，熵的计算公式如下

$$entropy(DS_i) = -\sum_{j=1}^{k} P_i(c_j) \log_2 P_i(c_j) \tag{11}$$

$$entropy_{total} = \sum_{i=1}^{k} \frac{|DS_i|}{|DS|} entropy(DS_i) \tag{12}$$

其中，$P_i(g_j)$ 是 c_j 类数据在 DS_i 中所占的比例。纯粹度的计算公式如下

$$purity(DS_i) = \max_j P_i(g_j) \tag{13}$$

$$purity_{total} = \sum_{i=1}^{k} \frac{|DS_i|}{|DS|} purity(DS_i) \tag{14}$$

实验采用交叉验证的方式，选取 SC-EM 算法作为 baseline，把 SC-EM 算法与 WNS 相结合的结果作为最终结果，交叉验证的平均结果如表 2 所示：

表 2　产品属性归并结果

训练集比例	SC-EM			SC-EM+WNS		
	准确率	熵	纯粹度	准确率	熵	纯粹度
20%	53.74%	1.64	0.53	56.36%	1.59	0.57
30%	60.51%	1.47	0.62	64.73%	1.42	0.61
40%	65.47%	1.36	0.65	68.42%	1.28	0.66
50%	68.92%	1.22	0.67	72.36%	1.16	0.69

如表 2 所示，把 WNS 算法融入到 SC-EM 算法中后，提高了准确率，得到更小的熵值和更大的纯粹度，在产品评论中，属性和情感词经常搭配使用，针对不同的属性，评论者会用不同的情感词来表达情感，而同类属性描述的是产品的同一个部分，可以用相同的情感词来修饰，利用属性和情感词的搭配，以及属性和情感词的互信息形成的二部图，用 WNS 算法利用二部图的信息算出各个属性的相似度，并与 SC-EM 算法中的贝叶斯分类器进行结合，所得的结果要好于只用贝叶斯分类的结果，两者结合的方式既考虑了同类属性周围有相同的词共现，又考虑了评论句中属性与情感词的搭配，所以得到了更好的属性归类结果。

随着训练集比例的增大，实验结果增长的幅度越来越小，这一方面是因为有些属性通过较小的训练集就能得到比较准确的分类，另一方面是因为 SC-EM 算法通过同义词和含有相同词的属性

优化了初始化过程，新增加的训练集有可能已经出现在优化的过程中，所以这个优化过程，比较有用，可以用较小的训练集得到更好的结果。

实验过程中也会有一些因素，影响实验结果，分词过程会产生一定的偏差，从而会影响产品属性周围词的抽取，也会影响属性与情感词搭配对的抽取。有些属性在《同义词词林》中未登录，这也会用影响半监督学习中初始化的优化过程。

6 结论

本文主要把产品评论中同类属性的不同描述进行归类。同类属性虽然有不同的描述，但是在句中却和相同的情感词搭配使用。本文首先抽取评论句中属性和形容词的搭配关系，形成一个二部图，然后用权重标准化 SimRank 计算不同属性之间的相似度，并把所得的结果与半监督学习中的贝叶斯分类器进行融合，得到了更好的分类结果。

本文进行产品属性归类时，仍存在一定的不足。在利用属性与情感词搭配构成的二部图计算属性之间的相似度时，难免会引入一些噪音搭配，本文主要通过互信息来降低噪音搭配中属性与其他属性的相似度，能更精确的发现搭配，会得到更好的分类结果。下一步的工作可以研究属性与情感词之间的搭配关系，在比较全面发现搭配的同时，保障搭配的精度。

参 考 文 献

[1] Carenini G, R. Ng and E. Zwart. Extracting knowledge from evaluative text[C]. Proceedings of International Conference on Knowledge Capture, Banff, Canada, 2005: 8-15.

[2] Lee L. Measures of distributional similarity[C]. Proceedings of ACL. Maryland, USA, 1999: 25-32.

[3] Guo H., H. Zhu, Z. Guo, X. Zhang and Z. Su. Product feature categorization with multilevel latent semantic association[C]. Proceedings of CIKM. Hong Kong, 2009: 1087-1096.

[4] Andrzejewski D., X. Zhu and M. Craven. Incorporating domain knowledge into topic modeling via Dirichlet forest priors[C]. Proceedings of ICML. Montreal, Quebec, Canada, 2009: 25-33.

[5] Zhongwu Zhai, Bing Liu, Hua Xu and Peifa Jia. Grouping Product Features Using Semi-Supervised Learning with Soft-Constraints[C]. Proceedings of the 23rd International Conference on Computational Linguistics (COLING-2010), Beijing, China, August 23-27, 2010: 1272-1280.

[6] Y. L. Ma, H. F. Lin, and S. Jin. A revised simrank approach for query expansion[C]. Proceedings of the 6rd Asia Information Retrieval Societies Conference(AIRS 2010). Springer, December, 2010: 564–575.

[7] Minqing Hu and Bing Liu. Mining and summarizing customer reviews[C]. Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery & Data Mining (KDD-2004), Seattle, Washington, USA, Aug 22-25, 2004: 168-177.

[8] Lei Zhang and Bing Liu. Extracting and Ranking Product Features in Opinion Documents[C]. Proceedings of the 23rd International Conference on Computational Linguistics (COLING-2010), Beijing, China, August 23-27, 2010: 1462-1470.

[9] 徐琳宏, 林鸿飞, 潘宇等. 情感词汇本体的构造[J]. 情报学报, 2008, 27(2): 180-185.

[10] HIT-IRLab-同义词词林（扩展版），哈尔滨工业大学信息检索研究室: http://ir.hit.edu.cn/

[11] G. Jeh, J. Widom. SimRank: A measure of structural-context similarity [C]. In Proceedings of SIGKDD. Edmonton, Alberta, Canada, 2002: 538-543.

[12] IT168 网站 http://pinglun.it168.com.