DOI:10.16836/j.cnki.jcuit.2015.04.010

文章编号: 1671-1742(2015)04-0355-06

基于协同过滤和文本相似性的 Web 文本情感极性分类算法

张朝龙， 许源平， 郑皎凌
(成都信息工程大学软件工程学院,四川 成都 610225)

摘要: Web 文本情感极性分类算法在网络舆情监控方面具有重要的研究价值。针对传统文本分类算法依赖于情感词典的弊端,以及不能很好的应用于不规则的 Web 文本分类的局限性,提出基于协同过滤和文本相似度的 Web 文本情感极性分类算法。先统计分析网络文本高频词汇覆盖情况,进而根据统计结果,基于协同过滤和余弦相似度计算提出一种新的 Web 文本情感极性分类算法,其利用余弦相似度方法计算出 Web 文本的相似性,判断文本的情感极性。对于无法直接判断情感极性的文本,该算法设计了协同过滤中的情感词评分以及 Top-N 情感词推荐机制,且通过对情感词的评分与推荐输出进行多次迭代相似度计算来判断未知 Web 文本情感极性。最后使用中文情感挖掘语料(ChnSentiCorp)进行实验。结果表明,算法具有较高的查全率和查准率,在不规则的 Web 文本下也表现出较好的分类效果,可较实用地解决 Web 文本情感极性分类问题并应用于网络舆情监控。

关 键 词: 计算机应用技术; 智能信息处理; 文本情感分类; 舆情监控; 协同过滤; 余弦相似度; Web 文本
中图分类号: TP391.1 **文献标志码:** A

0 引言

在大数据时代,互联网已成为非常重要的舆论平台,现今人们不仅是网络的消费者,同时也是网络信息的生产者,越来越多的人愿意在网络上表达自己的观点、态度和想法[1]。网络上每天都会有海量的舆论话题产生,一个舆论话题可以通过微博、论坛、朋友圈迅速传播与扩散,网络舆情会直接影响社会舆情的走向,进而产生重大的社会影响[2]。通过舆情监控可有效识别网络舆情,过滤垃圾、有害、恶意的信息,针对发布的虚假信息进行溯源,对维护互联网的正常秩序具有重要的意义[3]。文本情感分类算法可有效识别网络文本所表达的情感,有助于及早发现恶意、有害的信息,因此,Web 文本的情感极性分类算法在网络舆情监控应用方面具有重要的研究价值。

在文本的情感极性分类方面,前人已经做了大量的相关研究工作,并取得了较好的研究成果。目前文本情感分类方面主要有两种研究思维: 基于情感词典和基于机器学习的算法[4-5]。

基于情感词典的方法如 Tong[6] 通过人工选择建立了专用情感词库,Hu[7] 通过判断新词的方法解决了 Tong 建立的词库只能用于特定领域的问题。中文方面李纯等[8] 从语言学角度出发,提出中心词概念来进行情感极性分类。

在另一方面,基于机器学习的方法中常用的算法有 Rocchio 分类算法、SVM(支持向量机)、Bayes(贝叶斯算法)、KNN(K-近邻算法)、互信息(Mutual Information)等。Pang 等[9] 学者分别应用 Native Bayes、ME(Maximum Entropy)、SVM 方法对电影评论语料进行了情感分类。孟迪等[10] 学者提出的 H-C 算法改进了传统互信息方法情感项可信度的计算方法,在一定程度上改善了概率偏向的问题。

当前主要的情感极性分类方法存在以下问题: (1) 大都是以单词为基础,忽略了句子级和篇章级对文本情感的影响; (2) 情感词库一般都是针对某一特定领域,不具有通用性; (3) 情感词库中情感词的情感只有正面和负面两种,忽略了不同情感词对句子和篇章的影响权重; (4) 由于网络文本的随意性,结构复杂,语法错误,使得传统机器学习方法的效果大大降低。例如,H-C 等机器学习算法大都难以处理不规范的文本[10],特别是 Web 文本。

在前人的基础上,针对传统情感分类方法存在的缺点,综合协同过滤和文本相似度计算,提出了一种新型高效的 Web 文本情感极性分类方法。协同过滤算法广泛应用于推荐系统[11],其可被分类为基于用户(User-Based)的算法和基于项目(Item-based)的算法。在文本情感极性分类算法中,可把文本看作用户(Us-

收稿日期: 2015-07-29
基金项目: 国家自然科学基金资助项目(61203172、61202250); 四川省应用基础计划资助项目(2012JY0111)

er)，把文本中的情感词看作项目(Item)。其基本原理是利用已知情感的文本(用户)获取情感词(项目)信息，再预测未知情感的文本的情感极性。该算法具有以下优点：(1)不是简单地统计单词，而是从句子级和篇章级来整体分析文本；(2)对多个领域的文本都能有效的分类，具有较强的通用性；(3)不依赖于情感词库；(4)基于文本相似度计算，对网络不规范的文本也具备较好地处理能力。

课题设计并实现的算法包含：文本-情感词矩阵表示、相似度计算、协同过滤情感词评分和Top-N情感词推荐。

实验表明，基于协同过滤和文本相似性的情感极性分类算法相比于传统的基于情感词典和基于机器学习的算法具有更高的查全率和准确率。

1 文本情感词覆盖率情况统计实验

中文文本中常用字出现的频率非常高，国家语言文字工作委员会和国家教育委员会发布的《现代汉语常用字表》中常用字只有2500个，次常用字1000个。山西大学计算机科学系统计了200万字的语料，检测常用字的使用频率和覆盖率，其结果是：常用字覆盖率达97.97%[12]。

中文词汇方面也有类似的特点，北京语言学院语言教学所对200万字现代汉语语料进行了统计，列出了使用度最高的前8000词词表。可见中文文本使用的词汇大都是出自高频词汇。

基于这个前提，对网络文本进行了统计。从凤凰新闻、新浪新闻、天涯论坛及地方论坛共提取了2670篇文档，对其进行词频统计，其中1500篇用于统计词频，1170篇用于词频覆盖率统计。统计结果如表1所示。其中平均覆盖率和80%覆盖率所占的比率增长图1如所示。

表1 文本词组覆盖率统计

文档数	统计词数	最高覆盖率/%	最低覆盖率/%	平均覆盖率/%	80%覆盖率所占的比率/%
100	10305	96.61	24.63	79.01	47.69
300	22174	100	32.87	89.77	97.44
500	26706	100	34.27	92.15	98.63
700	30501	100	33.58	93.57	99.23
900	33678	100	36.61	94.34	99.32
1100	37426	100	38.02	94.99	99.32
1300	41464	100	38.48	95.55	99.66
1500	43819	100	38.48	95.88	99.74

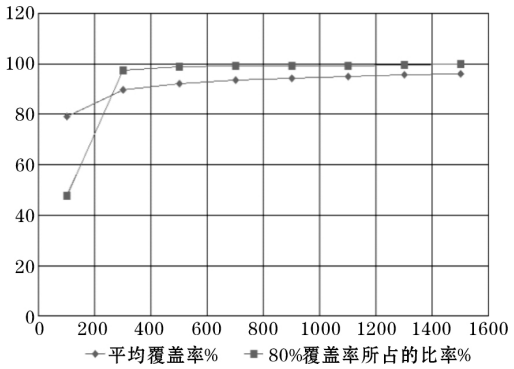

图1 文本词组覆盖增长率统计图

考虑影响文本情感极性的词组的词性包含形容词、名词、动词、状态词、代词[10]。其中，形容词大多修饰某种情感，所以目前大多数文本情感分类算法都以形容词为基础。但是，根据语料分析，发现引入形容词的同时将名词、动词、状态词、代词一并考虑会大大提高情感分类的准确性(实验结果见表7)，因为虽然大多数名词、动词、状态词和代词本身并没有情感倾向，但拥有相同情感极性的文本中通常会重复出现某些词性为名词、动词、状态词和代词的词汇，比如在文中使用的情感挖掘语料[13]中关于酒店的评论中包含的“房间”这个名词在负面评论中出现的频率非常高，如“房间很旧，有霉味，居然还有蟑螂”、“房间地毯还有沙发都很脏”，这里的“房间”、“蟑螂”都是名词，本身也没有极性，但它们出现在文本中都是为了说明房间不好，即为负面情感服务的。因此，把形容词、名词、动词、状态词、代词都识别为情感词。

利用上述文本针对情感词进行了词频统计，统计结果如表2所示。其中平均覆盖率和80%覆盖率所占的比率增长图如图2所示。

表2 文本情感词覆盖率统计

文档数	统计词数	最高覆盖率/%	最低覆盖率/%	平均覆盖率/%	80%覆盖率所占的比率/%
100	8260	94.34	16.53	74.34	24.01
300	17726	100	24.45	89.77	92.22
500	21343	100	26.06	90.73	96.41
700	24116	100	26.98	92.47	97.69
900	26495	100	28.01	93.42	98.63
1100	29340	100	29.62	94.26	99.15
1300	32528	100	29.85	94.98	99.32
1500	34274	100	29.85	95.39	99.49

从表1、表2、图1、图2可以看出，文档数从100篇增加到300篇时，词汇的覆盖率和情感词的覆盖率快速上升(全部词汇的平均覆盖率从79.01%上升到89.77%，情感词的覆盖率从74.34%上升到89.77%)，词汇80%覆盖率所占的比率急剧上升(全部词汇从

47.69% 上升到97.44%，情感词从24.01% 上升到92.22%）。当文档数达到 500 篇时，所包含的词汇和情感词在测试文档中均能达到90% 以上，且覆盖率达80% 以上的文档所占的比率在90% 以上，覆盖率趋于稳定状态。这和北京语言学院语言教学所做的统计结果一致。实验结果表明，少量文档中出现的词汇基本能覆盖大多数文档中的词汇，这类词汇就是常用高频词汇。

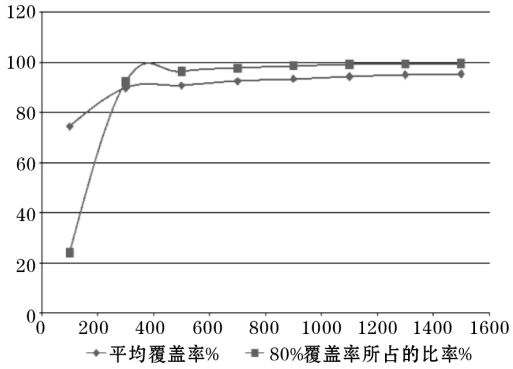

图 2　情感词覆盖增长率统计图

2　协同过滤在文本情感极性分类的应用

2.1　文本 - 情感词矩阵表示法

从前一章的统计词汇覆盖率的相关实验结果可以得出大部分文本使用的词汇都为常用高频词汇。一篇文本包含多个情感词，一个情感词也会出现在多篇文本中。使用协同过滤算法，可把文本看作用户（User），把文本中的情感词看作项目（Item），从而得到一个二维表，其中每行表示为一个文本（User），每列表示一个情感词（Item），而值是文本中出现相应情感词的得分情况，这样就生成了文本-情感词评分矩阵如表 3 所示。

表 3　文本-情感词评分矩阵 $\boldsymbol{R}(m, n)$

	Item_1	…	Item_k	…	Item_n
User_1	$\boldsymbol{R}_{11}$	…	$\boldsymbol{R}_{1k}$	…	$\boldsymbol{R}_{1n}$
…	…	…	…	…	…
User_i	$\boldsymbol{R}_{i1}$	…	$\boldsymbol{R}_{ik}$	…	$\boldsymbol{R}_{in}$
…	…	…	…	…	…
User_j	$\boldsymbol{R}_{j1}$	…	$\boldsymbol{R}_{jk}$	…	$\boldsymbol{R}_{jn}$
…	…	…	…	…	…
User_m	$\boldsymbol{R}_{m1}$	…	$\boldsymbol{R}_{mk}$	…	$\boldsymbol{R}_{mn}$

2.2　文本相似度计算

相似度的计算方法很多，如常用的有相关相似性（Correlation Similar）[14]、余弦相似性（Cosine Similarity）[15] 和修正、余弦相似性（Adjusted Cosine Similarity）[14]，和基于云模型的相似度[14]。相关相似性也称皮尔森（Pearson）系数相关，可用于度量用户之间的相似性。余弦相似性是通过计算用户间的向量夹角的余弦值来度量用户之间的相似性。修正的余弦相似性是对余弦相似性的一种改进，主要改善了不同用户评分标准不一致所带来的缺陷。云模型把每个用户看作一朵云，通过逆向云算法计算两朵云的相似度[15]。

不同的 Web 文本，因为文本长度不同，向量规模也不同，比如一篇 10000 字的文本比一篇 500 字的文本的向量维度规模大很多，比较规模没有太大意义。余弦相似度计算通过计算两个向量的夹角来度量它们的相似度，在比较过程中，向量的规模不予考虑，仅仅考虑向量的方向，可有效避免因文本规模不同而产生的差异。通过计算向量的夹角可有效地判断文本的相似度。因此，本文采用余弦相似性计算两个文本的相似度，计算公式为

$$\text{sim}(i,j) = \cos(i,j) = \frac{\vec{i}}{\vec{j}} = \frac{\sum_{c=1}^{n} R_{i,C} R_{j,c}}{\sqrt{\sum_{c=1}^{n} R_{i,c}^2}\ \sqrt{\sum_{c=1}^{n} R_{j,c}^2}} \quad (1)$$

其中，$R_{i,c}$ 和 $R_{j,c}$ 分别表示文档 i 和文档 j 对情感词 $c(\text{Item}_c)$ 的评分。

2.3　文本情感极性输出

根据文本相似性可计算出待预测情感极性的目标文档的最近邻居集合。为输出文本情感极性，还需应用的基本步骤包括: 求出文档对任意情感词的评分和文本 Top-N 的情感词推荐。

文档对任意情感词的评分。对于文档 u_k，最近邻居集合 $U = \{u_1, u_2, \cdots, u_p\}$，$u_k$ 不属于 U，从 u_1 到 u_p，$\text{sim}(u_k, u_i)$ 从大到小排列。可利用相似邻居进行中心加权求和的方法来给情感词 i 评分

$$P_{uk,i} = \overline{R_u} + \frac{\sum_{j \in U} \text{sim}(u_k, u_j) \times (R_{u_j,i} - \overline{R_{u_j}})}{\sum_{j \in U} |\text{sim}(u_k, u_j)|} \quad (2)$$

其中，$P_{u_k,i}$ 表示文档 u_k 对情感词 i 的评分，$\overline{R_{u_j}}$ 表示文档 u_j 在已经打分的情感词的平均分，$\text{sim}(u_k, u_j)$ 是余弦相似度。

Top-N 的情感词推荐。分别统计待预测“最近邻居”集中的文档对不同情感词的评分的加权平均值，其中 N 个排在前面且不属于 I_i（I_i 表示用户 i 评分的情感词集合）的情感词作为 Top-N 推荐集。在得出 Top-N 推荐集后，文档获得了新的情感词推荐，对相似度低于阈值的文档使用新的文本 - 情感词向量矩阵计

算文本相似度。通过迭代使用相似性计算和协同过滤算法,可进一步判断待预测文本的情感极性。

3 情感极性分类算法实现

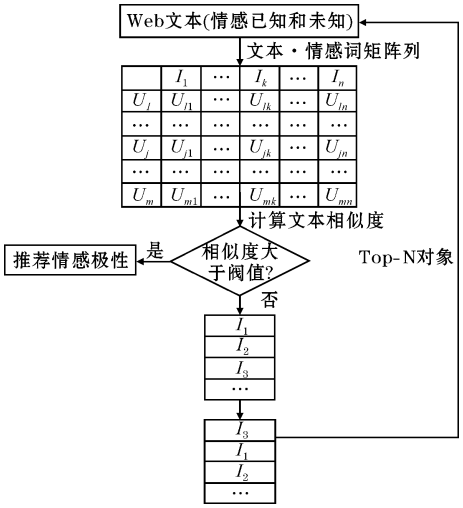

图3 基于协同过滤和文本相似性的文本情感极性分类流程

表4 文档-情感词表

文档	情感词	情感极性
1	高兴、幸福、友好、真挚	正面
2	高兴、友好、清甜	未知
3	高兴、幸福、清甜、真挚	未知
4	友好、差、无语、吐槽	负面
5	高兴、友好、无语、吐槽	未知

表5 文档-情感词评分矩阵

文档	高兴	幸福	友好	清甜	真挚	差	无语	吐槽
1	1	1	1	0	1	0	0	0
2	1	0	1	1	0	1	0	0
3	1	1	0	1	1	0	0	0
4	0	0	1	0	0	1	1	1
5	1	0	1	0	0	0	1	1

图3总结了该基于协同过滤和文本相似性的文本情感极性分类算法的总体处理流程。先对已知情感极性和待预测情感极性文本做预处理,提取文本中的情感词,形成一个文档-情感词二维矩阵。再使用余弦相似性计算方法,式(1)计算文本间的相似度,当待预测文档与已知情感极性的文档的相似度大于阈值时,可直接对该文本推荐情感极性,其情感极性为相似度最大的已知情感极性的文档的情感极性。对于低于阈值的文档,需要继续使用协同过滤算法计算文档对情感词的评分和Top-N情感词推荐,并进一步计算文本相似度。

假设有5篇文档,文档所包含的情感词和情感极性如表4所示,其中文档1和文档4的情感已知,其它文档为待预测的文档。先把文档和情感词矩阵化,得到文档-情感词评分矩阵,如表5所示。再使用余弦相似性计算式(1)分别计算文档之间的相似度,计算结果如下

$$\mathrm{sim}(1,2)=\cos(1,2)=\frac{1}{\sqrt{4}}=0.5$$

$$\mathrm{sim}(1,3)=\cos(1,3)=\frac{3}{4}=0.75$$

$$\mathrm{sim}(1,4)=\cos(1,4)=\frac{1}{4}=0.25$$

$$\mathrm{sim}(1,5)=\cos(1,5)=\frac{1}{\sqrt{4}}=0.5$$

$$\mathrm{sim}(2,3)=\cos(2,3)=\frac{1}{\sqrt{3}}=0.577$$

$$\mathrm{sim}(2,4)=\cos(2,4)=\frac{1}{\sqrt{4}}=0.5$$

$$\mathrm{sim}(2,5)=\cos(2,5)=\frac{1}{4}=0.25$$

$$\mathrm{sim}(3,4)=\cos(3,4)=\frac{0}{\sqrt{4}+\sqrt{4}}=0$$

$$\mathrm{sim}(3,5)=\cos(3,5)=\frac{1}{4}=0.25$$

$$\mathrm{sim}(4,5)=\cos(4,5)=\frac{3}{4}=0.75$$

设置相似角度阈值为$45°\left(\cos(45°)=\frac{\sqrt{2}}{2}=0.707\right)$,则文档1和文档3的相似度超过阈值,可直接推荐文档3的情感极性和文档1一致,即文档3也为正面。文档4和文档5的相似度也超过阈值,推荐文档5的情感极性为负面。

文档2和文档3相似度低于阈值,需要迭代计算。其中文档2的最近邻居为$U=\{u_1,u_3,u_4,u_5\}$,由公式2计算文档2的情感词“幸福”、“真挚”、“差”、“无语”和“吐槽”的评分,结果如下

$$P_{u,2}=1$$
$$P_{u,5}=1$$
$$P_{u,6}=0$$
$$P_{u,7}=0$$
$$P_{u,8}=0$$

再根据Top-N推荐情感词,可认为文档2也包含“幸福”和“真挚”这两个情感,再次计算余弦相似度为$\mathrm{sim}(1,2)=\cos(1,2)=\frac{1}{\sqrt{6}}=0.82$,此时相似度已超过阈值,可推荐文档2的情感极性为正面。

4 实验结果分析

4.1 实验数据

文中的实验数据是由数据堂提供的中文情感挖掘

语料-ChnSentiCorp[13]，包含对酒店、笔记本电脑（简称电脑）和书籍 3 个领域的评论语料，每个领域包含负面文档和正面文档各 2000 篇。分别对这 3 个语料进行测试，从每个语料库中选取 500 篇负面文档和 500 篇正面文档作为已知情感极性文档，剩余 3000 篇作为待预测文档。

4.2　文本情感极性分类结果及分析

分别对酒店、电脑和书籍 3 个领域的文本进行测试，从查全率、查准率和 F 值评估算法的性能指标。查全率、查准率和 F 值的定义如下。

查全率和查准率。在情感极性分类中，以负面情感为例，负面文档并分类为负面的文档数为 a，负面文档被分类为正面的文档数为 b，正面文档被分类为负面的文档数为 c。则查全率为正确判断为负面的文档数与所有负面文档总数之比，即 $R=a/(a+b)$。查准率为正确判断为负面的文档数与所有判断为负面的文档数之比，即 $P=a/(a+c)$。

F 值为查全率 R 和查准率 P 的函数

$$F=2P\times R/(P+R) \tag{3}$$

文中与文献[10]的 H-C 算法、文献[16]的基于情感词典和朴素贝叶斯的算法进行了对比。通过测试并调整阀值，得出较好的结果，如表 6 所示。

表 6　实验结果分析比较

语料集	H-C			基于情感词典和朴素贝叶斯			文中方法		
	R/%	P/%	F/%	R/%	P/%	F/%	R/%	P/%	F/%
书籍	81.15	80.67	80.93	79.34	78.02	78.67	82.08	81.86	81.96
酒店	82.54	81.09	81.65	78.54	79.86	79.19	84.55	83.662	84.08
电脑	82.68	81.96	82.01	80.35	81.44	80.89	83	84.5	83.7

从表 6 可知，相比于 H-C 方法、基于情感词典和朴素贝叶斯的方法，基于协同过滤和文本相似性的算法具有更高的查全率、查准率和 F 值，在这 3 个领域的语料集中都有明显的提高。由于 Web 文本的用词随意性以及语法的不规范性，同时，在一篇负面文档中也会出现正面句子，正面文档也出现正面句子，及一篇文本有一个主要情感，而在某一小的方面予以不同于主情感的肯定或否定，使得 H-C 算法不能很好的处理这类文本。而文中算法基于文本相似度计算方法，且训练集和测试集来源相同，因而在不规范的 Web 文本也能表现较好的分类效果。传统使用情感词库的方法，情感词的情感极性大都是非正即负，忽略了不同词汇对情感的影响权重和文本间的相似性，也忽略了句子和篇章对文本情感极性的影响，这些弊端使得基于情感词典的分类算法分类效果不理想。而文中的情感分类方法，使用文本相似性计算有效地避免了传统情感词库的弊端，通过文本相似性的分析并调整相关阀值，使得分类效果具有显著的提高。

为对比不同词性种类作为情感词对文本情感极性分类效果的影响，文中对只选取形容词作为情感词和选取本文定义的情感词两个方法应用基于协同过滤和文本相似性的算法进行对比，实验结果如表 7 所示。

表 7　形容词和本文定义情感词的实验结果对比

情感词词性	形容词			形容词、名词、动词、状态词、代词		
	R/%	P/%	F/%	R/%	P/%	F/%
书籍	79.6	78.2	78.9	82.08	81.86	81.96
酒店	80.4	77.8	79.1	84.55	83.662	84.08
电脑	81.5	82.8	82.1	83	84.5	83.7

从表 7 可知，与只把形容词作为情感词的方法相比较，把形容词、名词、动词、状态词、代词都识别为情感词时，算法在这 3 个领域的语料集中的查全率、查准率和 F 值都有一定的提高。因为相同情感极性的文本会重复出现类似的词汇，这类词汇的词性不仅包含形容词，还包含了名词、动词、状态词、代词，使得这类词虽然从字意上看没有情感倾向，但在整个句子甚至整个篇章中都会隐含地表现出某种情感的倾向。所以相对于单纯的以形容词为情感词的方法，加入名词、动词、状态词、代词后，分类效果具有明显的提高。

5　结束语

提出一种基于协同过滤和文本相似性的 Web 文本情感极性分类算法。方法使用余弦相似性计算文本的相似度，通过协同过滤推荐文本情感极性和 Top-N 情感词推荐，能够更精确地判断文本的情感极性。该算法改进了传统词库忽略句子和篇章以及传统互信息方法概率偏向的弊端。实验表明，方法在 Web 文本情感极性分类应用上表现出较好的结果，相比传统互信息方法具有更高的查全率和查准率（实验结果见表 6）。该算法在海量网络文本舆情自动化监测方面具有较好的实用性。

从实验中可以看出，阀值的选择对算法的效果会有很大的影响，并且不同领域的语料集会需要不同的最佳阀值，而最佳的阀值是难以计算得出。在将来的工作中，还需要进一步研究自适应阀值选择，同时改进相似度计算方法，实现多个特征值提取，以进一步提高

分类精度和效率。

参考文献:

[1] 于帅. 中文 Web 文本情感倾向性分析技术的研究 [D]. 哈尔滨: 哈尔滨工程大学,2013.

[2] 王平,谢耘耕. 突发公共事件网络舆情的形成及演变机制研究 [J]. 中国传媒大学学报,2013,(3):63 -69.

[3] 杨志国. 基于 Web 挖掘和文本分析的动态网络舆情预警研究 [D]. 武汉: 武汉理工大学,2014.

[4] 李光敏,许新山,熊旭辉. Web 文本情感分析研究综述 [J]. 现代情报,2014,(5):173 -176.

[5] 王洪伟,刘勰,尹裴,等. Web 文本情感分类研究综述 [J]. 情报学报,2010,29(5):931 -938.

[6] Tong,R M. An operational system for detecting and tracking opinions in on-line discussions [C]. Working Notes of the ACM SIGIR 2001 Workshop on Operational Text Classification. New York, NY: ACM,2001:1 -6.

[7] Minqing Hu,Bing Liu. Mining and Summarizing Customer Reviews [C]. Association for Computing Machinery (ACM) Special Interest Group on Knowledge Discovery and Data Mining(SIGKDD) International Conference on Knowledge Discovery and Data Mining; 20040822 -20040825; Seattle, WA; US,2004.

[8] 李钝,曹付元,曹元大,等. 基于短语模式的文本情感分类研究 [J]. 计算机科学,2008,(4).

[9] Pang B,Lee L,Vaithyanathan S. Thumbs up Sentiment Classification using Machine Learning Techniques [J]. Proceedings of Emnlp, 2002:79 -86.

[10] 孟迪,李立宇,于津. 基于情感项区分极性可信度的文本情感分类 [J]. 汕头大学学报: 自然科学版,2014,(3).

[11] 郭艳红. 推荐系统的协同过滤算法与应用研究 [D]. 大连: 大连理工大学,2008.

[12] 郑泽之,王强军,张普,等. 基于大规模 DCC 语料库的《现代汉语常用字表》、《现代汉语通用字表》收字情况统计分析 [J]. Advances,2003.

[13] 数据堂. 中文情感挖掘语料-ChnSentiCorp [EB/OL]. http://www.datatang.com/data/14614.

[14] 徐翔,王煦法. 协同过滤算法中的相似度优化方法 [J]. 计算机工程,2010,(6):52 -54.

[15] 张光卫,李德毅,李鹏,等. 基于云模型的协同过滤推荐算法 [J]. 软件学报,2007,18(10):2403 -2411.

[16] 杨鼎,阳爱民. 一种基于情感词典和朴素贝叶斯的中文文本情感分类方法 [J]. 计算机应用研究,2010,27:3737 -3739.

The NovelSentiment Classification Algorithm for Web Texts based on Collaborative Filtering and Text Similarity

ZHANG Chao-long, XU Yuan-ping, ZHENG Jiao-ling

(College of software Engineering,Chengdu University of Information Technology,Chengdu,610225)

Abstract: The sentiment classification algorithm has important research value for applications of web texts based network monitoring public opinion. To overcome the limitations that traditional sentiment classification algorithms depend heavily on their built sentiment word bases and they are not suitable for nonstandard web texts, we proposed a novel sentiment classification algorithm for nonstandard web texts based on the collaborative filtering and text similarity theories. This paper starts with a comprehensive evaluation of the coverage of high-frequency words in web texts. And based on the evaluation results, we proposed a novel collaborative sentiment classification algorithm based on the innovation theories of the collaborative filtering and text similarity computing. It calculates the similarity among huge large amounts of web texts by using the cosine similarity equation, and then automatically judge sentiments for corresponding web texts. For texts unable to judge sentiments directly, this algorithm application of sentiment word score and Top-N sentiment word recommendation on collaborative filtering, and judge sentiment of web texts by similarity computing using iterative way. Finally, the devised algorithm has been tested and evaluated by using the ChnSentiCorp data from internet. Experiments show that this algorithm has high recall and precision, and also better result for nonstandard web texts. It can solve nonstandard web text classification problem better and practically applied to applications of network monitoring public opinion.

Key words: technology of computer application; intelligent information processing; sentiment classification; public sentiment monitoring; collaborative filtering; cosine similarity; web texts