DOI:10.16707/j.cnki.fjpc.2017.04.007

基于改进贝叶斯算法的电商在线评论倾向性研究

李宛真[1]，王兴芬[2]
（1 北京信息科技大学 计算机学院 北京 100101；
2 北京信息科技大学 信息管理学院 北京 100192）

【摘 要】本文用采集器抓取上万条在线评论语料，通过数据清洗、分词以及语义标注进行主题词抽取并确定评论语料针对商品、店家服务以及物流三大主题的好评、中评、差评。将训练集导入朴素贝叶斯算法中进行训练，为避免算法在本项目中出现下溢，本实验采用自然对数处理，优化后的算法避免下溢或浮点数舍入导致的对结果的不精确判断。将改进后的算法用于电商网站在线评论数据集的好评率分类，实验结果表明，对乘积取对数，将连乘变为连加，更适用于电商评论的分类问题。

【关键词】在线评论；贝叶斯算法；情感倾向

1 引言

据 CNNIC 报告，截止 2015 年 12 月，我国网络购物用户规模达 4.13 亿，同比增加 5183 万，增长率为 14.3%，高于 6.1%的网民数量增长率。2015 年中国网络购物市场继续保持快速发展。当年度全国网络零售交易额达 3.88 万亿元，同比增长 33.3%。然而，伴随其快速的发展，网络零售市场不乏存在网络诈骗，网络售假，物流、售后等服务欠佳等诸多问题[1]，严重制约其健康持续的发展。因此，本文试图从第三方的视角，通过对消费者的在线评论情感倾向的挖掘与验证，得出一种较准确的判定电商网站卖家真实可信度高低的方法。

2 国内外研究进展

通过查阅近 10 年有关 C2C 电商网站的相关文献，发现信任匮乏是电子商务的症结所在，对信任的概念，可总结为 经济交易的一方认为另一方是可靠的并且能够履行自己的承诺[2]。而这种承诺通过在线评论，即消费者基于个人使用经验创造出来的一种代表性的电子口碑直观表现出来，因此在线评论的有用性与可信度是研究领域近年来讨论最多的话题。McKnightetal 认为只有当消费者认为所接收到的信息具有可信度时，他们将会更有信心去采纳该电子口碑并且使用它来帮助自己做出购买决策[3]。学者秦良娟以中国 C2C 电子商务平台中的卖家为研究对象，从结构保障视角研究了卖家对电子商务平台制度信任形成的影响因素，通过问卷调查，并使用 SPSS19.0 对数据进行了信度和效度检验，采用最大似然估计方法对模型中的潜变量进行了验证性因子分析[4]。对中文词语而言，北京大学的王治敏等人[5]基于《人民日报》基本标注语料库的真实文本实例进行统计归纳，得到词语的情感倾向。物理学中的 Spin 模型也被用来估计单个词语的倾向性极性[6]，多数学者以消费者的在线评论着手，从应用看，商用商品信息反馈系统 OpinionObserv-er[7]利用网络上丰富的顾客评论资源生成反馈信息。姚天昉等[8]选择汽车评论作为语料，构建了一个汽车评论情感分析系统。Pang 等人[9]比较了不同类型的分类器在倾向性分类问题上的性能，使用了与传统文本分类相似的特征(Unigram，Bigram 等)，同时也最早在电影评论语料中提出电影情节和电影评论之间的耦合对倾向性分类的影响。然而针对在线评论是由“主题”和“倾向性”两个不同维度，在实际分类任务中会给分类器的设计带来区别。在文献[10]中首先研究了 structural correspondence learning(SCL)在跨领域倾向性分类上的应用。中国科学院计算技术研究所的吴琼等提出了基于图(Graph based)的迁移方法[11]，谭松波等人则采用了基于朴素贝叶斯分类器的迁移学习方法[12]。

综上所述，我们从电商网站在线评论入手，通过对多个电商网站、多个店家上万条在线评论的爬取、清洗以及预处理，借助贝叶斯分类算法和词频统计分析编写判断评论情感倾向性的分类器，挖掘出 C2C 电商网站店家的真实可信度。通过语料搜集，将在线评论分为三大主题，分别为针对商品的评论（好评、中评、差评）、针对店家的评论（好评、中评、差评）、针对物流（好评、中评、差评）的九类评论，继而通过分词以及类别标注训练出分类器，通过测试用例判断其有效性，最终分别爬取淘宝网信用高低不同的若干店铺的在线评论，利用分类器计算出每家店铺分别对商品，店家和物流的好评占比，并对每一个主题按照好评率进行信用排名，并与网站提供的搜索结果进行对比，最终得到一个较好的分类模型。

3 评论倾向性挖掘模型

消费者在网购平台购买商品以后，平台提供消费者对该订单的评价功能。消费者对网购的评价内容主要集中在产品的认可度上，部分涉及店家的服务水准，以及对物流的评价。部分平台明确要求对某些要素，比如店家或物流，进行满意度打分。而目前多数平台则是基于用户对购买行为的评价进行满意度评分。这种机制有利于消费者分享购买感受，为其他消费者提供参考，改进店家服务，其后端还可为生产商、物流供应商提供参考。为更好地挖掘网购评价的价值，有必要对其描述的关键信息进行更细致地提炼。

评论倾向性挖掘模型如图 1 所示：

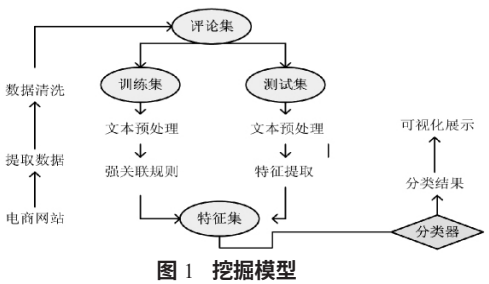

图 1 挖掘模型

基金项目：国家自然科学基金资助项目“网络零售交易风险动态评估及预警研究”（71571021）

3.1 在线数据采集

我们从淘宝网入手，实验中所使用的数据通过淘宝网搜索 iPhone6 从得到的店铺里面抓取而获得，利用八爪鱼数据抓取程序[13]针对各个店家建立抓取任务，提取 Ajax 页面加载评论，设置评论循环及翻页循环，最后提取要提取的数据字段，即“累计评价”，采集各店家的在线评论数据。可以选择单机采集或者云采集方式，并设置每周定时采集。

3.2 数据预处理

从上述数据集中随机选出 10000 条在线评论作为本实验的原始训练数据，首先进行数据清洗，将重复的数据、小于五个字符的数据以及非字符的无关数据清洗掉[14]，之后将干净数据按照特征进行语义标注[15]，主要分为三大类，如表 1 所示。与语法标注不同的是，语义标注的过程及标注内容均建立在语义 Web 和关联数据技术上，即通过规则处理技术进行语义标注，然后按类别抽取出对应主题句，当然此时的针对三大类别的评论倾向既有好评又有差评，因此，我们邀请到 9 个标记员进一步对三类评论进行倾向性标记，最终得到分类结果如表 2 所示。

表 1 评论文本按主题词特征分类

类别	类别特征	评论示例
商品	产品外观、组件、功能	手机好漂亮，屏幕也大，非常好用。
	东西颜色、价钱等	机子没有任何问题，物超所值，已经第二次购买了。
		正品，但就是信号不太好。
店家 (服务)	老板、售后、发票	店家服务蛮好的。
	服务态度等	给老板点个赞，非常耐心，服务杠杠的。
物流	物流、快递、运费	物流很给力，第二天就到了。
	快递小哥、快递公司名称等	快递小哥真心不错，当面验货，服务周到。

表 2 评论文本倾向性分类

类别	倾向性	评论示例
商品	好评	机子非常棒，漂亮大气，用起来也很好，愉快的购物。
	中评	感觉一般，价格也没有想象中的便宜，再用用看。
	差评	机子差评，有大问题，已发短信就卡机，垃圾。
店家 (服务)	好评	店家服务超级好啊，售后专业，客服细心，好评。
	中评	卖家还说的过去，售后一般般吧。
	差评	老板态度也太差了吧，客服半天没人回复，很生气。
物流	好评	物流超快，外省两天就到了。
	中评	快递一般般吧，差不多说的过去，中评。
	差评	这个送货速度真心不敢恭维，太失望了，实在渣。

3.3 实验论证

我们从消费者对网购的评价入手，首先按信用度排名搜索出销售 iphone6 的各官方旗舰店，如三际数码官方旗舰店、淮坊联通官方旗舰店、迪信通官方旗舰店、能良官方旗舰店、环球机库官方旗舰店、神通数码官方旗舰店。分别统计出各旗舰店下的如实描述、服务态度以及物流服务的分数，如表 3 所示。统计完毕之后，抓取上述各店家的在线评论，通过数据清洗等预处理之后作为实验的测试集，导入到朴素贝叶斯算法分类器之后，运行结果通过 flask 可视化展示。

表 3 按信用度排名的店铺如实描述得分统计

	店家名称	如实描述	服务态度	物流服务	抓取评论条数
1	官方旗舰店 1	4.82	4.79	4.78	1449
2	官方旗舰店 2	4.80	4.77	4.76	5896
3	官方旗舰店 3	4.86	4.74	4.71	7080
4	官方旗舰店 4	4.85	4.79	4.73	1617
5	官方旗舰店 5	4.78	4.71	4.73	1331
6	官方旗舰店 6	4.79	4.70	4.72	200

根据贝叶斯定理计算后验概率公式(1)为：

$$P(y_i \mid x)=\frac{P(x \mid y_i)P(y_i)}{P(x)} \quad (1)$$

($x=\{a_1,a_2,\cdots,a_m\}$，a_i 为待分类的特征属性；$C=\{y_1,y_2,\cdots,y_m\}$ 为有类别集合)

分母对于所有类别来说都是为常数，因此只需计算分子即可。又因为各个特征具有相同的地位且各自出现的概率相互独立，所以得(2)：

$$P(x \mid y_i)P(y_i)=P(a_1 \mid y_i)P(a_2 \mid y_i)\cdots P(a_m \mid y_i)P(y_i)=P(y_i)\prod_{j=1}^{m}P(a_j \mid y_i) \quad (2)$$

将表 3 中六家旗舰店的评论文本测试集导入到训练好的分类器中进行运算，计算出倾向分别为好评，中评和差评的百分比，由于各分词概率连乘，使得结果小至 10 的数十次方分之一，导致数据下溢，计算机精度丢失，都近似计为零了，而几个类别所占比例相同，故结果就只能取第一个好评率为百分之百，中评率和差评率均为零。这种最初的算法明显有悖于我们对原始数据的分析判断。

4 改进贝叶斯算法的实验与分析

在贝叶斯定理计算后验概率的公式中 P(ajlyi)是关键，其含义就是样本各个特征在各个分类下的条件概率，计算各个划分的条件概率是朴素贝叶斯分类的关键性步骤，也就是数据训练阶段的任务所在。实际项目中，概率 P 往往是值很小的小数，连续的微小小数相乘容易造成下溢出使乘积为 0 或者得不到正确答案，比如当本实验在线评论稍微长一点，分出的关键词概率连乘之后结果会非常小，比如有 10 个词，每个词出现的概率为千分之一，那么计算结果就是 1000 的 10 次方之一，就是 10 的 30 次方分之一，由于计算机精度限制，只能取近似值零，继

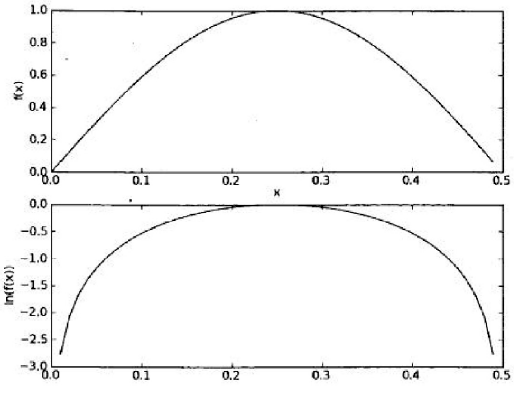

图 2 f(x)和 ln(f(x))函数的对比曲线

而被作为误差忽略掉。因此本实验的解决办法就是对乘积取自然对数，将连乘变为连加，ln(AB)=lnA+lnB。采用自然对数处理不会带来任何损失，可以避免下溢出或者浮点数舍入导致的错误。下图给出了 f(x)和 ln(f(x))函数的曲线，对比可以发现 相同区域内两者同增或者同减、在相同点取得极值，因此采用自然对数不会影响最终比较结果。因此上式的计算可以转变为其对数的计算，概率的比较也可以转换为概率对数的比较。

已知数学公式：

$$\ln\left(P(a_1 \mid y_i)P(a_2 \mid y_i)\ldots P(a_m \mid y_i)\right) = \sum_{j=1}^{m} \ln P(a_j \mid y_i) \tag{3}$$

由公式(1)(2)(3)得：

$$P(y_i \mid x) = \frac{P(y_i)\sum_{j=1}^{m} \ln P(a_j \mid y_i)}{\mathrm{P(x)}} \tag{4}$$

综上所述，公式(4)即为我们改进后的公式，用它来构造朴素贝叶斯分类器，我们通过改进后的分类器将评论倾向性好评率、中评率、差评率计算出来，结果如表 4 所示，使用 python 编写的轻量级 web 应用框架—Flask[17]可视化展示，将上面分析结果可视化表示出来，将这三个排序分别与各店家的对应的商品描述，店家服务、物流服务的官方得分进行对比，并画出折线图 3，图 4，图 5，通过改进后的贝叶斯算法训练的分类器，完全可以通过对在线评论的挖掘，正确的判定消费者对卖家商品、店家服务以及物流评价的倾向性。

表 4 算法改进后的实验结果

		旗舰店 1	旗舰店 2	旗舰店 3	旗舰店 4	旗舰店 5	旗舰店 6
商品	好评率	88.83%	88.74%	88.49%	88.15%	85.82%	73.33%
	中评率	1.71%	1.40%	0.47%	0.94%	1.16%	6.67%
	差评率	9.47%	9.86%	11.04%	10.91%	13.02%	20.00%
店家	好评率	91.09%	90.73%	90.57%	90.79%	88.27%	86.67%
	中评率	1.13%	1.73%	0.77%	1.47%	1.62%	6.66%
	差评率	7.78%	8.55%	8.66%	7.73%	10.12%	6.67%
物流	好评率	77.51%	75.60%	72.18%	78.02%	74.75%	73.33%
	中评率	0.57%	1.15%	0.37%	0.90%	0.83%	0.05%
	差评率	21.92%	23.25%	27.45%	21.08%	24.42%	6.62%

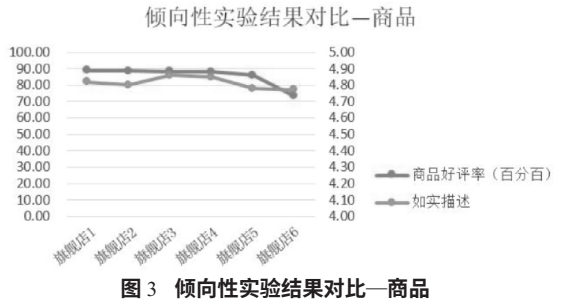

图 3 倾向性实验结果对比—商品

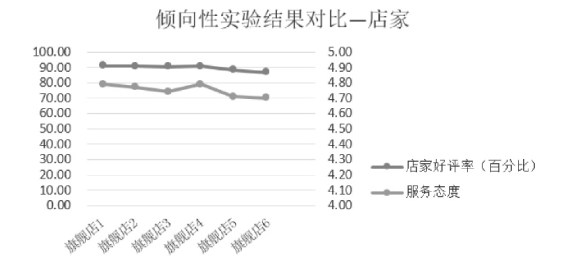

图 4 倾向性实验结果对比—店家

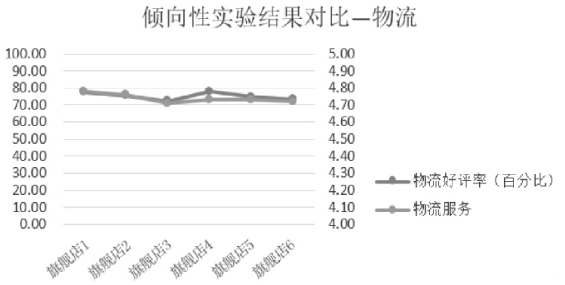

图 5 倾向性实验结果对比—物流

5 不足与展望

为了能更好的支持决策，还可对结果进行深入分析。比如，当系统分析出用户对某一产品的“服务”方面不满意，可找到“服务”这一分类词表中所包含的主题词，根据用户对主题词的倾向占比确定用户是对该产品的售后服务还是网购配送服务不满意，以帮助企业做出更好的决策。

参考文献：

[1] 潘煜,张星,高丽. 网络零售中影响消费者购买意愿因素研究——基于信任与感知风险的分析 [J]. 中国工业经济,2010,(07):115- 124.

[2] 余世欣,巫孝君. 我国电子商务中信任问题研究综述 [J]. 中小企业管理与科技(上旬刊),2010,(02):232- 233.

[3] McKnight D H，Choudhury V，Kacmar C． The impact of initial consumer trust on intentions to transact with a Web site: a trust building model [J]． Journal of Strategic Information Systems，2002，11 (3/4): 297- 232.

[4] 秦良娟,徐琦琦,曹淑艳. C2C 电子商务平台中卖家制度信任的影响因素研究——基于结构保障视角 [J]. 国际商务(对外经济贸易大学学报),2014,(06):116- 125.

[5] 王治敏,朱学锋,俞士汶. 基于现代汉语语法信息词典的词语情感评价研究 [C]//Recent advancement in Chinese Lexical Semantics, Proceeding of 5th ChineseLexical Semantics Workshop (CLSW- 5),2004, Sin- gapore.

[6] Hiroya Takamura, Takashi Inui, Manabu Okumura Extracting semantic orientations of words using spinmodel [C]// Proceedings of ACL 2005.

[7] LiuB,HuMQ,ChengJS.OpinionObserver:AnalyzingandComparingOpinionsontheWeb [C]. In:Proceedingofthe14thInternationalConferenceonWorldWideWeb(WWW'05). NewYork:ACM,2005:342－351.

[8] 姚天昉. 一个用于汉语汽车评论的意见挖掘系统 [A]. 中国中文信息学会. 中文信息处理前沿进展——中国中文信息学会二十五周年学术会议论文集 [C]. 中国中文信息学会:,2006:22.

[9] Bo Pang, Lillian Lee.Shivakumar Vaithyanathan Thumbs up Sentiment Classification using Machine Learning Techniques [C]//Proceedings of the Confer- ence on Empirical Methods in Natural Language Pro- cessing (EMNLP),2002.

[10] John Blitzer ,Mark Dredze,Fernando Pereira,etal.Boom boxes and Blenders:Doma in Adaptation for Sentiment Classification [C]//Proceedings of the As sociation for Computational Linguistics(ACL)2007.

[11] 吴琼. 基于图排序模型的跨领域倾向性分析算法 [A]. 中国中文信息学会.中国计算机语言学研究前沿进展(2007- 2009) [C]. 中国中文信息学会:,2009:6.

[12] Songbo Tan,Xueqi Cheng,Yuefen Wang,Etal.Adapting Naïve Bayes to Domain Adaptation for Sentiment Analysis [C]//Proceedings of 31th EuropeanConference on Information Retrieval(ECIR)2009:337- 349.

[13] 崔玉洁,廖坤. 借助八爪鱼采集器实现过刊网刊元 (下转第 22 页)

3 数据处理分析

Excel 作为当今普及性的、不用编程就能处理数据使用工具，每一步操作都能即时见到结果，且操作简单方便，可以提高工作效率，在相似处理数据过程中只需处理一次，可以把计算过程固定下来，录入对应数据即得到相关运算结果 还可以保留计算数据便于后期核对和验算。将 Excel 运用到实验数据的处理中可以帮助使用者容易发现其中的规律和特点。

由于补偿电容 C1 的取值的随机性，实验者选取的补偿电容值不尽相同，我们可以借助 Excel 中的数据排序功能，以补偿电容为主要关键字，升序排列，即可得到图 5 所示的补偿电容 C 和功率因数 cosφ 的对照表，通过表 5 可以让实验者观测补偿电容和功率因数的变化规律。若表 5 的规律不是很明显，还可以利用 Excel 的图表生成功能绘制出功率因数 cosφ 随补偿电容 C 变化曲线。如图 6 所示。

补偿电容 C(uF)	功率因数 COSΦ
0.00	0.370
2.30	0.563
3.60	0.752
4.00	0.822
4.80	0.950
4.90	0.962
5.00	0.973
5.10	0.982
5.20	0.989
5.50	1.000
5.90	0.986
8.00	0.664
9.10	0.523
9.50	0.483

图 5 补偿电容和功率因数对照表

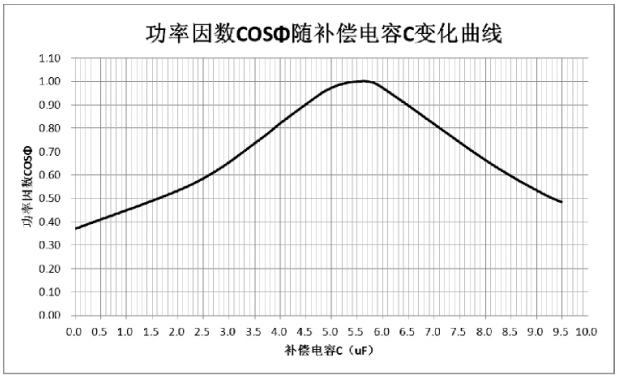

图 6 功率因数随补偿电容变化曲线

从图 5 的补偿电容和功率因数对照表和图 6 的功率因数随补偿电容变化曲线我们可以明显的看出并补偿电容在其有效变化范围内改变时，功率因数几乎呈正态分布 ;当补偿电容为 5.5uF 左右时，日光灯等的功率因数达到几乎为 1 最大值；而当补偿电容过大或过小时，日光灯的功率因数反而降低了。可见对于不同的日光灯感性电路要根据实际情况具体分析，选择恰当的补偿电容，以进一步提高其功率因数，在当今提倡的节约型社会和能源日益匮乏的现状有着十分重要和实际的意义。通过基于 Excel 和 Multisim 使得日光灯功率因数提高在学生在方便、安全的操作过程中，直观形象的掌握其重点和难点。

4 结语

基于 Multisim 和 Excel 的日光灯功率因数提高仿真分析，让学生对功率因数提高和实验数据处理分析及涉及到的相关概念在虚拟仿真的过程有所掌握和巩固，激发了学生的学习兴趣和主动性，可以看到把 Multisim 和 Excel 软件引入电路基础课教学，能够帮助学生牢固地掌握电路的基础知识，为以后所学的专业知识打下坚实基础。

参考文献：

[1] 王银. 基于 Multisim10 的电子技术课程实验仿真研究 [J]. 延安大学学报(自然科学版), 2010.
[2] 朱其祥,谢道平,陈劲松. 基于 Proteus 的提高电路功率因数仿真分析 [J]. 实验室科学, 2014.
[3] 刘亚兰,张雅娟. Multisim10 在高职数字电子技术教学中的实践 [J]. 时代教育, 2015.
[4] 杨奇,赖康荣,刘红. 基于 Multisim10 的电工学仿真—日光灯功率因数的提高与改善 [J]. 电脑知识与技术, 2013.
[5] 张建红,郑文,贺琳. 基于 Multisim 10 的日光灯电路及功率因数提高的仿真分析 [J]. 无线互联科技, 2015.
[6] 张立萍,柴万东. 利用 Multisim10 软件提高数字电路的教学质量 [J]. 赤峰学院学报(自然科学版), 2015.
[7] 江有永. 基于 Multisim 和 Excel 的二极管特性仿真实验 [J]. 实验技术与管理,2011..
[8] 顾菊平,包志华,钱骏. 功率因数提高的教学探讨和实践 [J]. 电气电子教学学报,2002.

作者简介：

闵卫锋(1977－),男 (汉) ,陕西省渭南市人,讲师,在读硕士,主要研究方向为电子技术仿真实验教学。

(上接第 19 页)

数据的自动提取 [J]. 编辑学报,2016,(05):485- 488.
[14] 郭志懋,周傲英. 数据质量和数据清洗研究综述 [J]. 软件学报,2002,(11):2076- 2082.
[15] 章昉,颜华驹,刘明君,赵中英. 基于词项关联的短文本分类研究 [J]. 集成技术,2015,(03):69- 78.
[16] Wegener D ,Mock M, Adranale D,etal.Toolkit- based high- 213.performance data mining of largedata on mapreduce clusters . [C] //IEEE International Conference on Data Mining Work- Shops,2009:296- 301
[17] 叶锋. Python 最新 Web 编程框架 Flask 研究 [J]. 电脑编程技巧与维护,2015,(15):27- 28.

作者简介：

李宛真 (1990-) ,女 (汉) ,河南平顶山市鲁山县人,学生,硕士,主要研究方向为电子商务与 WEB 安全 ;王兴芬 (1968-) ,女(汉),黑龙江哈尔滨人,教授,博士,主要研究方向为物流管理、电子商务与 WEB 安全。