

一种基于威尔逊区间的商品好评率排名算法

徐林龙¹, 付剑生², 蒋春恒¹, 林文斌²

(1. 西南交通大学 数学学院, 四川 成都 610000;

2. 西南交通大学 物理科学与技术学院, 四川 成都 610000)

摘要:传统基于商品好评率的排名算法在处理小样本的评价数据时,存在着明显的缺陷问题,例如小样本的排名准确性问题等。为了解决这个问题,文中通过引入威尔逊置信区间估计的概念,提出了一种利用置信区间下限值来代替好评率的改进算法。该算法综合考虑了商品好评率与评论数,能有效解决好评率排名存在的小样本准确性问题。通过真实数据上的实验表明,无论是小样本数据还是大样本数据,改进算法都能避免以上问题并提供更为可信的排名结果。

关键词:电子商务;商品排名;好评率;威尔逊区间估计

中图分类号:TP301.6

文献标识码:A

文章编号:1673-629X(2015)05-0168-04

doi:10.3969/j.issn.1673-629X.2015.05.040

A Product Ranking Algorithm Based on Wilson Interval of Users' Positive Ratings

XU Lin-long¹, FU Jian-sheng², JIANG Chun-heng¹, LIN Wen-bin²

(1. School of Mathematics, Southwest Jiaotong University, Chengdu 610000, China;

2. School of Physics Science and Technology, Southwest Jiaotong University, Chengdu 610000, China)

Abstract: When dealing with products with few rating information from users, the product ranking results using traditional ranking algorithm based on positive ratings are not reasonable. To solve the problem, introduce the concept of Wilson confidence interval estimation and propose an improved ranking method of using the confidence interval lower value instead of positive ratings. This algorithm comprehensively considers both the positive ratings and the total number of ratings from users, which can effectively solve the small sampling accuracy problem existed in positive raking. Experiments on real data set indicate that this improved method produces better ranking results not only for the small sampling data but also for big sampling data.

Key words: e-commerce; product ranking; positive rating; Wilson interval estimation

0 引言

随着电子商务的迅速发展,如何针对用户提交的关键词,检索出关联商品并给出相应排名,已经成为一个重要的研究课题^[1-3]。大部分购物网站都会提供多种商品排名列表,比如销量排名、价格排名、好评率排名等,以便于用户根据自身需要选择不同类型的排名列表。不同的购物网站采取的排名算法也不同,一个优秀的排名算法可以为用户提供舒适便捷的购物体验,并增加网站和商家的营业收入。文中从购物网站的好评率排名列表入手,分析了传统基于好评率的排名算法存在的小样本上的评价不可信问题,综合考虑了商品好评率与评论数,引入威尔逊置信区间估

计^[4-5]的方法,实现对商品排名进行合理评价。

1 基于好评率的排名算法

大多数网络购物网站都提供商品评价与评分功能,即在用户购买商品后,对商品做出相应评价。对于商品评价,存在多种类型的评价系统,从简单的好评和差评,到复杂的五分评分制,等等,据此收集到的大量用户对于商品的反馈信息,将构成一个庞大的数据集。如何充分利用这个数据集对商品进行排名,成为了当下电子商务研究领域的一个热门话题。

针对购物网站中的用户评价,为简单起见,文中假定:

收稿日期:2014-07-04

修回日期:2014-10-07

网络出版时间:2015-04-22

基金项目:教育部新世纪优秀人才支持计划项目(NCET-10-0702)

作者简介:徐林龙(1989-),男,硕士,研究方向为数据挖掘;林文斌,教授,研究方向为高性能计算、数据挖掘。

网络出版地址: <http://www.cnki.net/kcms/detail/61.1450.TP.20150422.1005.016.html>

(1)用户对商品的评价分为三个等级:好评、中评、差评;

(2)每个用户对商品的评价是一个独立事件。
一般而言,网站评价系统会根据用户的评价信息,计算每件商品的好评率,依此作为基本的评价指标。假设商品获得的好评数为 a ,中评数为 b ,而差评数为 c ,则总评论数为:

$$n = a + b + c$$

定义如下形式的商品好评率:

$$p = (a + 0.5b)/n \times 100\%$$

该好评率的定义考虑了中评,即将中评视为半个好评。由于很多网站采用的是两个等级的评价体系,即只有好评与差评,则有 $b = 0$ 。此时,商品好评率即是简单的“好评数占总评论数的比例”。

根据好评率直接对商品进行排名是绝大多数网络购物网站采纳的排名策略:好评率越高的商品排名越靠前。这种排名方法存在明显的缺陷,它忽视了总评论数对好评率的影响。这种影响更多地体现在好评率的可信程度。表1示例可直观地反映这一问题。

表1 基于好评率对商品(A-E)排名

排名	商品名称	好评率/%	评论数
1	A	100	1
2	B	100	26
3	C	99	33
4	D	90	100
5	E	89	300

表格的第一列代表商品排名,第二列为商品名称,第三列为商品的好评率,而最后一列为商品的总评论数。比如,商品A的好评率为100%,评论数为1;商品C的好评率为99%,评论数为33。如果基于传统的好评率评价方法,商品A的排名在商品C之前。但根据统计抽样理论,可以断定使用好评率进行评价时,商品C相比商品A更为可信。此外,比较商品A和商品B,两者的好评率均为100%,但商品B的评论数要远大于商品A,如果基于传统的好评率评价方法,无法对两者进行区分。同样地,根据统计抽样理论,可以断定商品B相比商品A更为可信。因此,对于任意两个商品,如果两者好评率相同,还应当考虑其各自的好评数,进行综合评判。如果两者好评率不同,评论数也会对最终排名造成一定程度的影响。

好评率的可信问题对于多种类型的商品都存在。表2是京东商城在线商品数据的统计结果。为了能够体现商品差异性,选择六种典型的商品(iphone5、手表、笔记本、耳机、运动鞋和键盘)进行检索,并统计检索结果中评论数不超过10个且好评率为100%的商

品(简称“置信存疑商品”)数目。

表2 置信存疑商品统计数据

关键词	关联商品数量	置信存疑商品数量	置信存疑商品所占比例/%
iphone5	81	23	28.4
手表	506	150	29.6
笔记本	112	26	23.2
耳机	94	10	10.6
运动鞋	706	170	24.1
键盘	152	22	14.5

根据检索结果,六种商品中置信存疑商品比例最少的是10.6%,最高接近于30%,由此可见,置信存疑商品是一个普遍存在的问题。而基于好评率的商品排名方法在处理该类商品时存在严重不足。文中引入威尔逊置信区间改进传统的基于好评率的排名方法。

2 基于威尔逊置信区间的商品排名算法

2.1 置信区间

假设 θ 是总体的一个参数,其参数空间为 $\Theta, x_1, x_2, \cdots, x_n$ 是来自该总体的样本,对给定的一个 $\alpha(0 < \alpha < 1)$,若有两个统计量 $\theta_L = \theta_L(x_1, x_2, \cdots, x_n)$ 和 $\theta_U = \theta_U(x_1, x_2, \cdots, x_n)$,使得对任意的 $\theta \in \Theta$,有

$$P_{\theta}(\theta_L \leq \theta \leq \theta_U) \geq 1 - \alpha$$

则称随机区间 $[\theta_L, \theta_U]$ 为 θ 的置信水平为 $1 - \alpha$ 的置信区间, θ_L 和 θ_U 分别成为 θ 的置信下限和置信上限^[6]。

例如,某一商品的好评率为80%,但是这个值不一定可信,根据统计学知识,只能说有95%的把握可以断定,好评率在75%到85%之间,即置信区间是 $[75\%, 85\%]$ 。

置信区间是指由样本统计量所构造的总体参数的区间估计,其实质就是进行可信度的修正,以弥补样本量过小的影响。在根据好评率对商品进行排序的时候,如果该商品评论数较多,就说明结果比较可信,不需要很大的修正;相反,如果商品数量较小,必须进行较大的修正,否则就会出现上面提到的问题。

置信区间的计算方式有很多种,常见的是轴枢量法,但它只适用于大样本情形,对于小样本估计的准确性较差。为此,文中引入威尔逊置信区间,以解决小样本估计准确性和可信度问题。

2.2 威尔逊置信区间的性质

威尔逊区间是美国数学家Edwin Bidwell Wilson在1927年提出的一个针对二项分布的置信区间问题的修正公式,已应用于Reddit社区评论排名问题^[7]。

威尔逊置信区间的计算依赖于物品的好评率、评论数,它的计算公式如下:

$$\frac{p + \frac{1}{2n}z_{1-\alpha/2}^2 \pm z_{1-\alpha/2}\sqrt{\frac{p(1-p)}{n} + \frac{z_{1-\alpha/2}^2}{4n^2}}}{1 + \frac{1}{n}z_{1-\alpha/2}^2}$$

其中, p 表示好评率; n 表示评论数; $z_{1-\alpha/2}$ 表示对应某个置信水平的 z 统计量, 在 95% 的置信水平下, z 统计量为 1.96。

威尔逊区间的下限值为:

$$\frac{p + \frac{1}{2n}z_{1-\alpha/2}^2 - z_{1-\alpha/2}\sqrt{\frac{p(1-p)}{n} + \frac{z_{1-\alpha/2}^2}{4n^2}}}{1 + \frac{1}{n}z_{1-\alpha/2}^2}$$

根据威尔逊置信区间的计算公式, 可以确定置信区间的宽度为:

$$\frac{2z_{1-\alpha/2}\sqrt{\frac{p(1-p)}{n} + \frac{z_{1-\alpha/2}^2}{4n^2}}}{1 + \frac{1}{n}z_{1-\alpha/2}^2}$$

此外, 还可以确定该置信区间的均值为:

$$\frac{p + \frac{1}{2n}z_{1-\alpha/2}^2}{1 + \frac{1}{n}z_{1-\alpha/2}^2}$$

当 n 足够大时, 该均值趋向于 p , 即表明当商品评论数足够大时, 置信区间越窄, 它的下限值接近于好评率, 此时商品的排名主要根据好评率高低; 相反, 当 n 小于一定值时, 该均值远小于 p , 置信区间越宽, 它的下限值与好评率的差也越大, 此时好评率对商品排名的影响削弱。

通过上面的分析, 可以得到以下结论: 当评论数 n 足够大时, 可以直接依据商品的好评率 p 作为商品排名的标准。但对于小样本情形, 置信区间跨度较大, 评论数 n 越小, 威尔逊区间置信下限与好评率的差越大, 如果使用好评率进行排名其可信度就存在问题。因此, 采用威尔逊区间的下限值来代替好评率 p 。

2.3 威尔逊置信区间的改进

z 统计可以用来检验两个平均数之间差异的显著程度, 适用于大样本情形(一般地, 样本数要大于 30)。对于小样本情形, 人们通常使用 t 统计进行检验。给定置信水平, z 统计量和 t 统计量都可以通过查表得到, 均为常数。为了解决传统基于好评率的排名方法在处理小样本问题时的局限, 利用 z 统计和 t 统计的上述性质, 对威尔逊区间作如下改进: 当评论数不大于 30 时, 使用 t 统计量代替式中的 z 统计量。

2.4 排名算法

本节根据威尔逊置信区间给商品排名, 主要包括以下三个基本步骤:

Step1: 计算商品的好评率 p ;

Step2: 基于威尔逊置信区间估计, 依据如下规则计算好评率的置信下限:

· 当 $n > 30$ 时, 取置信水平 95%, 查表可知 z 统计量为 1.96, 使用如下公式计算商品排名分值:

$$\text{Score} = \frac{p + \frac{1}{2n}z_{1-\alpha/2}^2 - z_{1-\alpha/2}\sqrt{\frac{p(1-p)}{n} + \frac{z_{1-\alpha/2}^2}{4n^2}}}{1 + \frac{1}{n}z_{1-\alpha/2}^2}$$

· 当 $0 < n \leq 30$ 时, 取置信水平 95%, 查表可知自由度为 30 的 t 统计量为 2.042 3, 使用如下公式计算商品排名分值:

$$\text{Score} = \frac{p + \frac{1}{2n}t_{1-\alpha/2}^2 - t_{1-\alpha/2}\sqrt{\frac{p(1-p)}{n} + \frac{t_{1-\alpha/2}^2}{4n^2}}}{1 + \frac{1}{n}t_{1-\alpha/2}^2}$$

· 当 $n = 0$, 即商品没有评论, 则记 $\text{Score} = 0$ 。

Step3: 根据商品排名分值 Score 对商品进行排名, 分值越高则排名越靠前, 反之亦然。

3 实验结果分析

文中使用网络爬虫技术, 从京东商城抓取大量的在线商品数据, 解析并保存用户对商品的评价数据。在对数据进行预处理过后, 选取大量具有代表性的关键词检索商品数据库, 将传统基于好评率的排名方法得到的结果, 与文中提出的基于威尔逊置信区间的改进排名算法得到的结果进行对比分析^[8-9]。

任选其中一个检索词, 表 3 为两种排名算法的对比结果。忽略了商品的具体名称, 第一列是基于商品好评率的排名, 第二列表示商品的好评率, 第三列为商品评论数, 最后两列分别表示基于改进排名算法计算得到的商品评分与排名。由表 3 可以看到, 在基于好评率排名中排名靠前的置信存疑商品, 通过改进算法的适当调整, 排名出现下滑。改进算法综合利用商品的好评率与评论数刻画用户对于商品的态度和偏好。

表 4 为文中改进算法的排名结果, 第一列是改进算法排名, 第二列表示商品的好评率, 第三列为商品评论数, 第四列表示改进算法的商品评分。

通过表 4, 可以发现在排名靠前的商品中, 并没有出现上文提到的置信存疑商品, 相比于基于好评率的商品排名结果, 改进算法的排名结果有效地解决了评论数过小时的好评率可信度维恩难题, 显得更为合理可信。

当前, 大多数网络购物网站的商品排名算法都属于商业机密, 并没有对外公布。文中提出的改进排名算法有效地降低了小样本评论数对商品排名的不利影

响,获得的商品排名结果更为合理和可信^[10-11]。

表3 两种排名算法的对比结果

基于好评率 的商品排名	<i>p</i>	<i>n</i>	改进算法商品评分	改进算法的 商品排名
1	100%	11	0.725 1	21
2	100%	10	0.705 7	26
3	100%	9	0.683 3	29
4	100%	6	0.589 9	38
5	100%	5	0.545 2	40
6	100%	5	0.545 2	41
7	100%	3	0.418 4	52
8	100%	2	0.324 1	61
9	100%	1	0.193 4	64
10	100%	1	0.193 4	65
11	97.6%	42	0.876 8	3
12	97.3%	37	0.861 8	8
13	96.9%	32	0.842 6	10
14	96.7%	60	0.886 4	2
15	96.4%	14	0.718 9	22
16	96%	50	0.859 7	6
17	96%	50	0.859 7	7
18	95.8%	12	0.684 3	31
19	95.5%	22	0.771 8	14
20	95%	10	0.640 1	35

表4 改进算法的排名结果

改进算法的 商品排名	<i>p</i>	<i>n</i>	改进算法的 商品评分
1	94.8%	134	0.893 3
2	96.7%	60	0.881 4
3	97.6%	42	0.870 3
4	92.6%	162	0.872 4
5	91.2%	261	0.869 3
6	96.0%	50	0.859 7
7	96.0%	50	0.859 7
8	97.3%	37	0.854 6
9	88.5%	379	0.847 5
10	96.9%	32	0.834 6
11	93.0%	64	0.834 9
12	88.2%	250	0.833 9
13	90.7%	81	0.820 3
14	95.5%	22	0.771 8
15	89.4%	47	0.767 8
16	89.2%	37	0.745 7
17	89.7%	34	0.744 3
18	94.4%	18	0.731 0
19	100%	11	0.725 1
20	96.4%	14	0.718 9

4 结束语

传统基于好评率的商品排名算法对于小样本情形,无法给出合理的商品排名结果。文中引入威尔逊

置信区间的概念,综合考虑商品好评率与商品评论数,使用威尔逊置信区间估计下限值代替单纯的好评率,并以 *t* 统计量的使用改进基于威尔逊区间的商品排名算法。通过抓取京东的实际商品数据,对比结果发现,改进算法提供的排名结果更为合理、可信。文中仅限于商品好评率排名的研究,未来工作中,将提取出更多的商品特征(如商品关键词、商品价格、商家信誉、运费等)^[12-14],从不同角度分析各因素对商品排名的影响,设计出一个更健壮、更精确的商品排名算法。

参考文献:

[1] Feng Q,Hwang K,Dai Y. Rainbow product ranking for upgrading e-commerce[J]. IEEE Internet Computing, 2009, 13(5):72-80.

[2] Zhang K,Narayanan R,Choudhary A. Voice of the customers: mining online customer reviews for product feature-based ranking[C]//Proceedings of the 3rd conference on online social networks. [s.l.]:USENIX Association,2010.

[3] Mohanty B K, Passi K. Web based information for product ranking in e-business:a fuzzy approach[C]//Proceedings of the 8th international conference on electronic commerce:the new e-commerce:innovations for conquering current barriers, obstacles and limitations to conducting successful business on the internet. [s.l.]:ACM,2006:558-563.

[4] Wilson E B. Probable inference,the law of succession and statistical inference[J]. Journal of the American Statistical Association,1927,22:209-212.

[5] Agresti A,Coull B A. Approximate is better than ‘exact’ for interval estimation of binomial proportions[J]. The American Statistician,1998,52(2):119-126.

[6] 茆诗松,程依明. 概率论与数理统计教程[M]. 北京:高等教育出版社,2004.

[7] Miller E. How not to sort by average rating[EB/OL]. 2009-02-06. <http://www.evanmiller.org/how-not-to-sort-by-average-rating.html>.

[8] 鲍琳,牛军钰,庄芳. 基于中位数的用户信誉度排名算法[J]. 计算机工程,2014,40(3):63-66.

[9] Robertson S E. The probability ranking principle in IR[J]. Journal of Documentation,1977,33(4):294-304.

[10] 陈伟柱,陈英,吴燕. 基于分类技术的搜索引擎排名算法 CategoryRank[J]. 计算机应用,2005,25(5):995-997.

[11] 闫友彪,陈元琰. 机器学习的主要策略综述[J]. 计算机应用研究,2004(7):4-10.

[12] 王实,高文,李锦涛. Web数据挖掘[J]. 计算机科学,2000,27(4):28-31.

[13] 李慧,李存华,王霞. 基于特征选择的网页排名算法[J]. 计算机工程,2010,36(13):37-39.

[14] 李太勇,王会军,吴江,等. 基于稀疏贝叶斯学习的个人信用评估[J]. 计算机应用,2013,33(11):3094-3096.

一种基于威尔逊区间的商品好评率排名算法

作者：[徐林龙](#)，[付剑生](#)，[蒋春恒](#)，[林文斌](#)，[XU Lin-long](#)，[FU Jian-sheng](#)，[JIANG Chun-heng](#)，[LIN Wen-bin](#)

作者单位：[徐林龙, 蒋春恒, XU Lin-long, JIANG Chun-heng \(西南交通大学 数学学院, 四川 成都 , 610000\)](#)，[付剑生, 林文斌, FU Jian-sheng, LIN Wen-bin \(西南交通大学 物理科学与技术学院, 四川 成都, 610000\)](#)

刊名：[计算机技术与发展](#)

英文刊名：[Computer Technology and Development](#)

年，卷(期)：[2015 \(5\)](#)

引用本文格式：[徐林龙](#). [付剑生](#). [蒋春恒](#). [林文斌](#). [XU Lin-long](#). [FU Jian-sheng](#). [JIANG Chun-heng](#). [LIN Wen-bin](#) 一种基于威尔逊区间的商品好评率排名算法[期刊论文]-[计算机技术与发展](#) 2015 (5)