

基于电子商务同类商品的推荐算法研究

张春生¹, 图雅¹, 翁慧², 李艳¹

(1. 内蒙古民族大学 计算机科学与技术学院, 内蒙古 通辽 028043;

2. 内蒙古民族大学 化学化工学院, 内蒙古 通辽 028043)

摘要: 个性化推荐算法是电子商务系统的研究热点。文中给出一种基于同类商品的推荐算法, 使用户在购买商品时, 快速得到性价比高的同类产品, 提高系统的服务能力。算法针对同类产品, 将供应商名称、商品价格、购买人数、收货人数、用户评论5个参数作为推荐指标, 在充分论证的基础上, 确定了各个指标的权重, 在此基础上进行了数据建模。为验证该模型的正确性, 抓取了2014年2月1日淘宝网(SAMSUNG/三星 s7898)的产品列表, 根据销售情况, 选取前67个商家的销售情况进行实验。结果表明该模型客观、有效。目前, 关于个性化的推荐算法较多, 但针对同类商品的推荐算法的研究成果相对较少, 该推荐算法的实施可减少用户查找满意商品的难度, 提高系统的服务水平。

关键词: 电子商务; 同类商品; 推荐算法; 研究

中图分类号: TP311

文献标识码: A

文章编号: 1673-629X(2016)05-0017-05

doi: 10.3969/j.issn.1673-629X.2016.05.004

Research on Similar Products Recommendation Algorithm Based on Electronic Commerce

ZHANG Chun-sheng¹, TU Ya¹, WENG Hui², LI Yan¹

(1. College of Computer Science and Technology, Inner Mongolia University for Nationalities,
Tongliao 028043, China;

2. College of Chemistry & Chemical Engineering, Inner Mongolia University for Nationalities,
Tongliao 028043, China)

Abstract: Personalized recommendation algorithm is a hot issue in the study of the electronic commerce system. A recommendation algorithm based on similar products is presented in this paper, by which users in the purchase of goods can quick get cost-effective products and improve the service ability of the system. This algorithm aims at similar products, the five arguments including supplier name, commodity prices, the number of purchase, receiving the number, user reviews are selected as recommended indexes. On the basis of sufficient demonstration, the weight of each index is determined and data model is established. To test and verify the correctness of the model, a experiment is conducted according to the sales of the first 67 of the dealer of the list in Taobao product (SAMSUNG/SAMSUNG s7898) on February 1, 2014. The results show that the model is objective and effective. At present, there are more personalized recommendation algorithms, but the research achievements of recommendation algorithm proposed for the similar goods are relatively small, the implementation of the recommendation algorithm can reduce the difficulty of users finding satisfactory goods, and improve the service level of the system.

Key words: electronic commerce; similar products; recommendation algorithm; study

0 引言

随着电子商务的蓬勃发展, 给用户的日常生活带来了巨大影响, 同时也带来了“信息超载”问题, 从而降低了信息的利用效率, 这在很大程度上影响了消费者的购物体验 and 满意度。为解决这些难题, 电子商务

企业推出了个性化的推荐系统, 它是一种以网上购物环境为前提, 为消费者推荐符合其兴趣爱好的商品为目的的系统, 使消费者无需从海量信息中寻找自己想要的信息^[1-6]。目前, 已有的几个大型电子商务企业推出的推荐系统其推荐策略有很大不同, 且对这些推

收稿日期: 2014-12-03

修回日期: 2015-04-07

网络出版时间: 2016-05-05

基金项目: 国家自然科学基金资助项目(81460656); 内蒙古自然科学基金(2012MS0913); 通辽市与内蒙古民族大学合作项目(SXZD2012021)

作者简介: 张春生(1965-), 男, 教授, 研究方向为数据库技术、数据挖掘、软件理论及应用。

网络出版地址: <http://www.cnki.net/kcms/detail/61.1450.TP.20160505.0814.014.html>

荐策略没有一个统一的分类标准。

1990 年对推荐系统的研究才当作一个相对独立的课题被学者们提出。1995 年,由美国学者展示了 WebWatchet、LIRA、Letizia 这三大系统。2000 年,我国正式开始了个性化推荐系统的研究,并逐步从理论走向实践。2009 年 7 月国内首个推荐系统科研团队—北京百分点信息科技有限公司成立。2011 年 9 月,百度世界大会 2011 上,李彦宏将推荐引擎与云计算、搜索引擎并列为未来互联网重要的战略规划以及发展方向。百度新首页将逐步实现个性化,智能地推荐出用户喜欢的网站和常用的 APP。

在文献[7-9]中,作者给出了推荐系统的形式化定义。他是将推荐给用户的对象集合用 S 表示,所有用户集合用 C 表示。将对象 S 对用户 C 的推荐度用效用函数 $u(\cdot)$ 表示,即 u 可以表示为 $C \times S \rightarrow R$ (R 表示为指定范围中的全序非负实数)。寻找到最大的推荐度 R 所对应的那些对象 S 正是推荐系统需要研究的问题。

根据国内外研究者对推荐策略的研究,学者们普遍认可的推荐策略可分为以下几类:协同过滤、基于内容的推荐、混合推荐、基于用户—产品二部图网络结构的推荐^[10-11]。

(1) 协同过滤推荐。

协同过滤推荐策略最早被提出,且在推荐系统中应用研究最为广泛。其基本思想是利用目标用户 C 的历史信息,找到与 C 相似的其他用户 C_i ,利用 C_i 对其他产品的评价来预测 C 对特定产品的喜好程度^[12-13]。

(2) 基于内容的推荐。

基于内容的推荐是根据用户选择过的对象,推荐其他具有类似属性的对象。

(3) 基于网络结构的算法。

此算法不关心用户和产品的具体内容,只是将他们视为抽象的节点,有关算法的计算信息全部隐藏在他们之间的选择关系中。

(4) 混合推荐算法。

混合推荐的目的是通过组合各种推荐方法弥补各自推荐技术的不足,根据实际的数据研究发现混合后的推荐系统具有较好的推荐效果。

然而,从目前的研究成果来看,主要是根据用户个人的历史行为或与其相似用户的行为或与其消费的商品相似的商品等作为条件,推荐感兴趣的物品,对于用户查找的明确的同类产品,如何给出可信的、具有最优性价比的产品方面的研究还很少。文中从用户购买的同类商品出发,客观地给出各种评价指标,按商品供应者的可信性、性价比等方面给出推荐序列,供用户来选择。

1 同类商品推荐算法的基本理论

对于同类产品的推荐,与以上谈到的算法不同,用户购买目标已经明确,不用考虑个体因素、用户兴趣,而重点放在如何选择信誉可靠、物美价廉的产品上面。因此,只从产品本身的特征出发,通过对产品本身特征的评价,对同类产品排行,推荐给用户选择。

在用户选择目标明确的前提下,如何客观、科学地向用户推荐产品,关键在于对商品特征的选择。商品特征选择的好坏直接影响到推荐算法的科学性、真实性、合理性。在充分分析和评估的基础上,通过分析商品的所有特征,最后从供应商、商品价格、购买人数、收货人数、用户评论 5 个方面对推荐商品建模,给出推荐指标,供用户选择。

1.1 供应商可信度评价

供应商的信誉好坏直接影响用户的购买行为。企业的规模、企业的信誉等在一定程度上反映了用户对企业的信任程度。

文中把供应商分为 3 类:知名企业、一般企业、个体经营。分别给予不同的分值,根据经验并通过合理的测试,由专家给出 3 种类型企业的权值,代表用户对企业的信任程度。

评价函数为:

$$f_s(s_i) = \begin{cases} \delta_1 & s_i = \text{知名企业} \\ \delta_2 & s_i = \text{一般企业} \\ \delta_3 & s_i = \text{个体经营} \end{cases} \quad (1)$$

1.2 商品价格评价

商品价格是用户选择商品的重要评价指标,根据用户的消费观念和消费习惯,用户往往不会选择价格过高的同种商品,而价格过低,用户又怀疑其商品的质量和来源渠道,一般用户也不选择,而用户恰恰喜欢价格适中的产品。价格适中正是用户的真正需求,而用户的这个消费习惯正符合正态分布的特点。

对价格的评价函数,文中采用标准正态分布函数:

$$f_p(p_i) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{p_i^2}{2}\right) \quad (2)$$

其中, p_i 是价格。

1.3 付款人数

付款人数表示选择某一商品的用户数量,也就是用户下单数量。这里有一个概念,下单数量不等于交易数量,因为还存在着退款行为,但从某种角度来看,它在一定程度上代表了用户对该商品的认可程度。因此,文中将付款人数作为评价商品质量的一个指标。

$$f_o(o_i) = \frac{o_i}{\max(o_i, i = 1, 2, \dots, n)} \quad (3)$$

其中, o_i 是某供应商商品的付款人数。

1.4 收货人数

收货人数是最重要的评价指标,它表示用户购买了某种商品,代表了用户对该商品的认可程度,虽然用户可能存在退货行为,但大多数用户在收到商品后,在不合适的情况下,选择换货的可能性较大,而真正退货的较少。

$$f_i(t_i) = \frac{t_i}{\max(t_i, i = 1, 2, \dots, n)} \quad (4)$$

其中, t_i 是某供应商商品的收货人数。

1.5 用户评论

用户评论也是决定商品好坏的重要指标。用户的评论包括褒贬两个方面,精确的算法应该区分这两种不同的形式,但若区分这两种形式,就需要进行文本分析,使得算法过于复杂。事实上,对一个商品评论的多少(不分褒贬)也代表了用户的认可程度,用户不购买这个商品,不关注这个商品也就不可能对该商品进行评论。为此,文中用评论总数量作为商品的评价指标。

$$f_r(r_i) = \frac{r_i}{\max(r_i, i = 1, 2, \dots, n)} \quad (5)$$

其中, r_i 是某供应商商品的评论人数。

1.6 推荐指标

商品是否受用户的欢迎主要表现为用户的交易数量,文中体现为收货人数,但从人们的消费习惯来看,有群体行为的现象,也就是看到大家都买那个商品自己也买那个商品,但这个商品不一定最优。从电子商务角度,不应该完全按这个思路出发,而应该从供应商的信誉和商品价格等方面出发。为此文中给出两种推荐指标。

商品的总体推荐指标将融合供应商、商品价格、付款人数、收货人数、用户评论 5 个指标,通过加权组合在一起。首先将每个商品离散为一个向量:

$$x_i = (s_i, p_i, o_i, t_i, r_i)$$

分别代表:商品=(供应商类型,商品价格,付款人数,收货人数,用户评论数)。构建 5 个加权系数,分别代表 5 个评价指标的权重: $\omega_s, \omega_p, \omega_o, \omega_t, \omega_r$ 。

(1) 商品热销前推荐指标。

商品热销前,商品的付款人数和收货人数没有或很少,不能作为推荐指标;所以,文中只选择供应商类型、商品价格、用户评论数作为评价指标,此时的推荐指标为:

$$f(x_i) = w_s f(s_i) + w_p f(p_i) + w_r f(r_i) \quad (6)$$

(2) 商品热销后推荐指标。

商品热销后,商品的付款人数和收货人数代表了用户对商品的认可程度,同时也要考虑用户的群体心理;所以,此时的推荐指标为:

$$f(x_i) = w_s f(s_i) + w_p f(p_i) + w_o f(o_i) + w_t f(t_i) +$$

$$w_r f(r_i) \quad (7)$$

2 同类商品的推荐算法

如图 1 所示,同类商品的推荐算法描述如下:

(1) 开始,针对某一种商品的销售,通过网站提供的 API 或其他手段,从电子商务网站抓取用户消费数据,其中包括上述 5 个指标,形成初始数据集 D_s 。

(2) 梳理 D_s , 按 $x_i = (s_i, p_i, o_i, t_i, r_i)$ 的格式整理数据,形成目标数据集 D_d 。

(3) 对所有供应商进行遍历,按每个供应商进行计算。

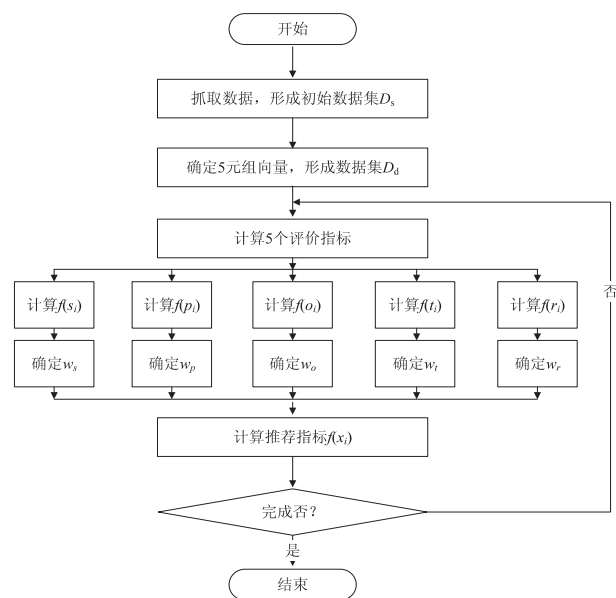


图 1 算法流程

(4) 分别计算 5 个评价指标,并确定 5 个指标的权重 $\omega_s, \omega_p, \omega_o, \omega_t, \omega_r$ 。

(5) 计算商品热销前推荐指标: $f(x_i) = w_s f(s_i) + w_p f(p_i) + w_r f(r_i)$ 。

(6) 计算商品热销后推荐指标: $f(x_i) = w_s f(s_i) + w_p f(p_i) + w_o f(o_i) + w_t f(t_i) + w_r f(r_i)$ 。

(7) 若遍历完成,转到(8),否则转到(4)。

(8) 结束。

3 实例

文中抓取 2014 年 2 月 1 日淘宝网(SAMSUNG/三星 s7898)的产品列表(见图 2),根据销售情况,选取前 67 个商家的销售情况进行实验。

(1) 数据预处理。

对供应商信息进行数值化,(知名企业、一般企业、个体经营)分别给予权值(1.0,0.6,0.2)。

对商品价格首先进行标准化处理,使得价格区间标准化为 $[-1, 1]$,这样可保证价格的正态分布效果。

对付款人数、收货人数、用户评论进行归一化

处理。

	A	B	C	D	E	F
1	1	三际数码官方旗舰店	¥ 628.00	1484人付款	1259人收货	3283条评论
2	2	天蛙数码专营店	¥ 608.00	410人付款	357人收货	1062条评论
3	3	赫之光数码专营店	¥ 595.00	182人付款	117人收货	102条评论
4	4	梦途释08公益宝贝	¥ 450.00	231人付款	166人收货	249条评论
5	5	百变手机117061	¥ 470.00	502人付款	420人收货	753条评论
6	6	lemoniusz	¥ 500.00	149人付款	118人收货	196条评论
7	7	绿森数码官方旗舰店	¥ 619.00	344人付款	307人收货	1129条评论
8	8	周成通讯公益宝贝	¥ 598.00	128人付款	108人收货	179条评论
9	9	轩达速通讯	¥ 618.00	388人付款	348人收货	310条评论
10	10	杭州晶杰数码专营店	¥ 628.00	88人付款	61人收货	38条评论
11	11	耘园科技8859	¥ 620.00	66人付款	60人收货	154条评论
12	12	环球机库官方旗舰店	¥ 598.00	36人付款	18人收货	15条评论
13	13	新灿科电子	¥ 568.00	41人付款	28人收货	30条评论
14	14	威廉购物网	¥ 480.00	146人付款	130人收货	232条评论
15	15	姚hxf	¥ 1,767.00	21人付款	16人收货	18条评论
16	16	佳沪数码专营店公益宝贝	¥ 589.00	11人付款	7人收货	201条评论
17	17	启能数码专营店	¥ 629.00	48人付款	42人收货	22条评论
18	18	天配数码专营店	¥ 605.00	54人付款	48人收货	229条评论
19	19	能良数码官方旗舰店	¥ 598.00	33人付款	30人收货	80条评论
20	20	陈媚虹2304	¥ 1,180.00	10人付款	8人收货	4条评论
21	21	万象科技通讯	¥ 700.00	31人付款	27人收货	1275条评论
22	22	三星天亚专卖店	¥ 599.00	8人付款	5人收货	8条评论
23	23	南昌健柏数码专营店	¥ 610.00	20人付款	14人收货	32条评论
24	24	lingainal9	¥ 1,528.00	3人付款	1人收货	1条评论
25	25	al5013517272	¥ 569.00	4人付款	2人收货	2条评论
26	26	金色_河_流	¥ 618.00	2人付款	2人收货	5条评论
27	27	海梓通讯	¥ 490.00	11人付款	10人收货	12条评论
28	28	谷米数码专营店	¥ 688.00	2人付款	0人收货	0条评论
29	29	百分信80	¥ 560.00	3人付款	2人收货	2条评论
30	30	新宏鑫商行	¥ 540.00	9人付款	9人收货	9条评论
31	31	迪信通河南数码专营	¥ 799.00	8人付款	6人收货	10条评论

图 2 原始数据

为了便于比较和观察变化趋势,保证变化曲线的总体走向,打破 67 个企业的自然顺序,文中对实验数据按收货人数降序排序。

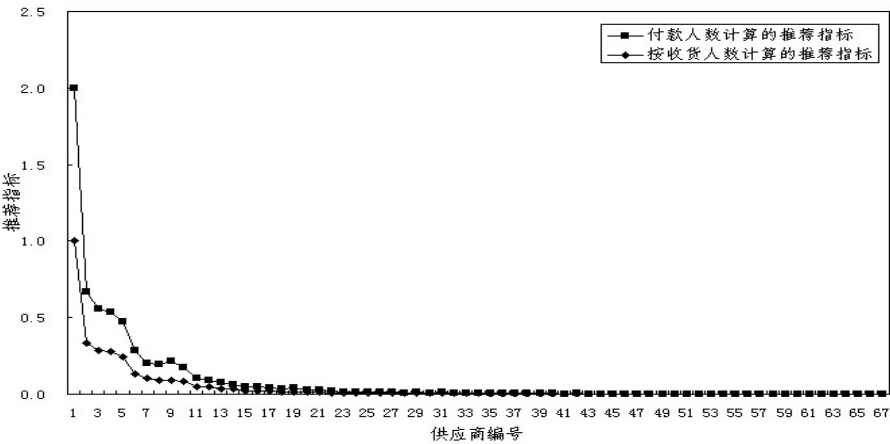


图 3 付款人数与收货人数推荐指标比较

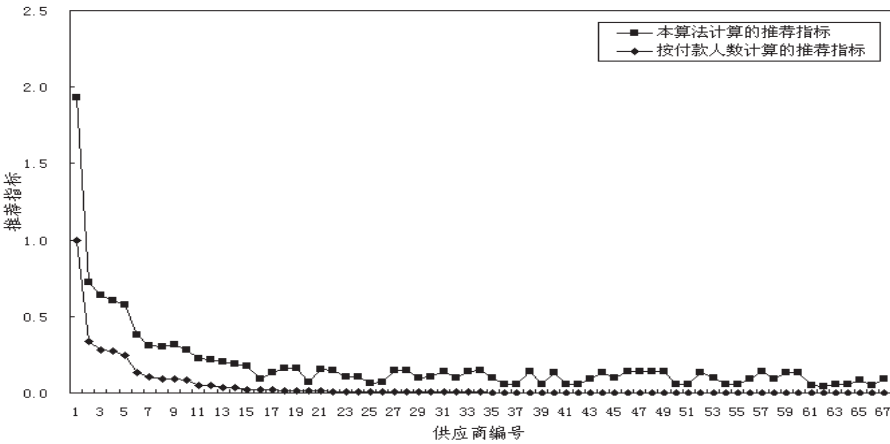


图 4 热销后推荐指标与付款人数计算的推荐指标比较

(2)按付款人数与收货人数进行比较。

图 3 是分别按付款人数和收货人数计算的推荐指标的比较结果。

从图中可以看出,付款人数与收货人数的趋势走向一致,计算推荐指标时,可选择其中之一。

(3)商品热销后推荐指标。

付款人数与收货人数的趋势走向一致,故计算推荐指标时,只采用付款人数。

($\omega_s, \omega_p, \omega_o, \omega_r$) 的加权值选择为 (0.1, 0.1, 0.7, 0.1)。

图 4 是对文中算法计算的推荐指标和按付款人数计算的推荐指标的比较结果。

从图中可以看出,文中算法计算的推荐指标与按付款人数计算的推荐指标走向一致。

(4)商品热销前推荐指标。

商品热销前推荐指标不考虑付款人数与收货人数 2 个因素,故文中算法计算推荐指标时采用供应商类型、商品价格、用户评论数三个因素。

($\omega_s, \omega_p, \omega_r$) 的加权值选择为 (0.2, 0.4, 0.4)。

图 5 是文中算法计算的推荐指标和按付款人数计算的推荐指标的比较结果。

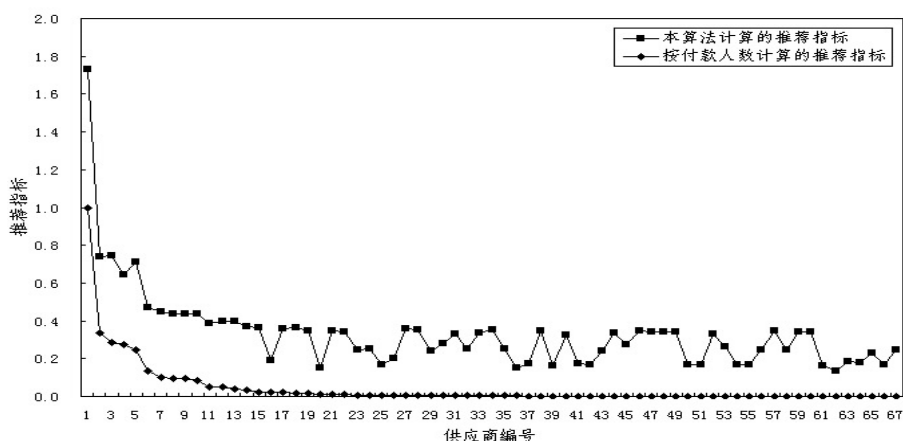


图5 热销前推荐指标与付款人数计算的推荐指标比较

从图中可以看出,商品热销前推荐指标的走向与付款人数的走向总体趋势一致,推荐指标比较合理。

从以上实验结果可以看出,在认可“付款人数计算的推荐指标”是实际商品销售指标的前提下,文中算法计算的推荐指标无论是商品热销前还是商品热销后都与实际商品销售指标走向一致。因此,文中算法无论是在商品热销前还是商品热销后都可作为简单有效的推荐算法对同类商品进行推荐。

商品热销前推荐指标的准确性是文中算法的最大特色,它可在商品热销前正确指导用户消费。

4 结束语

文中针对用户明确选择的某类产品的推荐研究成果较少的事实,选择了供应商、商品价格、付款人数、交易人数、用户评论5个评价指标,在充分分析的基础上,对推荐指标进行了建模,并给出了商品热销前和热销后两种评价指标。实验结果表明,购买人数与收货人数的趋势走向一致,计算推荐指标时,可选择其中之一。同时付款人数的趋势走向可作为真正的销售走向。商品热销后推荐指标的走向与付款人数的走向一致。商品热销前推荐指标的走向与付款人数的走向总体趋势一致,尤其是商品热销前推荐指标的准确性是文中算法的最大特色,它可在商品热销前正确指导用户消费。

文中建立的推荐指标可用于电子商务网站的同种商品推荐上,具有一定的应用价值。特别是针对于文中支持项目“蒙医方剂数据挖掘关键技术研究”,拟建立一个专门进行蒙药产品销售的系统,同时为了使广大用户认识和认可蒙药的独特疗效,可在取得其他厂商同意的基础上,根据功能与主治信息,寻找同类其他药物与蒙药进行同时销售。通过用户的购买行为、评论行为等进一步验证文中算法的合理性,同时也通过

用户对蒙药的认识和使用,发挥其独特的作用,为用户减少疾病的痛苦。

参考文献:

- [1] 许海玲,吴 潇,李晓东,等. 互联网推荐系统比较研究[J]. 软件学报,2009,20(2):350-362.
- [2] 刘凤霞,孙家蓉. 基于商品分类的电子商务推荐系统设计[J]. 计算机应用与软件,2014,31(5):37-41.
- [3] 刘建国,周 涛,汪秉宏. 个性化推荐系统的研究进展[J]. 自然科学进展,2009,19(1):1-15.
- [4] Liu R R, Jia C X, Zhou T, et al. Personal recommendation via modified collaborative filtering[J]. Physica A, 2009, 388(4): 462-468.
- [5] 王国霞,刘贺平. 个性化推荐系统综述[J]. 计算机工程与应用,2012,48(7):66-76.
- [6] 王巧荣,赵海燕,曹 健. 个性化服务中的用户建模技术[J]. 小型微型计算机系统,2011,32(1):39-46.
- [7] Weng S S, Lin B S, Chen W J. Using contextual information and multidimensional approach for recommendation[J]. Expert System with Applications, 2009, 36(2): 1268-1279.
- [8] Yan Duanwu. Research on knowledge service oriented intelligent recommendation system[D]. Nanjing: Nanjing University of Science and Technology, 2007.
- [9] Zhang Z, Lu L, Liu J G, et al. Empirical analysis on a keyword based semantic system[J]. The European Physical Journal B, 2008, 66(4): 557-561.
- [10] 郑 鑫,张初志. 一种基于模糊C均值聚类的协同过滤推荐算法[J]. 济南大学学报:自然科学版,2016,30(1):55-59.
- [11] 谭文安,沈腾腾,孙 勇. 基于偏好相似度的混合信任推荐模型[J]. 太原理工大学学报,2016,47(1):62-67.
- [12] 王全民,王 莉,曹建奇. 基于评论挖掘的改进的协同过滤推荐算法[J]. 计算机技术与发展,2015,25(10):24-28.
- [13] 徐新瑞,孟彩霞,周 雯,等. 一种基于Spark时效化协同过滤推荐算法[J]. 计算机技术与发展,2015,25(6):48-55.