协同过滤推荐中基于用户分类的邻居选择方法*

张 尧, 冯玉强
(哈尔滨工业大学 管理学院, 哈尔滨 150001)

摘 要: 为了提高推荐系统的推荐结果质量,找到目标用户恰当的邻居是协同过滤算法中非常关键的一个环节。网络中的用户可以分为专家型用户、可信用户与兴趣相似用户三个维度,由于不同类型的邻居对用户的影响及用户对不同邻居的依赖倾向的不同,因此利用岭回归分析估计用户对于这三类用户的主观倾向,即邻居选择权重,由此获得目标用户邻居集合,进而产生推荐,通过利用标准 *F*1 方法与传统推荐方法对比实验分析表明,推荐结果的质量显著提高;同时利用 K-means 方法对用户作聚类分析及类别之间的方差齐性分析,并与行为研究结果相对比,验证了推荐结果的可信性。

关键词: 协同过滤; 邻居选择; 邻居权重; 用户分类; 岭回归; K-means 聚类
中图分类号: TP311 **文献标志码**: A **文章编号**: 1001-3695(2012)11-4216-04
doi:10.3969/j.issn.1001-3695.2012.11.054

Approach of neighbor selection based on user classification in collaborative filtering recommendation

ZHANG Yao, FENG Yu-qiang
(*School of Management*, *Harbin Institute of Technology*, *Harbin* 150001, *China*)

Abstract: In order to improve the quality of recommendation results, selecting proper neighbors would be the important link in collaborative filtering. The user might be divided into three types, which was expertise, trustworthy and similarity, since there were differences between neighbors, the important of them could be differentiated from target users. As the target user, the importance and weight of these three types of neighbors would be analyzed by the method of ridge regression. As a result, the proper neighbors might be found for the target user. According to comparative experiments based on *F*1 method, it shows that the accuracy has been improved significantly. Meanwhile, through the K-means cluster analysis and LSD(least-significant difference), the result coincides with behavior research's.

Key words: collaborative filtering; neighbor selection; neighbor weights; user classification; ridge regression; K-means cluster

0 引言

在互联网与信息技术高速发展的时代,随之而来的就是大量的信息,因此,就需要一种可以对信息进行过滤,留下有用信息的信息技术。其中,推荐系统就是一种帮助找到需要信息的信息技术。推荐系统在电子商务网站中扮演虚拟店员(virtual salespeople)的角色,它向客户提供商品信息和建议[1]。

协同推荐是迄今为止最成功的个性化推荐技术,被应用到很多领域中,协同过滤推荐能针对任何形态的内容进行过滤,其产生推荐的基础是用户间共同评价项目集,利用该项目集找到目标用户的兴趣相似的邻居,并通过相似邻居的购买历史产生推荐。

这种基于用户购买历史相似的算法,由于不可信用户的存在,所以无论这种不可信的邻居如何与目标用户相似,他的推荐都是不可信的。因此,很多学者提出基于可信用户的推荐系统[2~5],但对于可信用户的寻找或定义方面学者们有着不同侧重点。一种为显性的定义可信性,如基于信任网的推荐方法[6]。在该方法中,由用户之间接触关系构建信任网,每一个用户的信任值来源于与其接触过的其他用户的评分,通过该网络寻找目标用户的可信用户,进而产生可信推荐。而另一种是通过用户的某些特征分析其可信性,如将可信用户定义为有能力进行推荐的用户[7]。文献[8]按照用户在某一领域的购买历史,将用户分为有推荐能力和无推荐能力用户,一方面可以降低算法的时间复杂度,另一方面可以提高算法的推荐准确性。另外还有一些研究从用户的特征和兴趣角度对系统过滤算法进行了研究[9],取得了较好的推荐效果。

但对于目标用户来说,不同用户对于可信用户或有推荐能力的用户的依赖程度是不同的。例如,在选择电影时,一些用户愿意听取朋友或兴趣相似的邻居的建议,而一些更愿意听从专家的建议。甚至对于同一目标用户来说,在不同的项目推荐环境下其依赖倾向也是不同的。在消费者心理学和营销学领域认为,如果用户已经拥有相关产品的知识,那么他可能很少

收稿日期: 2012-04-17; **修回日期**: 2012-05-22 **基金项目**: 国家自然科学基金资助项目(71172157);国家自然科学基金海外合作基金资助项目(71028003)

作者简介: 张尧(1981-), 男,吉林磐石人,讲师,博士研究生,主要研究方向为电子商务、电子推荐、网络信任(zhangyaocc@yahoo.com.cn);冯玉强(1961-),女,教授,博导,主要研究方向为商务谈判、数据挖掘、企业信息化、客户满意度.

依赖专家的建议,而对于没有相关知识的用户来说,其更愿意听取专家的建议[10]。

综上,由于用户之间的异质性(如兴趣相似用户、可信用户、专家用户)和用户自身在不同项目推荐环境的依赖倾向的异质性,所以在选择目标用户的邻居时应该因人而异,以符合目标用户的依赖特征。因此本文在 B2C 电子商务中首先对用户的相似性、专家型、可信性、三个维度进行定量描述,然后利用回归分析得到其对经验用户、可信用户、相似用户的主观态度,即权重,进而得到 N 个最有价值的用户,最后产生个性化推荐,并通过对比实验分析算法的准确性,即有效性。通过对实验结果的定性分析,并与行为研究相对比以验证算法的可信性。

研究框架如图 1 所示。

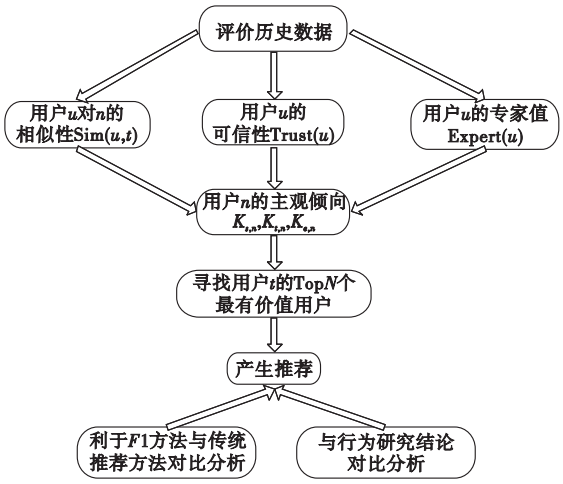

图1　研究框架

1　算法设计

1.1　用户相似度计算

相似性是指用户之间对于共同项目的评分相似,主要反映了用户之间的偏好相似,计算方法通常使用皮尔逊相关系数(pearson correlation),具体为

$$\text{userSim}(u,n)=\frac{a'\sum_{i\subset CR_{u,n}}(r_{u,i}-\bar{r}_u)(r_{n,i}-\bar{r}_n)}{\sqrt{\sum_{i\subset CR_{u,n}}(r_{u,i}-\bar{r}_u)^2}\sqrt{\sum_{i\subset CR_{u,n}}(r_{n,i}-\bar{r}_n)^2}}\tag{1}$$

其中:$CR_{u,n}$表示用户 u 与邻居 n 所共同评价过的项目的集合,$r_{u,i}$为用户 u 对项目 i 的评分,$r_{n,i}$为用户 n 对项目 i 的评价。

一般情况下,如果用户之间共同评价的数量越多、同时用户之间越倾向于相似,如表 1 所示。

表 1　用户评价项目集举例

用户	a	b	c	d	e
用户 A	3	3	3	3	3
用户 B	3	3	3	3	—
用户 C	3	3	3	—	—

在传统算法基础上引入系数 α,但在实际电子商务系统中用户间的共同评价项目并不多,所以在此基础上提出改进措施,用共同评价的数量占各自全部评价数量的比重代替共同评价数量,共同评价项目占各自评价历史比重越高,则用户之间的兴趣度越相似,因此得到 α',计算方法为

$$\alpha'=\frac{\sum_{i\in CR_{u,n}}1}{\sqrt{R_u R_n}}\tag{2}$$

其中:R_u 为用户 u 评价的项目总数,R_n 为用户 n 评价的项目总数。

1.2　用户可信度计算

在这里可信用户指如果评价者能够真实反映个人评价意愿,从对商品的感知给出的真实评分,则认为此用户是可信的。在这里根据文献[11],如果用户的评分与商品的平均评分相似,那么就可以认为评价者真实地反映了个人的意愿,而且无论是正相关还是负相关。在这里本文认为,只要他们之间的相关关系越强,则用户给出真实评价的可能性越大,所以本文将相关系数取绝对值,使值的范围在[0,1]上。另外,在两个用户相似度相同的情况下,如果用户评价项目数越多,用户越倾向于给出真实的评价。因此在皮尔逊相关系数基础上引入系数 β 为

$$\text{userTrust}(u)=\beta\left|\frac{\sum_{i\subset m}(r_{u,i}-\bar{r}_u)(r_{\text{avg},i}-\bar{r}_{\text{avg}})}{\sqrt{\sum_{i\subset m}(r_{u,i}-\bar{r}_u)^2}\sqrt{\sum_{i\subset m}(r_{\text{avg},i}-\bar{r}_{\text{avg}})^2}}\right|\tag{3}$$

其中:m 为用户 u 所有评价的项目集合,$r_{\text{avg},i}$为项目 i 的平均评分,$\bar{r}_{\text{avg}}$为 m 个项目平均评分的平均分。如果 $m>50$,$\beta=1$,否则 $\beta=m/50$。这里的阈值 50 是通过对数据的预分析得到的一个估计值,通过引入 β,可以提高推荐算法查准率,但对于稀疏程度的不同数据集,应该采用不同的阈值。

1.3　用户专家值计算

根据用户对于网络购物的知识可以将用户分为 familiarity(熟练)和 expertise(专家)两种用户[12]。熟练用户是拥有大量与产品有关的经验的用户;专家用户是具有可以成功完成产品相关工作的能力的用户。专家用户需要在不断实践中积累经验,形成有关网络购物的知识结构,反过来再应用到实践中。所以,在这里本文将专家用户定义为在某一领域有较高度的活跃度且有准确预测评分能力的用户。在本文所使用的数据集中,数据具有如下分类导航特征,本文使用 brandclass3 作为领域划分依据,如图 2 所示,共划分为 11 个领域。

goodsID	name	brand	brandclass1	brandclas	brandclass3	branc
140308	银狼SW338	银狼	安全/应急	汽车用品	家用电器、	汽首页
151837	格来德WKF-	格来德	电水壶/热水	厨房电器	家用电器、	汽首页
239413	硕美科SM-(	硕美科	麦克风	时尚影音	手机数码	首页
1000343	as安都柔软	安都	T恤	女装	服饰鞋帽	首页
266038	诺基亚C6-(	诺基亚	手机	手机通讯	手机数码	首页
313994	高赛尔万足	高赛尔	金银藏品	礼品	礼品箱包、	钟首页
357272	山多力OE-5	山多力	改装/配件	汽车用品	家用电器、	汽首页
171330	卡西欧esf2	卡西欧	电子词典	时尚影音	手机数码	首页
202696	伊丽莎白·Elizabe		女士香水	香水SPA	个护化妆	首页
255617	娇韵诗身体	娇韵诗	纤体瘦身	香水SPA	个护化妆	首页
245057	金奇仕鳕鱼	金奇仕	营养品	营养辅食	母婴、玩具	首页
366756	康佳LC32HS	康佳	平板电视	大 家 电	家用电器、	汽首页
187678	欧莱雅醒肤	欧莱雅	乳液面霜	面部护理	个护化妆	首页
164974	奔腾PFF40E	奔腾	电饭煲	厨房电器	家用电器、	汽首页

图2　商品信息表截图

用户 u 的专家值为

$$\text{Expertise}(u)=\frac{1}{t}\times\sum_{i=1}^{n}m_i\times\text{userExpertise}(u,i)\tag{4}$$

用户 u 共评价了 t 个项目,共分为 n 个领域,每个领域评价了 m_i 个项目,$\text{userExpertise}(u,i)$为用户 u 在领域 i 上的专家值。

用户 u 在领域 i 上的专家值为

$$\text{userExpertise}(u,i)=\varphi(u,i)\times\frac{\sum_{j\subset i}\text{userCorrect}(u,j)}{\sum_{j\subset i}1}\tag{5}$$

其中:j 为用户 u 在领域 i 上的评价项目集合;$\varphi(u,i)$为用户 u 在领域 i 上的活跃度,如果评价次数 $\text{count}(u,i)$大于所有其他

用户的平均值 avg(i),其值为1,否则为 count(u,i)/avg(i);userCorrect(u,j)为用户 u 在对项目 j 的预测准确性测量。

用户 u 预测的准确度体现为该用户的评价与预测值之间的相似程度。根据文献[4]有

$$\text{userCorrect}(u,j)=\begin{cases}0 & \text{if } |r_{u,j}-\text{pred}(u,j)|>\varepsilon\\ \lambda & \text{if } |r_{u,j}-\text{pred}(u,j)|\leqslant\varepsilon\end{cases} \tag{6}$$

其中:$r_{u,j}$为用户 u 对项目 j 的实际评分;pred(u,j)为用户 u 对 j 的预测评分,如式(7)所示;ε 为1.8[4];当项目 j 除 u 外所有评价数量 n 小于10时,λ 为 n/10,否则为1。由于在用户购买历史数据集中,用户的平均购买的商品类型数量近似为10,所以这里阈值取值为10。

$$\text{pred}(u,j)=\bar{r}_u+\frac{\sum_{n\subset\text{neighbors}(u)}\text{userSim}(u,n)\cdot(r_{n,j}-\bar{r}_n)}{\sum_{n\subset\text{neighbors}(u)}\text{userSim}(u,n)} \tag{7}$$

1.4 用户主观意愿估计

构建回归方程

$$V_{n,u}=K_{s,n}\cdot\text{userSim}(u,n)+K_{t,n}\text{userTrust}(u)+K_{e,n}\text{userExpertise}(u)+\varepsilon_n \tag{8}$$

式(8)中:$V_{n,u}$为用户 u 对于用户 n 的效用;$K_{s,n}$、$K_{t,n}$、$K_{e,n}$分别为用户 n 对于相似性、可信性、专家性的主观意愿;ε_n 为调节因子。

如果用户 u 与 n 的共同评价评分越接近,则用户 u 对于 n 的效用越高。同时,根据行为研究[13],无论哪种类型的商品评价的深度(在文章中用评论的字数表示,字数越多评价深度越深,反之越浅)。$V_{n,u}$为

$$V_{n,u}=\theta_u\times(C-\text{MAE}_{n,u})=\theta_u\times\left(C-\frac{\sum_{i=1}^{m}|R_{n,i}-R_{u,i}|}{m}\right) \tag{9}$$

其中:$r_{u,i}$为用户 u 对 i 的评分;θ_u 为用户 u 的评论贡献度,其体现为,如果其评论平均字数 word(u)大于其他用户的平均值 avgword,则系数为1,否则为$\frac{\text{word}(u)}{\text{avgword}}$;$C$ 为常数,由于 $\text{MAE}_{n,u}$的范围在(0,4)之间,所以 C 的值为5。

在回归计算过程中,由于相似度、可信度、专业度可能存在多重共线性,因此利岭回归方法进行分析。岭回归分析是一种专门用于共线性数据分析的有偏估计方法,它对最小二乘法进行了改进,通过忽略最小二乘法的无偏性及部分精确度,以寻求更符合实际的回归过程。岭回归分析过程采用统计分析软件 SPSS 11。

1.5 产生推荐

在对用户主观意愿估计后,利用式(8)进行效用计算,得到对用户 a 效用最高的前 w 个邻居的购买集合,计算每种商品出现的次数,并对商品集合按照出现次数降序排列,从集合中取前 n 个商品作为推荐项目。

2 实验分析

2.1 数据准备

本文以京东网上商城为研究背景,利用 MetaStudio、DataScraper 开源工具对相关网页进行抓取,抓取的数据项包括用户(用户名、评价评分、评论内容)、商品(商品编号、商品名称、商品价格、五种商品类别)两大类。由于抓到的数据是以 XML 文件形式存储,所以本文利用 Visual Basic 与 SQL Server 2000 数据库相结合对 XML 文档进行数据预处理。经过整理,实际抓取用户数29 519名,对195 156件商品进行了评价。抽取其中80%作为训练数据集,20%作为测试数据集,测试数据集为用户购买历史中的一部分商品。

2.2 评价方法

为了对提出的算法进行验证,本文将从有效性(准确性测量)及可信性(与行为研究结果对比)两个方面对算法进行验证。

1)有效性验证

这里利用标准 $F1$ 方法对推荐结果进行测量[14],这种方法是对查全率与查准率两种测量方法的综合,如式(10)

$$F1=\frac{2\times\text{Recall}\times\text{Precision}}{(\text{Recall}+\text{Precision})} \tag{10}$$

查全率

$$\text{Recall}=\frac{|\text{test}\cap\text{top-}N|}{|\text{test}|}$$

查准率

$$\text{Precision}=\frac{|\text{test}\cap\text{top-}N|}{|\text{top-}N|}$$

其中:test 为测试数据集,top-N 为推荐结果集,$|\text{test}\cap\text{top-}N|$表示系统给出的正确的推荐结果数量。

由于邻居数量对推荐准确性有较大的影响,因此首先对比分析了不同的邻居数量对于传统 User-based CF 算法与本文 MU-based CF(multiple-user based collaborated filtering)算法推荐效果的影响。

如图3所示,在对用户综合评价的基础上所引入的推荐用户显著提高了推荐的质量。但是对构成 $F1$ 值的查准率与查全率的影响程度是不同的,通过对由回归分析得到的 K_s、K_t、K_e 的描述性统计分析可以看出(表2),在所有数据中,用户对于 K_e 的态度是最强烈的,表明用户对专家型用户依赖程度较强,用户在产品的选择上是谨慎的,这同时也符合了数据集的特征——与体验型商品相比,搜索型产品占大多数。

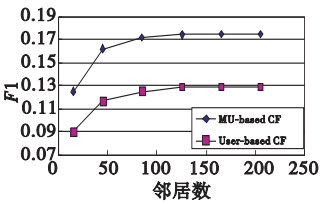

图3 邻居数量对$F1$值的影响

表2 用户的 K_s、K_t、K_e 值描述性统计

类别	N	极小值	极大值	均值	标准差
K_e	29 519	.03	6.81	1.015 0	.507 41
K_t	29 519	-2.07	4.27	.289 30	.379 54
K_s	29 519	.05	7.05	.757 10	.437 62
有效的 N	29 519				

因此,通过对比本文提出的推荐方法在查准率与查全率的变化趋势上可以看出(图4),由于目标用户对专家用户的态度较强,使得算法得到的最相似邻居集合中,邻居专家值权重较大,所以查准率随着邻居数量的增加很快就达到了极值。但另一方面,邻居相似值权重相对降低,所以查全率随着邻居数量

的增加达到极值的速度较慢。这样的结果是与数据集中用户的态度特征相吻合的。

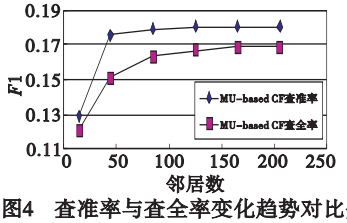

图4　查准率与查全率变化趋势对比分析

2)可信性验证

根据 Duhan 等学者的研究,用户自身所拥有关于产品的信息量越多,其对于专家用户的依赖越弱[15]。这里的专家是指其专业程度及对商品的熟悉程度均较高。而本文提取的专家值同样也是以在某一具体领域内用户的活跃度和预测准确性为基础,从大量数据中抽取的,所以可以与行为研究中的专家定义近似相同。

利用 K-means 方法对所有用户在 K_s、K_t、K_e 上的属性值进行聚类分析,经过反复测试,聚类数量 4 效果较好,并且类与类之间利用方差齐性 LSD 方法两两对比,在 0.01 上显著,聚类结果如表 3 所示。

表 3　用户态度 K-means 聚类

类别	cluster1	cluster 2	cluster 3	cluster 4	*F*	sig.
标准化 K_e	6.741 6	-.415 7	.571 9	1.351 6	873.629	.000
标准化 K_t	6.598 2	-.207 3	.305 7	-.618 3	105.341	.000
标准化 K_s	.365 7	-.132 6	.097 1	.752 4	457.582	.000
用户数 *N*	56	14 183	11 308	372		

用来聚类的数据是已经标准化的数据,通过 *F* 值可以看出,K_e 值对聚类的影响较大,因此对于每一类用户按照 K_e 值降序排列为 cluster1→cluster4→cluster3→cluster2。

然后再对每一类按照用户的专家值进行描述性统计分析,如表 4 所示。

表 4　专家值在不同类别中的描述性统计

类别	*N*	均值	标准误差	极小值	极大值
cluseter1	56	.063 5	.002 77	.02	.31
cluseter2	14 183	.397 1	.008 79	.04	.93
cluseter3	11 308	.250 3	.008 54	.04	.90
cluseter4	372	.178 4	.004 63	.02	.93

将类别再按照专家值均值升序排列:cluster1→cluster4→cluster3→cluster2,其结果恰好与按用户对专家的态度降序排列的结果相吻合。因此算法的可信性得以验证,如果用户本身是专家则对专家的依赖很少。

3　结束语

由于电子商务系统中用户的异质性,所以对用户的推荐也应该因人而异。因此,本文通过形式化描述相似用户、可信用户与专家用户,并利用岭回归分析目标用户对于这三类用户的主观态度,以选择对目标用户最有价值的邻居集合,从而产生推荐。

通过对实际的数据进行分析可看出,本文提出的方法显著提高了推荐的质量。同时,根据本文所使用的数据集的特点,用户对于专家型用户的依赖程度较高,因此目标用户的最有价值邻居集合中,专家型用户所占的比重较大,所以算法的查准率会快速提高,而查全率由于相似性用户占有邻居比重降低,因此提高得比较慢。

在不同的推荐环境下,用户对推荐项目的准确性、新颖性、可信性依赖程度是不一样的,对于因采纳推荐而付出成本较低的情况来说,新颖性可能是用户比较看重的,如电影、音乐的推荐;而对于可能付出较多成本的项目或者项目本身是体验类型的情况来说,推荐的准确性、可信性应该是用户希望得到的,如消费品推荐,尤其是价格较高的消费品。

本文的研究数据大多集中在 B2C 电子商务环境下的电子产品类,所以算法的通用性、算法中的相关参数设置及在不同的数据环境下的推荐效果还需要进一步通过数据进行验证。

参考文献:

[1] RESNICK P, VARIAN H R. Recommender systems[J]. Communications of the ACM, 1997,40(3): 56-58.

[2] 蔡浩,贾宇波,黄成伟. 结合用户信任模型的协同过滤推荐方法研究[J]. 计算机工程与应用,2010,46(35):148-151.

[3] MASSA P, AVESANI P. Trust-aware recommender systems[C]// Proc of ACM Conference on Recommender Systems. New York: ACM Press,2007:17-24.

[4] O'DONOVAN J, SMYTH B. Trust in recommender systems[C]// Proc of the 10th International Conference on Intelligent User Interfaces. New York:ACM Press,2005:167-174.

[5] 张富国,徐升华. 基于信任的电子商务推荐多样性研究[J]. 情报学报,2010,29(2):350-355.

[6] MASSA P, AVESANI P. Trust-aware collaborative filtering for recommender systems[C]//Proc of OTM Confederated International Conferences, CoopIS, DOA and ODBASE. Berlin: Springer,2004:492-508.

[7] O'DONOVAN J, SMYTH B. Is trust robust? An analysis of trust-based recommendation[C]//Proc of the 11th International Conference on Intelligent User Interfaces. New York: ACM Press, 2006: 101-108.

[8] 李聪,梁昌勇,马丽. 基于领域最近邻的协同过滤推荐算法[J]. 计算机研究与发展,2008,45(9):1532-1538.

[9] 严冬梅,鲁城华. 基于用户兴趣度和特征的优化协同过滤推荐[J]. 计算机应用研究, 2012,29(2):497-500.

[10] ERIC J, JONHNSON J, EDWARD R. Product familiarity and learning new information [J]. Journal of Consumer Research,1984,11(1): 542-550.

[11] CHO J, KWON K, PARK Y. Implicit user credibility extraction for reputation rating mechanism in B2C e-commerce[J]. International Journal of Intelligent Information and Database Systems,2007, 1(3):247-263.

[12] JACOBY J, TROUTMAN T. Experience and expertise in complex decision making[J]. Advances in Consumer Research, 1986, 13(1):469-472.

[13] SUSAN M, MUDAMBI D S. What makes a helpful online review? A study of customer reviews on amazon. com[J]. MIS Quarterly, 2010,34(1):185-200.

[14] YANG Yi-ming, LIU Xin. A Re-examination of text categorization methods[C]//Proc of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. New York:ACM Press,1999: 42-49.

[15] DUHAN D F, JOHNSON S D, WILCOX J, *et al*. Influences on consumer use of word-of-mouth recommendation sources[J]. Journal of Academy of Marketing Science, 1997,25(24):283-295.