文章编号:1001 - 9081(2006)04 - 0966 - 03

基于商品属性隐性评分的协同过滤算法研究

陈冬林,聂规划
(武汉理工大学 经济学院,湖北 武汉 430070)
(chendl@ mail. whut. edu. cn)

摘 要:在分析目前电子商务推荐系统及算法存在问题的基础上,提出了一种准确的、实时的、基于 Web 日志的 Internet 电子商务推荐算法。基于客户浏览行为,设计了 CGA(Customer-Good-Attribute)模型,综合考虑客户浏览路径和时间、商品属性及其在网页中的分布等因素,研究了客户对商品属性的隐性评分函数,给出了算例说明,讨论了基于商品属性的协作过滤算法。该算法已成功应用于电子商务智能模拟系统中。

关键词:商品属性;隐性评分;协同过滤;电子商务
中图分类号: TP391 **文献标识码**:A

Research on collaborative filtering algorithm based on item's attribute implicit rating

CHEN Dong-lin, NIE Gui-hua
(*School of Economics*, *Wuhan University of Technology*, *Wuhan Hubei* 430070, *China*)

Abstract: After analyzing the previous work of recommendation system and algorithm, a recommendation algorithm based on Web daily record was presented which was accurate and efficient. Based on the behavior of customers, a CGA(Customer-Good-Attribute) model was designed which considered synthetically such factors as times, trails, commodities attributes and those distribution in Web pages. The function to the customers' implicit rating of the attributes of commodities was discussed, and a case was given. Then a filter algorithm was put forward based on attributes of goods. An experience result indicated that the algorithm is effective.

Key words: item's attribute; implicit rating; collaborative filtering; E-commerce

0 引言

电子商务推荐技术是指收集和统计客户站点访问信息(如客户的浏览历史,购买商品等),通过分析活动客户的浏览和购买行为来进行智能推荐[1]。推荐系统安装在商业站点可以帮助客户更好更快地访问有用的信息,站点管理者也可利用推荐系统进行交叉销售以提高销售量、满足电子商务推荐需要[2]。

电子商务推荐技术主要有基于商品内容的推荐技术和协同过滤技术。协同过滤推荐技术是当前研究的热点,其最大优点是不需要分析对象的属性,对推荐对象没有特殊要求[3]。推荐系统使用的算法主要有神经网络方法、关联规则方法和聚类方法[4~6],其中关联规则方法和聚类方法具有较好的推荐效果。关联规则法利用 Apriori 算法通过挖掘客户浏览历史记录的关联来进行推荐。这种方法有以下缺点:如果支持度和置信度选取不恰当,会造成计算时间太长或较差的推荐性能;一般电子商务网站的网页数目巨大,如果用关联规则方法进行推荐,会使系统很复杂,效率比较低。文献[6]提出用事务聚类的方法来构筑推荐系统,并取得了较好的效果。但是这种方法在聚类时事务的表示太简单,如只采取访问页面的布尔量,即用 0、1 表示客户是否访问了某网页。很显然,这种方法没有反映客户对某个网页访问的次数的影响。

商品评分是推荐技术和算法的信息输入。目前采用的模型是显式评分模型和隐式评分模型。显式评分模型通过对客户的在线调查和反馈等方式,来获得客户对商品的评分;而隐式评分则通过记录客户在线浏览行为等因素,分析出客户对商品的评分[8,9]。

影响客户在线访问行为并发现客户兴趣的因素相当复杂。从客户角度分析有客户心理、消费环境、工作任务、访问次数、访问时间、反馈情况(包括购买、咨询、再次访问)等;从网站角度分析有商品属性、网页大小、网站风格等。目前的研究多局限于对商品评价,没有考虑到商品属性对商品评价值的综合影响。如文献[7]综合考虑到客户访问次数、访问时间和页面大小综合建立 Web 站点客户访问矩阵;文献[9]建立了 Probabilistic Latent Semantic Analysis(PLSA)分析网站内容与客户工作任务之间相关性。

1 基于 Web 的商品属性隐性评分算法

1.1 CGA(Customer-Good-Attribute)模型

在电子商务环境下,客户随心所欲在网站上浏览,寻找中意的商品。网站可以通过客户的浏览记录捕捉客户的行为,从而分析出客户的购买倾向和兴趣。这里定义客户—商品—属性三维模型 CGA 来描述客户、商品和属性关系。

定义 1 客户—商品购买矩阵 $A = \{a_{ij}\}$,i 代表客户,j 为表示商品,$a_{ij} = 0$ 或 1,1 表示订购,0 表示没有订购。

定义 2 客户—属性评分矩阵 $R_i = \{R_{ijk}\}$,反映客户 i 对

收稿日期:2005 - 10 - 08;**修订日期**:2005 - 12 - 22 **基金项目**:国家自然科学基金资助项目(70572079)

作者简介:陈冬林(1970 -),男,湖北安陆人,副教授,博士,主要研究方向:电子商务、商务智能、知识网格; 聂规划(1957 -),男,河南人,教授,博士生导师,博士,主要研究方向:知识管理、人工智能、电子商务.

商品 j 第 k 个属性的隐性评分。

1.2　基于商品属性的商品综合评分方法

通过分析客户的浏览行为,计算其对商品某种属性的浏览时间占总浏览时间的比例,引入商品属性的权重矩阵,获取客户对商品的综合评分。如果客户多次浏览网站网页,将其浏览行为合并处理。采用平均法来求客户对同一网页中不同商品属性的浏览时间。详细计算步骤如下:

步骤 1:客户 i 第 s 次浏览行为中用于商品 j 的第 k 个属性的时间记为 $t_{sk}^{(ij)}$,按下式计算:

$$t_{sk}^{(ij)} = \frac{t_s^{(i)}}{\sum_{k=1}^{K_j} f_{sk}^{(ij)}} \times f_{sk}^{(ij)} \tag{1}$$

其中 i,j,k 分别代表客户编号、商品编号和商品属性编号;$f_{sk}^{(ij)} = 1$ 表示客户 i 第 s 次浏览的网页上有商品 j 的第 k 个属性,其他情况为 0;$t_s^{(i)}$ 为客户 i 第 s 次浏览的网页所用时间。

步骤 2:客户 i 对商品 j 的第 k 个属性的总浏览时间:

$$t_k^{(ij)} = \sum_{s=1}^{S_i} t_{sk}^{(ij)} \tag{2}$$

其中 S_i 为客户 i 的总浏览次数。

步骤 3:按下式计算客户对商品属性的隐性评分值。

$$R_{ijk} = \frac{t_k^{(ij)}}{\sum_{s=1}^{S_i} t_s^{(i)}} \tag{3}$$

步骤 4:商品 — 属性权重矩阵 $W = \{w_{jk}\}$,综合反映所有客户对商品 j 的第 k 个属性的兴趣程度,即:

$$w_{jk} = \frac{\sum_{i=1}^{I} t_k^{(ij)}}{\sum_{i=1}^{I}\sum_{k=1}^{K_j} t_k^{(ij)}} \tag{4}$$

其中 I 为客户的数量,K_j 为商品 j 的属性值数量。

步骤 5:客户 i 对商品 j 的综合评分表示为 R_{ij}:

$$R_{ij} = \sum_{k=1}^{K_j} R_{ijk} \times w_{jk} \tag{5}$$

1.3　算例

电子商务智能模拟实验系统中,记录了客户的对某品牌电脑浏览路径如下图 1 所示。

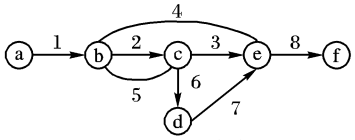

图 1　客户浏览路径

表 1　客户网页浏览信息表

浏览顺序	网页代码	商品属性							浏览时间
		f1	f2	f3	f4	f5	f6	f7	
1	a	1	1	0	1	0	0	0	15s
2	b	1	0	0	1	1	0	1	13s
3	c	0	1	1	0	0	0	1	17s
4	d	1	0	0	1	0	1	1	20s
5	b	1	0	0	1	1	0	1	7s
6	c	0	1	1	0	0	0	1	10s
7	e	0	0	0	1	0	0	0	9s
8	d	1	0	0	1	0	1	1	9s
9	f	0	0	1	0	1	0	0	10s

该电脑的属性集(CPU,内存,主板,价格,声卡,显卡,显示器),用向量表示为 $F(f1, f2, f3, f4, f5, f6, f7)$,其属性交叉分布在 6 个网页中,结合图 1 浏览路径可以得到表 1。

根据公式(1)、(2)、(3)计算,客户对该商品属性的隐性评分向量为(0.16,0.13,0.13,0.24,0.09,0.07,0.19)。

2　基于客户的协同过滤算法

协同过滤推荐根据其他客户的观点产生对目标客户的推荐列表,它基于这样一个假设:如果客户对一些商品的评分比较相似,则他们对其他商品的评分也比较相似。协同过滤推荐系统使用统计技术搜索目标客户的若干最近邻居,然后根据最近邻居对商品的评分预测目标客户对商品的评分,产生对应的推荐列表[1,2,6]。

客户评分数据用一个 $m \times n$ 阶矩阵 $\boldsymbol{R}(m,n)$ 表示,m 行代表 m 个客户,n 列代表 n 个商品,第 i 行第 j 列的元素 R_{ij} 代表客户 i 对商品 j 的评分数值。本文采用上节所述的基于 Web 的商品属性隐性评分算法来构造客户评分矩阵 $\boldsymbol{R} = \{R_{ij}\}$。

2.1　最近邻居查询

基于客户的协同过滤推荐系统的核心是为一个需要推荐服务的当前客户寻找其最相似的“最近邻居”集,即:对一个客户 C,要产生一个依相似度大小排列的“邻居”集合 $C = \{C_1, C_2, \cdots, C_t\}$,$u$ 不属于 C,从 C_1 到 C_t,$sim(u, C_k)$ 从大到小排列。

最近邻查询是整个基于客户的协同过滤推荐算法的核心部分,其效果和效率很大程度上决定了基于客户的协同过滤推荐算法的效果和效率。最近邻查询阶段实质上就是基于客户的协同过滤推荐算法的模型建立阶段。

2.2　相似度计算

度量客户 i 和客户 j 之间相似性的方法如下,首先得到客户 i 和客户 j 评分过的所有商品,然后通过不同的相似性度量方法计算客户 i 和客户 j 之间的相似性。度量客户间相似性的方法有多种,主要有余弦相似性、相关相似性和修正的余弦相似性[7],这里采用相关相似性。

相关相似性:设客户 i 和客户 j 共同评分的商品集合用 P_{ij} 表示,则客户 i 和客户 j 之间的相似性 $sim(i,j)$ 通过 Pearson 相关系数度量:

$$sim(i,j) = \frac{\sum_{p \in P_{ij}} (R_{i,p} - \overline{R_i})(R_{j,p} - \overline{R_j})}{\sqrt{\sum_{p \in P_{ij}} (R_{i,p} - \overline{R_i})^2} \sqrt{\sum_{p \in P_{ij}} (R_{j,p} - \overline{R_j})^2}} \tag{6}$$

其中 $R_{i,p}$ 和 $R_{j,p}$ 分别代表客户 i 和客户 j 对商品 p 的评分;$\overline{R_i}$ 和 $\overline{R_j}$ 分别表示客户 i 和客户 j 对商品的平均评分。

2.3　商品推荐产生

基于最相似性测量出来的最相似的客户,下一步就是查看目标客户的评分,来计算预测值。假设客户 u 对商品 P 的评分预测为 $PR_{u,p}$ 表示为:

$$PR_{u,p} = \overline{R_u} + \frac{\sum_{i=1}^{I} (R_{i,p} - \overline{R_i}) \times sim(u,i)}{\sum_{i=1}^{I} sim(u,i)} \tag{7}$$

这里 $R_{i,p}$ 表示客户 i 对商品 p 的评分,$\overline{R_u}$ 和 $\overline{R_i}$ 分别表示客户 u 和客户 i 对商品的平均评分,$sim(u,i)$ 表示客户 u 和客户 i 的相似度,I 为客户的数量。

通过上述方法预测客户对所有未评分商品的评分,然后选择预测评分最高的前若干个商品作为推荐结果反馈给当前客户。

3 实验系统及结果

在我们开发的电子商务智能模拟实验系统中,采用基于商品属于隐性评分的协同过滤算法。该系统从机房管理信息系统获取学生的上网浏览的日志,以校园卡平台为基础有效识别上机的学生身份,并记载上机浏览网页日志。而电子商务的销售和商品等信息存储在 AS400 小型机 DB2 数据库中,系统采用 J2EE 结构开发,建立了推荐模型库、在线数据挖掘模型库和 CGA 模型库等。此系统已取代了原有的电子商务模拟系统,成为电子商务及相关专业的课程实验系统。

通过近半年 900 多人次的电子商务模拟实验,系统自动记载学生浏览行为,综合考虑浏览路径和时间、商品属性及其在网页中的分布等因素,来获得学生对商品属性的隐性评分值,以此为基础,给出协作过滤的商品推荐,学生的模拟购买率和反馈率高达 86%,证明系统的推荐质量较高。

4 结语

在分析现有基于客户评分协同过滤技术存在的缺陷问题情况下,提出了一种准确的、实时的、基于 Web 日志的 Internet 电子商务推荐算法。该算法从四个方面提升推荐系统的功能:(1)从商品相似分析扩展为商品属性相似分析;(2)从客户对商品评价扩展为商品各属性的评价;(3)从客户相似性扩展为客户偏好结构的相似性;(4)通过客户浏览行为在线挖掘客户对商品属性的评分。但在本方法没有考虑到不同客户(VIP 客户、主要客户、普通客户等)之间的差异,他们的浏览行为对商品的属性权重有不同影响,另外页面的信息量也影响着客户浏览时间,这些方面有待进一步研究。

参考文献:

[1] GAUL W, SCHMIDT - THIEME L. Recommender systems based on navigation path features[A]. International Conference on Knowledge Discovery and Data Mining[C]. San Francisco, 2001.

[2] FU X, BUDZIK J, HAMMOND KJ. Mining navigation history for recommendation[A]. Proceedings of 2000 international conference intelligent user interfaces[C]. New Orleans, LA: ACM, 2000. 106 - 112.

[3] 余力,刘鲁,罗掌华. 我国电子商务推荐策略的比较分析[J]. 系统工程理论与实践,2004,24(8):96 - 101.

[4] 黎星星,黄小琴,朱庆生. 电子商务推荐系统研究[J]. 计算机工程与科学,2004,26(5):7 - 10.

[5] 余力,刘鲁,李雪峰. 用户多兴趣下的个性化推荐算法研究[J]. 计算机集成制造系统,2004,10(12):1610 - 1615.

[6] 王太雷. 基于相似模式聚类的电子商务网站电子商务推荐系统研究[J]. 计算机工程与应用,2005,41(6):152 - 157.

[7] 王勋,凌云,费玉莲. 基于 Web 日志和缓存数据挖掘的电子商务推荐系统[J]. 情报学报,2005,24(3):324:328.

[8] MOBASHER B, DAI H, LUO T, *et al*. Discovery of aggregate usage profiles for Web personalization[A]. International Conference on Knowledge Discovery and Data Mining[C]. Boston, 2000. 210 - 221.

[9] JIN X, ZHOU Y, MOBASHER B. Web usage mining based on probabilistic latent semantic analysis[A]. Proceedings of the Tenth ACM SIGKDD Conference[C]. 2004.

[10] SARWAR B, KARYPIS G, KONSTAN J, *et al*. Item-based collaborative filtering recommendation algorithm[A]. Proceedings of the Tenth International World Web Conference[C]. 2001. 285 - 295.

[11] FU X, BUDAZIK J, HAMMOND KJ. Mining navigation history for recommendation[A]. Proceedings of the International Conf on Intelligent User Interfaces[C]. New Orleans, LA: ACM, 2000. 106 - 112.

(上接第 960 页)

由于地址式样是 1391201 *,所以 BE_A 继续发送 AccessRequest 给 BE_D;

5) BE_D 查询得到 Td 的呼叫地址信息后,通过 AccessConfirmation 消息反馈给 BE_A;

6)由 BE_A 转发 AccessConfirmation 给子域的 BE_C;

7) BE_C 发送 LCF 将地址确认信息传给 GK_C。终端收到 ACF 后发起呼叫。根据 GK 配置的不同,可以选择 GK 路由模式呼叫,这里不再赘述。

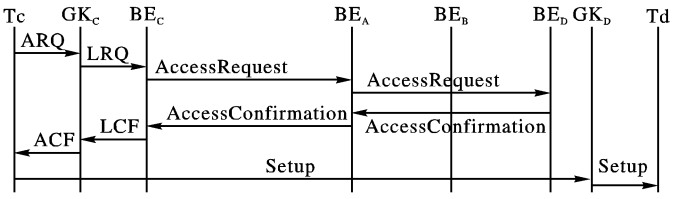

图 6 子管理域间呼叫流程 Tc 呼叫 Td

4 结语

本文提出的管理域层次域间通信方案和其他方案相比有以下优点:1)便于大规模的组网。GK 组网中就有父子关系的组网模式,有利于网络的扩展;2)全网格结构中一旦某个地址描述符改变就要更新除他自己之外其他所有邻域 BE 的地址模板配置,而本方案中只需要改变上一级 BE 的配置,避免了全网格中广播更新消息可能造成网络拥塞。3)有利于运营商之间的管理和计费。由于域间通信主要用于运营商之间,这种层次域间通信方案的 BE 可以作到地址解析、授权、认证等功能的顶层会聚,具有实际应用价值。不过这样对 BE 的稳定性和处理能力要求较高。在以后的工作中,我们将进一步研究 Annex G 域间通信中 BE 的稳定性、高效的全局地址解析方法和运营商得到精确的“使用”信息的途径。

参考文献:

[1] ITU-T Recommendation H.323(Version 4). Packet-Based Multimedia Communications Systems[S]. 2000.

[2] ITU-T Recommendation H.225.0. Call signalling protocols and media stream packetization for packet-based multimedia communication systems[S]. 2002.

[3] ITU-T Recommendation H.225.0 Revised Annex G (Version 2), Communication between and within Administrative Domains[S]. 2002.

[4] ITU-T Recommendation H.501, Protocol for Mobility Management and Intra/inter-domain Communication in Multimedia Systems [S]. 2002.

[5] 仇佩亮,杨永康, 赵志峰. IP 电话系统和呼叫路由技术[J]. 中国数据通信,2003,11(5):128.

[6] LI R, YU ZHW, WANG H. Research on Multi-Zone and Hierarchical Routing for Video Gatekeeper Based on Agent[A]. IEEE Proceedings of the 2005 Systems Communications (ICW'05)[C]. Canada. 2005. 159 - 164.