商业站点推荐策略的研究

The Research on the Recommendation Strategies to Commerce Website

谢 中 邱玉辉

(重庆西南师范大学计算机与信息科学学院 重庆400715)

Abstract Under the new style of electronic commerce, it is very important to analyze and learn customers deeply and create the modes about them. In the current researches the modes about item-to-item correlation or user-to-user correlation are always discussed, but here the modes about user-to-item correlation is presented. In this essay web usage mining and web content mining technologies are combined to build the user-item modes which indicate how a user is interested to a item. At last the recommendation strategies to current user based on the previous created modes are given.

Keywords Recommendation strategies, Commerce website

1 引言

进入21世纪后,随着网络的普及,电子商务的发展越来越引起研究者们的关注,期望能够在这种新型的商务模式下,利用它诸多的优点,获得更多的客户以提高收益。但是这种新的商务模式在获得更多的客户以提高收益的同时,面临着巨大的考验。因为它突破了传统商务空间上和时间上的限制,客户只需要简单的几个点击操作就可能流失到竞争者那里。因此对商务站点上的企业来说,从各方面获得关于客户的知识,了解客户的需求和兴趣偏好,然后构建一个推荐系统向客户提供个性化的服务,使他们能够方便快捷地找到感兴趣的商品和得到符合他们各自偏好的服务,已经越来越重要,并且成为计算机领域一个新的研究热点。

目前,对于推荐系统的构建国内外已经有了相关的研究。例如,Minnesota 大学的 J. ben Schafer 等人运用协作过滤方法产生推荐[9];Stanford 大学 Kwong Hiu Yung 等人的在线售书推荐系统,运用到了多种数据挖掘技术[11];意大利 Paolo Buono 等人研究的推荐系统中,从显式和隐式两方面获得用户模式及用户评价[12],以此作为推荐的依据。在这些推荐系统的研究中,采用的推荐算法从最初的对数据的简单检索(raw retrieval)到现在普遍采用的基于物品和物品关系的推荐(item-to-item correlation)和基于用户和用户关系的推荐(user-to-user correlation),都没有考虑用户和物品的关系(user-to-item correlation)。实际上考虑用户和物品的关系,知道某一类用户或某一个用户对某个物品的兴趣模式,向用户提供推荐服务时才更有针对性。论文针对这一点,主要讨论了如何建立用户和物品关系的模式(用户-物品模式),表明用户对某个物品的兴趣度。并通过建立的这些模式,结合 item-to-item correlation 和 user-to-item correlation 两种方法讨论针对当前用户的推荐策略。

论文首先运用统计学相似度计算的方法得到两个物品间相似度模式,然后讨论了用户的聚类和从页面中如何获得商品信息以及商品相对于页面的权值,接下来根据前面获得的用户聚类模式和物品页面模式建立用户-聚类模式,最后是基于这些模式的推荐策略。

2 物品之间相似模式的建立

物品相似度是指物品 i 和物品 j 它们在被顾客购买时存在的一种普遍性的相互关联的一个尺度,用 sim(i,j)表示。对应在实际的商业运作中,sim(i,j)表示对在站点上购买过东西的用户而言,物品 i 和物品 j 总在一起被购买的几率,这种几率越大,我们就说这两个物品越相似。物品间相似模式的建立可以从分析用户购买数据库获得,可以把用户购买数据库看成如图1所示的矩阵:

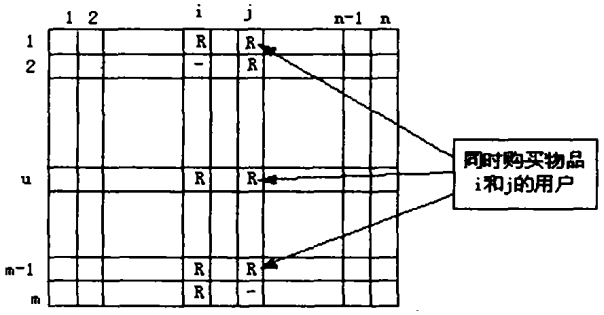

图1

矩阵中的行代表购买过物品的用户,列代表物品空间上的物品,从列的方向上把物品看成是关于用户的矩阵,则可以用统计学中基于余弦的相似度计算方法计算任意两个物品 i 和 j 的相似度,计算公式为:

$$sim(i,j)=\cos(\rho_i,\rho_j)=\frac{\rho_i * \rho_j}{|\rho_i| \times |\rho_j|}$$

当两个都不经常被购买的物品在某个用户处产生重叠时,上述方法的计算可能会使这两个根本不相干的物品它们的 sim(i,j)值可能会很高,但是它们只是恰好被某一个用户同时购买。为了处理这种情况,我们需要把计算出来的 sim(i,j)乘与一个因子 k,使得 sim(i,j)能减少错误。经过分析,确定 $k=|U_{i,j}|/|U|$。$|U_{i,j}|$表示既购买物品 i 又购买物品 j 的用户数,$|U|$表示数据库中总的用户数。经过这样的处理,使得只有当物品 i 和 j 同时被购买的情况对大多数用户来说都是成立的,则 sim 值才可能较高。

3 用户聚类模式

用户聚类的目的是把具有相似浏览行为的用户分归为一类,假定具有相似兴趣的用户会产生相似的访问行为,因此以浏览行为归类用户相当于把具有相似兴趣偏好的用户归类。

这里分析的是 Web 服务器上的日志文件，采用的是 Web usage 分析方法。完成用户聚类工作，事务识别是重要的一环。事务识别就是依据一定的事务识别标准，在一个用户会话中，分割用户会话中全部页面的参照序列为一个个的事务单位。事务可以是一页，也可以是全部页。常用的识别方法有文[1]中提到的两种：时间窗法和最大向前参照法。本文采用的是 Bamshad Mobasher、R. Cooley 研究的时间窗法。以 IP 和用户 ID 为准在服务器日志中把用户对站点的访问划分成为一个个的用户访问事务 t，L 为服务器访问日志的记录集，则用时间窗法识别出来的某个事务表示为：$t=\langle ip_i, uid_i, \{(l'_1.url, l'_1.time), \cdots, (l'_m.url, l'_m.time)\}\rangle$。其中 $1\leqslant k\leqslant m$，$l'_k\in L$，$l'_k.ip=ip_i$，$l'_k.uid=uid_i$，并且 $l'_{k+1}.time-l'_k.time\leqslant\Delta t$，$t$ 是某个事务；ip_i 为这个事务用户的 IP 地址；uid_i 为这个事务用户的用户标识，非注册用户的 uid_i 为空值；l'_k 是这个用户在事务中产生的第 k 个访问日志记录；$l'_k.time$ 表示第 k 个日志记录的资源被访问的时间。对服务器日志进行事务识别后产生 m 个事务，组成事务集 $T=\langle t_1, t_2, \cdots, t_m\rangle$，$t_i\in T$，而且 t_i 是站点上所有页面的一个子集。

因为这些识别出来的事务是用户在某一次对站点访问的页面序列的片断，这个片断就体现了当时那一刻用户对站点的访问情况。聚类这些事务就等于聚类用户对站点的访问情况，则具有相似访问情况的事务根据一定的聚类算法就可以产生一定数量的事务聚类。对于某个事务聚类，代表了某一类用户，这个聚类中的用户对站点有相似的访问，因此用户的聚类最终是事务的聚类。事务的聚类，采用数据挖掘中最广泛运用的 K-means 算法[2]。首先必须把各个事务表示成 K-means 能够运算的数值型向量。这里把事务映射在页面空间上，表示成页面空间上的多维向量。设站点所有 web 页面的集合为 $P=\{p_1, p_2, \cdots, p_n\}$，则事务 $t\in T$ 可以表示成页面空间 P 的多维数值向量 $t=\langle w(p_1,t), w(p_2,t), \cdots, w(p_n,t)\rangle$，其中：

$$w(p_j,t)=\begin{cases}\gamma & \text{页面 } p_j \text{ 在事务 t 中出现的次数}\\ 0 & \text{如果页面 } p_j \text{ 没有在事务 t 中出现}\end{cases}$$

$1\leqslant j\leqslant n$，$\gamma>0$

识别出来的事务都转化成页面空间 P 上的多维向量后，利用 K-means 算法对事务进行聚类运算。通过 K-means 聚类运算，则把这些用户的事务分成一个个的事务组，根据一定尺度的相似度计算或距离计算，事务组中的每个事物之间最相似或者距离最短，由此得到关于用户的事务聚类 $TC=\{c_1, c_2, \cdots, c_m\}$，m 是聚类的总数，每一个 c_i 就是一个事务集 T 的子集。TC 中的每一个聚类，就代表了一类具有相似访问模式的用户。并且由此聚类 TC 可以得出某个页面 p 在某个聚类 c 或 u 中的权值：

$$weight(p,c)=\frac{1}{|c|}\times\sum_{t\in c}w(p,t)$$

4 物品-页面模式

这里讨论的是物品的推荐。因为如果进行页面的推荐，页面内容一旦改变，页面间原来存在的关系就不成立了。这对于页面内容经常改变的商业站点很不合适。因此我们需要进一步找出页面涉及的物品内容，对用户进行以物品为基础的推荐。我们需要用到内容数据挖掘中信息抽取方法分析和识别出页面所涉及的物品信息。信息抽取的目的是对站点上的 HTML 或 XML 页面运用信息抽取的方法，从 HTML 或 XML 中抽取出相关的物品信息可运用机器学习理论[3,4]或传统的 Web 内容挖掘技术。

通过对 HTML 和 XML 进行信息抽取，可以获得每个页面涉及的物品集以及这些物品相对这个页面的权值 w (i,p)，i 表示物品空间中的某个物品，p 是页面空间中的某个页面。权值的定义有多种方法，这里我们定义权值为物品 i 在页面 p 中出现的频率，除此之外还可以用文[5]中的权值定义方法。设 Web 站点涉及的所有物品组成的空间为 IS(Item Space)，空间大小为 n。于是在进行信息抽取后，每个页面 p 可以表示成物品空间 IS 上的 n 维向量，建立物品-页面模式 $p=\{w_1^p, w_2^p, \cdots, w_n^p\}$，其中：

$$w_i^p=\begin{cases}w(i,p) & \text{物品 i 在页面 p 中的权值}\\ 0 & \text{如果页面 p 不涉及到物品 i}\end{cases}\quad 1\leqslant i\leqslant n$$

实际上 w(i,p)权值的确定除了可以用上述的方法获得外，还可以在页面被创建的时候由站点分析人员人为地指定，直接确定当前建立的这个页面涉及了什么物品，并且这些物品关于该页面的重要程度如何。

5 建立关于用户-物品模式

通过 Web usage 挖掘产生了具有相似访问行为的用户聚类，通过 Web 内容挖掘的信息抽取产生了页面在物品空间 IS 上的向量表示。以这两者为基础，产生用户-物品模式。用户-物品模式是某类用户对物品空间 IS 上的某个物品兴趣度的表示，记为 φ(i,c)。用户-物品聚类模式的求取公式为：

$$\varphi(i,c)=\frac{1}{|c|}\sum_{p\in c}w(i,p)\times weight(p,c),$$

其中 w (i,p)是物品 i 相对页面 p 的权值，weight (p,c)是页面 p 相对聚类 c 的权值。n 是物品空间 IS 的大小。于是对于每一个聚类 c 可以表示为物品空间中的物品相对于这个聚类的权值的 n 维向量，建立物品的用户聚类模式：$c=\langle\varphi(1,c), \varphi(2,c), \cdots, \varphi(n,c)\rangle$。由 $c=\langle\varphi(1,c), \varphi(2,c), \cdots, \varphi(n,c)\rangle$可知，属于某个聚类 c 中的客户对于物品空间中的每个物品感兴趣的程度。有了这个模式，在用户浏览站点并确定用户的分类后，就可以有针对性地选择推荐的物品。

6 推荐策略

针对当前用户的推荐主要是通过计算某个物品 i 相对于当前用户的推荐值 Rec(i)是多少而确定，所采用的推荐方法是 item-to-item correlation 和 user-to-item correlation 相结合的方法。设物品 i 是用户曾经购买过的物品，物品 j 是正在计算是否应该推荐给当前用户的物品，则推荐时从策略上主要考虑以下因素：

①用户对曾经购买过的物品 i 的评价值 $R_{u,i}$；

②用户曾经购买过的物品 i 和正在计算推荐值的物品 j 的关联相似度 sim(i,j)；

③该用户在曾经对网站的浏览中，对于正在计算推荐值的物品 j 的物品-用户聚类模式 φ(j,u)。

于是推荐器所采用的推荐计算公式为：

$$Rec(j)=\varphi(j,u)+\sum_{i\in u}sim(i,j)^*(R_{u,i}+1)。$$

其中，i 是用户 u 在某个时期里注册登记在数据库中的购买过的物品，或者是正在购物车中的物品。$R_{u,i}$是该用户对曾经购买过的物品的评价值。这里使用 $R_{u,i}+1$作为相乘因子，是为了避免 $R_{u,i}$等于0时，造成该物品在公式中其它参与计算的值失效。

（下转第120页）

user { $topic_1$, threshold, limited;
$topic_2$, threshold, limited;
…
$topic_n$, threshold, limited}

简记为$\langle p_1, p_2, \cdots, p_n\rangle$，其中 limited 表示用户感兴趣话题 $topic_i$ 的外延限制，threshold 表示用户阈值。阈值 threshold 用于判别用户对一篇文档是否感兴趣，属于对内涵的要求。

相应的有下面的判定规则：

规则1

对文档 D 与用户 u，若 D 与话题 $T_i(i\in N)$的相关值 $R>=user_u:threshold_{T_i}$，并且 D 符合用户 u 对话题 T_i 的外延限制，则称 D 是用户 u 所需要的文档。

文档与话题都有一个相关值，这个值是统一的，与用户无关的，在0和1之间。Threshold 也取0和1之间，但每个用户依据自己的需要其 threshold 值各不相同，形成了个性化需求。

3.5 用户 profile 的维护

一定时期内经过用户端的信息流是一个信息集合，记为 Q；在 Q 中符合用户兴趣需求的子集记为 U；其他不属于用户兴趣范围的信息构成子集 M；显然有：

$$Q=U+M$$

从 Q 中依据表达用户需求的向量$\langle p_1, p_2, \cdots, p_n\rangle$而生成的信息集合记为 U'。用户兴趣向量$\langle p_1, p_2, \cdots, p_n\rangle$中元素选取的优劣、数量多少直接影响到 U' 与 U 的接近程度和信息过滤算法的效率。也是评价信息过滤效果、进而调整用户 profile 文件的重要依据。向量$\langle p_1, p_2, \cdots, p_n\rangle$中元素的选取应该既是较少的以便使计算量较少，同时又是较优的以便能准确描述用户的需求。

用户代理通过观察、记录、分析用户对 U' 的行为，将 U' 中用户不感兴趣(没有阅读等操作)的信息的特征从向量$\langle p_1, p_2, \cdots, p_n\rangle$中去除或是修改其特征值；将新发现的用户感兴趣的信息的特征加入到向量$\langle p_1, p_2, \cdots, p_n\rangle$中。从而不断动态地调整、修正用户的兴趣 profile 文件，使其能更准确地表达用户动态变化的兴趣需求，使 U' 能逐渐逼近 U。

小结 基于关键词的用户需求模型只注重了用户需求的个性，而未注重用户需求的共性，不便于进行协同过滤和对用户群的过滤，虽可减轻网络流量，但却使系统要处理的用户模型的数量随用户规模的扩大而成正比扩大，不适合于大规模信息过滤系统。此外，尽管用户的话题需求是较为长期、稳定的，但由于不断地进行阅读活动，用户对话题的外延的要求却是随时间变化的。因此本文从话题的角度并加进了外延限制而建立的用户兴趣模型，既能反映用户需求的共性，又能反映用户的个性化特征，使系统可以较为灵活地处理用户需求，在为单个用户过滤信息的同时，还可以通过改变外延限制而使系统只需以较少的模型数量来统一地表示用户群的信息需求，高效地为用户群过滤信息，增强了系统的可伸缩性，适合于大规模信息过滤系统。

参 考 文 献

1 Allen R B. User models: Theory, Method, and practice, Int. J. Man-Machine Studies, 1990, 32: 511～543
2 Stadnyk I, Kass R. Modeling User's Interest in Information Filters. Communication of the ACM, 1992, 35(12): 49～50
3 Lee D L, Chuang H, Seamons K, Document Ranking and the Vector Space Model. IEEE Software, 1997, 14(2)
4 Foltz P W, Dumais S T. Personalized Information Delivery: An Analysis of Information Filtering Methods. Communications of ACM, 1992, 35: 51～60
5 tauritz D. Adaptive Information Filtering as a Means to overcome Information Overload, M. Sc. thesis, Department of Computer Science Leiden University, the Netherland, sep. 1996
6 Resnick P, Varian H R. Recommender Systems. COMMUNICATIONS of the ACM, 1997, 40(3): 56～58
7 Rodriguez-Mula G, et al. Collaborative value filtering on the Web. Computer Networks and ISDN Systems, 1998, 30: 736～738

(上接第122页)

最后由站点的分析员定义阈值，对推荐结果进行排序、筛选、合并，确定给用户的推荐商品集为：USER(u)←{j|Rec(j)>α, j∈IS}。阈值的确定可以根据需要灵活使用不同的方法，例如可以确定 $\alpha=\frac{1}{|J|}\sum_{j\in J}\mathrm{Re}\,c(j)$。这部分产生的推荐主要可以使站点提高交叉销售收益。

此外，对当前用户 u 还有另一方面的推荐集。该用户对站点访问的 session 对象中的最后一个页面 p 可以认为代表用户的最新兴趣，用户之所以访问这个页面 p，就是因为该页面所涉及的物品内容。所以运用信息抽取方法[3]，找出该页面涉及的物品有 $k_1, \cdots\cdots, k_z$，$1\leqslant z\leqslant|IS|$，而且有 $w(k_1, p), \cdots, w(k_z, p)$。则此时衡量物品 m 是否可以推荐给用户 u 的关于物品 m 推荐值的计算公式为：$\mathrm{Re}\,c(m)=\sum_{i=1}^{z}w(k_i, p)\times sim(k_i, m)$。由此公式可以得到当前对于用户 u 的另一个推荐集为：USER(u)←{m|Rec(m)>β, m∈IS}。

小结 在电子商务的新形势下，为了提高客户的忠诚度我们需要深入分析、了解客户，建立关于客户的兴趣偏好等知识模式，然后以这些建立的知识模式为依据向客户提供个性化的服务。本文则在当前已有的研究基础上采用 item-to-item correlation 和 user-to-item correlation 相结合的方法，确定针对当前用户的推荐策略，并根据需要建立了物品间相似模式、用户-物品模式等。

参 考 文 献

1 Cooley R, Mobasher B, Srivastava J. Grouping Web Page References into Transactions for Mining World Wide Web Browsing Patterns. In: Proc. of ICTAI 1997
2 Han Jiawei, Kamber M. data mining
3 Craig A. Knoblock. Accurately and Reliably Extracting Data from the Web: A Machine Learning Approach. IEEE Data Engineering Bulletin, Volume 23
4 Customer Data Quality. http://WWW.firstlogic.com
5 Balabanovic M, Shoham Y, Yun Y. An Adaptive Agent for Automated Web Browsing. http://WWW.citeseer.com
6 [美]Berson A, Smith S, Thearling K 著，贺奇等译. 构建面向CRM的数据挖掘应用
7 Schafer J B, et al. E-Commerce Recommendation Applications. Data Mining and Knowledge Discovery, 2001, 5(1/2): 115～153
8 Nahm U Y, Mooney R J. Text Mining with Information Extraction. In the proceeding IJCAI-2001
9 Buchner A G, et al. An Internet-enabled Knowledge Discovery Process. http://www.citeseer.com
10 Karypis G. Evaluation of Item-Based Top-N Recommendation Algorithms: [Technical Report CS-TR-00-46]. Computer Science Dept., University of Minnesota. 2001