

一种基于网络书签的个性化信息推荐方法^①

王世云, 王志军, 李运杰

(辽宁工程技术大学 电子与信息工程学院, 葫芦岛 125105)

摘要: 针对目前 Internet 上信息的爆炸式增长, 提出了一种基于 Web 挖掘的个性化信息推荐方法, 介绍了个性化推荐的流程和实现模型, 并借鉴复杂网络中的社团结构划分方法, 提出了基于网络书签的个性化信息推荐方法, 分析了社团内基于协作过滤和社团间基于“信息桥”的个性化信息推荐, 并通过实验验证了此方法在个性化信息推荐中是有效的。

关键词: 个性化推荐; 网络书签; 社团结构; 协作过滤; 信息桥

A Personalized Information Recommendation Method Based on Network Bookmark

WANG Shi-Yun, WANG Zhi-Jun, LI Yun-Jie

(School of Electronic and Information Engineering, Liaoning Technical University, Huludao 125105, China)

Abstract: For the present Internet information detonation type grew, this paper proposed a kind personalized information recommendation method based on the Web mining. It introduced the personalized recommendation's flow and realizes model, and profits from the complex network the mass organization structure division method. It proposed personalized information recommendation method based on the network bookmark, analyzed in the mass organization based on the cooperation filtration and the mass organization based on “the information bridge” personalized information recommendation. Through the experiment this method is proved to be effective in the individualized information recommendation.

Keywords: personalized recommendation; network bookmark; mass organization structure; cooperation filtration; information bridge

随着 Web 的迅猛发展, 如何快速、准确地从浩瀚的网络资源中获得有价值的信息, 如何在这些海量信息中不“迷失自己”。为了满足用户的特殊需求, 个性化信息推荐成为互联网时代商家克敌制胜的法宝。要想在新的一轮信息化竞争中站稳脚跟, 个性化信息推荐已经成为其必不可少的手段, 本文提出了一种基于网络书签的个性化信息推荐方法。

1 个性化信息推荐

个性化信息推荐就是根据 Web 用户的自身需求、兴趣或行为模式, 将 Web 用户感兴趣的信息推荐给 Web 用户。传统的大众化信息服务模式已不能满足

Web 用户复杂多变的信息需求, 迫切需要针对不同的 Web 用户提供个性化信息服务。个性化信息服务不仅能够提供结合 Web 用户的个性特征、使用习惯, 针对 Web 用户提出的明确要求提供高效的信息服务, 而且还根据 Web 用户的知识结构、心里倾向等方面充分激发 Web 用户的潜在信息需求, 促进 Web 用户对信息的有效利用。个性化信息推荐是一种典型的“信息找人”模式, 优于以信息检索为代表的传统“人找信息”的模式^[1]。

2 基于Web挖掘的个性化信息推荐流程

个性化信息推荐的实现需要以 Web 用户兴趣模型

① 基金项目:国家自然科学基金(50874059);辽宁工程技术大学第三届研究生科研立项(Y201000503)

收稿时间:2010-09-10;收到修改稿时间:2010-10-16

为基础产生面向特定 Web 用户的信息推荐。Web 挖掘的功能之一就是找出大量看似无关的信息之间的联系和规律,从而构建 Web 用户兴趣模型。首先 Web 站点服务器日志记录了 Web 用户的行为轨迹,Web 使用挖掘能够从中发现有用的知识和模式,是构建 Web 用户兴趣模型的关键。其次,Web 内容挖掘能够实现 Web 页面特征提取、基于内容的 Web 页面聚类、Web 页面之间内容的关联规则发现,它们都可以用于构建 Web 用户兴趣模型。最后,Web 结构挖掘可用于 Web 页面分类,并由此获得有关不同 Web 页面间相似度及关联度的信息,同样可以用于构建 Web 用户兴趣模型。基于 Web 挖掘的个性化信息推荐流程如图 1 所示^[2]。

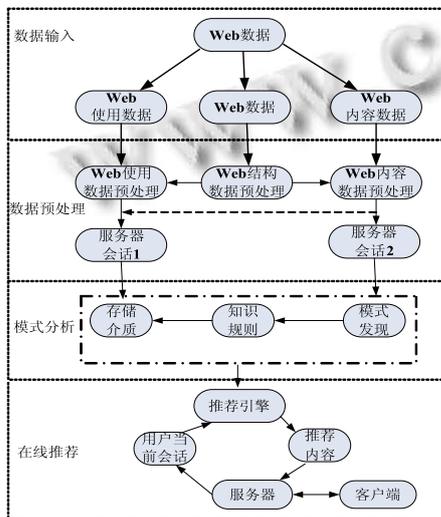


图 1 基于 Web 挖掘的个性化信息推荐流程

3 基于网络书签的个性化信息推荐方法

3.1 网络书签

网络书签提供一项叫做“社会化书签”的网络服务,Web 用户把喜爱的超链接网址随时加入自己的网络书签中,并可以用多个社会化标签(Tag)来标示和整理自己的网络书签,并与其他人共享。网络书签作为一种隐性社会化软件,其出现和产生也经历了一个从量变到质变的过程,即网络化、协同化、社会化三个阶段^[3]。最初,网络书签的前身--网页收藏夹是在 Web 用户的本地浏览器中实现的,Web 用户的收藏行为只影响本地系统。经过网络化阶段,便可以把网页收藏夹发展到网络上,由此出现了网络书签,方便 Web 用户在不同地点访问网页收藏夹。

3.2 网络书签系统模型

如图 2 所示,在网络书签中,存在着 Web 用户和 Web 资源两类对象,由此可以建立网络书签系统模型。

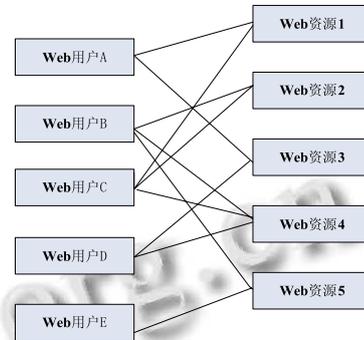


图 2 网络书签系统模型实例

在网络书签中,Web 用户之间可能并不认识,但是如果他们均收藏了同一 Web 资源,那么彼此之间就可以建立连接关系。而且,他们共同收藏的 Web 资源数量越多,彼此的关系也就越强,同时意味着兴趣的关联度越高。在实际应用过程中,有网络书签系统导出的可能是几个彼此分割的 Web 用户关系网络。这些属于不同 Web 用户关系网络的 Web 用户之间没有收藏过同一 Web 资源,同时也可以认为这些不同 Web 用户关系网络之间的兴趣的差异度非常大。

4 网络书签系统中基于社团结构的个性化信息推荐

4.1 社团结构

社团结构已成为复杂网络领域的热点研究方向之一,网络中的社团结构常用的定义是基于相对连接频数的:网络中的顶点可以分为组,组内连接稠密而组间连接稀疏。还可以定义为是以连通性为标准定义社团,称之为派系。一个派系是指有 3 个或 3 个以上的顶点组成的全连通子图,即任何两点之间都直接相连^[4]。

通常个性化信息推荐过程可以划分为离线环节和在线环节。其中 Web 用户关系网络的社团结构划分是离线结构环节的核心任务,而在线环节的基本任务就是依据 Web 用户关系网络中的社团结构完成实时的个性化信息推荐。在实时的个性化信息推荐中,可以借鉴协作过滤的思想实现社团内的个性化信息推荐。同时,Web 用户关系网络中的社团结构可能是一种有重

叠的社团结构, 存在一个节点属于多个社团的情形, 因此可以利用这些扮演“信息桥”角色的重叠节点实现社团间的个性化信息推荐, 网络书签系统中基于社团结构的个性化信息推荐模型, 如图 3 所示。将网络书签系统中唯一的 Web 用户关系网络划分为 3 个 k-派系社团, 其中社团 1 的浅灰色节点已经收藏了①②③⑤…Web 资源, 同属社团 1 和社团 3 的黑色节点已经收藏了②③④⑤…Web 资源, 社团 2 的深灰色节点与社团 3 的黑色节点之间的连接关系属于强联系。图 3 分析的是网络书签系统只导出一个 Web 用户关系网络的情形, 即网络书签系统中的所有 Web 用户都划分到同一个 Web 用户关系网络。如果由网络书签系统导出的是多个分割的 Web 用户关系网络, 那么就需要针对每个 Web 用户关系网络进行社团结构划分, 图 3 所示的模型来实现不同的 Web 用户关系网络中的个性化信息推荐。

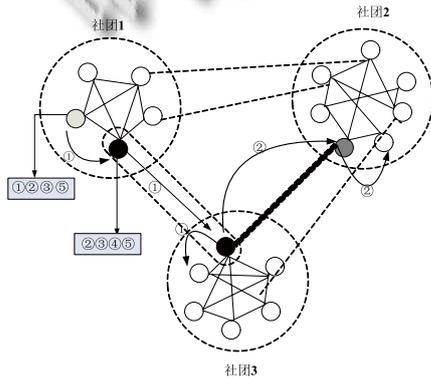


图 3 网络书签系统中基于社团结构的个性化信息推荐系统

4.2 社团内基于协作过滤的个性化信息推荐

以社团 1 为例, 社团内基于协作过滤的个性化信息推荐的基本原理可以描述为: 如果浅灰色节点和黑色节点之间的关系较高, 可以依据浅灰色节点的兴趣来预测黑色节点的个性化推荐对象, 如 Web 资源①。社团内基于协作过滤的个性化信息推荐是以社团内的节点以及这些节点收藏的 Web 资源为分析对象, 从而大幅度降低个性化信息推荐的时间复杂度。如果没有进行社团结构划分, 个性化信息推荐系统需要考虑 Web 用户关系网络中的所有 Web 用户及其收藏的全部 Web 资源, 无疑大幅度增加了复杂程度。

4.3 社团间基于“信息桥”的个性化信息推荐

社团内基于协作过滤的个性化信息推荐以社团为

分析对象, 所推荐的 Web 资源仅限于社团内各节点已收藏的 Web 资源, 存在一定的局限性。就需要利用“信息桥”实现社团间的个性化信息推荐, 以弥补社团内基于协作过滤的个性化信息推荐的不足。具体来说, 可以利用社团之间的重叠节点以及不同社团节点之间存在的强联系来构建“信息桥”。

在图 3 中, 社团 1 和社团 3 拥有一个重叠节点—黑色节点。在社团 1 中, 黑色节点接收了浅灰色节点推荐的 Web 资源①之后, 它起到连接社团 1 和社团 3 “桥梁”地位, 将 Web 资源①推荐给社团 3 中的节点, 从而实现社团间的个性化信息推荐。同样, 在图 3 中, 如果社团 2 中的深灰色节点和社团 3 的黑色节点之间的连接关系大于预先设定的阈值, 可以认定它们之间存在强联系, 可以将 Web 资源②推荐给深灰色节点, 使得 Web 资源②能够从社团 3 共享到社团 2, 最后深灰色节点采用类似社团内基于协作过滤的个性化信息推荐方法将 Web 资源②推荐给社团 2 中的相关节点。

5 实验分析

为了评价基于社团结构的个性化信息推荐系统方法的优劣, 实验中采用平均绝对误差 MAE 作为衡量的标准。MAE 通过计算预测的用户评分与实际的用户评分之间的偏差来衡量预测的准确性。假设在测试集中对 X 个项目预测的偏好值向量表示为 $\{p_1, p_2, \dots, p_x\}$, 实际的用户评价集合为 $\{q_1, q_2, \dots, q_x\}$, 则 MAE 的计算公式如下式所示^[5]:

$$MAE = \frac{\sum_i |p_i - q_i|}{X} \quad (1)$$

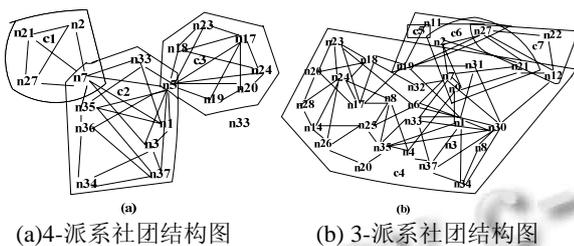
其中, $|p_i - q_i|$ 为 p_i 和 q_i 之间的绝对误差值; MAE 越小说明推荐质量越高。由于, Web 用户对 Web 文档的平均值只有 1 (已收藏) 和 0 (没收藏) 两中情形, MAE 的取值区间是 [0,1]。

5.1 社团内基于协作过滤的个性化信息推荐实验

本实验数据来源于基于 Web2.0 的某高校图书馆知识服务平台网站的“个性书签”系统数据。以选修《数据结构》课程的学生登录该平台网站, 82 个学生共收藏了 1053 篇 Web 文档, 为了避免数据稀疏性等问题, 通过数据预处理, 选择收藏大于 15 篇的 37 位学生、共收藏的 734 篇文档作为实验数据。

根据网络书签系统模型, 建立网络邻接矩阵 $G_{37 \times 37}$, 节点之间权重的最大值为 7, 即双方收藏的 Web 文档有 7 篇是相同的, 大部分关系权重在 0~4。利用

网络分析软件 UCINET,可以得到包含全部 37 个节点的 Web 用户关系网络,即 n1---n37 个节点,利用分析互相重叠的社团结构的派系过滤算法(CPM)软件 CFinder 对 Web 用户关系网络结构划分,如图 4 所示,当 k-派系社团 k=4,可以找到 3 个社团(C1,C2,C3),共 18 个节点;当 k=3,可以找到 4 个社团(C4,C5,C6,C7),共涉及 31 个节点,随着 k 值的增大,社团会越来越小,其结构越来越紧凑。



(a)4-派系社团结构图 (b) 3-派系社团结构图
图 4 社团结构关系网络划分

以图 4 实验中的 7 个社团为对象进行测试,可以看出 7 个社团内基于协作过滤的个性化信息推荐的 MAE 变化情况,如图 5 所示:

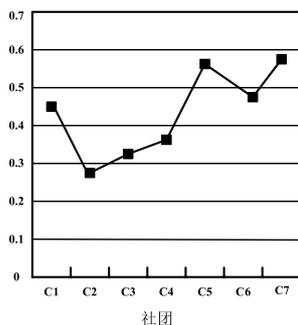


图 5 不同社团内的个性化信息推荐的 MAE 比较

从图中可以看出,首先,对于节点较多的 C2、C3、C4 个性化信息推荐质量较高,且趋于稳定。其次,4-派系社团结构(C1、C2、C3)的个性化推荐质量总体上高于 3-派系社团(C4、C5、C6、C7)的个性化信息推荐质量。最后,C1 是 C6 的子集,C2 和 C3 也是 C4 的子集,而 C1 的 MAE 低于 C6 的 MAE,C2 和 C3 的 MAE 也低于 C6 的 MAE,这表明,大社团划分为小社团,小社团内部结构更加紧密,个性化信息推荐质量更高。

5.2 社团间基于重叠节点的个性化信息推荐实验

在社团间基于重叠节点的个性化信息推荐实验中,4-派系社团结构中 C1 和 C2 的重叠节点 n7、C2

和 C3 的重叠节点 n5,以及 3-派系社团 C4 和 C5 的重叠节点 n19、C4 和 C6 的重叠节点 n7 和 n9、C4 和 C7 的重叠节点 n12、C6 和 C7 的重叠节点 n27 进行测试。如图 6 所示。

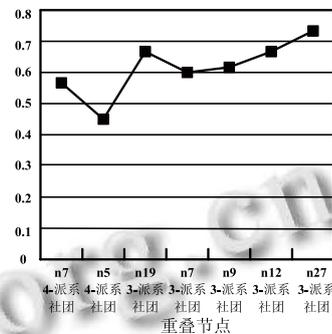


图 6 不同社团间个性化信息推荐的 MAE 比较

从图中可以看出,n7 和 n9 都是 3-派系社团结构中 C4 和 C6 的重叠节点,但由于 n7 与社团中其他节点有着更密切的联系,所以基于 n7 的社团间个性化信息推荐质量略高一些,这说明,社团间基于重叠节点的个性化信息推荐质量主要取决于重叠节点在社团中的结构特征。

5.3 两种推荐方法的比较实验

如图 7 所示是两种个性化信息推荐方法的 MAE 的变化:

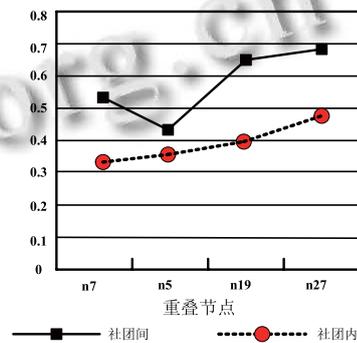


图 7 两种个性化信息推荐方法的 MAE 比较

从图中可以看出,第一,社团间个性化信息推荐的 MAE 高于社团内的个性化信息推荐的 sMAE,说明社团内个性化信息推荐质量较高,第二,C2 内基于 n7、C3 内基于 n5、C4 内基于 n19、C6 内基于 n27 的个性化信息推荐质量均低于图 5 中对应的社团内基于协作过滤的个性化信息推荐质量。

(下转第 173 页)

中存储的光子进行了辐照度预计算, 当光子数目较小时, 本文算法无论在速度上, 还是在渲染的结果上, 相对辐照度缓存算法更具优势, 但随着光子数目的增加, 辐照度缓存算法在速度上更具优势。

4.2 总结

本文利用光子映射和辐照度缓存的各自优势, 通过对全局光子图中存储的光子的位置来生成光照缓存记录, 即实现了视点无关的辐照度缓存算法。另外, 对光子进行辐照度的预计算, 这样辐照度缓存算法在采样间接光照时, 只需根据离交点最邻近光子的辐照度计算即可, 实现了快速而准确的全局光照计算。当光子数目较小时(对于一些简单的场景光子数目已经足够), 本文的算法渲染的图片在速度和质量更优于辐照度缓存, 但随着光子数目的增加, 辐照度缓存算法在速度上更具优势。由于本文实验采用的场景比较简单, 在复杂场景下, 本文的算法在在计算速度方面的优势应会更加明显。我们接下来会针对复杂场景对现有的算法进行改进, 同时尝试在 GPU 上实现现有的算法。

参考文献

- Jensen HW. Global illumination using Photon Maps. *Rendering Techniques'96*. Eds. X. Pueyo and P.Schr der, 1996.
- 陈皓. 基于光子映射的虚拟现实真实感渲染算法研究[博士学位论文]. 合肥: 合肥工业大学, 2008.
- Ward GJ, Rubinstein FM, Clear RD. A ray tracing solution for diffuse interreflection. *Computer Graphics (Proc. SIGGRAPH '88)*, 1988, 22(4):85-92.
- Jensen HW. Realistic Image Synthesis using Photon Mapping. ISBN:1-56881-140-7. A K Peters, July 2001.
- Christensen PH. Faster photon map global illumination. *Journal of Graphics Tools*, 2000, 4(3):1-10.
- Anson O, Sundstedt V, Gutierrez D, Chalmers A. Efficient selective rendering of participating media. *ACM International Conference Proceeding Series*, 2006. 153:135-142.
- Purcell T, Donner C, Cammarano M, Jensen HW. Photon Mapping on Programmable Graphics Hardware. *Proc. of the ACM SIGGRAPH/ EUROGRAPHICS conference on Graphics hardware*. 2003:41-50.
- Günther J, Wald I, Slusallek P. Realtime caustics using distributed photon mapping. *Proc. of the Eurographics Symposium on Rendering*, 2004:111-121.
- Zhou K, Hou QM, Wang R, Guo BN. Real-Time KD-Tree Construction on Graphics Hardware. *SIGGRAPH Asia 2008*.
- Ward G, Heckbert P. Irradiance gradients. *Proc. of EGWR*, 1992:85-98.
- Krivanek J, Gautron P, Bouatouch K, Pattanaik S. Improved radiance gradient computation. *Proc. of SCCG*, 2005: 155-159.
- Gautron P, Krivanek J, Bouatouch K, Pattanaik S. Radiance cache splatting: a gpu friendly global illumination algorithm. *Proc. of EGSR*. 2005:36-39.
- Jaroszw W, Donner C, Zwicker M, Jensen HW. Radiance caching for participating media. *ACM Trans. Graph.* 27, 1, 2008:1-11.

(上接第 183 页)

6 总结

本文主要研究了基于网络书签的个性化信息推荐方法, 主要依据 Web 用户收藏的 Web 资源之间的关系建立 Web 用户关系网络, 利用派系过滤算法进行社团结构划分, 来实现社团内基于协作过滤的个性化信息推荐和社团间基于“信息桥”的个性化信息推荐, 并通过实验验证了此方法在推荐中是有效的。

参考文献

- 余力, 刘鲁. 电子商务个性化推荐研究. *计算机集成制造系统*, 2004, (10):1306-1313.
- Mobasher B, et al. Integrating Web Usage and Content Mining for More Effective Personalization. *Proc. of the EC-WEB Conference*. Springer, 2000:165-176.
- 张树人. 从社会性软件、Web2.0 到复杂适应信息系统研究[博士学位论文]. 北京: 中国人民大学, 2006.
- Palla G, et al. Uncovering the overlapping community structure of complex networks in nature and society. *Nature*, 2005, 7043:814-818.
- 曾庆辉, 邱玉辉. 一种基于协作过滤的电子图书推荐系统. *计算机科学*, 2005, (6):147-150.