doi:10.3969/j.issn.1000-436x.2013.z2.033

适用于校园网的视频推荐系统的设计与实现

丁欣，马严，吴军
（北京邮电大学 信息网络中心，北京 100876）

摘 要：针对校园网 P2P 视频分享的特点，对校园网络视频推荐 FP-CNVR(campus network video recommendation based on FP-growth)系统进行原型设计与实现。提出了基于顾客细分思想的数据预处理方法 CS-DP(data preprocessing based on customer segmentation)，并对所使用的 FP-growth 算法中 FP 树的结构做出了优化。实验表明，与传统推荐系统相比，引进了 CS-DP 方法的 FP-CNVR 系统的推荐结果类型更为丰富，推荐结果的召回率提高了一半并保持了准确率基本稳定。

关键词：推荐系统；数据清洗；用户行为；校园网
中图分类号：TP393 **文献标识码**：A **文章编号**：1000-436X(2013)Z2-0175-05

Design and implementation of a video recommendation system in campus network

DING Xin, MA Yan, WU Jun
(Institute of Network Technology, Beijing University of Posts and Telecommunications, Beijing 100876, China)

Abstract: According to the feature of video-sharing with P2P in campus network, a video recommendation system in campus network called FP-CNVR system was designed and implemented; a novel data cleaning method based on customer segmentation was proposed; the structure of FP-tree in the process of frequent pattern mining with FP-growth algorithm was improved. Experimental results show that the novel recommendation system can provide more reliable recommended results, and also the novel data cleaning method can improve the accuracy of the recommendation results. The research may have some inspiration to related subjects in campus network.

Key words: recommendation system; data cleaning; user behavior; campus network

1 引言

随着互联网的迅速发展，网络视频尤其是电影电视等精品视频在线平台也在飞速发展，成为当下人们休闲娱乐的一个重要组成部分。然而，在繁杂的网络视频数据中，如何能将用户所青睐的部分推荐给用户以得到更好的用户体验和更高的点击率，这仍是当前业界与学术界共同关注的一个研究热点。

在推荐系统中，从使用的不同推荐技术角度上一般可分为 2 类：一种是基于内容的系统(content- based system)，这类系统主要考察的是推荐项的性质；另一种是协同过滤系统(collaborative filtering system)，这类系统通过计算用户或/和项之间的相似度来推荐项。经过广泛实践获知，单凭以上技术本身的推荐结果并不足够，一些新的算法或一些新的综合性系统设计被证明在推荐系统中十分有效。

针对各高校校园网 IPv4 流量带宽压力很大而 IPv6 资源丰富亟待发展的现状，对校园网视频分享方式提出了一种 IPv6 环境下的 P2P 解决方案。在实验过程中对校园网视频分享平台中各高校的用户数据研究发现，来源相似的用户其年龄相近、居住区域相似、作息时间等特征也有很大共性，其行为偏好具有一定的相似性，由此引发是否能够针对这一特点提出一个更为便捷的视频推荐方案的思考。

收稿日期：2013-09-10

2 FP-CNVR 系统方案设计

推荐系统通常包含数据预处理、数据分析、结果智能过滤等几个关键部分[1]。其中用户行为挖掘方法中，关联规则揭示用户与项之间隐含的关联关系。其中，FP-growth 算法[2]不需多次遍历数据库也不会产生大量候选集，是当前已发表的挖掘频繁项集算法中最有效的算法之一。然而 FP-growth 算法在挖掘频繁模式时，需要递归生成大量的条件 FP 树，同时当所挖掘的数据库很大时，占用内存问题严重，执行效率也不够高。由此，本文主要在数据预处理中引入一种基于顾客细分的 CS-DP 方法，即先根据具有相似性来源的用户进行细分，形成更小规模的、数据更密集的用户组。在用户行为挖掘时，根据新的模型对所使用的算法进行相应的改进。根据需求，FP-CNVR 系统原型设计框架如图 1 所示。

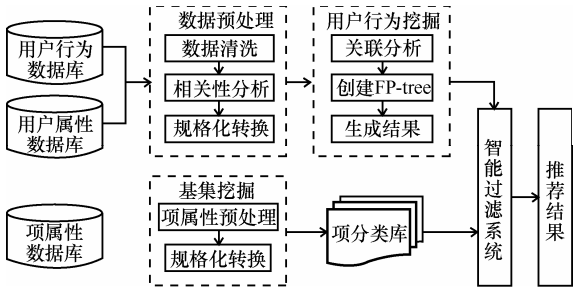

图 1 FP-CNVR 系统设计框架

由图 1 可知，数据预处理、用户行为/项挖掘和智能过滤是 FP-CNVR 系统中重要的 3 个部分。

1) 数据预处理：通过一段时间的信息采集和研究发现，原始数据存在噪声，比如一些空缺值、重复数据和失效数据，这对用户行为分析将造成不容忽视的影响。通过预处理的几个步骤可以令挖掘结果更为可靠。

2) 用户行为/项挖掘：视频推荐系统中用户行为主要记录用户浏览视频信息记录、用户的观看记录和用户的对视频的评分记录。由于校园网 BT 模式下用户的操作特点，系统主要考察用户的视频浏览行为。

3) 智能过滤：输入用户行为挖掘结果和当前热门视频 TOP10，由于推荐结果离线生成，过滤系统无需实时性满足。所以根据用户行为挖掘结果和用户/项属性，替换或部分替换默认推荐信息即可。

3 FP-CNVR 系统的实现

3.1 CS-DP 数据预处理

用户行为数据的获取往往通过编写脚本或者采集 Web 日志文件等方式抓取保存的。将用户行为的原始数据整理到事务数据库中，本文设计表单格式如表 1 所示。

表 1 用户浏览/观看视频行为数据表单

用户 ID	视频 ID	访问时间	来访 IP
12	297	1 366 594 952	2001:DB8::/32

针对校园网用户相似性的特点，数据预处理分为 3 个步骤。

1) 数据清洗：去除或减少原始数据噪声，去掉或补全空缺值，去除重复数据（同一用户访问同一视频只保留最近一次的访问记录）和失效数据等。

2) 相关性分析：根据用户属性和用户来源将用户细分，将其浏览行为记录分组，将满足相似阈值条件的用户浏览行为分为一组，按时间顺序列出。

3) 规格化转换：根据下一步骤对用户行为挖掘的需要，将以用户浏览行为表单存储的数据转化为“用户—浏览记录”的形式，如表 2 所示。浏览记录为该用户（组）浏览过的视频 ID 集合，按时间顺序排序。

表 2 用户视频浏览记录规格化示例

用户 ID	浏览记录
12	1 10 14 25 100 111 127 237 242 255 258 269 272 273 274 275 276 277 278 282 283 284 285 286 287 288 289 291 293 294 295 297 300 302 304 309 310 311
13	181 258 260 268 271 288 302 303 317 319 320 321 322 325 326 329 333 336 338 339 340 342 344 346 347 352

3.2 用户行为挖掘

用户的视频浏览行为具有大量且存在冗余等特点，而关联规则具有单一而高效的特点，可以很好地用于寻找用户视频浏览行为中有趣的规则。经研究，FP-CNVR 系统的用户浏览行为采用 FP-growth 挖掘算法[3,4]。此算法只需扫描两次事务数据库，不使用候选集，通过算法生成的频繁模式树生成关联规则结果。

3.2.1 频繁模式挖掘问题

假设 $V=\{v_1,v_2,\cdots,v_m\}$是一个包含 m 个视频的项目集，v_i用视频 ID 表示，是唯一的。D 表示用户视频浏览记录的事务集，事务的每一条目 T 对应一个用户

观看的视频集。每条事务由一个 TID(transaction identifier)唯一标识。对于一个项目集 X，如果 $X \subseteq V$ 且 $k=|X|$，则 X 被称作 V 一个 k 项集。X 的支持度是事务集 D 中包含 X 的事务计数与事务集 D 全部个数的比值，即 $sup(X)=sup\_count(X)/|D|$。本文确定一个合理的最小支持度阈值 *min_sup*，当项集 X 支持度大于或等于确定的 *min_sup* 时，则 X 是频繁的。当前问题是已知事务集 D 的情况下，如何确定一个合适的阈值 *min_sup*，使得到频繁模式（即视频的关联规则）既满足推荐系统的时效要求，又符合推荐效果指标要求的范围，最终将结果再通过智能过滤后推送给用户。

3.2.2　FP-tree 的构造算法

FP-tree（frequent pattern tree）结构将事务集 D 中的频繁项信息存储在树的枝干中，将挖掘频繁项集的问题转化成 FP-tree 的挖掘。FP-tree 结构每个节点中至少包含：节点所代表的视频 ID、频度以及指向下一个节点的指针。算法实现时，在节点中添加父节点指针可使算法更加高效。下面详细说明 FP-tree 的构造过程。

1)扫描一次事务数据库 D，计数视频的频繁项，按其支持度降序排列，生成频繁项表单 L。

2)二次扫描事务数据库 D，对事务做如下处理：①按表单 L 中的次序排列每个事务中的视频 ID，同时删除支持度小于设定阈值 *min_sup* 的项；②设置根节点为 NULL，将事务递归插入到树中，同时同步计数。

表 3 为事务数据库 D，最小支持度阈值设定为 0.2，图 2 表示根据 FP-tree 算法构造的 FP-tree。

表 3　　一个事务数据库例子

用户 ID	视频浏览记录
1	v_5　v_{13}　v_{14}
2	v_1　v_3　v_7　v_9
3	v_4　v_8　v_{11}
4	v_2　v_4　v_{10}
5	v_4　v_5
6	v_1　v_3　v_5　v_9
7	v_1　v_3　v_5　v_6　v_9
8	v_1　v_5　v_7
9	v_1　v_3　v_5　v_9
10	v_3　v_5　v_7

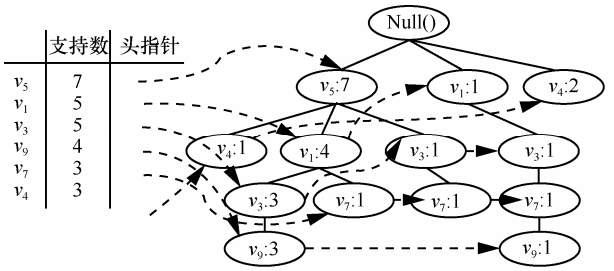

图 2　基于事务数据库 D 的 FP-tree 示例

3.2.3　挖掘方法—FP-growth 算法

算法从 FP-tree 中挖掘频繁模式，输入为基于 FP-tree 构造算法所生成的 FP 树，输出为完整的频繁模式集。算法从长度为 1 的频繁模式开始，构造其条件模式基，然后构造其条件 FP 树，递归地在该树上进行挖掘。

具体实现过程如下：

Procedure FP-growth(*Tree*,*a*)

{

1)如果 *Tree* 包含单路径 P，那么对于路径 P 中节点的每个组合(记作 b)，产生支持度 *sup* 为 b 中节点的最小支持度频繁项目集 $b \cup a$；

2) 如果 *Tree* 不包含单路径 P，那么对于每个在 *Tree* 头部的 v_i 做如下处理：

① 产生支持度 *sup* 为 v_i 的支持度的模式 $b=v_i \cup a$；

② 构造 b 的条件模式基，然后构造 b 的条件 FP 树 $Tree_b$；

③ 如果 $Tree_b$ 不为空，则递归调用 FP-growth ($Tree_b$,b)；

}

则对于满足最小支持度要求的事务数据库 D 中的每一项，可以得到如表 4 中的结果。

表 4　　基于事务数据库 *D* 的频繁模式

项	条件模式基	条件 FP-tree
v_5	null	null
v_1	{(v_5:4)}	{(v_5:4)}|v_1
v_3	{(v_5:3,v_1:3),(v_5:1),(v_1:1)}	{(v_5:4,v_1:3)}|v_3
v_9	{(v_5:3,v_1:3,v_3:3),(v_1:1,v_3:1,v_7:1)}	{(v_5:3,v_1:3,v_3:3)}|v_9
v_7	{(v_5:1,v_1:1),(v_5:1,v_3:1),(v_1:1,v_3:1)}	{(v_5:2)}|v_7
v_4	{(v_5:1)}	null

3.3　智能过滤系统

由于 FP-CNVR 系统基于用户进行推荐，而不再针对每一影片单独个性化地推荐同系列结果，同时为了保证推荐信息的实时显示，系统每天对用户

的推荐进行一次更新（仅查询更新当天登录用户的推荐结果）。对于每一个用户，智能过滤系统具体实施步骤如下：

如果该用户为新用户（无视频浏览记录），则默认推荐近期热门影片 TOP10 给用户；

如果该用户有视频浏览记录，则获取用户行为挖掘结果中该用户记录所包含的频繁模式，并按频繁支持度排序，随后去除重复 ID 和记录中已存在的浏览记录等冗余部分。取前 10 个推荐结果将默认的热门影片推荐替换。若不满足情况时，则部分替换。

Web 端通过前端代码获取当前用户 ID，在数据库进行查询，得到数据库推荐结果表单对应项后，从 10 个结果中随机显示 5 个推荐结果。

4 实验及结果分析

实验环境：视频分享平台及 FP-CNVR 系统搭载的服务器环境为 Ubuntu10.10、Apache2.X、PHP5.2.5

实验结果：读取数据库中当前用户的全部视频浏览记录，进行 CS-DP 预处理后，设定频繁模式挖掘的阈值为 0.2 进行挖掘，最后通过智能过滤得到的结果存在数据库 pi_recommend_data 表单中，推荐结果表单示例如表 5 所示。

表 5 推荐结果数据库表单示例

ID	结果
1	55 318 317 325 315 270 289 294 327 335
6	2 3 14 162 166 168 169 205 209 210
7	2 3 151 161 167 174 175 177 178 179
8	2 3 151 160 161 168 173 177 178 179
123	55 318 317 325 315 270 289 294 327 335
133	2 3 12 151 158 160 173 174 176 179
137	144 160 164 167 168 174 175 177 178 179
141	2 151 160 164 167 168 174 175 177 178

与传统推荐系统仅推荐同类型热门影片的结果相比，从推荐列表 ID 即可直观地反映出 FP-CNVR 系统的一个特点，就是丰富了推荐结果的影片类型。

由于 FP-CNVR 系统的改进部分具有明确的二分喜好特点，因而适合分类准确度指标来衡量[5]。目前最常用的分类准确度指标有准确率(precision)、召回率(recall)和 F_1 指标。假设 M 为测试用户数，N 为系统推荐视频总数目，u_n 为系统推荐且用户喜欢的视频数目，则系统整体的准确率 $P(N)$可表示为

$$P(N)=\frac{1}{M}\sum\frac{u_n}{N} \tag{1}$$

设 u_N 表示用户所喜欢的所有视频数目，则召回率 $R(N)$表示为

$$R(N)=\frac{1}{M}\sum\frac{u_n}{u_N} \tag{2}$$

而常用更直观的综合指标可表示为

$$F_1(N)=\frac{2P(N)R(N)}{P(N)+R(N)} \tag{3}$$

用 $F_1(N)$来反映随系统推荐数目 N 变化而变化的系统表现。经实验，三项指标随变量 N 的变化走向如图 3 所示。

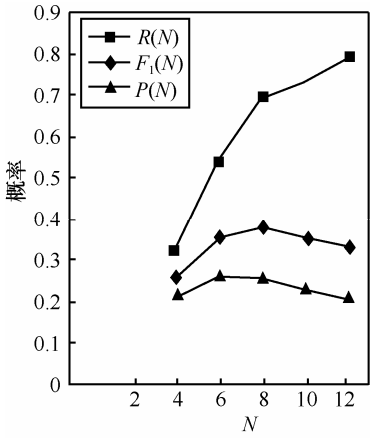

图 3 三项指标随变量 N 的变化趋势

由图 3 可知，当 N 在 6~10 范围内时，综合指标较佳，而 N 在 8~10 时，又有较优的召回率，综合系统推荐的多样性，在 FP-CNVR 系统中取 N=10。

不使用 CS-DP 预处理方法的 $P(N)\approx0.212$，$R(N)\approx0.391$；使用改进后的系统实验得到：$P(N)\approx0.229$，$R(N)\approx0.727$。与不使用 CS-DP 预处理方法相比，召回率 $R(N)$提高了一半以上并保证了准确率 $P(N)$的稳定。

对于此种现象，一方面是由于用户浏览数据存在很大的噪声及大量稀疏数据带来的干扰，另一方面准确率 $P(N)$的稳定初步推断与视频分享机制有很大关系。校园网用户视频分享的端与端多来自同学校或邻近用户（内部分享），用户又往往更偏爱有共享资源、质量有保证的资源，因而校园网用户的视频分享具有局域相似性的特点。

5　结束语

本文通过对校园网视频分享平台用户数据的研究，通过用户来源属性进行用户细分，根据用户视频浏览记录实现 FP-CNVR 系统，通过实验发现，本文推荐系统可以获得较好的推荐效果。在 CS-DP 预处理分类和用户行为挖掘过程中，采用不同阈值其处理效果不同，还需要在实践中继续检验。另一方面，推荐系统还可以将项属性、用户分享/评分记录进行深度挖掘以优化推荐结果。

参考文献：

[1] SHI M Y, MA H M, PENG Y B. Personalized Internet advertising recommended system based on RSS[J]. Journal of WUT(Information & Management Engineering), 2009, 31(4):569-572.

[2] HAN J W, PEI J, YIN Y W. Mining frequent patterns without candidate generation[A]. Proc of 2000 ACM-SIGMOD Int'l Conf on Management of Data[C]. Dallas, 2000. 1-12.

[3] PENG H L, SHU Y X. A new FP-tree-based algorithm MMFI for mining the maximal frequent itemsets[A]. Computer Science and Automation Engineering(CSAE), 2012 IEEE International Conference[C]. Zhangjiajie,2012.61-65.

[4] FAN M, LI C. Mining frequent patterns in an FP-tree without conditional FP-tree generation[J]. Journal of Computer Research and Development, 2003, 40(8):1216-1222.

[5] ZHU Y X, LÜ L Y. Evaluation metrics for recommender systems[J]. Journal of University of Electronic Science and Technology of China, 2012,41(2):163-173.

作者简介：

丁欣（1988-），女，黑龙江齐齐哈尔人，北京邮电大学硕士生，主要研究方向为计算机网络、数据挖掘等。

马严（1955-），男，北京人，北京邮电大学博士生导师，主要研究方向为下一代互联网关键技术。

吴军（1978-），男，江西乐平人，北京邮电大学讲师，主要研究方向为下一代网络、P2P 网络等。