一种基于兴趣度知识的推荐系统框架

郭 浩

(南京大学 社会学院,江苏 南京 210093)

摘 要：随着 Web Mining 技术的应用,基于 Web Mining 技术的推荐系统得到了迅速发展。本文就此系统作了一些改进,并提出了工作框架 RESIK。

关键词：推荐系统;Web Mining;兴趣度知识

中图分类号：TP393 **文献标识码**：A

A framework about recommendation system based on interest knowledge

GUO Hao

(School of Social and Behavioral Sciences, Nanjing University, Nanjing 210093, China)

Abstract: In these years, recommendation system has got a new development when Web mining technologies is used. This paper developed this system and build a new framework called RESIK.

Key words: recommendation system; Web mining; interest knowledge

随着网络应用的不断普及,越来越多的公司将注意力从传统商务转向了电子商务,这在方便了用户浏览和购买产品的同时,也带来了如何让用户尽快地从上百万件产品中找到所需产品的难题。为了解决这个问题,提出了推荐系统技术。

推荐系统被电子商务站点用来向用户提供信息以帮助用户选择产品,它根据统计结果或者用户以前的浏览和购买记录来预测用户未来的行为,向用户推荐产品。由于基于传统技术的推荐系统有很多缺陷[1-4],所以能够克服这些缺陷的基于 Web Mining 的推荐系统近来得到了迅速发展,其主要的工作流程如图 1 所示。

一般而言,推荐系统由两部分构成:离线部分和在线部分。离线部分对数据进行处理,生成相应的模型;在线部分应用离线部分的处理结果,根据用户的当前会话,向用户推荐个性化的信息。推荐系统所提供信息的个性化程度分为三类:

(1)非个性化信息,在同一个点上站点提供给所有用户的信息都是相同的(一般是由管理员或其他人编辑好,然后提供给用户)。

(2)浅度个性化信息,站点根据浏览路径和浏览行为的不同向用户提供不同的信息。

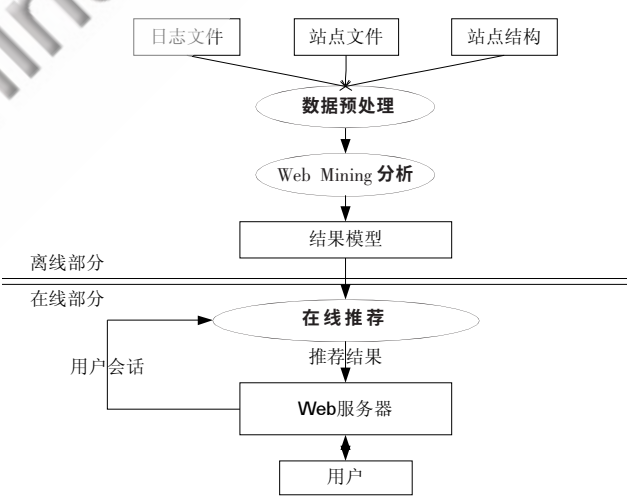

图 1 基于 Web Mining 的推荐系统的工作流程

(3)深度个性化信息,即使不同用户具有相同的浏览路径和浏览行为,站点也会根据历史兴趣的不同向他们提供不同的信息。

推荐系统一般提供的是浅度个性化和深度个性化信息。

基于 Web Mining 的推荐系统也有其自身的缺陷，本文就此系统作了一些改进，并提出了工作框架 RESIK (Recommendation System based on Interest Knowledge)。

1 RESIK 框架的提出

基于 Web Mining 的推荐系统的缺陷主要表现在[5]：

(1)不正确的推荐。对于推荐给用户的页面，有可能是用户不感兴趣的信息，下次推荐时就不应该再向该用户推荐相关内容的页面。而推荐系统主要是根据用户会话进行推荐，如果下次该用户以同样的浏览顺序访问网站时，则推荐系统将会把用户不感兴趣的信息再次推荐给用户。

(2)新加入的信息。对于新加入的页面，由于没有任何浏览记录与之相关，所以在线推荐时，无法将其推荐给用户。更有甚者，对于一个网页来说，如果经常得到推荐，则其浏览次数也会增加，下次该网页得到推荐的机会也将增加，这显然是不合理的。

本文基于以上的缺陷，提出了一个推荐系统的工作框架 RESIK。

RESIK 框架与基于 Web Mining 的推荐系统一样，也分为离线和在线两部分，所不同的是，RESIK 在线推荐时，不仅使用离线部分的处理结果，而且还根据要推荐的信息对该用户的兴趣度进行判断，只有当兴趣度超过设定的阈值，才认为要推荐的信息对该用户是有用的。

RESIK 的工作流程如图 2 所示。

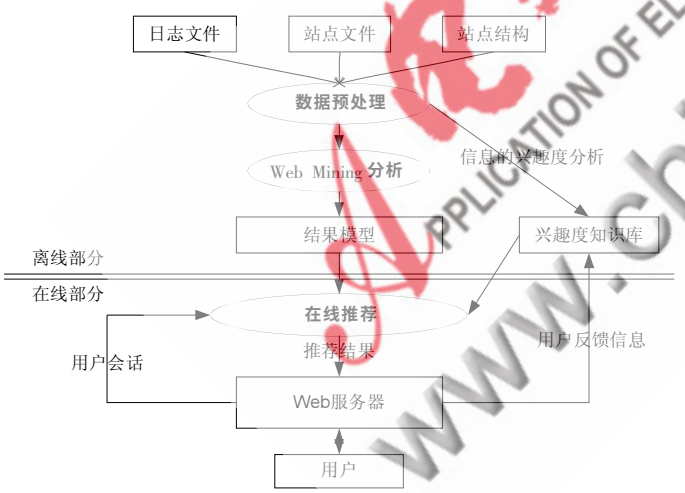

图 2 RESIK 的工作流程

图中，兴趣度知识库存储的是经过兴趣度分析得到的网页与用户的相关兴趣度，在线推荐时，不仅将离线所生成的结果模型推荐给用户，还要将与该用户相关兴趣度高的新加入的页面推荐给用户。因为兴趣度知识库是在离线部分生成的，这样在线推荐时只需要增加很小的开销就能解决新加入信息的缺陷。

对于多次将用户不感兴趣的同一信息推荐给用户的缺陷，也可以通过兴趣度知识库来解决。在线推荐时，根据要推荐的页面对兴趣度知识库进行查找，只有该页面对用户的相关兴趣度超过设定的阈值时，才将其推荐给用户。

另外，兴趣度知识库也可以由管理员向其中人工添加规则。例如，将某些重要信息设置为对所有用户都有很高的兴趣度，这样在用户访问网站时，都会得到该信息的推荐。

2 RESIK 的详细处理过程[2,4-5]

2.1 数据收集与预处理

RESIK 工作所需要的数据主要有三类：日志文件、站点文件和站点结构。日志文件存储了用户访问站点的信息，包括浏览路径、浏览时间等；站点文件包括页面，用户注册信息等；站点结构即拓扑结构，包含了页面的链入链出信息。

在进行具体的挖掘和分析之前，需要对采集的数据进行预处理，以将它们转换成符合挖掘所需要的高质量数据。这些预处理包括内容预处理和使用预处理。

内容预处理为站点文件建立挖掘所需要的特征表示，根据 TFIDF 对文件抽取关键词并建立 VSM 模型，即对关键词集合

$D=\{d_1,d_2,\cdots,d_n\}$

$V(p)=\{<d_1,w(p,d_1)>,<d_2,w(p,d_2)>,\cdots,<d_l,w(p,d_l)>\}$

其中 $w(p,d_i)=tf(p,d_i)\times\log\{N/n_i\}$，$tf(p,d_i)$ 是 d_i 在 p 中出现的频率，N 是所有的文档数，n_i 是内容出现了 d_i 的文档数。

使用预处理的任务是将采集的用户访问信息加工成可靠的事务文件，包括以下步骤：

(1)数据净化：Web 访问日志内存储的大部分信息对大多数挖掘而言，都是没有利用价值的，所以必须对日志进行净化处理。

(2)用户识别：对于已经注册的用户，这一步很简单；对于没有注册的用户，将日志文件按 IP 分割，每个 IP 对应 1 个用户群，对同一个 IP 的用户群，根据请求 Agent 的不同进一步将请求切分到单个用户。最终得到每个用户的访问记录。

(3)会话识别：对用户识别得出的单个用户的访问记录，以相邻访问发生的时间间隔是否大于 30 min 来进行会话识别。如果大于 30 min，就可以认为该用户在两个访问的中间点又开始了一个新的会话。最后得到各个会话的访问记录。

(4)帧页面识别：站点常常使用由多个页面组合而成的帧页面。在用户行为里，帧页面是一个整体，而在日志文件中，帧页面却被分解成了多个组合页面，这种不一致往往会对挖掘结果产生消极的影响。所以需要在会话识别的基础上处理日志记录中的组合页面，进行帧页面识别：顺次检查会话的访问记录，如果有请求网页内容含有“Frame”的标签，则以此网页组合为初始点使用

帧页面识别算法开始一个系列帧页面的识别过程,否则认为请求网页独自构成了1个帧页面。

(5)路径补缺:路径补缺的任务是处理缓存导致的请求缺失。

(6)事务识别:挖掘技术处理的粒度是用户的一个行为,所以要进行事务识别。事务识别得到用户的访问事务集。

数据预处理可以改进数据的质量,从而有助于提高其后的挖掘过程的精度和性能,因此在离线处理部分占有很大比重。

2.2 Web Mining分析

Web Mining所采用的分析技术主要有由数据挖掘技术演化而来的关联规则、聚类技术和序列模式以及一些统计学知识,其处理的对象为预处理之后的文档和事务集合,生成结果为可用于在线推荐的结果模型,模型的表示与所采用的分析技术有关。

2.3 信息的兴趣度分析

兴趣度分析以站点用户的注册信息为依据,对站点文件进行分析。首先对注册用户进行访问日志的分析,对其建立UP(User Profiles):

$$\boldsymbol{UP}=\{<d_1,w\ (UP,d_1)>,<d_2,w\ (UP,d_2)>,\cdots,<d_l,w\ (UP,d_l)>\}$$

式中,d_i为关键词集合中的元素,$w(UP,d_i)$为d_i关于某个用户的权重。

然后利用内容预处理的结果对每一个站点文件计算到各个用户的距离,并以此作为用户的一种兴趣度度量,称为软兴趣度知识。另外,兴趣度分析也接受来自Web服务器的用户反馈信息,根据用户对推荐系统所推荐页面的反应动作来做为用户的另外一种兴趣度度量,称为硬兴趣度知识。

2.4 在线推荐

推荐系统在线推荐时,使用Web Mining分析和信息兴趣度分析的结果得到推荐页面,具体推荐过程如下:

(1)使用推荐系统的一般方法从Web Mining分析的结果中得到要推荐的页面。

(2)将要推荐的页面依次和信息兴趣度分析的结果进行比较。如果和硬兴趣度知识发生冲突,则该页面绝对不能推荐给用户,如果和软兴趣度知识发生冲突,则由管理员预先制定的规则来处理。

(3)将软兴趣度知识中有较高兴趣度的页面也加入到要推荐的页面集合中,得到最终的推荐结果。

本文的下一步工作将在如下几个方面展开:

(1)将此工作框架应用到实践当中,以检验其效率和准确度。

(2)对于度量用户对站点文件的兴趣度,希望能够找到其他更准确合理的度量算法。

(3)希望找到将Web Mining分析和信息兴趣度分析的结果综合在一起的更好的方法。

本文简要介绍了基于Web Mining技术的推荐系统及其工作流程,并指出其缺陷,同时提出了工作框架RESIK来处理这些缺陷,详细描述了RESIK的工作流程,最后提出了下一步的工作方向。

随着Web的飞速发展,推荐系统在站点和用户之间扮演着越来越重要的角色。相信随着技术的发展,推荐系统也将得到越来越广泛的应用,更好地为Web应用服务。

参考文献

[1] SCHAFER J B, KONSTAN J A, RIEDL J. E-commerce recommendation applications [M]. Data Mining and Knowledge Discovery, 2001.

[2] ADOMAVICIUS G, TUZHILIN A. Recommendation technologies: survey of current methods and possible extensions [R]. Working paper, Stern School of Business, New York University, New York. 2003.

[3] NAKAGAWA M, MOBASHER B. Impact of site characteristics on recommendation models based on association rules and sequential patterns[C]. IJCAI'03. 2003.

[4] MOBASHER B. WebPersonalizer: a server-side recommendation system based on Web usage Mining [R]. Technical Report #01-004, DePaul University, School of CTI, 2000.

[5] LI J, ZAIANE O R. Combining usage, content, and structure data to improve Web site recommendation[C]. EC-Web 2004,2004:305-315.

(收稿日期:2009-05-08)