基于 Web Mining 的推荐系统

唐 哲 丁二玉 骆 斌 陈世福

(南京大学计算机软件新技术国家重点实验室 计算机科学与技术系 南京 210093)

摘 要 推荐系统(Recommender System)被电子商务站点用来向顾客提供信息以帮助顾客选择产品,其基本思想是以统计结果或者顾客以前的行为记录为依据,推测顾客未来可能的行为并给出相应的推荐。本文对基于传统技术和 Web mining 技术的推荐系统进行了简要综述,同时描述了基于 Web mining 技术的推荐系统的工作流程,重点分析了应用于推荐系统的各种具体 Web mining 技术及其算法比较。

关键词 推荐系统,Web mining

Recommender System Based on Web Mining

TANG Zhe DING Er-Yu LUO Bin CHEN Shi-Fu

(State Key Laboratory for Novel Software Technology, Nanjing University, Nanjing 210093)

(Department of Computer Science and Technology, Nanjing University, Nanjing 210093)

Abstract The recommender system is used by E-commerce sites to supply information to the customer in order to help them decide their choices. It's primary idea is to predict the potential behavior of the customer and give related recommendations, based on statistical data and behavior logs of the customer in the last. This paper introduces briefly the recommender system based on traditional technologies and Web mining technologies. It describes the processes of the recommender system based on Web mining technologies, analyzes the Web mining technologies used in recommender systems and the comparison between their algorithms. This paper is a survey.

Keywords Recommender system, Web mining

1 引言

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,这在方便了顾客浏览和购买产品的同时,也带来了如何让顾客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顾客提供信息以帮助顾客选择产品,它根据统计结果或者顾客以前的浏览和购买记录来预测顾客未来的行为,向顾客推荐产品。推荐系统主要分为两类,一类是手工的推荐系统,另一类是自动的推荐系统[1]。本文主要讨论后一种推荐系统。

本文将在第 2 节介绍基于传统技术的推荐系统,在第 3 节介绍基于 Web Mining 技术的推荐系统,第 4,5 节详细介绍应用于推荐系统的 Web Mining 技术及其相互比较。最后是小结。

2 基于传统技术的推荐系统

推荐系统最早可以追溯到近似系统、信息检索和预测理论,在 20 世纪 90 年代中期发展成为一门独立的研究领域[2]。

在传统技术的领域中,推荐系统可以被描述为如下的形式:设 C 为用户集合,S 为项目集合,令 u: $C\times S\rightarrow R$ 为度量用户 $c(\in C)$ 对项目 $s(\in S)$ 有用程度的函数,其中 R 为有确定范围正数,对于每个用户 $c(\in C)$,选出对其最有用的项目,即:

$$\forall c\in C, s_c=\arg\max u(c,s) \quad (s\in S)$$

但是,u 的定义域并不是整个 $C\times S$ 空间,而只是其中的一部分。这意味着需要对 u 进行扩展,这个扩展的过程就是推荐的过程[2]。目前一些主要的传统推荐技术有:

(1)基于内容的过滤(content-based filtering):根据某个用户以前选择过的项目来进行推荐,所推荐的项目与用户以前选择过的项目有很高的相似度。

(2)协同过滤(collaborative filtering):根据其他用户以前选择过的项目来进行推荐,其中其他用户与该用户有着相似兴趣和爱好。

(3)基于规则的过滤(rule-based filtering):以决策树作为技术基础,要求用户回答一系列的问题,然后进行决策树推导,最后将结果提供给用户。

传统的推荐技术有其无法克服的缺陷[1~4]:

(1)巨量数据:各种传统技术大多是实时处理的技术,在面对大规模数据时往往会力不从心。

(2)稀疏数据:对于推荐系统来说,用户所选择的项目只占项目集合中很小的一部分,这就造成了异常稀疏的数据,传统技术对于这类数据的处理往往很困难。

(3)精度有限:运用传统技术的先决条件是高质量的数据,而传统技术对于原始数据没有预处理或只有简单的处理,因此一些错误的数据会影响推荐结果的精度。

(4)需要交互:传统技术需要与用户进行交互,这客观上给用户增加了负担,同时也会让一些潜在用户离开该站点。

3 基于 Web Mining 技术的推荐系统

近年来,Web Mining 技术的出现和发展为传统技术缺陷的解决提供了有效的手段。Web Mining 技术被广泛应用于推荐系统之中。

(1)Web Mining 技术是数据挖掘技术在 Web 上的发展,所以能够很好地处理巨量数据。

(2)Web Mining 技术能够通过调整阈值来控制其行为捕

唐 哲 硕士研究生,研究方向为数据挖掘,Web 挖掘。**丁二玉** 硕士研究生,研究方向为数据挖掘、Web 个性化。**骆 斌** 教授,博士,研究方向为数据库,人工智能。**陈世福** 博士生导师,研究方向为人工智能。

捉粒度，所以能够解决稀疏数据的问题。

(3)Web Mining 技术通过数据预处理来消除错误数据的影响，而且可以依靠 Web 内容挖掘技术为新信息提供足够可靠的参照数据。

(4)Web Mining 技术不需要以用户的输入作为先决条件。

所以，近来学术界越来越关注 Web Mining 技术在推荐系统中的应用，其中以 Web 使用挖掘技术为主，除此以外还有 Web 内容挖掘技术、Web 结构挖掘技术以及它们之间的结合。

基于 Web Mining 的推荐系统工作流程如图 1 所示。

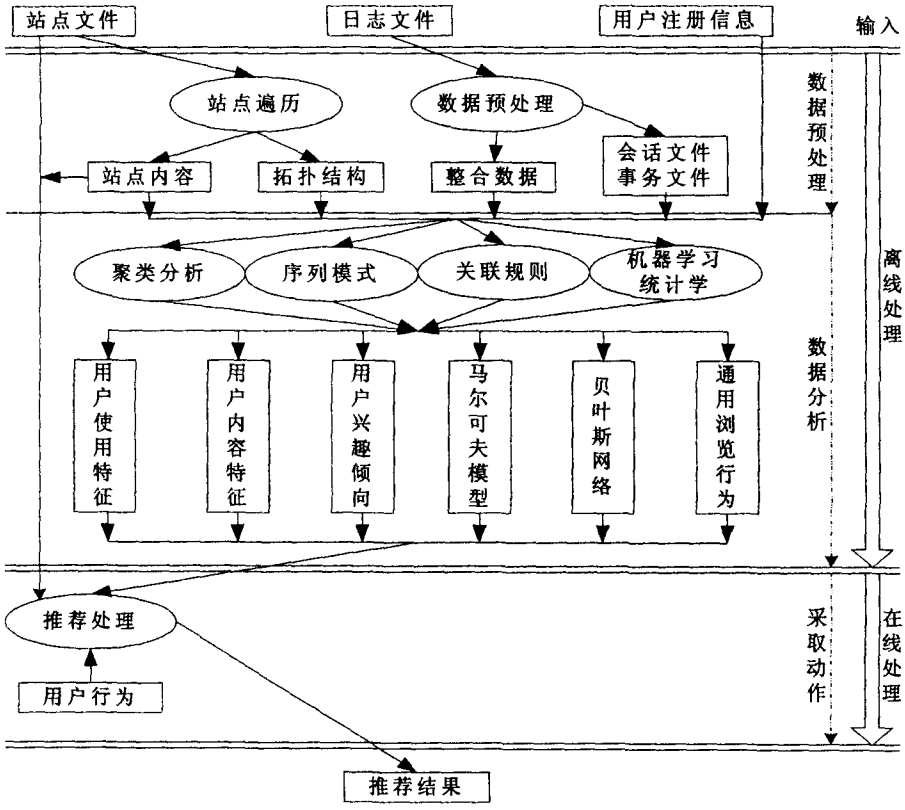

图 1　基于 Web Mining 的推荐系统工作流程

整个工作流程分成两个基本部分：离线部分和在线部分。离线部分较为复杂，所以周期性地在后台离线执行。它处理基本的站点数据：对其进行挖掘，获得需要的知识，建立相应的模型。在线部分相对简单，它利用了离线部分的处理结果，可以很快地在线执行，不会影响到用户的浏览速度。

基于 Web Mining 的推荐系统工作所需要的数据输入包括日志文件、站点文件和用户注册信息。日志文件含有站点的使用信息，进行处理后生成会话文件、事务文件或者某种挖掘技术所需要的数据格式。站点文件被遍历后，生成站点的拓扑结构和文件内容，它们可以单独提供给 Web 内容挖掘和 Web 结构挖掘，也可以配合使用数据提供给 Web 使用挖掘。对于用户注册信息，因为隐私或用户填写配合度的问题，所以很多推荐系统不考虑用户注册信息。

基于 Web Mining 技术的推荐系统较为常用的算法有：聚类分析、序列模式、关联规则和相关的统计学技术等。它们被用来分析提供的数据，然后根据挖掘得到的知识，建立相应的模式结构和模型。现在广泛使用的模式结构和模型主要有：

(1)使用向量机制描述的用户使用特征和用户内容特征，它们也是被应用较多的一种形式。

(2)使用一些特殊技术建立的用户兴趣倾向模型，比如表示用户各个兴趣倾向有无的二元向量模型。

(3)借鉴概率和统计方法建立的贝叶斯网络或马尔可夫模型。

(4)类似知识库的通用浏览行为和通用浏览规则积累模式，例如一些网页关联规则，几条频繁访问路径等。

4　应用于推荐系统的 Web Mining 技术

4.1　Web Mining 技术简介

Web Mining 技术是一个综合了数据库、信息检索、人工智能、机器学习和自然语言处理的新兴边缘学科，它使用数据挖掘技术从 Web 文档和服务中自动地发现并抽取信息。一般工作过程为信息收集，预处理，概化，分析四个阶段[5]。

按照处理的 Web 对象的不同，Web Mining 技术可以分为三类[5,6]：

Web 内容挖掘(Web content mining)：从 Web 内容中发现有用信息的过程。其处理的 Web 对象包括很广的范围，从非结构化数据，如文档；到半结构化的数据，如 HTML；到结构化的数据，如表或数据库的数据。Web 的内容绝大多数是半结构或非结构的。

Web 结构挖掘(Web structure mining)：从 Web 站点的结构中发现隐含模型的过程。其处理的 Web 对象为 Web 页面所表现出来的结构信息。

Web 使用挖掘(Web usage mining)：从用户会话和行为中提取有意义的数据的过程。其处理的 Web 对象为站点与用户之间的交互信息，如日志(包括 Server，Proxy，Client)，user profiles，注册信息，会话，事务，cookies 等。

4.2　推荐系统中使用的 Web Mining 技术

在基于 Web Mining 技术的推荐系统中，以上三类技术都有应用，其中 Web 使用挖掘所处理的对象为结构化的数

据，可以比较容易地使用数据挖掘的常用算法来处理数据，而且使用数据隐含了用户的各种个性化信息，与推荐系统的目的相一致，因此是基于 Web Mining 技术的推荐系统最常用的方法[4,7]。由于 Web 使用挖掘发现的知识仅具有历史经验上的一致性，具有语义缺陷，使得 Web 使用挖掘无法处理新加入的信息，且无法揭示已有经验的语义联系，而 Web 内容挖掘可以弥补 Web 使用挖掘的语义缺陷，因此 Web 内容挖掘在推荐系统中也起到了越来越重要的作用。此外，Web 结构挖掘在推荐系统中也有少量应用[8]。

在算法上，基于 Web Mining 技术的推荐系统使用关联规则、聚类分析、序列模式、分类和统计学等算法，其中以关联规则、聚类分析和序列模式为主。具体算法内容如下：

1. 关联规则(Association rule) 其挖掘能够发现一些潜在项目之间的联系，是 Web 挖掘最重要的技术之一，也是推荐系统应用中最常用的技术。关联规则的工作过程如下：

(1)数据预处理：对于项目集合 $P=\{p_1, p_2, \cdots, p_n\}$，经数据预处理得到事务集合 $T=\{t_1, t_2, \cdots, t_n\}$，其中 $t_i \subseteq P$，对于每个 $t \subseteq P$，有 $t=\{(p_1, w(p_1,t)), (p_2, w(p_2,t)), \cdots, (p_m, w(p_m,t))\}$，其中 $w(p_i, t)$ 为 p_i 在事务 t 中的权重[3,4]。

(2)规则挖掘：对事务集合 T，应用相关算法得到频繁项目集 I，其中 $I=\{I_1, I_2, \cdots, I_k\}$，$I_i \subseteq P$。对 I 挖掘关联规则集 $R=\{r_1, r_2, \cdots, r_s\}$，其中 r_i 形如 $X \rightarrow Y(\sigma_r, \alpha_r)$，$\sigma_r$ 为支持度，α_r 为置信度。

(3)推荐过程：对用户会话 $US=\{P_i, \cdots, P_j\}$，寻找关联规则 $r: X \rightarrow Y(\sigma_r, \alpha_r)$，$X \subseteq US$；计算各个规则中页面的推荐指数 $score(US,p)=\Sigma(\sigma_r \times \alpha_r)$，其中 $p \in Y$，然后根据设定的阈值 ρ 得到最终的推荐项目 $REC(US)=\{p \mid score(US,p) \geqslant \rho\}$[3]。

文[3]对 T 采用 Apriori 算法以发现 I，然后将 I 组织成频繁项目集图 G(Frequent itemset Graph)，在发生用户访问时，根据会话 US 对 G 进行深度优先搜索，就可以实现网页间关联规则和推荐网页集的实时发现；文[9]则将规则发现的焦点缩小集中到和具体用户或网页相关的数据上，应用以类 Apriori 算法为内循环的两层嵌套循环算法 ASARM 进行频繁项目集和关联规则的发现，同时实现了规则过滤阈值的动态调整，在提供持久个性化信息的同时解决了因具体用户和网页之间规则支持度和置信度不同而导致的规则不能有效发现的问题。

2. 聚类分析(Clustering Analysis) 能够从潜在的数据中发现共同的行为特征，是 Web Mining 技术中最成功的一种。聚类分析的工作过程如下：

(1)数据预处理：对于项目集合 $P=\{p_1, p_2, \cdots, p_n\}$，经数据预处理得到事务集合 $T=\{t_1, t_2, \cdots, t_n\}$。

(2)聚类分析：给定事务集合 T，应用相关的算法得到聚类集合 $TC=\{c_1, c_2, \cdots, c_k\}$，其中 $c_i \subseteq T$。对于每一个 $c_i \in TC$，计算其中的项目(如 URL，页面等)聚类 PR_C：

$$PR_C=\{(item, weight(item, PR_C)) \mid weight(item, PR_C) \geqslant \mu\}$$

其中 $weight(item, PR_C)$ 为 $item$ 在聚类 c 中的权重，μ 为阈值。

(3)推荐过程：对于用户会话 $US=\{S_1, S_2, \cdots, S_n\}$，计算其与聚类 c 的相似度 $match(US,c)=\frac{\sum(weight(item,PR_c) \times weight(item,S))}{\sqrt{\sum weight(item,PR_c)^2 \times \sum weight(item,S)^2}}$。然后对于每个 $item$，计算其推荐率 $rec(US,c,item)=\sqrt{weight(item,PR_c) \times match(US,c)}$ 最后得到推荐项目为 $REC(US)=\{item \mid rec(US,c,item) \geqslant \rho\}$[4]。

按照聚类的粒度(item)不同，主要分为两类：(1)群体聚类：以用户群体为处理粒度，分析和改进都从用户群体属性出发，目的在于发现群体的特征属性，算法的特点是粒度大，计算量小，效率高，可以实时处理，但是粒度过大，很难提供深度的推荐信息。(2)用户聚类：以单个用户为处理粒度，不但计算量较小，而且没有丢失用户本身的个性化信息，但是这种方式往往要面对一些焦点问题，很难实现理论上的效果。

文[4]使用群体聚类的方法，应用标准 K 阶聚类算法进行事务聚类得到 T，并定义权重计算方法为 $weight(item, PRC)=\frac{1}{|c|}\sum_{t \in c} w(item, t)$，同时将站点的拓扑结构考虑进来，用 $rec(US,c,item) \times ldf(US,item)$ 表示推荐率，其中 $ldf(US,item)=\log(dist(US,item))+1$，实现了 *WebPersonalizer* 系统；文[10]使用用户聚类的方法，对日志进行处理得到 access patterns，再根据日志的不同属性进行聚类，最后通过计算会话和聚类的相似度来进行推荐，实现了 system L-R 系统。

3. 序列模式(Sequential Pattern) 能够发现用户经常访问的页面顺序。序列模式的工作过程如下：

(1)数据预处理：对于项目集合 $P=\{p_1, p_2, \cdots, p_n\}$，经数据预处理得到事务集合 $T=\{t_1, t_2, \cdots, t_n\}$，其中 t_i 有序。

(2)规则挖掘：给定事务集合 T，应用相关算法得到频繁序列集 S，$S=\{(S_1, w(S_1)), (S_2, w(S_2)), \cdots, (S_k, w(S_k))\}$，$S_i$ 为 t_i 的子序列，$w(S_i)$ 为权重。其中事务 $t=\{p_1, p_2, \cdots, p_m\}$ 中的子序列 $s=\{s_1, s_2, \cdots, s_n\}$ $(n \leqslant m)$ 是指存在 n 个正数 $a_1, a_2, \cdots, a_n$ $(1 \leqslant a_1 < a_2 < \cdots < a_n \leqslant m)$，且 $s_i=p_{a_i}$ $(1 \leqslant i \leqslant n)$。

(3)推荐过程：对用户会话 $US=\{P_i, \cdots, P_j\}$，寻找频繁序列集 $RS=\{s \mid US$ 为 s 的子序列 $\wedge s \in S\}$，并计算各个页面的推荐指数 $score(US,p)=\sum w(s)$，其中 $s \in RS \wedge p \in s$，然后根据设定的阈值 ρ 得到最终的推荐页面 $REC(S)=\{p \mid score(US,p) \geqslant \rho\}$[3]。

文[3]对 T 采用类 Apriori 算法以发现 S，然后将 S 组织成频繁序列集 FST(Frequent Sequence Trie)，在发生用户访问时，根据会话 US 对 FST 进行深度优先搜索，就可以实现网页间序列和推荐网页集的实时发现；文[11]应用在 Apriori 基础上改良的算法 WM_0 进行 S 的发现，WM_0 将 S 以有序频繁项目集的形式表现出来，这种形式保证了 WM_0 预测时能够以非自然连续的系列进行父子序列匹配，能够在更广泛的意义上建立基于用户浏览模式的推荐系统。

4. 其他算法 除了上面几种主要的算法外，有些推荐系统也采用了其他 Web Mining 相关算法，如分类分析[10,12]和概率及统计学[13,14]等。

单纯应用以上几种基本算法对单一的 Web 对象进行挖掘会有一些难以克服的缺陷，因此有些推荐系统采取将几种算法相结合或者对多种 Web 对象进行挖掘的方式，如文[15]对 Web 内容数据和使用数据进行处理，在领域知识基础上定义实体词典 $F=\{f_1, f_2, \cdots, f_k\}$ 作为网页的内容特征描述 $p=\{(f_1, w(f_1,p)), (f_2, w(f_2,p)), \cdots, (f_k, w(f_k,p))\}$，将事务集 T 处理成内容特征增强型事务集 $TF=\{t_1, t_2, \cdots, t_m\}$，其中 $t=\{(f_1, w(f_1,t)), (f_2, w(f_2,t)), \cdots, (f_k, w(f_k,t))\}$，然后对 TF 应用各种挖掘技术，就可以发现具有丰富语义的知识；文[16]对 Web 内容数据、使用数据和结构数据进行处理，将日志中的会话依照页面内容处理成 mission 而不是 transaction，然后应用标准聚类方法进行聚类，再利用站点的拓扑结构生成增强型 Navigational Patterns，最后用于在线

推荐。

5 Web Mining主要算法比较

在对推荐系统的评价中，主要有两个评价标准：(1)准确率(Precision)，指在推荐系统做出的所有预测中预测准确次数所占的比率；(2)覆盖率(Coverage)，指在所有的用户访问或用户请求当中，系统能够给出预测结果的次数所占的比率。

$$\text{Precision}=\frac{\text{预测准确的次数}}{\text{作出预测的次数}}$$

$$\text{Coverage}=\frac{\text{作出预测的次数}}{\text{用户请求的次数}}$$

在以上几种主要算法中，序列模式往往能够达到较高的准确率，聚类分析的准确率较低，关联规则一般介于两者之间。而它们在覆盖率上的表现却是完全相反，聚类分析较高，序列模式较低，关联规则介于两者之间。从理论上分析，这个现象主要是基于以下原因：聚类分析仅仅要求信息之间具有某种联系，而关联规则还要考虑在一定空间(事务或会话)内发生联系，序列模式则除了空间限制外，还要求信息之间遵循一定的时间限制(即有先后顺序)。总的来说对同样的信息，在所考察的信息量上：序列模式 > 关联规则 > 聚类分析。所以序列模式的准确率最高，因为在更多的信息量上做出的预测往往是更准确的[17]。同时因为能够满足高信息量要求的信息相对会较少，所以序列模式的覆盖率也是最低的。其他二者同理。

6 发展与展望

现有的基于Web Mining技术的推荐系统过于依赖使用数据，具有一定的局限性[18]，而新技术的发展为现有技术的拓展提供了方向。

语义知识的引入给基于Web Mining技术的推荐系统带来了新的发展方向。因为现有的基于Web Mining技术的推荐系统在推荐具有很多属性的复杂对象时力不从心，而且也无法揭示用户模型的内在含义，而语义知识的引入能够较好地解决这些问题。基于语义的Web Mining技术是将domain knowledge融合进Web Mining过程中，包括domain ontology的获得、知识库的建立和知识增强型模式的发现[19]。文[20]曾提出一种基于本体的文本聚类方法COSA，基本思想是用一个简单的核心本体来限制相关文档集的特征以自动产生好的集成；文[21]把超链接的文字内容和超链接的环境相结合进行挖掘得到更为精确的结果。将语义知识融合进Web Mining过程中是今后基于Web Mining技术的推荐系统的发展方向之一。

基于Web Mining技术的推荐系统的另一个发展方向是多特征模型。网页信息具有很多的特征，例如用户浏览时间、链入链出结构、内容类别等，在建立推荐系统时，就是利用这些特征和访问日志建立会话和用户特征文件。但是大多数的系统往往只考虑了信息的一个或两个特征，这就遗漏了其它特征可能有的贡献，使得系统比较依赖于具体应用。文[22]比较早地关注到了这个问题，提出在进行序列模式分析时，要将路径所具有的多种特征都考虑在内；文[23]也同时考虑了网页的多个特征，在每个特征上都为网页建立一个向量模型P，然后根据一定的页面比重算法s匹配日志建立用户特征的多个模型$UP=\sum sP$，最后根据定义的各个模型权重计算用户特征的相似度，并进一步实现网页的分析和预测：$d(UP_i, UP_j)=\sum w\times cos(UP_i, UP_j)$；文[24]提出了具有开放性的特征包容模型FM，并通过相似度衡量方法PPED进行动态聚类和推荐。多特征模型也是基于Web Mining技术的推荐系统的一个发展方向。

结束语 本文介绍了推荐系统的概念及目前所采用的主要技术，其中详细叙述了基于Web Mining的推荐系统技术，描述了这些技术的工作过程，并对其技术特征进行了对比。最后介绍了目前基于Web Mining技术的推荐系统的发展方向。

随着电子商务的飞速发展，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，基于Web Mining的推荐系统也将得到越来越广泛的应用，更好地为电子商务应用服务。

参考文献

1 Schafer J B, Konstan J A, Riedl J. E-Commerce Recommendation Applications. Data Mining and Knowledge Discovery, 2001
2 Adomavicius G, Tuzhilin A. Recommendation Technologies: Survey of Current Methods and Possible Extensions: [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. IJCAI'03, 2003
4 Mobasher B. WebPersonalizer: A Server-Side Recommendation System Based on Web Usage Mining: [Technical Report # 01-004]. DePaul University, School of CTI, 2000
5 Kosala R, Blockeel H. Web Mining Research: A Survey. ACM SIGKDD, 2000
6 Michele F, Facca, Lanzi P L. Mining interesting knowledge from Weblogs: a survey. Data & Knowledge Engineering, 2004
7 Srivastava J, Cooley R, Deshpande M, et al. Web Usage Mining: Discovery and Applications of Usage Patterns from Web Data: [PhD thesis]. Dept. of Computer Science, University of Minnesota, 2000
8 Ramakrishnan N. PIPE: Web Personalization by Partial Evaluation. IEEE Internet Computing, 2000
9 Lin Weiyang, Alvarez S A, Ruiz C. Efficient Adaptive-Support Association Rule Mining for Recommender Systems. Data Mining and Knowledge Discovery, January 2002
10 Ishikawa H, Ohta M, Yokoyama S, et al. On The Effectiveness of Web Usage Mining for Page Recommendation and Restructuring. In: 2nd Annual International Workshop of the Working Group, 2002
11 Nanopoulos A, Katsaros D, Manolopoulos Y. Effective Prediction of Web-user Accesses: A Data Mining Approach. WebKDD' 01, 2001
12 Baglioni M, Ferrara U, Romei A, et al. Preprocessing and Mining Web Log Data for Web Personalization. AI * IA 2003 (Advances in Artificial Intelligence), 2003
13 Deshpande M, Karypis G. Selective Markov Models for Predicting Web-Page Accesses. In: 1st SIAM Data Mining Conf. 2001
14 Anderson C R, Domingos P, Weld D S. Relational Markov Models and their Application to Adaptive Web Navigation. In: Proc. of the Eighth ACM SIGKDD, 2002
15 Mobasher B, Dai H. A Road map to More Effective Web Personalization: Integrating Domain Knowledge with Web Usage Mining. IC'03, 2003
16 Li Jia, Zaiane O R. Combining Usage, Content, and Structure Data to Improve Web Site Recommendation. EC-Web, 2004. 305~315
17 Mobasher B, Dai H, Luo T, Nakagawa M. Using Sequential and Non-Sequential Patterns in Predictive Web Usage Mining Tasks. ICDM'2002, 2002
18 Zhou Yanzan, Jin Xin, Mobasher B. A Recommendation Model Based on Latent Principal Factors in Web Navigation Data. In: WWW 2004 Conf. New York, 2004
19 Dai Honghua, Mobasher B. Integrating Semantic Knowledge with Web Usage Mining for Personalization. Draft Chapter in Web Mining: Applications and Techniques, Anthony Scime (ed.), 2003
20 Hotho A, Maedche A, Staab S. Ontology-based text clustering. In: Proc. of the IJCAI-2001 Workshop "Text Learning : Beyond Supervision", August, Seattle, USA, 2001
21 Chakrabarti S, Dom B, Gibson D, et al. Automatic resource compilation by analyzing hyperlink structure and associated text. In: Proc. of the 7th World-wide Web conference, 1998
22 Gaul W, Schmidt-Thieme L. Recommender systems based on navigation path features. WEBKDD' 01, San Fransisco, CA, 2001
23 Heer J, Chi G H. Separating the Swarm: Categorization methods for User Access Sessions on the Web. In: Proc. of ACM CHI 2002 Conf. on Human Factors in Computing Systems, 2002
24 Shahabi C, Banaei-Kashani F. Efficient and Anonymous Web-Usage Mining for Web Personalization. INFROMS Journal on Computing, 2003