WEB 挖掘在农业信息网站个性化服务中的应用

陈基漓,牛秦洲　(桂林工学院电子与计算机系,广西桂林 541004)

摘要　针对个性化推荐及个性化检索服务,给出了基于 WEB 挖掘的农业信息网站个性化服务系统框架,对系统中涉及的用户兴趣建模、WEB 使用挖掘、WEB 内容分类等关键技术进行了讨论。
关键词　个性化服务;用户兴趣模型;WEB 挖掘
中图分类号　S126　　**文献标识码**　A　　**文章编号**　0517 - 6611(2008)35 - 15735 - 03

Application of WEB Mining in Personalized Service of Agriculture Web Site

CHEN Ji-li et al　(Department of Electronics and Computer Science, Guilin University of Technology, Guilin, Guangxi 541004)

Abstract　For personalized recommendation and retrieval service, a personalized service system frame based on WEB mining of agriculture web site was given. Key technologies were discussed, including user's interested model, WEB usage mining, and WEB content classification.
Key words　Personalized service;User's interested model;WEB mining

随着网络技术的发展,网络已经成为人们获取信息的一个重要途径。各地各级农业部门相应建立了农业信息网站,为广大农民及农村工作者提供信息服务。农业信息网站提供了大量的信息,为农业生产、农产品供求、农业技术推广服务、农业咨询等提供了方便。如何让用户在访问网站时更准确、更快捷地获得自己需要的信息,是网站发展面临的一个问题,农业信息网站也不例外。个性化服务是农业信息网站发展的一个新方向。个性化服务是指为用户提供符合其个性特征的信息,使不同用户在访问网站时能获取不同的信息。个性化服务的关键是描述用户的个性特征及兴趣偏好。笔者介绍了个性化服务涉及的主要技术——WEB 挖掘,针对个性化检索、个性化推荐 2 种典型的服务,提出基于 WEB 内容挖掘及基于 WEB 使用记录挖掘的个性化服务,并讨论了个性化服务在农业信息网站中的具体应用,在此基础上,给出了农业信息网站个性化服务系统的结构框架,并对涉及的关键技术进行了说明。

1　WEB 挖掘与网站个性化服务

为用户提供符合个性特征的个性化服务,是提高网站服务质量的一个有效途径。个性化服务包括个性化推荐、个性化检索、建立虚拟用户社区等[1]。实现个性化服务,首先要获取用户的个性特征以及兴趣偏好。获取这些信息的途径分为静态和动态 2 种。静态是指根据用户注册时填写的一些基本信息以及对页面的反馈获得,需要用户主动参与;动态获取则是根据用户对页面的访问情况进行统计获得,不需要用户主动参与,更为客观方便,这些信息主要通过 WEB 挖掘获得。WEB 挖掘是个性化服务的主要技术之一。

1.1　WEB 挖掘　WEB 挖掘就是从 Web 文档和 Web 活动中抽取感兴趣的潜在的有用模式和隐藏信息。WEB 挖掘分为 WEB 内容挖掘、WEB 结构挖掘和 WEB 使用记录挖掘 3 类[2]。WEB 内容挖掘是指对 Web 页面内容(包括文本、图像、视频、音频等)进行分析,挖掘出用户感兴趣的或有价值的内容。WEB 结构挖掘是对构成 Web 超链接的拓扑结构进行分析,挖掘出有用的信息模式,改善网站结构。WEB 使用记录挖掘是对 Web 日志记录进行分析,发现用户访问 Web 页面的规律和模式。该文重点讨论 WEB 内容挖掘和 WEB 使用记录挖掘在农业信息网站个性化服务中的应用。

1.2　农业信息化网站个性化服务　个性化推荐是将用户最感兴趣的资源(包括新闻、商品供求、技术服务、咨询服务等)提供给用户。个性化检索是指根据用户的兴趣特点进行检索,并返回与用户需求相关的检索结果,在用户输入检索的关键词之后,将检索结果按用户的兴趣程度排序后提供给用户。农业信息网站向用户提供了大量的信息。这些信息是动态变化的。将信息针对用户的兴趣偏好进行个性化提供,会给用户提供更大的方便,避免用户花费较多的时间进行信息筛选,使用户在更短的时间内更准确地获得自己真正感兴趣的信息。

2　基于 WEB 挖掘的农业信息个性化服务系统

2.1　系统基本框架　个性化服务系统工作的主要步骤是:首先分析用户对 Web 访问的规律,根据用户兴趣,在对 WEB 内容挖掘的基础上,将符合用户兴趣的结果推荐给用户,或将用户检索的结果进行 2 次处理后,将用户最有可能关注的信息提供给用户。系统框架见图 1。

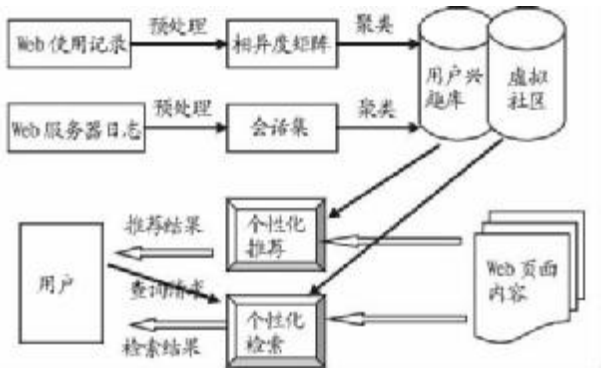

图 1　基于 WEB 挖掘的农业信息个性化服务系统框架

Fig.1　Personalized service system frame based on WEB mining

2.2　用户兴趣及虚拟社区建立　用户兴趣的描述主要依据 Web 使用记录。WEB 使用记录挖掘则是对 Web 的访问记录进行分析,从中发现用户的访问规律。WEB 使用记录挖掘通常可从 2 个方面考虑:①针对 Web 网站,对 Web 服务器的访问日志进行挖掘;②针对单个用户,对其访问 Web 资源的行为进行挖掘。

基金项目　广西区科技攻关项目(桂科攻 0428002-1)。
作者简介　陈基漓(1972 -),女,广西玉林人,硕士,副教授,从事信息检索、数据挖掘方面的研究。
收稿日期　2008-10-27

用户的 Web 访问行为与其兴趣密切相关。用户兴趣根据用户对 Web 进行访问的各种浏览行为进行挖掘而得。一些典型的操作,如用户下载、较长时间的浏览、添加至收藏夹等行为,反映出用户对相关内容感兴趣。用户兴趣采用向量空间描述,形如{(*A*1,*W*1),(*A*2,*W*2),…,(*An*,*Wn*)}。*Wi* 取值范围为[0,1],*Wi* 值越大,说明相应的兴趣度越高。Web 上的内容异常丰富,针对农业信息建立用户兴趣模型可将范围缩小至只考虑与农业相关的领域。按照农业信息分类,大类分成农机、水果、经济作物、粮食作物等若干类,每一大类下又可分为若干小类,如农机大类下分为拖拉机、农用柴油机、播种机、培土机、收割机等,水果大类下分为苹果、香蕉、柑橘等。如,某用户的兴趣向量空间为{(苹果,0.8),(小麦,0.7),(玉米,0.2)},说明该用户在苹果关键词上的兴趣值为 0.8,而在玉米关键词上兴趣值为 0.2,兴趣度较低。用户兴趣的向量空间中,每个关键词对应的权重是动态变化的。当用户进行相关内容的下载、长时间浏览等操作时,权重增加(增加至 1 时不再递增);若长时间未进行相关内容的浏览操作,则权重值减少。设定一个阈值,当某一关键词对应的权重低于该阈值时,将相应分项从向量空间中去除,同样,当某一新增关键词的权重高于该阈值时,要在向量空间中增加对应分项,使向量空间反映出用户兴趣的实际变化情况。

Web 服务器日志记录的是所有用户对该网站的访问情况,经过数据清理、用户识别、会话识别、路径补充、事务识别等一系列预处理后,对记录进行聚类处理,将具有相同或相近兴趣的用户形成一个虚拟用户社区,为同一社区的用户提供个性化服务,同社区内的用户可进行交流。

2.3 个性化推荐服务 得到单个用户的兴趣向量空间并且建立虚拟用户社区后,可实现个性化推荐服务。针对单个用户,将符合用户兴趣的新增的信息资源(包括新闻、供求信息、技术服务、专家答疑等)推荐给用户,使用户能及时获取自己感兴趣的最新资源。通常采用的推荐方法为用户登陆时以页面的形式给出推荐页面,也可将推荐内容发送到用户邮箱中。针对用户群建立的虚拟用户社区,可将相同的信息推荐给同一社区中的所有用户。

个性化推荐服务以 WEB 内容挖掘为支撑,首先对农业信息网站上以及其他网站上出现的农业信息进行分析,提取出关键词;根据关键词确定该资源所属的类别,对用户兴趣中对应类别的权重达到设定阈值的用户进行推荐。

2.4 个性化检索服务 很多农业信息网站都提供了检索服务,加入个性化功能,使检索结果更具针对性。当用户输入某个查询词后,通常会有很多满足条件的查询结果,若全部提供给用户,则用户可能只浏览前面的部分结果,而忽略了真正有用的信息。个性化检索服务的工作主要是对检索结果进行 2 次处理。对检索结果进行 2 次处理包括 2 个方面:①将页面内容按标题提取关键词后,根据关键词将内容归为某一类,然后根据用户的兴趣,将检索结果按与用户兴趣匹配程度从大到小排序后,再提供给用户;②检索的结果是不考虑时间性的,而用户可能只对最新出现的资源感兴趣,在以前的检索结果中已经出现过的资源用户可能会忽略,对检索结果按时间排序,将最新出现的资源首先提供给用户。

3 系统关键技术

3.1 WEB 使用记录数据的处理 WEB 使用记录是用户兴趣及虚拟社区建立的关键。它所包含的内容主要来源于 2 个方面,一是 Web 服务器日志记录,另一方面是用户在客户端操作的记录。前者可直接从服务器日志文件中获得,但数据量庞大,需要经过数据清理、转换、会话识别等一系列预处理过程[3];后者则必须通过对用户的浏览操作进行跟踪记录,可在网页上增加对用户下载、保存等与兴趣程度相关操作的记录,用小型代理的形式实现。

用户兴趣挖掘中关键问题是用户身份的识别。日志文件提供的是用户的 IP 地址。采用用户 IP 地址作为用户身份的标识,只适用于直接连接在 Internet 上且具有唯一 IP 地址的计算机。对于通过代理服务器访问 Internet 的用户,一个 IP 地址对应一批用户[4]。对于使用代理服务器的用户,通过识别客户机的操作系统、浏览器类型等进行用户身份识别。更好的方式是利用 IP 地址与网站的拓扑结构,根据用户的浏览路径来识别用户。

3.2 WEB 内容挖掘中页面内容的表达与分类 在个性化推荐及个性化检索服务中,首先要对待处理的资源进行分类。若考虑整个页面的内容,则虽然能得到精确的内容表达,但对正文进行处理费时太多,所以采用对标题进行关键词提取,再根据关键词进行分类的方法。分词采用分词软件完成。页面分类工作流程如图 2 所示。

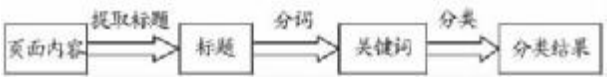

图 2 Web 页面分类工作流程

Fig. 2 Process of Web page classification

分类算法有决策树分类、贝叶斯分类、基于遗传算法的分类等。该系统采用 KNN 分类算法。分类过程中,将训练页面及测试页面经过标题提取及分词后,形成相应的矢量。KNN 分类算法包括训练和分类 2 个阶段[5],为每一类别选取若干训练页面,形成训练集合。训练过程为:对训练集合中每个页面计算 *TF-IDF* 值。对分类页面 *Y* 的分类过程如下:

(1)形成 *Y* 的测试页面向量 $Y(w_{y1},w_{y2},w_{y3},\cdots,w_{yn})$。

(2)计算该测试向量与训练集中每个页面的相似度,计算公式[6]为:

$$SC(Y,D_i)=\frac{\sum_{j=1}^{n} w_{yj}d_{ij}}{\sqrt{\sum_{j=1}^{n}(d_{ij})^2\sum_{j=1}^{n}(w_{yj})^2}} \quad (1)$$

式中,d_i 为第 i 类的中心向量。

(3)按照页面相似度,在训练集中选出与测试页面最相似的 k 个页面。

(4)在测试页面的 k 个近邻中,依次计算每类的权重,将页面 *Y* 分到权重最大的类别中。

3.3 虚拟用户社区的建立与维护 个性化服务系统中虚拟用户社区的建立采用聚类的方法,将用户聚合在不同社区中。虚拟社区的建立主要依据用户浏览行为的记录。常用的聚类算法有基于划分方法、基于层次方法、基于密度方法、基于网格方法等。该系统采用较为简单的 k 平均划分方法进行聚类处理,设聚类后簇的数目为 k,具体过程如下[2]:①

随机选取 k 个对象作为初始的 k 个簇的质心;②将其余对象根据其与各个簇质心的距离分配到最近的簇,再求新形成的簇的质心;③上述迭代过程不断重复,直至目标函数最小化。

选择 Web 服务器端日志作为数据的主要来源,经过预处理后,得到算法需要的数据。根据用户访问模式聚类,不仅考虑了注册用户,而且考虑了大量普通的网站访问者。他们的活动能较好地反映出对农业信息的需求情况。为简化算法,选择最能体现用户兴趣的网络访问活动进行分析。主要考虑以下用户行为:下载资源、浏览资源。将下载活动表示为(userid,KJ),其中 userid 为用户标识,KJ 为下载的资源号以及下载时间。以在一段时间内用户下载相同资源的相同程度为基础,构建相异度矩阵。如,有 10 个农业信息资源,编号分别为 1 ~ 10,在同一段时间内用户 A,B 下载资源的情况为:A 下载的资源集合为(1,2,6,7),B 下载的资源集合为(1,2,3),A 下载资源中与 B 相同的比例为 50%,B 与 A 相同的比例为 66%,综合考虑,得 A,B 2 个用户下载资源活动的相近程度为 58%,则相异度为 42%。经过处理后,得到用户下载情况的相异度矩阵。

用户的浏览行为与其兴趣的相关程度在很大程度上取决于浏览时间,即用户对某个页面浏览时间越长,说明该用户对页面的兴趣度越大。在以浏览行为为评价对象进行用户相似性聚类时,除了考虑用户浏览页面的相同程度之外,还应考虑浏览时间[4]。为方便处理,将浏览时间按长短分为若干个等级,如浏览时间在 1 min 之内,1 ~ 5 min,5 ~ 10 min,10 min 以上等。以用户在一段时间内访问相同页面时间长度等级的差异情况为主,构建相异度矩阵。

分别对上述 2 个相异度矩阵,采用 k 平均划分方法进行聚类,也可以将 2 个相异度矩阵合并,然后进行聚类处理。合并时可以加上一定的权重,如侧重浏览行为,则对浏览情况的相异值乘上一个较大的系数 $B(0 < B < 1)$,下载的相异值乘以 $(1 - B)$;反之,如侧重下载行为,则系数 B 取一个较小的值。

用户的兴趣是动态变化的。相应的虚拟社区应根据用户的兴趣变化而变动。若某个用户的兴趣发生变化,某类兴趣值下降至设定的阈值,则将该用户从相应的社区中删除;若用户某类兴趣值增加至设定的阈值,则将该用户加入到对应的社区中。对新的用户经过一段时间的浏览行为跟踪后,分配至合适的社区中。

4　结语

介绍了 WEB 挖掘在农业信息网站个性化服务中的应用,利用 WEB 使用记录挖掘获取用户兴趣,根据 Web 内容进行个性化推荐及检索服务,对系统中涉及的关键技术进行分析。下一步工作是在系统中增加用户反馈功能模块,进一步增强个性化服务的功能,提高农业信息网站的服务质量。

参考文献

[1] 吴丽辉. 个性化的 Web 信息采集技术研究[D]. 中国科学院计算技术研究所,2005.
[2] HAN J W,MICHELINE KAMBER. 数据挖掘:概念与技术[M]. 范明,孟小峰,等,译. 北京:机械工业出版社,2001.
[3] 冯是聪,单松巍,张志刚,等. 基于 Web 挖掘的个性化技术研究[J]. 计算机工程与设计,2004(1):4 -6.
[4] 管涛,罗建军,冯博琴. 基于用户浏览时间的模式聚类算法[J]. 计算机工程与应用,2003,39(15):104 -105,202.
[5] 高洁,吉根林. 文本分类技术研究[J]. 计算机应用研究,2004(7):28 -30.
[6] RICARDO BAEZA-YATES,BERTHIER RIBEIRO-NETO. Modern information retrieval[M]. Addison Wesley,1999.

(上接第 15728 页)

显示出的共同趋势是,市场区域得分均高,销售力量得分也较高,即这 2 个因素对组织成员增收水平贡献大,这一点符合现有合作组织多以销售服务为主的现实情况,也说明合作组织必然要通过销售获取经济效益;③ 3 种类型的另一个共同趋势是,引资联营均低于同一水平其他因素的分值,说明目前都在这方面做的不好,还是一个薄弱环节,需要加强。这一分析结果与前面的定性分析结果是一致的,即说明宁夏新型农村合作经济组织的发展主要受到企业家、资金、营销队伍和市场区域等因素的影响,今后应在这几个方面进一步努力。

表 3　3 个类型组织的得分值比较　　分

组织分类	知识经验得分	财务制度得分	引资联营得分	营销力量得分	市场区域得分	品牌投入得分
健康类(优强态和健康态)	6.666	5.666	5.000	7.333	7.333	6.666
合格类(合格态)	3.333	3.666	1.333	5.000	5.666	2.666
不健康类(病态和夭折态)	1.666	1,333	1.333	1.666	3.333	1.000

参考文献

[1] 王飒飒,刘鹏飞. 影响我国农民专业合作经济组织发展的因素综述[J]. 甘肃农业,2007(12):21 -23.
[2] 傅夏仙. 农业中介组织的制度变迁与创新[M]. 上海:上海人民出版社,2006:263.
[3] 程同顺,黄晓燕. 中国农民组织化问题研究:共识与分歧[J]. 教学与研究,2003(3):42 -47.
[4] 潘劲. 对两种类型农产品行业协会的比较研究——以霍山、温州茶叶产业协会为个案[J]. 调研世界,2004(1):22 -25.
[5] 王朝良. 特色农业发展管理评价指标与方法研究[J]. 农业科学研究,2008(1):1 -6.