一种 N层体系结构下的 Web挖掘应用

但 微 才书训
(东北大学软件学院 辽宁 沈阳 110004) (东北大学 秦皇岛分校 河北 秦皇岛 066004)

摘 要 Web应用的深入使 N层体系结构的系统逐渐被广泛使用,同时网上的海量信息也为 Web挖掘提供了一个广阔的应用领域。本文针对在 N层体系结构中应用 Web挖掘技术进行了研究;包括在 N层体系结构中实现网站用户访问分析、智能搜索引擎和个性化推荐等;对数据源的处理和转换、数据仓库的建立和使用以及业务处理过程的改进等都进行了论述。

关键词 N层体系结构 Web挖掘 用户访问分析 智能搜索引擎 个性化推荐

WEB MINING IN N-TIER ARCHITECTURE

Dan Wei[1] Cai Shuxun[2]
[1] (Software College, Northeastern University, Shenyang Liaoning 110004, China)
[2] (Northeastern University at Qinhuangdao, Qinhuangdao Hebei 066004, China)

Abstract With the development of the World Wide Web, N-Tier Architecture systems have been widely used. With the huge amount of information available online, the World Wide Web is a fertile area for Web mining. This paper researched the different characteristics of Web Mining technology using in N-Tier Architecture systems. Web Usage Analysis, Intelligent Search Engine and Personalized Recommendation will be outlined. Data source processing, Data warehouse building and using, Business rule changing will be discussed.

Keywords N-Tier architecture Web mining Web usage analysis Intelligent search engine Personalized recommendation

0 引 言

随着系统开发和应用的不断发展,其计算模式已经经过了主机模式、两层 C/S模式、B/S模式和三层 C/S模式、面向 Web的 N层模式等发展阶段,不同的体系结构对系统的功能实现非常关键。N层体系结构具有良好的性能,易于集成和扩展,逐渐成为面向 Web应用的首选结构。同时网上的海量信息也为 Web挖掘提供了一个广阔的应用领域,而且已经出现很多能够使用的算法和技术用于网站用户访问分析、智能搜索引擎和个性化推荐等使系统具有智能性。如何能让各种 Web挖掘技术适用于 N层体系结构计算模式的系统就成为目前需要解决的问题。本文将首先介绍 N层体系结构系统的特点,然后以一个典型的电子商务 N层体系结构系统为例来说明 Web挖掘功能的实现过程和各模块在 N层体系结构中的地位与作用。

1 N层体系结构系统的特点

现在常用的 N层体系结构系统一般将前端数据呈现、中间业务处理和后端数据服务划分到不同的层中进行实现[1],对于 Web应用不需要专门的客户端程序,只要有浏览器即可使用。

N层体系结构有很多优点。系统将应用程序合理地分块;浏览器和 Web服务层专门处理数据显示和用户交互;业务处理层(应用程序服务器)能够自动地协调和处理来自多个客户的请求。数据服务层负责数据的维护和更新,提供数据持久性服务,处理所有定义的数据集的细节以及与数据库的交互。N层体系结构还具有良好的性能,易于集成和扩展。在 N层体系结构下能够更好的支持用户的并发访问,增强了分布式数据处理能力。

与此同时,N层体系结构的应用也使 Web系统的规模日益扩大,用户信息以及商品信息和与业务相关的其它各种信息呈现出膨胀式的增长,对这些信息的分析和挖掘也逐渐成为 Web系统建设的重要部分。

2 Web挖掘的数据源和能够提供的功能

Web挖掘是针对包括 Web页面内容、页面之间的结构、用户访问信息等各种 Web数据源,在一定基础上应用数据挖掘的方法以发现有用的隐含的知识的过程。在电子商务系统中数据源一般包括商品信息、用户注册信息、订单信息和用户访问日志。通过对这些数据源的分析和挖掘能够帮助顾客更有效的完成商品浏览和搜索,接受个性化推荐等整个业务流程,并能够帮助商务决策人员分析商品销售和客户的信息。

应提供的具体功能如下:

(1) 能够分析找到顾客的来源信息,以利于对特定地区进行商业宣传活动。

(2) 能够进行购物篮分析,通过对购物篮和订单数据分析,

收稿日期: 2005 - 06 - 15。但微,硕士生,主研领域:电子商务及 Web数据挖掘。

找到顾客同时购买商品的组合。

(3) 能够识别用户的访问模式,并预测用户下一步的访问活动。

(4) 用户能够通过系统内提供的智能搜索引擎找到需要的商品,并且商品的列表能够按照用户感兴趣的程度进行排序。

(5) 用户访问系统时能够得到个性化的商品推荐。

3　电子商务 N层体系结构中 Web挖掘的实现

3.1　电子商务系统的基本结构

电子商务系统的基本功能是完成 3方面的业务流程:用户购买商品,与物流和财务等外部系统进行协作以及系统管理员进行系统管理。

系统管理员的操作响应和处理主要是在业务处理层完成的。与物流和财务系统等其它外部系统的交互也是业务处理层要完成的工作。后端数据是以关系数据库存储的一些表,包括用户的注册信息,用户和管理员的帐号密码,商品的信息,订单信息等。数据服务层对关系数据库中的数据进行映射和持久化实现用户并发访问和事务处理等功能。

电子商务系统的 N层基本结构可表示成图 1的形式。

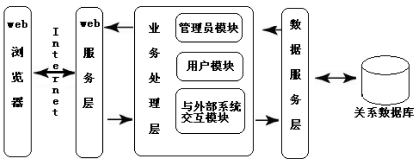

图 1　电子商务系统的 N层基本结构

3.2　建立数据仓库

数据仓库是由多个数据集市构成的支持管理决策过程的、面向主题的、集成的、稳定的不同时间的数据集合。数据仓库的建立是进行数据挖掘的一个必要的条件,必须针对海量的历史数据进行分析和挖掘才能形成一个完整的模式库。

为这个电子商务系统建立数据仓库时,数据源主要来自 Web服务器日志和关系数据库。在整个 N层体系结构中,数据仓库部分是相对独立的,周期性的将 Web服务器日志和关系数据库中的数据装入数据仓库中,然后进行分析和挖掘以形成新的模式库。根据上面第 2节涉及的 Web挖掘方面的需求,数据仓库应该完成的工作包括以下一些方面:

(1) 对购物篮和订单信息建立数据集;进行关联规则挖掘(比较常用的算法有 Apriori和 FP - tree算法等);就形成了同时购买商品模式库。商业决策人员可以根据这个模式库中的信息通过在相关页面中增加链接等方式提高销售效率。

(2) 对顾客订单和用户注册信息中的地区字段建立数据集,进行关联规则挖掘,形成顾客来源模式库。商业决策人员可以根据这个模式库中的信息在相关地区开展商业宣传活动。

(3) 从日志文件获得数据,数据一般包括用户 id,用户连接网页的时间变量,网页的标识;将用户访问记录组织成会话形式,包括会话的持续时间,在一个会话内点击的总次数和会话的开始时间;通过分析数据的众数和中值将异常点去掉;进行序列模式挖掘(常用算法有 PrifixSpan算法等),形成点击序列模式库。分析这个模式库可以进行网页拓片结构的改进。

(4) 将订单记录和用户注册信息建立数据集,对用户的注册信息使用决策树分类算法,得到该类用户对某商品是否购买的分类,形成用户信息模式库。在用户调用智能搜索引擎工作时,系统就可以根据用户信息模式库重新对得到的索引进行排序。也就是说如果跟当前用户注册信息相似的用户曾经购买了索引项中的某个商品,那这个商品的信息就会排在搜索结果序列的前面位置。

(5) 在进行个性化推荐时,系统必须在较短时间内得到推荐项,但是对用户访问的大量历史信息进行处理必然带来很大的系统开销,所以就必须首先准备好进行个性化推荐的一些模式库。协同过滤技术是应用最成功的技术,它是通过分析历史数据,生成与当前用户行为兴趣最相近的用户集,将他们最感兴趣的项作为当前用户的推荐结果,基于协同过滤技术的推荐过程可分为 3个阶段: 数据表述; 发现最近邻居;产生推荐数据集。首先要得到各用户对各项商品的评价值,如果是通过隐式方式得到,那就必须处理用户访问日志中的时间属性,用户在一个网页上停留的时间值与该网页信息量的比值可以作为该用户对这个网页的评价值,从而形成评价值模式库[2]。通过对用户评价的相似性进行聚类分析,就可以得到访问相关性模式库。这两个模式库的形成为个性化推荐引擎的工作提供了条件。

数据仓库建立后,周期性的装入日志和关系数据库中的数据,并且进行分析和挖掘,建立和不断更新模式库。这些模式库包括同时购买商品模式库、顾客来源模式库、点击序列模式库、用户注册信息模式库、评价值模式库和访问相关性模式库;商业决策人员可以使用专用工具直接查看模式库,进行相关的管理决策。智能搜索引擎和个性化推荐工作时也必须使用其中的几个模式库,下面就对业务处理层进行改进,来实现这些功能。

3.3　业务处理层的改进

数据仓库建立以后,在业务处理层就可以加入与 Web挖掘相关的各个模块。

3.3.1　智能搜索引擎　在用户使用电子商务系统时,除了按照商品目录进行浏览以外,往往是通过搜索引擎对系统内部的信息进行搜索。在这种情况下,搜索应用的范围仅仅限定在系统的内部,所以不需要象门户型搜索引擎那样面对复杂的 Internet环境,所以可以对这种内部搜索引擎进行性能优化并结合 Web挖掘技术使其具有智能性。

搜索引擎的工作主要由搜索器、索引器、检索器来完成[3]。针对电子商务系统的内部环境,由于商品信息本来就具有的目录结构,所以可以使用人工建立和维护索引库的方法[4],这种机制类似于以 Yahoo为代表的建立在分类基础上的门户型搜索引擎。

为了使搜索引擎具有智能,可以在两个方面进行改进。首先,当用户输入搜索字段时,可以使用一个同义词典对搜索条件进行扩展,找到更多的与用户兴趣相关的结果;其次,用户在使用搜索引擎时很可能是在登陆系统的最初阶段,所以还无法识别用户的点击模式,但是此时用户的注册信息是可以利用的资源,可以通过访问模式库中已经建立的用户注册信息模式库对得到的搜索结果进行排序,把该类用户最可能购买的商品排在结果列表前面位置。

3.3.2　个性化推荐　在模式库中已经存在评价值模式库和访问相关性模式库了,业务处理层工作步骤如下:

(1) 识别当前用户访问行为,得到用户对商品的评价值。这个步骤是通过读取服务器日志中对当前用户的记录完成的。

(2) 访问评价值模式库,对当前用户寻找最邻近集,最邻近

(下转第 96页)

户端实现,由流重定向实现流媒体的点组协作,不影响原有的流媒体构架。流重定向代理接收 rtsp流媒体协议的访问请求,从流数据缓冲区获取数据,发送给请求方。参与 P2P调度的机器上都存在两个服务,一个是 P2P重定向代理实现请求点功能,另一个是 P2P服务端代理实现服务点功能。当某个节点存储了一个流媒体分段数据时,启动 P2P服务端代理并加入到候选集中。当请求点向 P2P服务器请求候选集时,服务器根据请求点的点组特性进行过滤。选择服务集的算法遵循服务点最少原则,因为参与的服务点越少,说明服务点个体的性能越好,越能保证提供良好的服务质量。当服务子集由于数量少不能满足调度需要时,直接从流媒体服务器获取数据流并扩充候选集数量。获得服务点子集后,请求点与每个服务点间建立直接的对等通道,传输数据直至任务完成。请求点代理同时监控缓冲区的使用率,若发现在一段时间内缓冲区数据低于警告阈值,说明服务点集合传输速率降低,提出重新调度请求。图 5对一个包含 5个成员的点组进行应用性能分析,从监控数据可以看出,流媒体服务器在 P2P技术中 (右图,2个节点重定向)对带宽的要求变化较明显。

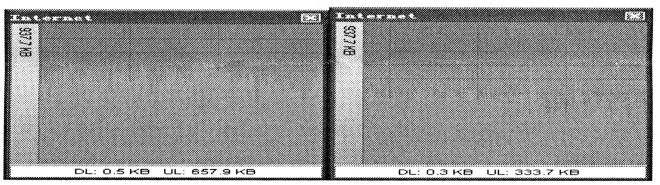

图 5 流媒体服务器带宽资源使用 (5个 realplayer结点)

5 结束语

本文将流媒体技术与 P2P技术相结合,讨论了在 P2P点组环境下的流媒体调度技术,并给出了基于 P2P多点调度的带宽分配算法,最后提出了一个 P2P点组网络下的流媒体数据调度模型,该模型适合在宽带区域如校园网内实现流媒体数据的共享。

参 考 文 献

[1] T. Nguyen and A. Zakhor. Distributed Video Streaming Over Internet [EB/OL]. http://www-video.eecs.berkeley.edu/papers/thinhq/spie2002.pdf.

[2] V. Padmanabhan, H. Wang, P. Chou, and K. Sripanidkulchai. Distributing Streming Media Content Using Cooperative Networking [EB/OL]. www.csee.umbc.edu/~pmundur/courses/CMSC691M-04/P2P2.pdf.

[3] D. Xu, M. Hefeeda, S. Hambrusch, and B. Bhargava. On Peer-to-Peer Media Streaming [EB/OL]. www.cs.purdue.edu/homes/mhefeeda/papers/icdcs02.pdf.

[4] S. Floyd, M. Handley, J. Padhye, and J. Widmer. Equation-based congestion control for unicast applications [EB/OL]. www.icir.org/tfrc/tcp-friendly.pdf.

[5] Jin B. Kwon Heon Y. Yeom. Multimedia Content Distribution over Peer-to-Peer Networks [EB/OL]. www.cse.msu.edu/icdcs/posters/final/01_s.pdf.

(上接第 91页)

集的产生是通过当前用户与评价值模式库中的所有用户的相似性计算得出的。

(3) 根据得到的最邻近集,读取访问相关性模式库中的信息,计算其对所有项的兴趣度,产生推荐项。

另外,由于系统建立初期协同推荐算法存在稀疏性[5]的问题,往往得到的推荐项非常少,此时可以根据同时购买商品模式库和点击序列模式库中的信息,将所有与协同推荐算法得到的推荐项相关的商品也推荐给当前用户。

3.4 数据服务层的改进

数据服务层的作用是为业务处理层提供统一的数据访问视图,并且对数据的访问规则进行控制。由于本系统的模式库一般是在数据服务层的后端以脱机方式建立和更新的,所以面对业务处理层只需要提供对模式库的只读访问;但是,用户的每一次点击操作基本上都会触发系统中 Web挖掘部分的响应,而且还要求系统的响应在短时间内完成;比如,当用户点击某种商品的信息时,个性化推荐模块就必须同时产生针对该用户的推荐。这样就使得模式库的数据成为系统中访问最频繁的资源,数据服务层可以通过提供缓存来提高模式库的访问效率,同时在部署数据服务器群集的时候,可以分别使用不同的机器来处理访问模式库、商品目录浏览和搜索的请求。

3.5 改进后的系统结构

给 N层体系结构的电子商务系统建立了数据仓库并对业务处理层和数据服务层进行改进以后,系统不但能够帮助用户更有效的访问系统,而且能够帮助商业决策人员分析整个系统的各种信息。系统此时的 N层结构如图 2所示。

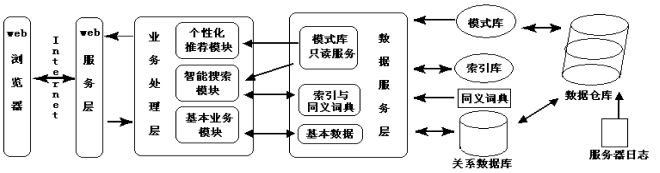

图 2 Web挖掘 N层结构图

4 小 结

N层体系结构逐渐成为面向 Web应用的首选结构,Web系统的规模也日益扩大,对信息的分析和挖掘也逐渐成为 Web系统建设的重要部分。本文以一个典型的电子商务 N层体系结构系统为例说明了 Web挖掘功能的实现过程和各模块在体系结构中的地位与作用。但是,本文提出的方法和思路还有很多地方远没有完善,这些都有待于进一步的研究。

参 考 文 献

[1] Deshpande, Y., Murugesan, S., Ginige, A., Hansen, S., chewabe, D., Gaedke, M., and White, B. A Software Architecture for Structuring Complex Web Applications. Journal of Web Engineering, 2002. Vol 1, No. 1: 003～017.

[2] 鲍玉斌、王大玲、于戈,"关联规则和聚类分析在个性化推荐中的应用",《东北大学学报(自然科学版)》,2003,24(12):1149～1152.

[3] Randolph Hock. Web search engines: features &commands. Online, 2000. 5/6: 17～25.

[4] 曾 春、邢春晓、周立柱,"个性化服务技术综述",《软件学报》,2002,13(10):1952～1961.

[5] 叶红云、陈华平、张文斌,"基于机群和多层软件体系结构的电子商务系统构造方法",《计算机应用与软件》,2003,20(10):77～79.