◎产品、研发、测试◎

# 基于 Ajax 与向量空间模型的个性化搜索引擎

李 蕾,周国民
LI Lei,ZHOU Guo-min

中国农业科学院 农业信息研究所,北京 100081
Agricultural Information Institute,The Chinese Academy of Agricultural Sciences,Beijing 100081,China
E-mail:colorful_helen@hotmail.com

**LI Lei,ZHOU Guo-min.Personalized search engine based on Ajax and VSM.Computer Engineering and Applications,2007,43(19):89-91.**

**Abstract**: There are there tough problems in building a personalized search engine system:first collecting user information,second auto-refreshing database and retrieval algorithm utilizing collected information.To solve the three problems,this paper creatively promotes user-tracing method based on Ajax,strategy of refreshing user-action database by session and personalized information retrieval model based on VSM plus user-document.Finally it describes the design and implementation of whole testing personalized system.

**Key words**: personalized search engine;Ajax;Vector Space Model(VSM);user profile

**摘 要**:针对个性化搜索的三个关键问题:用户信息搜集,用户信息库的动态更新与个性化检索算法,探索性地提出了基于 Ajax 用户行为跟踪方案,以会话为单位动态更新用户行为信息库策略与加入用户文档的向量空间检索模型,在此基础上设计并实现了个性化搜索引擎实验系统。

**关键词**:个性化搜索引擎;Ajax;向量空间模型;用户文档
**文章编号**:1002-8331(2007)19-0089-03 **文献标识码**:A **中图分类号**:TP391

## 1 引言

目前 www 上流行的搜索引擎普遍是基于关键词匹配的,缺乏对关键词语义的理解和对用户差异的考虑,对所有用户,只要输入关键词相同,返回结果就完全相同,检索效果不尽人意,在检索式“一词多义”的情况下问题更加严重。

例如:同样输入“苹果”,农业领域的用户需要的是苹果种植方面的信息,计算机专业的用户需要苹果公司方面的信息,简单关键词匹配,有可能使农业用户却得到苹果公司方面的信息,或者计算机用户却返回了苹果种植的结果。

针对此问题,本文提出了一个可以实时搜集、保存与更新用户行为信息,基于用户历史查询记录建立用户文档,从而提高用户适用性的个性化搜索引擎模型,并在现有农业专业搜索引擎“农搜”[1]的基础上开发出实验系统。

## 2 个性化搜索引擎的关键技术与算法

个性化搜索通常包括两个部分:搜集用户信息,挖掘并建立用户模型;利用用户文档实现个性化功能。

### 2.1 采用 Ajax 技术搜集用户行为信息

搜集用户信息策略分为隐式与显式两种。隐式信息搜集一般从用户访问行为中获取能够反映用户兴趣背景的数据,不需要用户参与,显式搜集需要用户参与,如填写表单。因为后者难以动态更新,容易引起用户的反感,所以本文选择了隐式搜集方式。传统的隐式搜集方式包括 Web 日志提炼、客户端程序与 cookies 等,它们都存在固有缺陷,Web 日志方式因为 IP 误解、本地缓存等问题,搜集到的信息准确度不高,而客户端程序升级和部署困难。在全面了解、分析比较后,本文借鉴了目前电子商务网站上广泛应用的 Ajax 技术,独创性地设计了针对特定网站用户行为跟踪方案,有效地避免了传统方式所存在的问题。

#### 2.1.1 Ajax 技术简介

Ajax 是 Asynchronous Javascript+XML 的缩写,这个名词的创造者 Adaptive Path 的咨询顾问 Jesse James Garrett 这样定义 Ajax[1]:Ajax 不是一种技术,实际上,它由几种蓬勃发展的技术以新的强大方式组合而成,Ajax 包含:

(1)基于 XHTML 和 CSS 标准的表示;

**基金项目**:国家科技基础条件平台建设项目(No.2005DKA31800)。
**作者简介**:李蕾(1983-),女,硕士研究生,研究方向:信息检索个性化技术;周国民(1969-),男,博士,研究员,研究方向:农业信息技术。

[1] “农搜”(http://www.sdd.net.cn)是由中国农业科学院农业信息研究所多媒体技术研究室独立开发的农业专业搜索引擎,具有基于 SDD 算法的语义检索与 lucene 全文检索两套检索引擎,详细信息参见 http://www.sdd.net.cn/intro.html。

(2)使用 Document Object Model 进行动态显示和交互;

(3)使用 XMLHttpRequest 与服务器进行异步通信;

(4)使用 JavaScript 绑定一切。

Ajax 出现之前,Web 站点强制用户进入提交/等待/重新显示的进程,用户的动作总是与服务器的“思考时间”同步。Ajax 的应用使得浏览器-服务器异步通信成为可能,用户不再需要等待服务器响应就可以继续在页面上进行其他操作,页面可根据服务器返回信息自动部分刷新。

其基本原理是:在网站页面内嵌入 Ajax 引擎,该引擎由事件处理函数、XmlHttpRequest 对象(简称 XHR,封装了发送与接收 Http 请求的方法和属性)与回调函数组成。当用户在页面上点击按钮或选择下拉菜单,将驱动 Ajax 引擎的事件处理函数,该函数创建 XHR 对象,设定请求目标 URL 与内容,将请求发送出去。得到服务器响应后,客户端回调函数从 XHR 对象获取返回信息,操纵页面 DOM 元素完成更新操作。

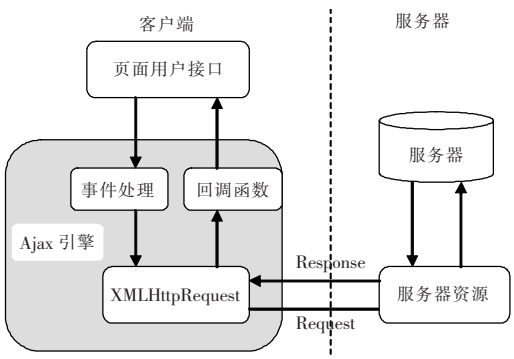


图 1 标准 Ajax 交互

### 2.1.2 Ajax 在用户行为跟踪中的应用

搜索引擎通常以 Web 应用的方式提供服务,页面的用户接口主要包括关键词提交表单和结果显示区两部分,用户行为分为检索特定关键词和查看搜索结果两类,系统搜集的目标数据,也即用以描述这两类行为的数据包括:用户 ID、动作时间、动作类型(查询/点击)和特征数据,其中特征数据是指用户输入关键词与点击结果的标题、链接,它们集中体现了用户兴趣背景,是实现个性化功能的关键。用户 ID 可以通过传统的注册登录的方式获取,关键词通过表单传送给服务器,但用户点击行为的相关数据是无法直接被服务器获取的。

本文利用 Ajax 异步通信能力简单有效地解决了这个问题,实现时只需要在页面内嵌入具有行为跟踪功能的 Ajax 引擎,在服务器端设置一个专门用于处理保存用户行为数据的脚本。数据搜集过程是一个单向 Ajax 交互过程,用户点击某条搜索结果时会触发点击事件处理函数,该函数将与此事件有关的数据(包括:用户动作类型、点击时间、点击结果标题、URL 地址)打包发送至服务器端特定目标脚本,该脚本将数据保存到用户行为信息库。通过试验证明此方案是切实可行的,在用户毫不知情的情况下搜集到准确有效的行为数据。

## 2.2 以会话为单位动态更新用户行为信息库

需要为用户行为信息库制定保存与更新策略,从可行性角度考虑,如果服务器每一次接收用户动作数据都与数据库连接、保存一次,随着用户数量增加,服务器的效率也会降低,反应速度减慢,而且随着用户信息库增大,维护的难度也越来越大,因此需要改变简单的“发送即保存”的模式,并对库的数量级进行控制。从必要性考虑,用户兴趣处于变化中,离其当前访问时间越远的行为,能够反映用户当前兴趣的程度越低,重要性越小,因此用户行为记录有必要淘汰更新,在这个问题上,本文基于 JSP/Servlet 的会话跟踪技术,提出了以会话为单位保存与更新数据库的策略。

### 2.2.1 会话跟踪与会话的定义

HTTP 是一种 Stateless 的协议,它只关心请求与响应的状态,当客户端发出请求,服务器才会建立连接,一旦客户端的请求结束,服务器便会中断与客户端的连接,服务器难以判断发出连续请求的是否为同一个用户。在 JSP 体系中,使服务器能够识别用户,并与用户保持一定时间连接的技术称为会话跟踪(Session Tracking),会话(session)是指单一客户端与服务器之间发生连续相关交互的时段,一段时间没有作用将会自动失效。JSP 提供了强大的 session API,使网站开发者可以设置会话的失效时间,可以为每一个访问用户创建一次会话,可以通过创建与会话绑定的对象在会话期限内缓存数据。

### 2.2.2 用户行为信息库保存与更新策略

(1)变“发送即保存”模式为“发送/缓存/保存”:服务器将连续接收到的用户行为数据先缓存起来,缓存期限为一次会话,会话失效后再一次性保存入库,这就大大缓解了服务器读写数据库的压力。

(2)以会话为单位维护与更新用户行为信息库:入库时每一次动作作为一条记录保存,记录字段包括:用户 ID、会话 ID、动作时间、动作类型(查询/点击)、特征数据(关键词/点击结果标题)和其他信息(点击链接地址)。其中会话 ID 是会话识别标志,属于同一会话的记录该字段取值相同,第一次会话取值为 1,以后按时序递增,每一次保存时先检查最近一次保存记录的会话 ID,如果需要控制用户记录数量在 20 次会话内,而最近一次保存的会话 ID 已达到或高于 20,即删除最早一次会话的所有记录,再保存新的会话,会话 ID 在最近一次会话的基础上加 1。由此解决了用户行为信息库维护难的问题,也实现了动态更新,为建立动态的用户模型打下基础。

## 2.3 加入用户文档的向量空间检索模型

个性化搜索功能可通过查询扩展、重构或者搜索结果过滤、重新排序实现,本文所提出的加入用户文档的向量空间检索模型[2]实质属于第一种策略,其原理是利用用户模型优化查询,使之更贴近用户语义,从而提高搜索引擎的查准率与个性化程度。

### 2.3.1 向量空间检索模型与算法

在经典的向量空间检索模型中,把索引中的每个词作为空间的一个维度,每一篇文档表示为词汇空间的一个向量,每一个查询同样以向量表示,通过计算文档和查询的内积或余弦来表示文档和查询的相关程度。本文借鉴了“农搜”现有的检索引擎[3,4],该引擎遵循向量空间检索模型,具体算法简述如下:

令 $D=\{\boldsymbol{D}_1,\boldsymbol{D}_2,\cdots,\boldsymbol{D}_n\}$ 表示由 $m$ 个词和 $n$ 个文档构成的文档集合,其中 $\boldsymbol{D}_j=(d_{1j},d_{2j},\cdots,d_{mj})^{\mathrm{T}}$ 是文档向量,$d_{ij}$ 表示词 $i$ 在文档 $j$ 中的词频权重,词-文档矩阵 $\boldsymbol{A}$ 定义如下:

$$\boldsymbol{A}=(\boldsymbol{D}_1,\boldsymbol{D}_2,\cdots,\boldsymbol{D}_n)=\begin{bmatrix} d_{11} & d_{12} & \cdots & d_{1n} \\ d_{21} & d_{22} & \cdots & d_{2n} \\ \vdots & \vdots & & \vdots \\ d_{m1} & d_{m2} & \cdots & d_{mn} \end{bmatrix}$$

其中 $d_{ij}=\dfrac{\log(f_{ij}+1)}{\sqrt{\sum_{k=1}^{k=m}(\log(f_{ij}+1))^2}}$

$\boldsymbol{Q}=(q_1,q_2,\cdots,q_m)^{\mathrm{T}}$ 表示查询向量，$\boldsymbol{q}_i$ 表示词 $i$ 在查询中出现的词频权重，同样按照 $d_{ij}$ 的公式计算。

检索过程中用户输入的关键词或文档，经过分析、分词等处理，成为一个 $k$ 维查询向量 $\boldsymbol{Q}_k$，按 $\boldsymbol{s}=\boldsymbol{Q}_k\times\boldsymbol{A}_k$ 计算相似度，据此给出搜索结果。需要补充说明的是系统在词-文档矩阵建立后利用 SDD 算法[5]对它进行了矩阵分解、降维，以强化语义关系提高空间效率，所以矩阵 $\boldsymbol{A}$ 转换成了 $\boldsymbol{A}_k$。

### 2.3.2　改进模型与算法

前面所述的检索模型按照用户输入构建查询向量，只考虑了文档与查询的相关性（也称系统相关性）而欠缺对用户相关性的考虑，因此本文试图在原有模型之上引入用户特征向量，通过查询扩展与重构建立一个兼顾查询相关性与用户相关性的个性化检索模型。

通过对所搜集到的用户个性化信息的分析发现：可以利用用户历次查询使用的关键词和查看结果标题构建用户特征向量，如果将这些具有文本特征的关键词与查询结果标题连接在一起就成了一篇虚文档，也是属于用户的个性化文档，此文档转换为基于词频权重的向量也就是用户特征向量。这看似是一个大胆的缺乏理论依据的方案，但仔细推敲还是有其可行性的：首先词频权重确实能够反映用户对某关键词的感兴趣程度；其二，在实践中已经得到多次验证，基于统计的方法往往比其他建模方法更为有效；其三，文档转换为向量的处理，包括：分词、词频统计、权重公式均可照搬原有机制，实现起来非常简单。

算法的改进只在查询向量部分，词-文档矩阵的定义与相似度计算公式没有变化。令用户输入关键词向量为 $\boldsymbol{Q}_k$，$\boldsymbol{Q}_k$ 与原算法的查询向量一致，用户特征向量经过处理后同样成为一个 $k$ 维查询向量 $\boldsymbol{U}_k$，新的查询向量 $\boldsymbol{N}_k=\alpha\cdot\boldsymbol{Q}_k+(1-\alpha)\cdot\boldsymbol{U}_k$，$0\leqslant\alpha\leqslant1$，$\alpha$ 用以调节相关度计算中查询相关性与用户相关性的比例。按 $\boldsymbol{s}=\boldsymbol{N}_k\times\boldsymbol{A}_k$ 计算相似度，据此给出搜索结果。

## 3　个性化搜索引擎设计与实现

### 3.1　系统设计

根据以上原理设计了个性化搜索引擎，该系统结构如图 2 所示，在原“农搜”引擎的基础上增加了用户行为跟踪与登录/注册两项功能，并改进了原查询处理流程。

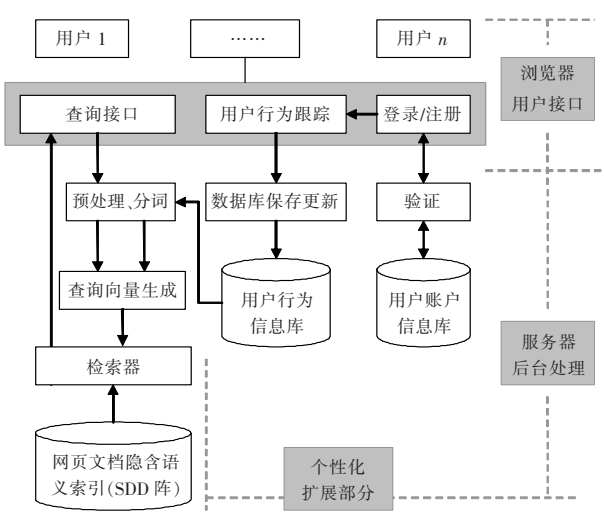


图 2　系统结构简图

系统运行时检索与数据搜集两条进程并行，相互影响，一方面用户登录验证成功后，系统启动跟踪引擎搜集并发送用户行为信息，服务器端脚本接收、缓存，在会话结束时保存至用户行为信息库，此次会话记录将影响该用户下一次会话搜索结果；另一方面用户通过查询接口提交关键词，服务器从用户行为信息库中检索出用户历史记录，提取特征数据字段连接成用户文档，对关键词和用户文档作处理、分词，生成个性化查询向量，提交给检索器，检索器计算查询文档相关度，然后返回搜索结果。

### 3.2　系统实现

本文在“农搜”现有的检索器、分词器与农业网页索引的基础上，采用 J2EE 动态网站开发技术，组合 Windows 平台、Tomcat 服务器与 MySQL 数据库实现了个性化搜索引擎的实验系统，对系统的初步测试表明，总体规划与设计是可行的，达到了预期的目标，对于同一个关键词，不同用户检索结果是不同的，即使同一个用户同一个关键词，不同时间搜索的结果也是变化的。图 3 显示了用户“李蕾”登录后输入关键词“农业”的返回结果页面，左侧为原搜索结果列表，右侧为个性化搜索结果列表，右上角显示登录后系统反馈信息。图 4 显示了此次会话结束后，查询后台用户行为信息库的结果，高亮显示为新增的记录。

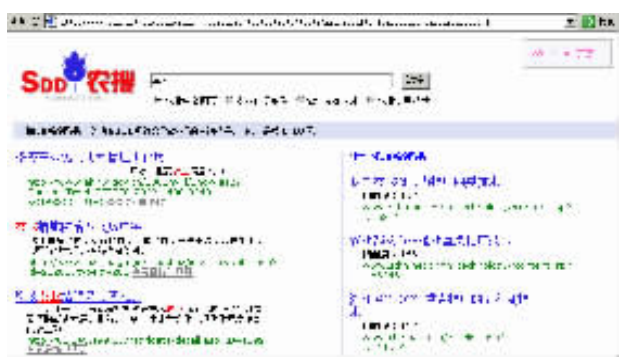

图 3　个性化实验系统搜索结果

图 4　用户行为信息库查询结果

## 4　结语

个性化搜索引擎是搜索引擎的发展趋势之一，是国际上的研究热点，本文为解决个性化搜索过程中的一些难题作了积极的探索，在深入研究与广泛涉猎的基础上，提出了解决方案，设计并实现了个性化系统，接下来的工作重点是个性化效果评测与系统完善。（收稿日期：2006 年 12 月）

（下转 114 页）