带反馈自适应 Web 搜索引擎研究*)

张卫丰[1,2,3] 徐宝文[2,3,4]

(南京邮电学院计算机科学与技术系 南京 210003)[1] (东南大学计算机科学与工程系 南京 210096)[2]

(江苏省软件质量研究所 南京 210096)[3] (武汉大学软件工程国家重点实验室 武汉 430072)[4]

摘 要 Web 搜索引擎是 Internet 上非常有用的信息检索工具,但是由于现有这些搜索引擎搜索出的结果只跟用户的搜索词条和它所采集的实际信息有关,用户对搜索结果的选择不能影响将来的搜索结果,这使得搜索引擎不能考虑大多数用户的兴趣状况。本文通过采集用户对搜索结果的访问序列来生成搜索引擎的反馈信号,以此来扩展原始查询串和影响搜索结果的生成,使得搜索引擎具有自适应能力。

关键词 WWW,搜索引擎,自适应

A Feedback-Based Self-Adaptive Web Search Engine

ZHANG Wei-Feng[1,2,3] XU Bao-Wen[2,3,4]

(Department of Computer Science and Engineering, Nanjing University of Posts and Telecommunications, Nanjing 210003)[1]

(Department of Computer Science and Engineering, Southeast University, Nanjing 210096)[2]

(Jiangsu Institute of Software Quality, Nanjing 210096)[3]

(The State Key Laboratory of Software Engineering, Wuhan University, Wuhan430072)[4]

Abstract The Web search engine is a very useful information service in the Internet. Because the current Web search engines give the search results which only relate to the search terms and the actual information collected by them andthe selection of the search results can not affect the future search results, the most people's interest can not be considered by the search engines. In this paper, the feedback signals which can extend the original query string and affect the search results, are produced by collecting the users' accessing lists, and then the self-adaptation abilities are provided by the search engines.

Keywords WWW, Search engine, Self-adaptation

1 引言

随着 Internet 和 WWW 的迅速发展,Internet 上的资源日趋丰富。基于 Internet 的各类信息检索服务应运而生并得到了迅速发展。信息搜索是互联网上最重要的网络应用之一,互联网搜索更成为"门户网站"必不可少的基础性服务,这一切都离不开搜索引擎[5]。实践证明 Web 搜索引擎是一种非常有用的信息检索工具,但是现有这些搜索对同样的检索请求所检索出的信息是相同的。其原因主要在于常用搜索引擎体系结构的固有不足。这些搜索引擎通过机器人程序不停地去访问站点,不断建立和完善关键词到 URL 的索引数据库。但这些机器人程序只是考虑不断扩充和完善索引数据库,而没有考虑到用户对搜索引擎的使用对索引数据库的影响。现有的搜索引擎如 Yahoo, Altavista, Infoseek 等为了弥补这方面的不足,一般通过个性化配置[1](利用 cookie 机制[7,8]、为用户建立配置文件等)来提高对用户搜索请求的精度和命中率[6]。这种方式一般需要用户在服务器上登记一些个人信息,但这可能造成用户某些隐私信息的泄露。那些在用户个人电脑上存放一些个人信息的方式也不利于用户隐私的保护,而且用户在使用其它机器时,其个人信息已经不存在了。还有一些搜索引擎如 Hotbot、ZDNET 等设置一些链接让用户给出一些反馈信息。这种方式被动地接受用户对搜索结果满意程度的反馈,收集的只是一部分反馈信息。

近年来,我们在对 Web 技术作了初步研究的基础上[5,7,8],又尝试将数据挖掘技术、Agent 技术和遗传算法等应用于 Web 技术的研究中,并取得了初步成果[11~15]。本文在此基础上进一步研究搜索引擎的自适应能力,给出了一个基于反馈的自适应搜索引擎。它能够主动采集众多用户对搜索结果的访问序列,挖掘这些访问序列中的隐含的用户兴趣信息,以

*)本研究得到国家自然科学基金(60373066)、国家重点基础研究发展规划 973 资助项目(2002CB312000)、国家自然科学基金青年科学基金(60303024)、江苏省自然科学基金(BK2001004)、江苏省科技攻关项目(BE2001025)、武汉大学软件工程国家重点实验室开放基金、江苏省计算机信息处理技术重点实验室开放基金(苏州大学)资助。张卫丰 博士,主要从事 Web 信息系统、人工智能、信息安全等方面的研究。

此来对用户的初始查询进行扩展和影响搜索引擎中的索引服务器[4]中记录的评价过程，从而使得搜索引擎的返回结果可以根据用户使用该搜索引擎的实际情况进行自适应调整。

本文将首先介绍搜索引擎自适应系统的基本原理，然后围绕搜索引擎自适应系统中各部件给出其在自适应搜索引擎中的实现过程，最后给出基于该思想的搜索引擎原型系统 FASE 和实验。

2 带反馈的自适应搜索引擎系统

如所知，所谓带反馈的自适应系统的基本原理是将系统的输出作为系统的反馈来影响系统以后的输出。为了解决常用搜索引擎单一的“输入-输出”响应模式，使得搜索引擎能够更好地适应用户，我们使用了带反馈的自动控制系统的基本原理对常用搜索引擎进行了改进(图 1 所示)。带反馈自适应搜索引擎系统不只接受用户的单一输入，而是把用户对搜索结果(输出)的选择情况作为反馈信号也输入到搜索引擎中，形成“输入输出-反馈”的响应模式。这样搜索引擎具有了对用户的适应能力。

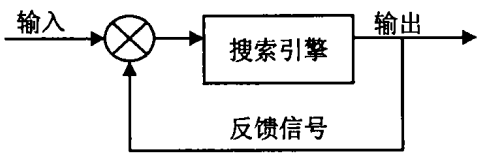

图 1 带反馈自适应搜索引擎系统原理图

在图 1 带反馈自适应搜索引擎系统原理图中，输入为用户输入的查询请求，输出为搜索引擎根据查询请求检索出的结果，反馈信号是用户对搜索结果的选择情况。这是一个具有反馈能力的自适应系统，随着系统的运行，搜索引擎不断调整自己的输出，使得自己的输出逐渐适应用户的需求。在该系统中，系统具有自适应能力主要在于反馈信号的存在。下文中将依据该系统自适应的原理，介绍反馈信号的采集与反馈信息库的生成、反馈信号与搜索结果的集成。

3 数据采集与反馈信息库的生成

3.1 数据采集

用户在访问搜索引擎的时候一般输入要查询的内容，然后提交给搜索引擎处理，搜索引擎将搜索的结果(一般是目标站点的 URL、该站点的注释等)反馈给用户，用户再根据这些信息跳转到对应的站点。搜索引擎的前端一般是 Web 服务器，Web 服务器中的日志一般只能记录用户访问的时间、用户的地址及用户在该站点所访问的 URL 等与 Web 访问有关的基本信息，而不能记录用户对搜索引擎的使用情况，因而不能根据 Web 日志中的信息来跟踪用户对搜索结果的选择情况。还有，一般的搜索引擎中固定的网页数是不多的，用户访问的主要是它所提供的主页面，然后将搜索请求发送给搜索引擎服务器上的搜索程序，得到动态的搜索结果页面。这样传统的 Web 日志记录的用户访问信息将是非常单一的，这些信息只能反映用户搜索的频数，而不能正确反映用户对搜索结果的兴趣(用户对搜索结果的选择情况表明了用户的兴趣)，因而不能像文[9]那样利用传统的 Web 日志来挖掘用户的兴趣。

常用搜索引擎的体系结构决定了搜索引擎不能从它的搜索输出中获取有用的信息。根据图 1 中自适应 Web 搜索引擎的要求，需要从搜索引擎的输出中获取反馈信息，因此，为了采集用户的兴趣信息，需要改进传统搜索引擎的体系结构以提供一种机制来跟踪用户对搜索结果的选择信息。这可以通过在搜索引擎返回的搜索结果中加入客户端脚本或者 Java 小程序实现，它使得在用户访问搜索结果中所链向的页面的同时，能够将用户的选择信息记录在搜索引擎服务器上。基于反馈的自适应搜索引擎的具体实现细节将在下文 FASE 中介绍。

为了区分用户的兴趣，还要对用户访问的事务进行有效的识别。由于用户客户端和代理服务器中缓存机制的存在，将很难区分用户的会话事务。例如在服务器记录的日志中，所有来自同一个代理服务器的请求拥有同样的标识，而实际上这些请求可能是由多个用户发出的。这可以通过 Cookie 机制来区分，但是由于用户保护隐私或者 Web 服务器有可能不具有支持 Cookie 的能力，因此利用 Cookie 机制来区分不同的用户也不是完全可行的。文[3]中给出了通过启发式算法来区分用户会话及在缺乏附加信息(如 Cookies)的情况下推导出日志中会话间关系的方法，这样可以区分出不同用户的事务。有些用户可能在对搜索结果进行选择的时候，多次选择同一个记录，这些重复选择的记录其实是多余的。在区分出用户的事务后，就可以方便地剔除一些无用的记录。在下文中，如不作特别声明，日志均指经过这样预处理后的日志。

3.2 反馈信息库的生成及其算法

在搜索引擎采集了用户的兴趣后，就可以从中提取信息作为图 1 中所示的反馈信号。这需要对记录下来的日志进行进一步的处理。为了便于讨论，下文中所述的查询串均指简单查询串，即查询串中不包含操作指示符。在描述反馈信号生成算法之前，先给出如下一些定义。

定义 1 词典 $T=\{t_1,t_2,\cdots,t_l\}$，其中 $t_1,t_2,\cdots,t_l$ 为日志中的所有词条，l 为词典的大小(日志中不同词条的个数)。

定义 2 搜索结果记录用二元组(hyperlink，

abstract)表示,简记为 *url*,其中,hyperlink 为一个指向其它页面的超链接,abstract 是有关该 hyperlink 所指向页面的简短说明。

定义 3 搜索结果记录集 $U=\{url_1,url_2,\cdots,url_m\}$,其中 $url_1,url_2,\cdots,url_m$,为日志中所有的 *url*,m 为日志中不同 *url* 的个数。

定义 4 兴趣用二元组(*item*,*url*)表示,其中 $item\in T$,$url\in U$,简记为 *iu*;兴趣的集合(兴趣集),简记为 *IU*。

定义 5 事务为用户对查询结果的一个访问序列,简记为 *tran*。

一个 *tran* 可以记录下用户在输入要查询的词条后,对搜索引擎返回结果的访问情况。如用户搜索"计算机软件",搜索引擎返回的结果按照一定的次序排序,如"URL1,URL2,URL3,URL4"返回给用户,而该用户对结果的访问顺序为"URL3,URL2,URL1",则在用户会话日志中记录下了〈SessionID,"计算机软件",(URL3,URL2,URL1)〉。为了形式化描述事务及平凡事务,先给出如下符号标记的定义:

SessionID:一个事务的唯一性标记;

Querystring:用户的查询串;

M:Querystring 中所包含的词条数目;

Q_j:Querystring 中第 j 个查询词条,$1<j<M$,$Q_j\in T$;

search(EngineID,SessionID,Querystring):为搜索引擎 EngineID、用户 SessionID 和 Querystring 的函数,它表示用户一次搜索结果记录的集合。

这样,事务 *tran*(SessionID,Querystring)可以表示如下:

tran(SessionID,Querystring)=〈SessionID,Querystring,urls(SessionID,Querystring)〉其中,urls(SessionID,Querystring)为在该事务中用户的访问序列:

urls(SessionID,Querystring)=($url(1)$,$url(2)$,…,($N_{SessionID}$))

这里,$url(1),url(2),\cdots,url(n_{SessionID})\in$search(EngineID,SessionID,Querystring),$n_{SessionID}$表示在 SessionID 所对应的事务中所记录的 URL 数目。

定义 6 平凡事务是满足一定条件的事务 *tran*(SessionID,Querystring),其中 Querystring$\in T$。平凡事务的集合(平凡事务集)记为 *Trans*。

事务 *tran*(SessionID,Querystring)中的 Querystring 可以划分为多个词条,根据 Querystring 中的词条可以分割为多个平凡事务。*tran*(sessionID,Querystring)的第 j 个平凡事务可以表示如下:

tran(sessionID,Q_j)=〈sessionID,Q_j,urls_sub(sessionID,Q_j)〉

其中 urls_sub(sessionID,Q_j)为访问序列 urls(sessionID,Querystring)中的子序列。

子序列 urls_sub(sessionID,Q_j)为从访问序列 urls(sessionID,Querystring)中筛选出的与词条 Q_j 相关的搜索结果记录序列。算法 1 为从用户事务中生成平凡事务序列的具体算法。

算法 1 平凡事务序列生成算法

```
function subserial_producing(int SessionID)
  for j=1 to M do //M 的大小由 SessioID 对应的
                  事务中的 Querystring 决定
    urls_sub(sessionID,Q_j)=(); //先置为空序
      列
    for 每一个 url∈s(sessionID,Querystring)do
      if url.abstract 中包含 Q_j then
        urls_sub(sessionID,Q_j)= urls_sub
          (sessionID,Q_j)+(url);
      end if;
    end for;
  end for;
end subserial_producing;
```

通过平凡事务序列生成算法对日志中的所有事务进行处理,就可以分割出所有的平凡事务。

通过对平凡事务中访问序列的统计就可以得到用户的兴趣及对该兴趣的对应评价值。所有的这些生成的兴趣及对应的兴趣评价值构成了反馈信息库。

对兴趣 $iu=(t,url)$,给出如下评价函数:

$$\text{value}(iu)=\sum_{i=1}^{k} number(i,t,url)\cdot rank(i) \quad (1)$$

在(1)式中,k 为 *url* 序列中所认为重要的的个数,由于用户一般只对搜索出来的前几个结果感兴趣,因此 k 一般不是很大,一般取小于 10 的值;$number(i,t,url)$为 i,t,url 的函数,它表示在平凡事务集合 *Trans* 中,*url* 被第 i 个访问的次数;$rank(i)$为一单调递减函数,它表示被第 i 个访问的权重,如它可以是 $k-i$,10^{k-i},等。

这样根据平凡事务集 *Trans*,可以计算出所有兴趣的评价。

3.3 基于粗糙集的兴趣简约

反馈信息库的大小与所采集的信息的数量密切相关,一般只是采集与保存一定时间范围内的用户反馈信息。根据定义 4 中的定义,字典 T 中的所有词条都有可能成为兴趣中一个属性,由于兴趣中另一属性,随着采集信息数量的增加,与词条组合的数量将不断增加;而且由于不同搜索引擎对同一个搜索结果在信息表达上的差异,使得对同一站点的概要介绍有所不同。基于上文中精确的用户兴趣表达

将带来如下问题:反馈信息库空间的爆炸式增长和很难将用户兴趣与反馈信息库中兴趣的精确匹配。这就需要将兴趣的概念进行一定程度上的近似。为此我们引入了粗糙集理论。

Pawlak 在 1982 年提出的粗糙集理论是标准集合理论的扩展,支持决策过程中的近似决策。它的基本思想是在论域的等价关系的基础上,通过一对近似操作算子(下近似和上近似)来刻画论域中的集合。许多需要近似计算的应用系统都可以利用这些算子。我们将在同义词关系的基础上构建兴趣粗糙算子。这样将兴趣构建在等价关系的基础上,有利于缩小兴趣的存放空间和解决搜索结果与兴趣的匹配问题。

设 R 为字典 T 上的同义词关系,一个非空的对象领域构造了一个近似空间 apr=(T,R)。对 T 的同义词划分可以记为 $T/R=\{C_1,C_2,\cdots,C_n\}$,其中 C_i 为 R 的一个等价类(即一组同义词)。对于 T 的任意子集 S,S 的下近似和 S 的上近似分别为:

$$\text{lower\_apr}(S)=\{x\in C_i \mid C_i\subseteq S\}$$

$$\text{upper\_apr}(S)=\{x\in C_i \mid C_i\cap S\neq\Phi\}$$

S 的两种近似实际上是在近似空间(T,R)中对集合 S 的一种近似描述。

有了基于同义词关系的粗糙算子,我们可以对反馈信息库进行简化和对搜索结果与兴趣进行近似匹配。下面将首先讲述反馈信息库的简化问题。

粗糙兴趣是二元组(C,*url*),其中 $C\in T/R$,$url\in U$,简记为 riu。粗糙兴趣的集合简记为 RIU。

公式(1)是基于兴趣的评价函数,通过改造公式(1)就可以得到公式(2)中的基于粗糙兴趣的评价函数。

对粗糙兴趣 $riu=(C,url)$,对应的评价函数为:

$$\text{VALUE}(\text{riu})=\sum_{\text{iu}\in C}\text{value}(\text{iu}) \qquad (2)$$

由粗糙兴趣和对应的评价值组成的反馈信号可以反馈给搜索引擎。搜索引擎可以利用该反馈信号对输出进行对应的处理,即反馈响应。

4 反馈响应过程

作为一个带反馈的自动控制系统,反馈信号用于调整输入而达到调节系统输出的目的。在基于查询反馈的自适应搜索引擎中,反馈信号存放在反馈信息库中。反馈信号的数据流动过程如图 2 所示。首先,在用户输入原始查询串后,由原始查询串和反馈信息库生成修改后的查询串;然后,将修改后的查询串传送给搜索引擎,搜索引擎将根据修改后的查询串搜索得到预搜索结果;预搜索结果与反馈信息库作用以后得到搜索结果返回给用户;用户在得到搜索结果以后进行反馈,由用户反馈和原始查询串来更新反馈信息库。

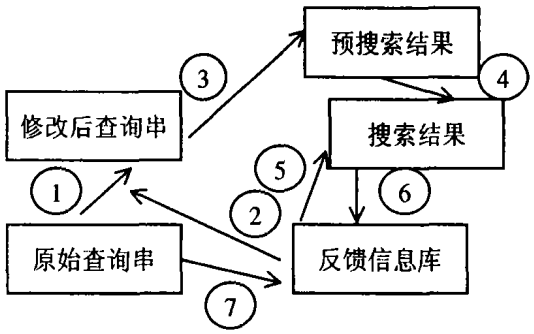

图 2 反馈信号生成及响应过程

上文中已经介绍了如何构造反馈信息库,这里将分两个方面来介绍反馈响应过程,首先给出如何将原始查询串在反馈信息库的作用下生成修改后的查询串;其次将介绍反馈信息库作用于预搜索结果生成搜索结果的过程。

4.1 基于用户信息反馈的查询扩展

用户反馈法是在查询再生成中使用得最多的方法。该方法的主要思想是:首先从那些被用户认为有用的文档中选择最重要的词条或者表达式,然后在新的查询串中增强这些词条或者表达式的权重。在这里我们将根据生成的反馈信息库来调整原始查询串中词条的权重。

对于查询串 q 中的第 i 个词条 $item_i$ 来说,其调整后的权重为:

$$\text{weight}(item_i)=\sum_{Q}VALUE(C,url_i)$$

其中,Q 为谓词 $C\subseteq\text{upper\_apr}(item_i)$。

在将查询串 q 中的词条权重标准化以后可以直接送给搜索引擎。

4.2 基于用户信息反馈的搜索结果排序

为了便于阐述反馈信号处理的具体过程,先给出如下几个概念。在图 2 的④中预搜索引擎搜索结果作为搜索结果处理的输入,它以一定的格式传送。带评分的预搜索结果的每条记录都可以表示为二元组,第 i 条预搜索结果记录形式化地表示为:

$$D_i=(url_i,R_{i,q})$$

其中,评分 $R_{i,q}$ 为搜索结果记录 url_i 与查询串的相关程度(以百分制表示)。

这样通过搜索引擎预先搜索到的带评分的结果集合可以记为 $D=\{D_1,D_2,\cdots,D_N\}$,$|D|=N$,N 为预搜索结果中记录的个数。

系统需要根据用户的反馈情况对预搜索到的结果进行重新排序。这主要是对预搜索结果中评分进行调整。在得到新的评分后,可以给出一个适当的阈值,以限制某些与用户相关度很小的记录返回给用户。

排序的过程主要是依据反馈信息库中的信息对预搜索结果集合 D 中的结果记录进行重新评价。在该过程中,使用了文[10]中给出的倒排文本频率函

数 IDF_j：

$$IDF_j=\log(N/\sum_{i=1}^{N}C_{i,j})$$

其中 $C_{i,j}=$

$$\begin{cases}1, & \text{如果 } upper\_apr(item_j)\cap url_i\cdot abstract\neq\phi\\ 0, & \text{如果 } upper\_apr(item_j)\cap url_i\cdot abstract=\phi\end{cases}$$

这样，对 D 中的第 i 条记录 D_i 进行处理后得到新的记录 D'_i，

$$D'_i=(url_i,R'_{i,q})$$

其中，$R'_{i,q}$为对预搜索结果记录 url_i 新的评价：

$$R'_{i,q}=R_{i,q}\cdot\sum_{item_j\in q}\sum_{Q}VALUE(C,url_i)\times IDF_j$$

其中，$Q=C\subseteq upper\_apr(item_j)$。

对预搜索结果集合 D 中的每条记录进行以上处理后，得到新的搜索结果集合 D'。对搜索结果集合中的记录按照 $R'_{i,q}$的值从大到小进行排序后得集合 D''，选取 D''中的前 K(K≤N)条记录返回给用户。这样的搜索结果中已经加入了反馈信息。用户对于这些搜索结果的选择信息又将被自动收集，这些被收集的信息将作为以后查询时候的反馈信息。这样该搜索引擎系统可以随着用户的选择而不断调整自己的搜索结果，使得返回的信息将更适合用户的需求。

5 带反馈的自适应搜索引擎系统原型设计与实验

根据以上反馈信号的采集和生成算法以及反馈响应算法，我们设计实现了一个实验性带反馈的自适应搜索引擎 FASE(参见图 3)，并通过实验对 FASE 的自适应能力进行了验证。

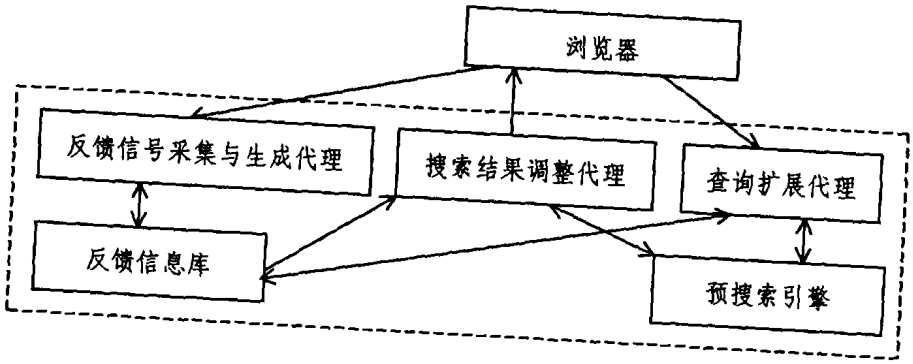

图 3 带反馈自适应搜索引擎 FASE

5.1 一个实验性带反馈自适应搜索引擎 FASE

FASE 自适应搜索引擎由反馈信号采集与生成代理[2]、反馈信息库、搜索结果调整代理、查询扩展代理和预搜索引擎组成。反馈信号采集与生成代理按照上文中的反馈信号采集与生成方法记录用户对搜索结果的选择情况，该代理以一定的周期(可以由用户设定)对这些记录进行处理以生成反馈信息，处理后的结果存放在反馈信息库中，也就是说，反馈信息库中信息的更新是有一定周期的，它的更新周期取决于用户对该搜索引擎的访问频率以及机器的负载能力。FASE 中的预搜索引擎与一般常用的搜索引擎相似，根据用户的查询请求从索引数据库[4]中检索出对应的信息，这些信息按照一定的数据格式发送给搜索结果调整代理。搜索结果调整代理主要负责最后搜索结果的集成，它将由搜索引擎送回来的预搜索结果和反馈信息库中的信息按照公式(1)对预搜索结果记录重新计算评价值，然后将这些记录根据新的评价值排序后返回给用户。查询扩展代理根据反馈信息库中的信息对用户的初始查询进行扩展以后送给预搜索引擎。在 FASE 中，反馈信息库利用微软公司的 SQL Server7.0 数据库。采用如表 1 所示的数据表结构，对表中的字段 term 和 url 建立 term+url 的索引。

表 1 反馈信息库数据结构

数据 / 字段	数据类型	注释
Term	Varchar	词条
url	varchar	搜索结果记录
Value	integer	评价指标

5.2 实验

带反馈的自适应搜索引擎 FASE 的自适应性体现在：用户对搜索结果的选择情况可以作为反馈信号输入到搜索引擎中，从而影响搜索引擎以后的搜索结果。反馈信号是依据使用 FASE 的所有用户对搜索引擎的使用情况而生成的。为了反映 FASE 的自适应能力，依据现有的实验条件我们设计了如下实验。在 FASE 中，选择新浪(www.sina.com.cn)作为预搜索引擎。FASE 为用户提供元搜索引擎服务。用户使用 FASE 提供的界面进行搜索，用户对搜索结果的选择情况被记录在 FASE 中的反馈信号采集与生成代理中。要获得实验结果需要对 FASE 中所记录的日志进行分析。通过分析比较在相邻时间区段(区段的长度为反馈信息库的更新周

期)内 FASE 所记录的日志,可以分析出 FASE 对用户的自适应能力。

为了便于对数据进行分析比较,必须对实验中的某些情况做一些假设。我们假设在我们选定的实验时间区段中所使用的新浪搜索引擎没有进行信息调整(考虑到新浪的信息调整是有一定周期的,也就是说,我们的测试过程是在新浪搜索引擎的信息稳定期内),而且用户在搜索的过程中没有使用它的个性化功能(即用户没有登陆到新浪服务器)。实验分为数据选择和数据分析两个阶段。数据选择主要是从 FASE 所记录的日志中筛选出对应时间区段和所选定的测试词条有关的日志。数据分析是通过对这些筛选出的数据进行分析比较得出 FASE 的运行规律。为了得到 FASE 的自适应趋势,主要比较相邻时间区段里对于同一个查询而言用户对搜索结果记录的访问次序。为了便于量化比较,我们只选择 FASE 日志中记录的用户对前 7 个搜索结果的选择情况。为了比较用户在相邻时间区段里,对同一查询所返回结果的选择差异,我们定义了一个相异度函数 XYD(T_i,T_{i+1},*item*)。

$$XYD(T_i,T_{i+1},item)=\sqrt{\frac{\sum_{j=1}^{n}(Location(T_i,item,R(item,T_i)_j)-Location(T_{i+1},item,R(item,T_i)_j))^2}{n}}$$

其中,T_i,T_{i+1}为两个相邻的反馈信息库更新时间区段,$0\leqslant i\leqslant K-1$,K 为一正数;*item* 为查询串;$R(item,T_i)_j$ 表示在时间区段 T_i 内对 *item* 搜索所得到的结果记录中的第 j 条记录;Location(T_i,*item*,$R(item,T_i)_j$)为记录 $R(item,T_i)_j$ 在时间区段 T_i 内对 *item* 查询所得的结果记录序列中的位置,很显然有 Location(T_i,*item*,$R(item,T_i)_j$)$=j$;n 为所考虑的结果记录的个数。

相异度函数 XYD(T_i,T_{i+1},*item*)表示在相邻时间区段 T_i,T_{i+1}内对同一查询所得搜索结果序列的相异程度。它反映了自适应搜索引擎中反馈信息库的更新对搜索结果的影响。通过对相邻时间区段内查询结果的相异度分析,可以知道一些有趣的结果。

我们固定相异度函数中的参数 K=7,n=10,时间区段的长度都为 T=3 小时,选择 29 个常用的查询串,分别计算它们在不同时间区段的相异度。在图 4 与图 5 中给出了相异度曲线及其趋势。

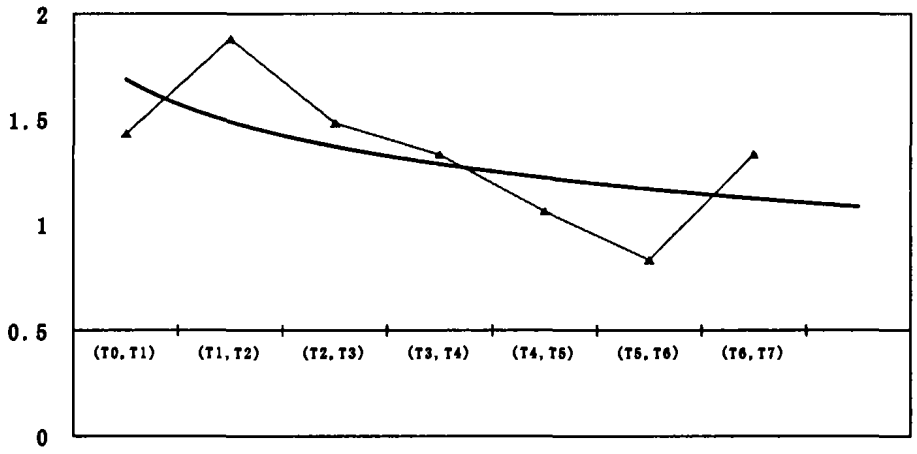

图 4 相异度曲线及趋势

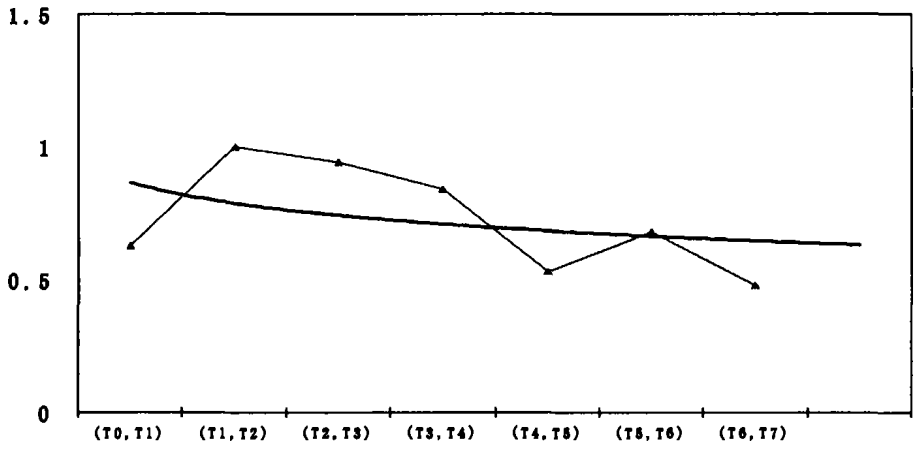

图 5 相异度总体曲线与总体趋势

(下转第 23 页)

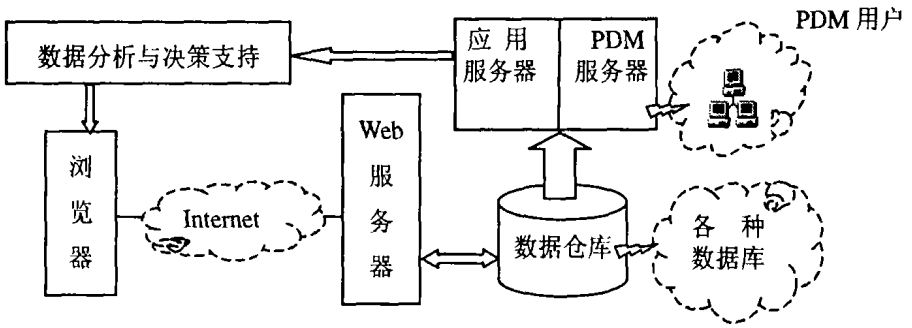

图3 基于Web的企业数据流模型

结束语 本文提出的将企业异构数据流从低成本的数据移植逐步建设成为综合的信息仓库，使得企业能够让多种业务应用系统、多种异构数据源并存，实现了数据动态展现、互访和综合利用，这样既保护了企业的原有信息化投资，又提供了应用系统由旧向新、系统平台由低向高平滑过渡的需要，同时基于Web的企业数据流挖掘不仅能满足阶段性、可扩展性信息系统建设的需要，又能实现所需信息服务目标，还能进一步加强企业PDM系统之间数据访问和产品应用的能力。数据仓库建设没有成熟的模式可循，具体实施时，应根据企业实际状况，不断探索和不断改进。

参考文献

1 陈义，宋执环，李平. 基于Web的流程企业数据仓库体系研究[J]. 计算机集成制造系统，2003，9(6)：493
2 北大高科网站. PDM与企业信息集成[Z]. 中国计算机报. http://www.pku-ht.com. 2004.3
3 熊海灵，伍胜，余建桥. 异构数据源的集成与访问[J]. 计算机科学，2003，30(5)：183
4 林行健. 精通oracle9i[M]. 机械工业出版社，2004.1
5 甘利人. 企业信息化建设与管理[M]. 北京大学出版社，2001.168～170
6 (加)Han Jiawei，Kamber M. Data Mining Concepts and Techniques[M]. 机械工业出版社，2001.44～46
7 朱全敏，蒲工. 基于Web的产品数据管理系统综述[Z]. http://www.e-works.net.cn/ewkArticles/Category33/Article14594.htm. 2003.6
8 oracle公司·oracle express server[Z]. http://www.oracle.com. 2004.4

(上接第8页)

在图4与图5中，横坐标表示相邻时间区段，纵坐标表示相异度，其中的小三角表示在该相邻时间区段间的相异度，图中的平滑曲线表示这些小三角的总体趋势。图4是从29个查询中随机抽取个而给出的相异度曲线及其趋势。为了反映整个系统的情况，我们分别计算这些相邻时间区段间这29个查询相异度的平均值，得到图5中的曲线。从图5中，我们可以看到相异度的总体趋势是渐近于某一值，但不是零。相异度的逐渐减小的趋势说明了FASE不断调整自己的反馈信息库，使得自己更能迎合大多数用户的兴趣。而相异度不会降为零正是FASE搜索引擎不断调整自己的动力所在。在面对众多用户的时候，FASE力求使自己的服务更好地适应大多数用户的兴趣。

结束语 本文根据带反馈的自动控制系统的基本原理，提出了自适应搜索引擎的概念。自适应搜索引擎首先通过一种机制来采集用户对搜索结果的访问序列，在此基础上生成自适应搜索引擎系统的反馈信号，在反馈信号生成过程中，我们采用了粗糙集理论来解决反馈信息库的空间限制和用户兴趣与反馈信息库中的兴趣的模糊匹配问题；然后通过一定的方式将反馈信息综合到搜索结果中去，从而使得系统具有自适应的能力。将来可以进一步探索反馈信号的生成方法和反馈信号与搜索结果集成的方法，以提高搜索引擎对用户的适应能力。

参考文献

1 Levy M R. Web Programming in Guide. Software，1998，28(15)：1581～1604
2 Genesereth M R，Ketch S P. Software Agents. Communication of the ACM，1994，37(7)：48～53
3 Bamshad M，Cooley R，Srivastava J. Data Preparation for Mining World Wide Web Browsing Patterns. Journal of Knowledge and Information Systems，1999，1(1)
4 Budi Y，Dik L. A world wide Web resource database system. IEEE Transactions on Knowledge and Data Engineering，1996，8(4)：548～554
5 张卫丰，徐宝文. Web搜索引擎框架研究. 计算机研究与发展，2000，37(3)
6 邹涛，等. WWW上的信息挖掘技术及实现. 计算机研究与发展，1999，36(8)：1021～1024
7 张卫丰，徐宝文，周晓宇. Web页面中的计数器研究. 小型微型计算机，2000，21(10)：1096～1099
8 张卫丰，徐宝文，周晓宇. Web页面中元素间交互技术研究. 计算机工程，2000，26(8)：62～64
9 Bamshad M，Cooley R，Srivastava J. Creating Adaptive Web Sites Through Usage-Based Clustering of URLs. In：Proc. of the 1999 IEEE Knowledge and Data Engineering Exchange Workshop (KDEX'99)，Nov. 1999
10 Weiyi M. Detection of Heterogeneities in a Multiple Text Database Environment. http://panda.cs.binghamton.edu/~meng/pub.d/coopis99.ps.gz
11 张卫丰，徐宝文，许蕾. Web页面安全性技术初探. 计算机工程与应用，2000，36(11)：158～161
12 张卫丰，徐宝文，许蕾. 利用Agent个性化搜索结果. 小型微型计算机[已录用]
13 Zhang Weifeng，Xu Baowen，Yang Hongji，Chu W C. A Genetic Algorithm Based General Search Engine. In：Proc. of IEEE MSE'2000
14 Xu Baowen，Zhang Weifeng，Chu W C，Yang Hongji. Application of Data Mining in Web Pre-Fetching. In：Proc. of IEEE MSE'2000
15 Zhang Weifeng，Xu Baowen，Chu W C，Yang Hongji. Data Mining Algorithms for Web Pre-Fetching. In：Proc. of WebSem'2000