AFDB 的设计与实现*)

徐洪丽[1] 苗 良[1] 张承明[1,2] 刘绍翰[4] 史 斌[3]

(山东农业大学信息科学与工程学院 泰安 271018)[1]

(山东科技大学地球信息科学与工程学院 青岛 266510)[2]

(泰安市公安消防支队 泰安 271000)[3] (南京航空航天大学信息学院 南京 210016)[4]

摘 要 主动模糊数据库系统(AFDBS),是由规则库(FEB)及相应的模糊事件监视器(FEM)、触发器评价器(TJ)、事件处理器(EP)组成。其中模糊规则库(FEB)由系统或用户定义的各种模糊事件驱动的模糊规则(模糊 ECA)组成。论述了主动模糊数据库的组成及相关理论,着重点为模糊 ECA 规则的执行,包括执行模式、执行方式、规则冲突解决。给出了设计实例,设计实例主要实现主动模糊数据库中模糊数据的定义、存储、录入以及模糊数据库的主动响应。

关键词 模糊 ECA,模糊事件监视器,模糊数据定义,模糊数据存储

Design and Realization of AFDB

XU Hong-li[1] MIAO Liang[1] ZHANG Chen-ming[1,2] LIU Shao-han[4] SHI Bing[3]

(Information Institute, Shandong Agricultural University, Taian 271018, China)[1]

(School of Geo Information Science and Engineering, Shandong Science and Technology University, Qingdao 266510, China)[2]

(Tai'an Fire Detachment, Taian 271000, China)[3] (Information Institute, Nanjing Aeronautics and Astronautics University, Nanjing 210016, China)[4]

Abstract Active Fuzzy Database System (AFDBS), is composed of fuzzy database (FDBS), fuzzy rules library (FEB), fuzzy event monitors (FEM), trigger judge(TJ) and event processor(EP). The FEB consists of the fuzzy event-driven rules. Deals with the AFDB composition, theory, focuses on the implementation of fuzzy ECA rules, including rules model, the implementation modalities, rules of conflict resolution. Based on the principle of active fuzzy database, designs and realizes AFDB, focuses on the definition of fuzzy data, how to input and store it, fuzzy incident and the fuzzy rules, how to realize its active feature.

Keywords Fuzzy ECA, Fuzzy FEM, Definition of fuzzy data , Fuzzy data store

AFDB (Active Fuzzy Database),即主动模糊数据库。它是将主动数据库技术和模糊数据库技术结合起来,一方面引入客观世界的模糊性和不确定性,把不完全性、不确定性、模糊性引入数据库系统中,从而形成模糊数据库[1],另一方面使数据库系统具有一定的主动性,即能根据一定事件的发生来主动地执行相应功能的动作。AFDB 是数据库研究的热点之一,本文基于主动模糊数据库的原理,设计与实现 AFDB,并着重就模糊数据的定义与录入,模糊事件和模糊规则、AFDB 的主动性方面做讨论。

1 主动模糊数据库的组成及特点

一个主动模糊数据库系统(AFDBS),实际上是由一个模糊数据库(FDBS)加上一个模糊规则库(FEB)及相应的模糊事件监视器(FEM)、触发器评价器(TJ)、事件处理器(EP)组成,即 AFDBS=FDBS+FEB+FEM+TJ+EP。

1.1 模糊规则库(FEB)

模糊规则库 FEB 由系统或用户定义的各种模糊事件驱动的模糊规则组成。这种规则称作模糊 ECA。模糊 ECA 规则的一般形式为:

RULE 模糊规则名[参数表]| ON 模糊事件

由于现实中大多数的事实和规则是模糊的,因此模糊逻辑通常处理在一定程度上为真的命题,它为解决模糊概念之间的关系推理,提供了一种表示不确定性的方法。所以可以把模糊逻辑集成到主动数据库的 ECA(事件-条件-动作)规则中。

基于事件-条件-动作的 ECA 规则应用系统执行操作和事件处理完成系统的主动性行为。ECA 规则具有很强的语义表达能力[2]。

ECA 规则系统可以表示为三元组:

事件(Events)集合是一组操纵被监视的数据操纵操作。

条件(Conditions)是一个在当前数据库状态和规则的过渡值上表示的谓词语句[3]。

动作(Actions)是一组操纵数据操纵的操作。

模糊 ECA 规则的一般形式为:

RULE 模糊规则名[参数表]| ON 模糊事件

IF(模糊条件 1 THEN(动作 1)

[WHERE 约束 1]

EXCEPTION(例外处理动作 1)

模糊规则执行模式模糊触发器允许" On Event IF Condition Then Action"中的 Event,Condition ,Action 都是模糊的。

规则的语义是:模糊规则由模糊事件驱动,模糊事件是以一定的发生度来发生的,即以一定的发生度作为发生的程度趋向,并不是完全发生。在实际系统中,通常对模糊事件的发生采用系统或用户定义的一个阈限值入。当模糊事件的发生度大于给定的阈限,系统就认为该模糊事件已经发生。模糊规则执行模式通常用触发器完成。

*)地理空间信息工程国家测绘局重点实验室经费资助项目(B2623);山东省水利科技专项资金(200357)。**徐洪丽** 硕士,讲师,研究方向为计算机应用;**苗 良** 教授,研究方向为计算机应用;**张承明** 副教授,主要研究方向为计算机应用、数字化技术;**刘绍翰** 博士,讲师,研究方向为数据挖掘和信息采集;**史 斌** 高级工程师,研究方向为计算机通讯。

1.2 模糊事件监视器(FEM)

FEM是一个随时监视FEB中的模糊事件是否已经发生的监视模块,它检查这些规则中包含的各种模糊事件是否已经发生,一旦发现某事件发生,就按FEB中指明的相应规则执行用户预先定义的动作。利用它可以实现主动模糊数据库的完整性约束、存取控制、例外处理、主动提出警告、自动产生报表以及周期性执行某些处理等。模糊事件监视器可以通过存储过程或自定义约束等实现。

1.3 触发器评价器和事件处理器

触发器评价器(TJ)负责判断触发条件。如果多个触发器被同一事件触发,则TJ按一定的优先级对多个触发器进行条件评价,同时也可按一定的优先级选择被触发的活动[4]。

TJ首先区分要进行时间事件评价还是进行非时间事件评价,然后创建条件评价器线程进行评价。情形评价器按优先级顺序评价所有被激活触发器(可能有多个触发器)中的情形。

事件处理器(EP)按照规定模式触发活动。

1.4 主动模糊数据库特点

主动模糊数据库融合了模糊数据库和主动数据库的特点,既能够处理模糊数据,又具有主动反映的能力,其特点如下:

模糊理论超越了简单的是非二值逻辑。它运用的是复杂的多值逻辑,从而导致数据的模糊性,即数据本身是模糊的,包括模糊数据变量,如"9:00左右"和模糊语言变量,如"一般"、"良好"、"很","非常"等。

数据操纵的模糊性及数据间关系的模糊性,包括数据定义、数据操纵和数据查询的模糊性,即数据间的联系以及依赖关系是模糊的。

数据库具有一定的主动性,系统根据自身情况,对数据库主动做出反映。这主要是通过一些主动模糊规则事先嵌入到模糊数据库系统的办法来实现。系统中提供了一个自动"监视"模块,它主动地不时地检查这些规则中包含的各种模糊事件是否发生,一旦发现某事件发生时就主动触发执行某个模糊动作。

2 理论支撑模糊数据库预备知识

2.1 模糊关系、模糊集合、隶属度

定义1(模糊关系) 设$D1,D2,\cdots,Dn$为n个论域,$F(Dn)$为模糊关系属性的值域,一个模糊关系定义为值域$F(D1)$,$F(D2)$,$F(D3)$,$\cdots$,$F(Dn)$上的笛卡儿$F(D1)\times F(D2)\times\cdots\times F(Dn)$上的子集合,$r=(x1,x2,\cdots,xn)$,$r$为模糊关系得一个元组[5],其中$xi\in Di(i=1,2,3,\cdots,n)$。

定义2(模糊集合,fuzzy sets) 首先定义U为某些对象的集合,称为论域,可以是连续的或离散的;u表示U的元素,记作$U=\{u\}$。论域U到[0,1]区间的任一映射mF,即$mF:U\,R[0,1]$,都确定U的一个模糊子集F,其中mF称为F的隶属函数(membership function)或隶属度(grade of membership)。也就是说,mF表示u属于模糊子集F的程度或等级。

但是隶属的程度各不相同:完全隶属度是1,完全不隶属度是0,部分隶属度则处于0和1之间。

关于隶属度解释:$vF(u)$表示论域U上的元素u隶属于模糊集合F的程度。例如"选择身高1.90的人",对于模糊集合高={0.6/180,0.7/185,0.8/190,1/195},180厘米属于高个头的程度为0.6,190厘米属于高个头的程度为0.8,而195厘米完全是高个子。可以对模糊算子做语气词转化,比如0.8转化为"非常",1转化为"很",则190厘米非常高,195厘米很高。

2.2 规则耦合方式

规则耦合方式(coupling mode)用来描述规则执行与触发规则的用户事务之间的时间关系。它分为三类:立即(immediate)、延迟(deferred)、分离(detached)。立即耦合方式的规则在触发时刻处理,触发事务暂时挂起,在被触发规则完成后继续,其净效果相当于将被触发规则嵌入到触发事务的触发点处执行。延迟耦合方式规则的执行被推迟到触发事务提交之前。分离耦合方式的规则在触发时刻启动,与触发事务并行,但要保证触发事务在先,被触发规则事务在后的串行化提交次序。

2.3 规则冲突图

定义3 R为主动规则集合,规则冲突图RCG(Conflict Graph)是由$\langle R,Ec\rangle$定义的二元组,其中:r属于R,为规则结点,无向边(r_i,r_j)属于Ec,表示r_i和r_j为相互冲突的规则。冲突图也称依赖图(dependency graph)[6]。

3 模糊规则库(FEB)中的模糊ECA规则

当触发器的条件和动作都为模糊命题时,称为模糊ECA。模糊ECA规则执行包括执行模式、执行方式、规则冲突解决。

3.1 模糊ECA规则执行模式的步骤

(1)Event信号:当一个事件被检测到的时候,把事件事实存入事件库,并按照规则库的规则选择或者ECA规则(通过触发器完成)。

(2)模糊化:评价规则条件的语言变量,然后把相应的资源(证据)数值模糊化。

(3)推理:也就是建立证据模糊集和结果模糊集之间的模糊矩阵(模糊关系)。用扎德方法来进行模糊推理,从模糊的数据输入得到模糊结论。

(4)选择相应的Action:把模糊集结果解模糊化生成一个精确数值即取其隶属函数的重心作为精确结果数值,对规则集的每一条规则推理。根据隶属度的排序,依次执行隶属度超过预定阈值(如0)的动作执行。

3.2 模糊ECA执行方式

模糊ECA执行方式包括迭代执行与递归执行。规则执行采用宽度优先规则调度-迭代方式,也可采用深度优先规则调度-递归方式。

采用宽度优先规则调度执行,系统保持一个被触发规则池(pending rule set),逐一取出池中规则处理,处理过程中新触发的规则被投放到规则池中,如此循环直至规则池为空。

采用深度优先调度,规则处理过程中触发新的规则,中断目前规则执行,转去处理新触发的规则,待后者执行完毕后继续原来被中断的规则。深度优先规则与立即耦合方式相对应。

3.3 模糊ECA规则冲突解决

模糊ECA不仅支持原子事件,还支持复合事件。原子事件包括数据库改变、事务边界、时序(Temporal)事件等。复合事件是由多个原子事件组合而成。

如果两个模糊事件中所包含的一系列动作存在冲突的可能,则称两个模糊事件具有动作相关性。当多个模糊事件具有动作相关性时,需要解决模糊事件所包含的动作之间存在的冲突,即规则冲突。

规则冲突的解决包括:一般性规则冲突的解决和传递规则冲突的解决。

一般规则冲突的解决:针对不同的耦合方式有不同的解决策略。对于 deferred 耦合方式的规则,若两个规则不冲突则颠倒二者的执行次序不会导致结果的不同;相反,若两个规则相互冲突则颠倒二者的执行次序可能导致不同的执行结果,所以需要设计者规定直接冲突规则对之间的优先级。而对于 immediate 和 detached 耦合方式的规则,除直接冲突之外,必须说明间接冲突规则对之间的优先次序,否则不能保证执行结果的唯一性。

传递规则冲突的解决:传递冲突(transitively conflict)定义:存在 r'_i, r'_j, conflict(r'_i, r'_j)、$p(r_i, r'_i)$、$p(r_j, r'_j)$ 即(1)规则 r'_i, r'_j 之间存在冲突,(2)r'_i 可能被 r_i 触发,r'_j 可能被 r_j 触发,且触发路径中包含有 immediate 或 detached 耦合方式的规则。若满足上述两个条件,则称规则 r_i, r_j 为传递冲突。

如图 1,实线表示触发关系,虚线表示直接冲突关系。

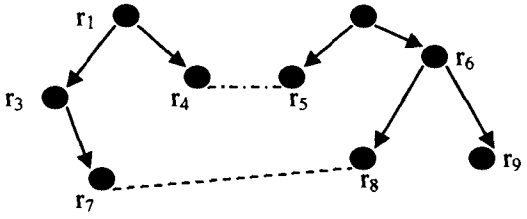

图 1 规则对冲突图

(r_4, r_5)、(r_8, r_9)是直接冲突规则对。只要除 r_9 之外的任何一个规则为 immediate 耦合方式,(r_1, r_2)就是一个间接冲突规则对,因为改变 r_1 和 r_2 的串行化次序将导致不同的数据库终止状态。但若除 r_9 之外的所有规则对为 deferred 耦合方式,则 r_1 和 r_2 不存在冲突。若 r_1 和 r_2 为 detached 耦合方式规则,即使二者不直接冲突,也是间接冲突的,因而也是冲突的,必须由用户规定相对次序。即当所有规则都为 deferred 耦合方式,则所有规则都被视为同一嵌套层次,只有直接冲突的规则对需要解除冲突。而对于 immediate 和 detached 耦合方式,直接冲突和间接冲突的规则对都需要解除冲突。

4 主动模糊数据库系统设计

主动模糊数据库中涉及到的规则管理器和条件评价器用 SQL SERVER 2000 的存储过程(sp)和扩展存储过程(xp)实现。模糊规则库用 SQL SERVER2000 触发器完成,触发器是由 SQL SERVER2000 自动激活或触发的特殊的存储过程。

4.1 要实现模糊数据库系统设计,最基本的一步为数据的录入

首先定义自定义数据类型(create type scoretype varying),绑定到数据库 xscj 中表 st_score 的 score 字段。界面如图 2 所示。

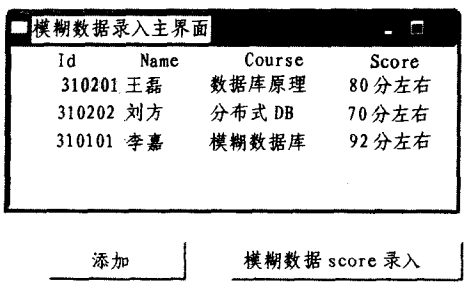

图 2 模糊数据录入主界面

成绩为模糊数据,当记录输入到成绩时,按"模糊数据成绩录入"按钮,转到模糊属性选择界面。这里有两种模糊数据类型可以选择,分别为对于模糊中心数、模糊区间数。如果用户选择的是模糊中心数,则可以直接在此界面上输入参数。当用户选择模糊中心数的界面:

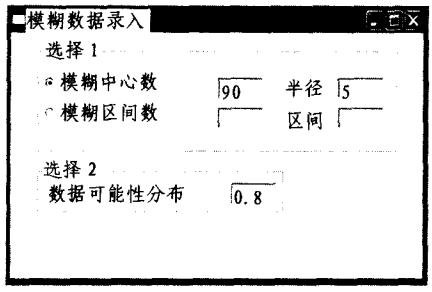

图 3 模糊中心录入的界面

说明:

模糊区间数:采用模糊区间数的方式来描述某一属性取值位于某一区间的可能性,记为[a, b] /p,表示该模糊数落在[a, b]中的可能性为 p;

模糊中心数:采用模糊中心数的方式来描述某一属性取值位于某一数值附近的可能性[7],记为(c, r) /p,表示该模糊数落在以 c 为中心,r 为半径的圆之中的可能性为 p。

例如:年龄=[20,30]/0.7 表示年龄在 20～30 之间的可能性为 0.7。成绩=(90,5)/0.8 表示成绩在以 90 为中心,2 为半径的圆之中的可能性为 0.8。

最终模糊数据在数据库表中的存储如表 1 所示,其中模糊数据类型为 1,表示模糊中心数方式录入,为 2 表示模糊中心数方式录入,id 对应图 1 中 id。

表 1 模糊数据在数据库表中的存储

表名	表名	id	模糊数据类型	参数 1	参数 2	可能性
St_score		1	1	90	5	0.8
St_score		2	2	80-90		0.9

4.2 主动响应部分

在 XSCJ 数据库的 score 表上创建一触发器,若对学号列和模糊数据列——成绩修改,则给出提示信息,并取消修改操作。

```
USE XSCJ
GO
CREATE TRIGGER update_trig
On xs_score
FOR update
AS
/*检查学号列(C0)和成绩列(C3)是否被修改,如果有某些列被修改了,则取消修改操作。*/
IF (COLUMNS_UPDATED() &9) > 0
BEGIN
RAISERROR ('不允许修改该列', 16, 1)
ROLLBACK TRANSACTION
END
```

以上触发器在表 xs_score 进行 Update 操作时被触发。当事件探测器收到该事件的触发消息后,根据消息中的参数可以确定该事件在事件知识库中已被定义,因为只有一个触发器 update_trig 与该事件相关联,所以在触发器的评价时该触发器默认为优先级最高,在进行 Update 操作的过程中,如果发生错误,触发器将执行事务回滚操作。

结束语 AFDB 是将主动数据库技术和模糊数据库技术结合起来,在传统数据库中引入表达和处理不确定信息(如模糊信息)的能力,并融合了主动数据库技术,无疑拓宽了数据

库的应用领域，在管理信息系统、专家系统、决策支持系统、群体工作环境系统中占有重要的位置。本文基于主动模糊数据库的原理，设计与实现 AFDB，并着重就模糊数据的定义与录入，模糊事件和模糊 ECA 进行了讨论。

参 考 文 献

[1] Bosc P, Pivert O. SQLf: a relational database language for fuzzy querying. IEEE Transactions on Fuzzy Systems, 1995. 3
[2] Peng T, Zuo W L, Liu Y L. Characterization of Evaluation Metrics in Topical Web Crawling Based on Generic Algorithm// 1th Intl Conf. ICNC, Changsha; Springer, LNCS. 3611. (SCI DBA23), 2005:690-697
[3] 左万利，刘居红. 关联图与主动规则集的终止性分析. 软件学报，2001(12):278-280
[4] 何新贵. 模糊关系型数据库的数据模型. 计算机学报，1989(2)
[5] 郝忠孝，熊中敏. 计算主动数据库中不可规约集合的有效算法. 计算机研究与发展，2006(1):285
[6] 关系型模糊数据库管理系统的设计与实现. 学位论文. 上海海事大学，2005,6
[7] 魏延. 主动模糊数据库中的事件与规则. 重庆师范学院学报，2002(12):175

（上接第 293 页）

$$\mathrm{Sim}(d_i,d_j)=\frac{\sum_{k=1}^{M}W_{ik}\times W_{jk}}{\sqrt{(\sum_{k=1}^{M}W_{ik}^2)(\sum_{k=1}^{M}W_{jk}^2)}}$$

K 值的确定是一个关键的问题。现在的一般做法是先选定一个初始值（几百到几千之间），在进行自动归类的过程中根据结果进行调整。接下来在新网页的 K 个邻居中，依次计算每一类的权重，计算公式为：

$$p(\vec{x},C_j)=\sum_{\vec{d}_i\in KNN}\mathrm{Sim}(\vec{x},\vec{d}_i)y(\vec{d}_i,C_j)$$

其中，$\vec{x}$ 为网页的特征向量，$\mathrm{Sim}(\vec{x},\vec{d}_i)$ 为相似度计算公式，而 $y(\vec{d}_i,C_j)$ 为类别属性函数，如果 $\vec{d}_i$ 属于类 C_j，那么函数值为 1，否则为 0。最后比较类的权重，将网页分到权重最大的那个类别中去。

2.2 网页自动聚类实现方法

网页的自动聚类一般包括四个步骤：

(1)网页表示：包括特征抽取和特征选择。特征选择是选择那些最具有区分性的特征，也就是最能把不同类别区分开来的特征，而不是大多数对象都具有的特征。

(2)相似度计算。主要根据网页表示的距离函数来定义。

(3)聚类：根据网页表示和相似度计算结果，按照规则将聚类网页分成不同的类。

(4)给出聚类的标识。在形成的每一类中抽取具有代表性的特征，作为该类的标识。

3 自动分类在搜索引擎中的应用策略

3.1 自动聚类和自动归类的应用

就目前的情况而言，自动聚类在搜索引擎中的实现要比自动归类容易一些，聚类的效果也比较显著。因此，可以考虑在搜索引擎中首先采用自动聚类。

如果要使用自动归类，首先就要考虑使用什么分类法。现在使用的分类法中既有传统的图书馆分类法，也有适应网络环境而生的网络分类法。二者各有千秋，传统的图书馆分类法系统性强，使用范围广，网络分类法比较灵活。如果条件许可的话，最好是两种类型的分类法都使用。对于熟悉图书馆分类法的用户就提供图书馆分类法的结果，对于一般用户则提供自编的网络分类法。在使用分类法的时侯，还要考虑分类的粗细问题，也就是分到几级类目。对于网页的分类，可能没有必要分得很细。下面主要论述自动聚类实现时涉及到的问题。

3.2 应用的时机

应用的时机是指自动聚类是在对网页数据进行索引的时候实施，还是在搜索引擎返回检索结果之后实施。前者可以利用网页的全文，后者一般只是使用网页的网址、标题和摘要等少量信息。一般而言，前者的结果要准确一些，但是综合考虑，后者的精确度虽然不如前者，但是成本比较低，实用性更强。它不需要对网页进行标引等预处理，工作量会大大降低，并且随着技术的发展，结果也会越来越令人满意。对于结果相关性的判断，既有客观因素，也有主观因素。机器只能够模拟人的思维而不能取代人的活动。自动聚类只是帮助用户进行相关性的判断而已，想靠它一劳永逸地解决相关性判断是不太现实的。

3.3 应用的对象

自动聚类可以应用到元搜索引擎或者单个搜索引擎中。单个搜索引擎的覆盖范围有限，且随着网络信息资源的迅速增长而不断下降，所以将自动分类应用于元搜索引擎返回的结果要比应用到单个搜索引擎的效果要明显一些。当然，元搜索引擎在对调用的搜索引擎进行选择时必须要遵循一定的原则，要选取质量比较高的，覆盖面比较广的，力争扩大检全率和检准率。对于单个搜索引擎返回结果，也没有必要全部包括在内，只需要前面的一部分就可以了（例如 50 条左右）。因为一般情况下，前面的结果与检索要求的相关度要高一些，这样做对于系统的精确性不会有太大程度的影响，但是可以将系统的成本大大降低，实用性更高。

3.4 用户界面

用户界面的设计是一个经常被忽略的问题，实际上用户界面的设计对于自动分类系统的使用效果有很大的影响。一个有关这方面的实验就证明了这一点。这个实验是 Hao Chen 和 Susan Dumais 完成的[20]。他们对七种检索界面的使用效果做了对比。这七种用户界面是：

(1)悬浮显示摘要的清单式界面，就是只有当鼠标移到返回的网页的标题时才显示出该网页内容的概要。

(2)内嵌摘要的清单是用户界面，将网页的摘要出现在返回网页的标题下面。

(3)显示类名的清单式界面，将返回网页的标题后面出现其所属的类目名称，同时给出网页的摘要。

(4)悬浮显示摘要的分类界面，首先给出类目的名称，然后显示出该类目下的网页标题，当鼠标移到该标题上的时候显示出该网页的摘要。

(5)内嵌显示摘要的分类界面，它与第四种界面基本上一样，除了是将网页的摘要显示在标题下面。

(6)无类名的分类界面，它将类目的名称和网页的摘要都去掉了。

(7)无网页标题的界面，只显示出类目供浏览。

结束语 综上所述，我们认为现阶段自动分类在搜索引擎中的应用主要应该考虑自动聚类在搜索引擎特别是元搜索引擎中的应用，将搜索引擎的结果进行自动聚类后返回给用户。采用类似于 Vivisimo 的用户界面，将类目的名称和网页的摘要明确地展现给用户，用户可以根据自动分类结果进行检索策略的修改。

参 考 文 献

[1] 冯是聪. 一种中文网页自动分类方法的实现及其应用. 计算机工程，2004(3):70-72
[2] 李晓黎. 基于向量机和无监督聚类相结合的中文网页分类器. 计算机学报，2001(1):62-68
[3] 秦兵，等. 可分性判据在中文网页分类中的应用. 微处理机，2002(1):26-28
[4] 范焱，等. 用 Naive Bayes 方法协调分类 Web 网页. 软件学报，2001(9):1386-1391
[5] Zamir O. A dynamic clustering interface to web search results// Eighth International World Wide Web Conference, 1999(5):11-14