信息捕捉器的研究与实现

中国科学院软件研究所 徐 进
北京理工大学计算机科学与工程系 刘明迪
中国科学院软件研究所 王振华

研究基于 Internet/Intranet 环境实用的网络信息获取技术，以便达到高效率的信息检索能力。运用Caching Web技术及模糊集理论、信息聚类等技术。改变了传统的信息检索方式，实现了基于教育领域的 Internet 网络信息捕捉器的系统原型。

随着信息技术的不断发展,万维网正在社会各个领域得到深层的运用，然而，现有的网络导航或网络搜索工具，都存有共同的缺陷，即在网上花费了很长时间搜索后，得到的并非直接的目标,而是缺乏总结性的一大堆信息，不能有效地滤掉不切题的内容。这不仅增加了网络使用费用，还会使人感到厌烦。怎样使用户从烦琐的信息查询浏览中解脱出来,以适应领域或行业范围的相对准确的信息需求，是研制的目标。信息捕捉器将改变以往的信息查询浏览方式,可对指定范围内的网上信息主动进行采集和分类，将基于词的内在含义的查询结果提交给用户，直接传送到用户的是具体的目标信息,而不是仅复合词表义的所有信息条目的罗列。

信息捕捉器技术概述

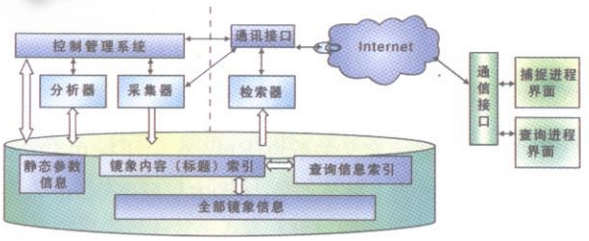
信息捕捉器区别于机器人、蜘蛛、漫游者、网虫等，前者是根据用户的旨意，在确定的范围内主动地、有效地采集/收集网上的资源；并经智能机制，自动地跟踪信息资源，进行内容筛选提取；动态地对提取后的信息组织数据库，进行有效的管理；以多种方式提供用户的信息访问。而后者主要采用信息的广泛漫游，依靠人工的抉择来达到信息的获取，是一种无论从时间和金钱上都很奢侈的信息获取方式。前者更强调主动性和准确性，是更经济的信息获取方式。

系统逻辑结构示意图如下:

捕捉器分为两大子系统：捕捉器控制管理子系统和查询服务子系统。

信息捕捉器抛开传统的词表义检索查询模式，采用对词的内在含义的查询检索模式。在此前提下，必须掌握每个词或组词的精确含义。然而由于领域的不同，每个词或组词其含义可能大不相同，如“工程”和“网络”这两个词，在计算机领域和建筑领域含义是不同的。捕捉器的控制管理子系统首先提供定义领域的范围的机制，不同领域的信息捕捉器可定义不同的领域，多种领域的信息捕捉器可定义多种领域。捕捉器控制管理系统还包括捕捉器的智能机制，利用专门领域的知识，通过推理机制和规则，建立领域的信息结构；控制采集并跟踪信息源机制；运用信息分类/聚类算法，分析整理信息的分析机制。

查询服务子系统，采用push技术，把用户寻找信息模式变为信息上门模式。提供多种信息服务方式；提供多种组合查询方式及收费、安全等管理机制。

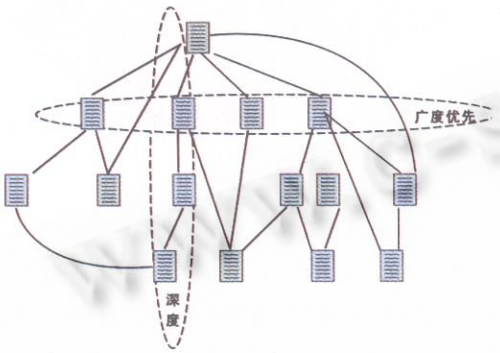


信息采集

(1)系统功能。该模块按系统定义的范围，采用Caching Web信息技术，到网上镜象指定网址的下属所有网页，然后建立“镜象内容（标题）索引库表”。并且按照指定的周期检查有无更改的页面，并对其镜象页面进行增、删、改处理。镜象的全部网页按文件方式存放。由于网上的各Web服务器内容不断地被更新，因此在镜象服务中的文

档，可能会很快变为陈旧或者过时。采集管理模块，根据所要采集/收集的专题属性，选择一时间周期，来寻找被更新的文档及时地更新镜象服务器中的内容，保持信息的一致性。

(2)镜象处理算法。通过遍历文档的超文本链路，采集器从万维网上获取文档。遍历文档的超文本链路有两种方法：广度优先和深度优先。我们采用广度优先搜索方法。广度优先搜索法是根据主页URL地址，搜索所有热键URL地址，层层深入，图示如下：



此外建立先进先出控制队列，管理搜索页面流和镜象处理流顺序。搜索页面流：采用广度优先算法，搜索页面全部页面热键URL填入控制队列。镜象处理流：采用先进先出算法取出各页面URL，到网上去镜象页面内容。示意图如下：

队列文件记录结构：

关系号
URL信息
页面编号

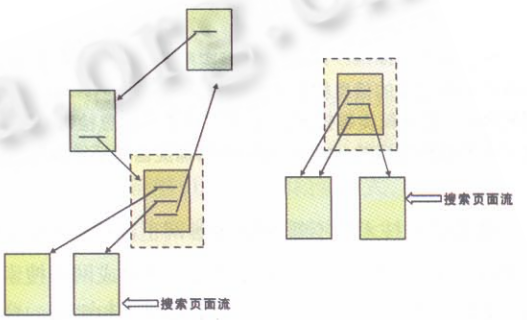
关系号规则：定义关系为Rm。n，下标m。n为关系号。其中m为上层节点号，n为本层顺序号。m为完整的上层节点号。

由于页面组织整体上看是网状连接形式，搜索页面流在队列文件中发现重复的URL，这实际上或是在上层曾经出现过的父节点页面；或是本层其他热键指向的同一页面。(参看示意图a、b)这时搜索页面流仍按规则生成关系号并登入队列记录，将URL信息中填入“空白”，并在页面编号中填入相同URL信息的实际页面编号。

信息分析

(1)系统功能。分析管理模块是信息捕捉器的核心部分。系统的分析部分是确定镜象文本的类属范围的过程。而查询部分是根据查询需求逐渐缩小类属范围，最终确定需求的过程。

分析管理模块的功能是：将镜象来得信息进行类属范围的确定，根据算法生成查询信息索引库表，这实际上是对查询数据库的分类概念模式自动生成的智能化转换过程，我们将这一过程的实现称为概念转换分类算法，包括分类算法和聚类算法。



a. 指向上上层内父节点　　b. 指向本层同一个节点

聚类算法是通过对镜象的文本内容进行关键词分析及隶属度计算来确定类属的。

(2)分析算法

①分类算法说明。分类算法与对应的查询需求算法及网页查询算法形成一个整体，是相互对应的过程。系统的分析部分是确定镜象文本的类属范围的过程。而查询部分是根据查询需求逐渐缩小类属范围，最终确定需求的过程。该方法的思想可归纳为：

·分类过程，是确定信息类属范围的过程。系统中的词根实际上是类属词。通过对信息标题的截词处理，与类属同义词匹配，来确定类属的方法进行分类。分类结果是建立类属与标题物理地址的对应关系表，系统中叫做“查询索引表”。

·查询过程，是经过需求语句分析，逐渐缩小类属范围的过程。通过对查询需求的截词处理，与查询索引表类属词匹配，得到相应关系的标题信息集。随着截取词的加长，其语义表示范围越来越缩小，匹配上查询索引表类属词的信息集也越来越少，最终得到满足全部需求的标题信息结果集。

②聚类算法说明

文档关键词表的基本结构为使计算机能够对文档进行聚类或分类，需要提取文档对象的特征，即对文档关

键词进行提取，并建立文档对象的特征类表。因此，先需要建立文档关键词类表，此表用于指明具有这些关键词的文档是属于哪一类文档。对于要分类的每一文档我们选取若干个关键词，如取第一关键词、第二关键词等。实际上在文档中出现的有些关键词术语并不唯一，如“Internet”，在中文文档中目前人们用“因特网”，但在该词规范化之前，有用“国际互联网”或“互联网”，这些词虽然属于同义词，但在这里仍按关键词来看待。另外，从管理上考虑还要有建立该关键词类的日期以及该关键词在实际使用中出现的频度。

文档关键类表的基本结构为:

父类	子类	关键词1	频度1 ...	关键词5	频度5	类别	日期	备注

将该类表每一行看作一个向量，向量的元素是该类表的某一行的一个字段。由于一个关键词可在属于不同类的多个文档中出现，为了确认该关键词所在的文档到底属于那一类，对该关键词属于某一类文档应有一个度量，即为隶属度，(也称作关联度)。如果关键词所在的文档对某一类具有大的隶属度，则该文档即属于这一类。

· 分类算法中隶属度的确定

设有N个文档F_1，F_2，...，F_N；这N个文档中含有M个不同的关键词K_1，K_2，...，K_M。这样，一个文档Fi可用M维向量来描述:

$F_i = (\phi i1, \phi i2, ..., \phi iM)$，

其中，ϕij =1，若关键词K_j属于文档F_i，否则 ϕij =0，i=1,2,...,N; j= 1,2,...,M 。

对每一关键词K_1，K_2，...，K_M，预先指定它是属于某一类或多类。统计K_j在第l类中出现的概率P_{jl}为:

$P_{jl}=L_{jl}/ L_l$, j = 1,2,...,M; l = 1,2,...,S,

此处，L_l是K_j在l类文档中出现的总次数；L_{jl}是K_j认定属于l类的次数，S为分类总数。文档F_i属于l类的隶属函数可由下式给出:

$u_l(F_i) = \Sigma_{j=1,M \phi lj} P_{lj} / \Sigma_{j=1,M \phi lj}$, l = 1,2,...,S; i = 1,2,...,N。

信息检索

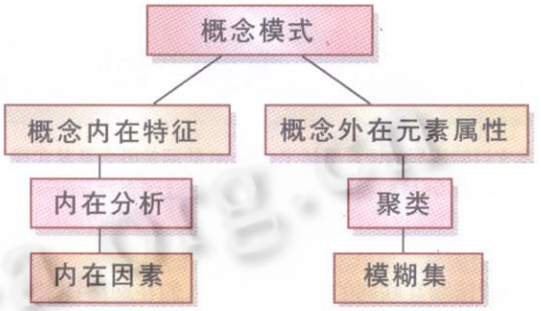
本系统网上信息检索思想，抛开了传统的对检索词表义进行查词的模式，注重对检索词的内涵意义(概念)，即词义进行查询检索。一个检索词的内涵意义(概念)可以从三个方面描述:

(1) 内在特征: 用概念内在特征来描述概念

(2) 外在元素: 用符合概念的全体元素来描述概念

(3) 模式: 用该概念与其他概念的相互关系来描述该概念（即概念模式）

内在因素用来描述概念的内在特性；而模糊集则用来描述元素属于概念的程度。



念的外在元素属性。

若（x）是元素x对概念A的属性，则A={x|x ∈ X，A（x）成立}

即是一聚类。即集合A是概念A的外在元素属性。其中X是给定的一个论域（特定领域），也称为全集。

若X是给定的全集，那么P（x）={A|A⊆(X}为X上的幂集，幂集是全集中所有可能的子集构成的集合。

给定集合A ∈ P（x），定义A的特征函数为如下映射: x_A: X → {0，1}

对于X中的一个元素x，有x =

那么聚类A可以用下式表示:
$$\begin{cases} 0 & \text{若 } x \in A \\ 1 & \text{若 } x \notin A \end{cases}$$

A={x(X | x_A（x）=1}

(x，xA（x)) ←→ A，∀ x ∈ X

一般聚类的运算实质上是全集中所有元素的隶属度的运算组合。

检索查询实质是对聚类的交、并运算。用户输入自己的查询需求，选择某种捕捉服务方式、查询的领域范围（如教育、体育、经济、政治、文化、艺术、娱乐、旅游、机构、服务、科研、医疗保健、社会、新闻出版、工商等）以及镜象内容的分类方式（按标题或文本内容）。检索系统则对查询需求进行逻辑分解、分类、然后查询信息索引库表，再对应镜象的标题或文本内容找到相应的信息。最后将捕捉到的信息直接显示给用户；或者将信息定期发送到用户的电子邮件箱中；或者将信息保存在用户的个人文件箱中，并通知用户信息已检索到；或者让用户自己去选择查看结果信息。■