文章编号:1000 - 582X(2004)07 - 0090 - 04

# 一种适合于专题式元搜索引擎的信息检索策略*

吕传宇[1],李 华[1],耿 虎[2]
(1. 重庆大学 计算机学院,重庆 400030; 2. 省泰兴市黄桥电视台,江苏 泰兴 225400)

**摘 要**:现有的元搜索引擎技术是基于关键词的信息检索,在查找某一专业知识时,简单的关键词组合不能真实地反映用户的检索意图,导致在检索中大量无关的信息被返回,专题式的元搜索引擎较好地解决了这一问题。本文通过对现有的元搜索引擎技术、领域知识库等研究,提出了一种适合于专题式元搜索引擎的信息检索策略,提高了检索的效率与精度。本文着重介绍了该检索策略的核心思想及其关键技术。

**关键词**:专题式;元搜索引擎;搜索引擎;检索策略
**中图分类号**:TP393 **文献标识码**:A

随着信息技术和互联网技术的发展,Internet 已经成为拥有 400 万站点和 3 亿页面的分布式信息空间,为了从纷繁芜杂的信息海洋中挖掘出有用的信息,出现了一批具有典型代表的搜索引擎,如 Yahoo、Google 等。由于各个搜索引擎实现技术不同,各自存在着检索性能、效率、质量上的优缺点,为了达到全面、准确的检索效果,元搜索引擎技术应运而生。元搜索引擎的特点是通过调用多个搜索引擎,并对检索结果进行分析处理得到期望的结果。但是,由于元搜索引擎检索的涵盖范围比独立的搜索引擎更广,用户只通过几个关键字布尔组合在巨大信息空间中进行查找,因此在对专题性、领域性目标进行检索时,就很难达到预期的效果,太多无关的信息被返回给用户[1]。

通过分析可以发现,现有元搜索引擎技术缺乏知识处理能力和理解能力,其核心采用的是"以词对网"的检索策略,即在拥有巨大信息量的互联网中,以关键词作为网络信息查询的入口进行检索。解决问题的根本和关键是构建专题式的元搜索引擎,即在现有的元搜索引擎技术基础上,通过引入领域知识库,把信息检索从目前的基于关键词的点提高到基于领域知识关联的面,实现对检索条件由点到面的转变,使系统能够"真正"理解用户的搜索意图,从而提高检索的效率与精度。

## 1 元搜索引擎的定义及运作机制

### 1.1 元搜索引擎的定义

搜索引擎是 Internet 上进行信息检索的工具,它向用户提供一个信息检索的接口,根据用户的检索请求,返回用户需要的信息,一般所说的搜索引擎是指的独立搜索引擎。而元搜索引擎是一种基于搜索引擎的搜索引擎,它由多个独立搜索引擎构成,在本文中将元搜索引擎内部的独立搜索引擎称为成员搜索引擎。

### 1.2 元搜索引擎的结构及运行机制

元搜索引擎主要由用户提问处理、搜索引擎调度及指令转化、检索结果排序处理及结果统一定制 4 个部分组成[2-3],如图 1 所示。结合图 1,分析元搜索引擎的运行机制如下:

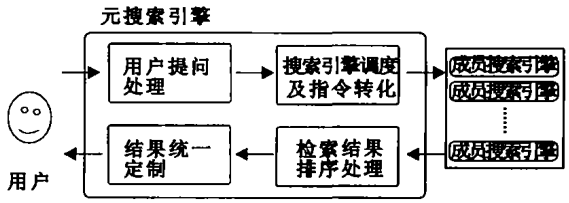


图 1 元搜索引擎结构图

用户的检索请求通过"用户提问处理"部件统一查询接口,在该部件中,用户可以指定成员搜索引擎,可以对检索结果进行约束和限制,该部件自动对用户的检索请求进行关键词切分。"搜索引擎调度及指令转化"部件,负责根据用户及系统对成员搜索引擎进

* 收稿日期:2004 - 03 - 10
作者简介:吕传宇(1979 - ),男,贵州兴义人,重庆大学硕士研究生,主要研究方向计算机网络、现代远程教育系统。

行统一的调度，并将检索请求转化为适合各个成员搜索引擎的特定指令。各个成员搜索引擎根据检索请求，返回相关信息，“检索结果排序处理”部件负责综合各个成员搜索引擎的搜索结果。“结果统一定制”部件负责将成员搜索引擎的检索结果统一呈现给用户。

## 2　专题式元搜索引擎检索策略的核心思想

一般的元搜索引擎在检索专业信息时，使用基于关键词的检索策略，会导致太多无关的信息被返回，使系统的查准率降低，为了提高检索专业信息的效率和精度，需要使用专题式的元搜索引擎。普通的元搜索引擎与专题式元搜索引擎在结构上最大的不同是，后者引入了领域知识库，知识库的核心是一个领域内的知识及其内在的联系的表示，笔者在研究知识库的基础上，提出了一种适合专题式元搜索引擎的信息检索策略，其核心思想如图 2 所示。

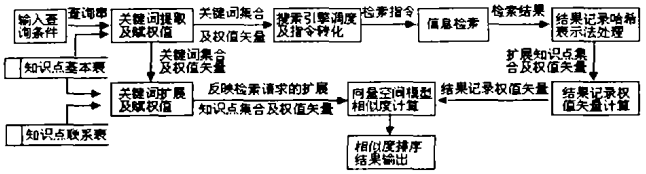


图 2　专题式元搜索引擎检索策略核心原理图

## 3　领域知识库中知识及其内在联系的表示

一本教科书的知识由若干知识点组成，同样，领域化、专题化的知识也可以通过知识点及其联系进行描述，因此，创建领域知识库必须以知识点及其内在的联系为基准。

### 3.1　知识点表示

根据知识点所处的层次进行划分[4]，知识点可以划分元知识点和复合知识点。

**定义 1**　元知识点　任何单独的一项知识称为元知识点，具有不可再分的特性。

**定义 2**　复合知识点　由相关的一组知识点及其内在联系组成的知识点。

**定义 3**　元知识点的结构，主要可分为标识、主题、内容、学科分类、检索词及知识点权值。标识是识别知识点的唯一标志，主题为知识点的主标题，是对本知识点的知识的一种高度概括的表述，内容是知识点的讲述内容，其中可含有任意的图片、音频、视频等多媒体资料，学科分类是本知识点的分类属性，可作为检索途径之一，检索词也是本知识点的检索途径之一，知识点权值表示知识点的重要程度，它的取值与知识点间关系有关。

在领域知识库中，为了描述知识点，可以创建知识点基本表，字段为{KB_ID，KB_TITLE，KB_EXPLAIN，KB_TYPE，KB_WORDS，KB_VALUE}，分别与定义的元知识点结构一一对应。

### 3.2　知识点间的关联关系

知识点之间的关系错综复杂，若干相关的知识点按其内在联系构成的网络称为知识点网络。网络的节点表示知识点，节点间的链接表示知识间的联系，其权值根据关系类型确定。由于两个知识点都不直接或间接相互支持，所以，知识点网络中任何两个节点都不会在一条有向环路上，因此知识点网络实质上就等价于一个带权的有向无环图。通过对有向图中的节点间的关系进行分析，可以将知识点间的联系分为两大类，绝对依赖关系和相对依赖关系。

**定义 4**　绝对依赖关系　两知识点之间形成一对一的关系，在有向图中表示为两节点间，一个节点只有一条出向链接，另一个节点只有一条入向链接。

绝对依赖关系可以细分为唯一支持关系和绝对游离关系。唯一支持关系指的是两知识点满足绝对依赖关系，且都属于领域内知识。绝对游离关系指的是两知识点满足绝对依赖关系，但其中一个知识点不属于领域内知识。

**定义 5**　相对依赖关系　$N$ 个($N>1$)知识点与一个知识点形成多对一关系或一对多关系，称为相对依赖关系。

相对依赖关系可以细分为多对一支持关系、一对多支持关系、推荐关系、相对游离关系。多对一支持关系是指要学习某一知识点与在学习该知识点前需要学习的知识点集合之间的关系；一对多支持关系指某一知识点与学习完该知识点后可以学习的知识点集合之间的关系；推荐关系是某一知识点与该知识点推荐学习的知识点集合之间的关系；相对游离关系指的是某一领域内知识点与其它领域的知识点集合之间的联系。

为了反映知识点对知识点支持的重要程度，笔者针对每一种关系引入了“关系权值”。在构建立知识点间联系时，对每一种关系都赋予默认权值，权值在 0 与 1 之间，领域专家可以在建库时对各类关系的“关系权值”进行初始化。

通过以上分析，描述知识间的联系需要的信息有关系标识、起知识点标识、终知识点标识，关系的类型、关系权值。因此，可以创建知识点联系表，它主要的字段有{KR_ID，KR_KID1，KR_KID2，KR_TYPE，KR_VALUE}，它们分别对应的是标识号，起知识点的标识号、终知识点标识号、关系类型、关系权值。

通过知识点联系表，可以构建知识点网络。例如，截取领域知识库中的知识点联系表的一个片断，结果可能如图 3 所示。

为了计算知识点的权值，设领域库中的知识点总

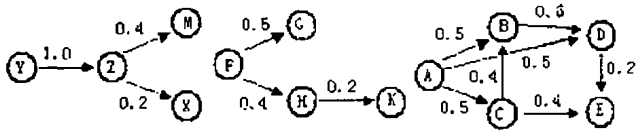


图 3 领域知识库片断形成的知识点网络图

数为 $n$，使用 $v_i$ 表示网络节点，$E(v_i,v_j)$ 表示知识点 $v_i$ 与知识点 $v_j$ 之间的“关系权值”，其中 $i,j=1,2,\cdots,n$。

**定义 6** 知识点权值 表示知识点的重要程度，使用 $V(v_i)$ 表示知识点 $v_i$ 的权值，其计算方法如下：

$$V(v_i)=\begin{cases}0.2, & \text{当 } v_i\in S,(i=1,2,\cdots,n)\text{ 时}\\ \max\{E(v_i+v_j)+V(v_j)\}, & \text{当 } <i,j>\in T,(i,j=1,2,\cdots,n)\text{ 时}\end{cases} \quad (1)$$

集合 $T$ 表示所有以 $v_i$ 为头的弧的集合，集合 $S$ 表示所有叶子节点集合。求解知识点权值的过程就是利用式(1)，从叶节点开始逆向递推至所有节点，从而求出所有知识点的权值。

在领域知识库中构建如上所述的知识点基本表和知识点联系表，较好地解决了知识点及知识点间联系表示的问题，为检索请求实现“由点到面”的扩展提供了有力的支持。

## 4 检索请求“由点到面”的扩展策略

对用户使用自然语言描述的检索请求，系统以知识点基本表中知识点的检索词为基准，使用字符串最大正向匹配法来提取关键词，初步得到代表用户检索请求的关键词集合 $Keys$，即 $Keys=\{key_1,key_2,\cdots,key_n\}$ 以及对应知识点的默认权值矢量，通过用户对每个关键词及权值进行确认和修改(注：权值在 0 与 1 之间)，最后得到关键词集合 $Keys$ 的检索权值矢量表示。检索权值是指在具体的检索中，某一关键词在该检索过程中的重要程度[5]。

然后，使用知识点联系表，将关键词集合进行“由点到面”扩展，其核心思想是通过知识点联系表按照两种方式搜索与关键词集合有直接关联关系的知识点，一种方式是以关键词集合中的关键词为链接终点进行检索，另一种是以关键词为链接起点进行检索，从而可以获得扩展知识点集合，并根据关键词的权值、扩展知识点自身的权值、以及关系权值等参数对扩展知识点的权值进行综合计算，最后得到扩展知识点集合检索权值。图 4 为扩展策略的实例图形表示。

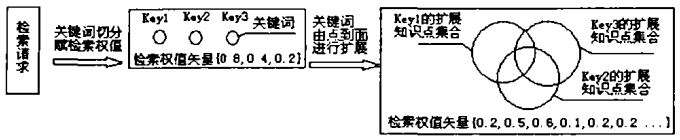


图 4 检索请求的扩展策略

### 4.1 计算扩展知识点集合

$Se(v)$ 表示以知识点 $v$ 为链接终点的扩展知识点集合，用 $Ss(v)$ 表示以知识点 $v$ 为链接起点的扩展知识点集合，$V'(v)$ 表示知识点 $v$ 作为扩展知识点时的检索权值，$Ses$ 表示以 $Keys$ 为终点的扩展知识点集合，$Sss$ 表示以 $Keys$ 为起点的扩展知识点集合。集合 $Se$ 和集合 $Ss$ 可以表示为：

$$Se(key_i)=\{Ekey_i1,Ekey_i2,\cdots,Ekey_ix\},key_i\in Keys \quad (2)$$

其中，$x$ 表示以 $Key_i$ 为链接终点的扩展知识点的个数。

$$Ss(Key_i)=\{Skey_i1,Skey_i2,\cdots,Skey_iy\},key_i\in Keys \quad (3)$$

其中，$y$ 表示以 $Key_i$ 为链接起点的扩展知识点的个数。通过 $Se$ 和 $Ss$ 的定义得到：

$$Ses=\cup_{i=1}^{n}Se\{Key_i\}-\cap_{i=1}^{n}Se\{Key_i\},key_i\in Keys \quad (4)$$

$$Sss=\cup_{i=1}^{n}Ss\{Key_i\}-\cap_{i=1}^{n}Ss\{Key_i\},key_i\in Keys \quad (5)$$

扩展知识点集合用 $Ekeys$ 表示，于是可以得到：

$$Ekeys=\{Ekey_1,Ekey_2,\cdots,Ekey_z\}=(Ses\cup Sss)-(Ses\cap Sss) \quad (6)$$

其中，$z$ 为扩展知识点集合中不同知识点的个数，即 $z=|Ekeys|$。

### 4.2 计算扩展知识点检索权值矢量

由于扩展知识点的检索权值与关键词的检索权值成正比，与扩展知识点的个数成反比，并与该知识点自身的权值以及该知识点与关键词之间“关系权值”有直接关系，因此，可以得到如下的计算规则：

$$V'(Ekey_ij)=w_i\times(V(Ekey_ij)+E(Ekey_ij,Key_i))/|Se(Key_i)|,j=1,2,\cdots,m \quad (7)$$

$$V'(Skey_ij)=w_i\times(V(Skey_ij)+E(Keyi,Skey_ij))/|Ss(Key_i)|,j=1,2,\cdots,y \quad (8)$$

其中，$w_i$ 表示关键词 $Key_i$ 的检索权值，$|Se(Key_i)|$，$|Ss(Key_i)|$ 表示集合中元素的个数，$V(Ekey_ij)$ 和 $V(Skey_ij)$ 表示知识点 $Ekey_ij$ 和 $Skey_ij$ 在领域知识库中的权值，$E(Ekey_ij,Key_i)$ 表示 $Ekey_ij$ 与 $Key_i$ 在知识库中的关系权值，$E(Key_i,Skey_ij)$ 表示 $Keyi$ 与 $SKey_ij$ 的关系权值。

由于在 $N(n\geq N\geq 2)$ 个不同的关键词的扩展知识点集合中有可能出现相同的知识点，于是通过式(7)和(8)的计算就会出现同一知识点对应 $N$ 个不同的检索权值的情况，解决冲突的办法是，选择最大的检索权值作为该知识点的最终检索权值。于是可以得到扩展知识点集合对应的检索权值矢量 $\boldsymbol{EW'}$。

$$\boldsymbol{EW'}=\{V'(Ekey_1),V'(Ekey_2),\cdots,V'(Ekey_z)\},\text{其中 } Ekey_i\in Ekeys \quad (9)$$

将 $\boldsymbol{EW'}$ 归一化处理后得到检索权值矢量 $\boldsymbol{EW}$。

通过“由点到面”的扩展策略，得到的结果有：代表检索请求的关键词集合 *Keys* 及其对应的检索权值矢量 $\boldsymbol{UW}$，以及其对应的扩展知识点集合及其对应的检索权值矢量 $\boldsymbol{EW}$，将二者进行合并，可以得到关键词及扩展知识点集合的并集 $V_1$ 及其对应的检索权值矢量 $V_2$，从而得到了代表检索请求的知识点网络。其中，$V_1$、$V_2$ 如下所示。

$V_1 = \{c_1, c_2, \cdots, c_m\} = Keys \cup Ekeys$，其中 $m$ 为集合中 $V1$ 中知识点个数。

$V_2 = \{w_1, w_2, \cdots, w_m\} = \boldsymbol{UW} \cup \boldsymbol{EW}$

## 5　结果记录的矢量表示与相似度计算

### 5.1　检索结果记录的矢量表示

为了将成员搜索引擎返回的结果记录进行矢量表示，对返回结果进行了分析。返回结果中的记录一般由标题文字和摘要说明及超级链接组成，为了分析每一条记录与检索请求的相似度，将对结果记录逐条进行处理，其过程如图 5 所示。

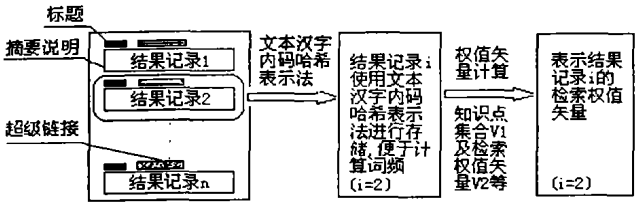


**图 5　结果记录的矢量表示处理过程**

经过“由点到面”的扩展策略得到了代表检索请求知识点矢量 $\boldsymbol{V}_1$ 和检索权值矢量 $\boldsymbol{V}_2$，为了计算结果记录的检索权值矢量，将知识点矢量 $\boldsymbol{V}_1$ 中的各知识点在结果记录中出现的频率定为权值。由于标题是对结果记录的高度概括，因此对标题和摘要说明分别处理。系统采用文本的汉字内码哈希表示法[6] 对标题和摘要说明进行存储，因此可以高效地计算出标题文字及摘要文字的检索权值矢量 $\boldsymbol{Ut}$ 及 $\boldsymbol{Ua}$，即 $\boldsymbol{Ut} = \{a_1, a_2, \cdots, a_m\}$，$\boldsymbol{Ua} = \{b_1, b_2, \cdots, b_m\}$，其中 $a_i, b_i$ 表示知识点 $c_i$ 出现的次数。

为了对结果进行归一化处理，引入摘要文字的权重之和 sum$A$ 和权重因子 $\sigma$。

其中，$\text{sum}A = \sum_{i=1}^{m} b_i$，$\sigma = \text{sum}A/m$。

由于标题是文本的高度概括，因此对标题部分所获得的权值矢量 $\boldsymbol{Ut}$ 进行强化处理即 $\boldsymbol{Ut}' = \sigma \times \boldsymbol{Ut} = \{\sigma \times a_1, \sigma \times a_2, \cdots, \sigma \times a_m\}$。

标题文字的权重之和 $\text{sum}T = \sigma \times \sum_{i=1}^{m} a_i$。

对 $\boldsymbol{Ut}'$ 及 $\boldsymbol{Ua}$ 进行归一化处理后得到归一化的标题部分知识点权值矢量 $\overline{\boldsymbol{Ut}'}$ 和摘要说明部分的知识点权值矢量 $\overline{\boldsymbol{Ua}}$。于是可以得到，结果记录的归一化的检索权值矢量 $\boldsymbol{U}$。

$$\boldsymbol{U} = \{y_1, y_2, \cdots, y_m\} = \overline{\boldsymbol{Ut}'} + \overline{\boldsymbol{Ua}} = \left\{\frac{\sigma \times a_1}{\text{sum}T} + \frac{b_1}{\text{sum}A}, \frac{\sigma \times a_2}{\text{sum}T} + \frac{b_2}{\text{sum}A}, \cdots, \frac{\sigma \times a_m}{\text{sum}T} + \frac{b_m}{\text{sum}A}\right\} \tag{10}$$

对检索请求得到的检索权值矢量 $\boldsymbol{V}_2$ 归一化处理，得到矢量 $\boldsymbol{V}$，令 $\text{sum}W = \sum_{i=1}^{m} w_i$，

$$\boldsymbol{V} = \{z_1, z_2, \cdots, z_n\} = \left\{\frac{w_1}{\text{sum}W}, \frac{w_2}{\text{sum}W}, \cdots, \frac{w_m}{\text{sum}W}\right\},$$

$$\text{其中，} 0 \leqslant \frac{w_i}{\text{sum}W} \leqslant 1 \tag{11}$$

### 5.2　向量空间模型进行相似度计算

通过前面的分析和计算，可以得到反映检索请求的矢量 $\boldsymbol{V}$ 和反映某一结果记录的矢量 $\boldsymbol{U}$，因此，两者的相似程度可以使用向量空间模型进行计算，即用两个向量间的夹角余弦值来度量，夹角越小说明相似度越高。相似度计算公式如下：

$$\text{Abstract\_sim}(V, U) = \cos(V, U) = \left(\sum_{i=1}^{n} zi \times yi\right) \Big/ \left(\sqrt{\sum_{i=1}^{n} (zi)^2} \sqrt{\sum_{i=1}^{n} (yi)^2}\right) \tag{12}$$

通过式(12) 可以计算出每个成员搜索引擎所返回的结果记录与检索请求之间的相似度。系统设置了一个合理的相似度阈值来区分相关记录与不相关记录，并将相关记录按照相似度由高到底进行排序呈现给用户，从而有效地提高了元搜索引擎的检索精度。

## 6　结束语

笔者所提出的专题式元搜索引擎的检索策略能够快速、准确地收集本专业内的专业信息，摒弃无关的信息，提高检索效率和精度。如果需要构建另一个专题的元搜索引擎，只需重新构建领域知识库即可，因此专题式元搜索引擎具有普遍的适用性。

**参考文献：**

[1] 刘丽，孙燕唐. 智能型元搜索引擎的设计与实现[J]. 计算机工程，2003，19(16)：118－121.

[2] 王芳，张晓林. 元搜索引擎原理与利用[J]. 现代图书情报技术，1998，(6)：18－22.

[3] 李广建，黄崑. 元搜索引擎及其主要技术[J]. 情报学报，2002，20(2)：175－179.

[4] 谢深泉. 知识点及其网络的特性分析[J]. 软件学报，1998，9(10)：785－789.

[5] 李振东，费翔林. 基于概念的信息检索模型研究[J]. 南京大学学报，2002，38(1)：108－109.

[6] 柳泉波，黄荣怀，何克搞. 智能答疑系统的设计与实现[J]. 中国远程教育，2000，8：43－48.

(下转第 129 页)

104 - 105.
[5] 包世华. 结构力学Ⅱ[M]. 北京:高等教育出版社,2001.
[6] 袁驷,张亿果. 常微分方程特征值问题的求解器解法[J]. 地震工程与工程振动,1993,13(2):94 - 102.
[7] 龚耀清,包世华,龙驭球. 半无限大弹性地基上变截面筒中筒高层建筑结构的自由振动[J]. 工程力学,1999,16(3):7 - 14.

# Analytic Method for the Free Vibration of Plane Frame Structure

*ZHOU Mao-sen, CHEN Zhao-hui*
(College of Civil Engineering, Chongqing University, Chongqing 400030, China)

**Abstract**: The analytic method for the free vibration of plane frame structure is presented. It is achieved by transforming the general eigenvalue problem of natural frequency and vibration mode of continuously distributed property system into typical boundary value problem of ordinary differential equation (ODE). A series of ODE, corresponding to the general eigenvalue problem, is formed, and then is solved by an general ordinary differential equation solver - COLSYS. Each component of a frame structure is regarded as an element in the analytic method. This makes it more efficient than that of finite element method which needs to increase elements to get the value of high - class natural frequency and vibration mode accurately. Some applications of the method show that it can solve the free bending vibration of plane - frame structure with various types of displacement constraint. The result indicates that this method is precise and efficient.
**Key words**: free vibration of plane frame structure; boundary value problem of ordinary differential equation; ODE solver

(编辑 姚 飞)

(上接第 93 页)

# Information Retrieve Strategy for Subject Meta Search Engines

*LV Chuan-yu*[1], *LI Hua*[1], *GEN Hu*[2]
(1. College of Computer Science, Chongqing University, Chongqing 400030, China;
2. Huangqiao TV Station, Taixing City, Jiangsu Province, Taixing 225400, China)

**Abstract**: Present meta search engines are key - based information retrieve. A plenty of irrelevant information is returned when user search some professional knowledge because Simple combination of key can not express accurately user's purpose. Subject meta search engines solve the problem. Based on the study of present technologies of meta search engines and domain knowledge base, an information retrieve strategy is put forward which is fit for subject meta search engines. It effectively improves efficiency and precision of information retrieve. The main content of this paper are to introduce the key idea and technologies of the retrieve strategy.
**Key words**: subject; meta search engines; search engines; retrieve strategy

(编辑 吕赛英)