面向领域资源的智能元搜索技术研究*)

苏 超[1,2] 蔡 铭[1] 姚玉荣

(浙江大学计算机学院 杭州 310027)[1] (杭州技师学院 桐庐 311500)[2]

摘 要 如何快速、有效地从互联网中获取特定领域资源信息,已成为当前研究热点之一。本文提出并构建了一种基于元搜索技术的领域资源检索系统,介绍了其中的关键技术,包括:领域资源模型构造、查询请求与调用、查询结果处理等。最后以计算机教学资源检索为例进行了实例验证。

关键词 元搜索引擎,分类系统,领域资源

Research of Intelligent Meta-search Technology for Domain Resources

SU Chao[1,2] CAI Ming[1] YAO Yu-Rong[1]

(College of Computer Science, Zhejiang University, Hangzhou 310027)[1] (Hangzhou Technician College, Tonglu 371500)[2]

Abstract The retrial of domain resources from internet is the hot spot of research currently. Based on meta-search technology, we develope a prototype system which is especially designed for the acquisition of domain resources. This paper mainly introduces the key technology about the system, including the domain resources model, inquiry requests transformation, remote retrial calling, and the result processing. Finally, an actual example of computer science resources retrieving for college teaching is given.

Keywords Meta-search engine, Classify system, Domain resources

1 引言

随着网络信息资源的迅速增加,综合性的搜索引擎在满足用户的专业搜索提问时面临新的挑战。就现有的搜索引擎来看,几乎没有任何一个综合性的搜索引擎能够很好地满足专业检索要求。为了满足用户的专业检索需求,必须设计一个针对某一学科或某一行业领域的搜索引擎,即专业元搜索引擎。

2 面向领域资源的智能元搜索引擎系统

目前一般的资源搜索引擎可以给资源检索者带来非常大的便利,但是由于单个搜索引擎所能索引处理的网页信息毕竟有限,而且由于技术上的难度,其搜索结果的查准率和查全率都很难在短时间内取得很大的突破。根据专家的评测,目前主要搜索引擎返回结果的相关比率不足45%,而且由于所采用的机制、算法与适用范围等不同,导致同一个检索请求在不同搜索引擎中的查询结果的重复率不足34%[1]。如果采用元搜索技术,综合运用多个成熟的搜索引擎,集成它们各自的技术优势和索引数据库,为用户提供一个统一的搜索接口,将领域资源集中在一起,将是一个比较好的解决方法,会有效地提高资源检索的查全率与查准率。

2.1 元搜索引擎简介

目前,搜索引擎已成为Internet研究的热点。元搜索引擎(meta-search Engine)被称作"搜索引擎之上的搜索引擎"或"搜索引擎之母"[2],在国外已得到了广泛研究,产生了像ProFusion、SavvySearch及MetaSeek等著名的元搜索引擎。元搜索引擎不同于传统的独立的搜索引擎,本身没有搜索引擎的网页搜寻机制,也没有自己独立的索引数据库,而是定制统一的检索界面,通过并行或串行地调用其他搜索引擎的检索功能来实现网络资源的查询,并将所有搜索引擎的查询结果集中起来,以统一的格式呈现到用户面前[3]。

2.2 系统结构

面向领域资源的智能元搜索引擎,其系统结构分为3个部分:搜索界面模块、搜索执行模块、搜索结果处理模块。系统结构如图1所示。

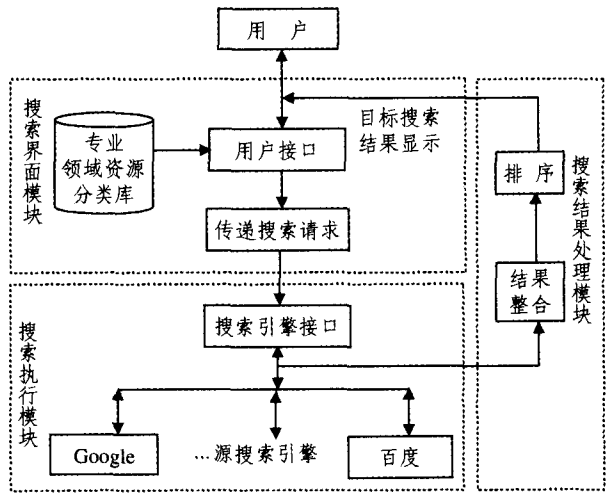

图1 面向领域资源的智能元搜索引擎系统结构图

2.3 系统主要功能模块

搜索界面模块:该模块是用户和搜索系统的中介,主要功能是按照专业领域资源的分类特点为用户提供一个友好的界

*)浙江省自然科学基金资助(项目编号:Y104435)。苏 超 讲师,主要从事计算机教学与研究,现浙江大学进修研究生;蔡 铭 工学博士,副教授,研究方向:语义网、ASP、嵌入式操作系统。

面，方便用户输入搜索需求；把搜索需求传递给下一个搜索执行模块；最后把经过处理的最终目标搜索结果再以友好的方式显示给搜索用户。

搜索执行模块：该模块首先获取搜索界面模块传递过来的搜索请求关键词，转换为各源搜索引擎的搜索格式，然后启动基于源搜索引擎的网络信息搜索蜘蛛（Spider），获取各源搜索引擎的搜索结果，并将获取的搜索结果存入系统数据库相应位置，备搜索结果处理模块使用。

搜索结果处理模块：按照系统实现的关键技术中介绍的四级结果集思想[4]来处理搜索结果。

3 系统实现的关键技术

3.1 领域资源模型

这里以计算机学科领域资源的分类为例，介绍一下领域资源模型的构造。我们采用《计算机科学与技术方法论》一书的附录中对计算机知识体的划分标准。该附录将计算机科学知识体分层组织成3个层次，即领域、单元、主题。最高一层是领域（area），代表一个特定的学科子领域，包括14个领域。每个领域用两个字母的缩写词来表示，比如OS代表操作系统。领域之下被分割成更小的单元（units），代表领域中单独的主题模块。每个领域都有若干单元，每个单元都用一个领域名加一个数字后缀表示，比如OS8表示文件系统的单元。各个单元又被细分成若干主题（topics），这是计算机科学知识体层次结构的最底层[5]。

我们可以用一个树状示意图来形象地表示这个分类体系，如图2。

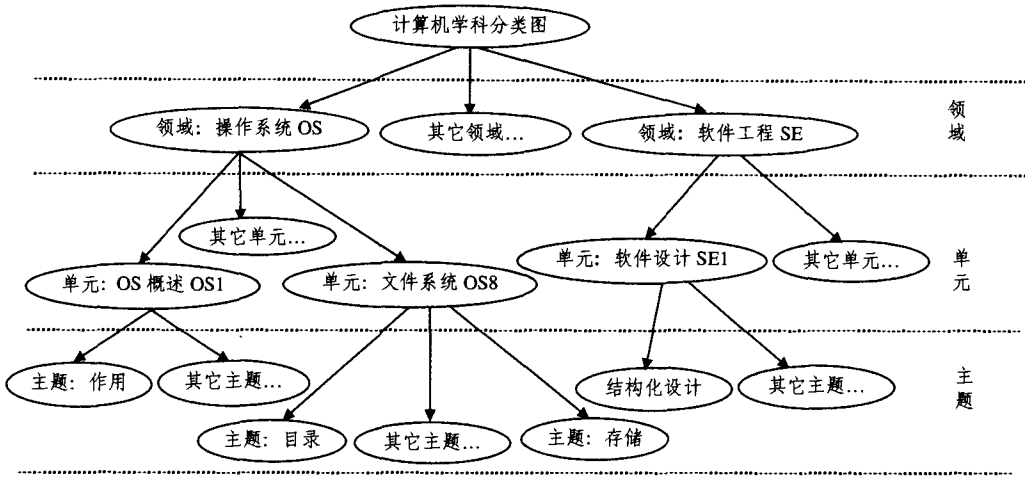

图2 计算机学科分类示意图

我们按照这样的分类标准，为用户提供一个统一的检索界面，在这个界面中为用户提供领域、单元、主题的选择框。同时按照计算机教学资源的用途，还提供了教案、课件、试卷、试题的类别选择。按照计算机教学资源的文件类型提供了资源文件类型的选择框，即Doc、PPT、Html、Gif等类型，同时允许用户不使用这些选择框而由自己输入关键字进行检索。

3.2 查询请求格式的处理及调用

3.2.1 查询请求的构成

这里的查询需求是指用于源搜索引擎搜索的最终搜索关键词（searchTerms），包括：搜索关键词（keywords）、领域（areaCategory）、单元（unitCategory）、主题（topicCategory）、资源用途类型（usageCategory）、资源文件类型（fileCategory），共6项。

用BNF描述查询请求（SearchTerms）的构成如下：

```
〈SearchTerms〉::=〈keywords〉〈areaCategory〉〈unitCategory〉〈topic-
                  Category〉〈usageCategory〉〈usageCategory〉〈file-
                  Category〉
〈keywords〉::=""|〈字符串〉
〈areaCategory〉::= ""‖〈字符串〉
〈unitCategory〉::= ""‖〈字符串〉
〈topicCategory〉::= ""‖〈字符串〉
〈usageCategory〉::= ""‖〈字符串〉
〈fileCategory〉::= ""‖"filetype:"〈字符串〉（注：当未选择资料文件
                  类型时取前者，选择时取后者）
〈字符串〉::=〈字母〉|〈汉字〉|〈字符串〉〈字母〉|〈字符串〉〈汉字〉
〈字母〉::= 〈大写字母〉 | 〈小写字母〉
```

3.2.2 请求格式的处理

请求格式的处理主要是进行查询请求的转换，需要根据不同的搜索引擎转换成可以进行检索的查询表达式，这里采用"Get"方式。就是将检索请求数据加到URL后，格式为"?关键词=数据"。这种方式的检索是元搜索引擎最容易实现的，需要根据源搜索引擎的搜索格式将最终搜索关键词转换成相应的格式，即基于源搜索引擎的格式化搜索请求。每个源搜索引擎都有其特有的搜索格式，要首先分析获取这种搜索格式，然后把最终搜索关键词转换为对应的格式。

以goolge、百度、雅虎中国为例：

```
www.google.com/search? q={searchTerms}
http://www.baidu.com/s? wd={searchTerms}
http://www.yahoo.com.cn/search? p={searchTerms}
```

3.2.3 基于源搜索引擎的Spider

这里的Spider是为了搜集各个源搜索引擎对于特定搜索关键词的返回结果，即面向源搜索引擎的，而不像一般Spider系统面向整个Web沿着遇到的各个链接去搜集所有的网页信息，所以它在遵循一般Spider设计准则的基础上更多地考虑了源搜索引擎的特点，即基于源搜索引擎的Spider[6]。

3.2.3.1 源搜索引擎返回结果页面分析

各个源搜索引擎的返回结果显然是不完全相同的，同时返回结果页面的格式（这里特指HTML标签）也是有差异的，需要通过一定的技术手段来解析各个源搜索引擎返回结果页面的格式，以便对结果页面进行分析处理。虽然网页HTML标签格式存在着差异，但各个源搜索引擎返回结果页面的内容模块却是大致相同的。搜索网页中都包括3类结果：即相关广告、返回结果条目、分页显示链接URL等。在抽取信息时，只抽取返回结果条目及分页显示链接URL两类结果信息。

(1)相关广告:无论是在页面左侧、右侧还是其它地方出现的相关广告,都是无用信息。我们针对具体搜索引擎的页面格式进行分析,根据 HTML 标签及某种识别规则(这里采用正则表达式)分别制定抽取规则来识别出广告信息,不抽取此类广告信息,这样就自然地过滤了广告。

(2)返回结果条目:这些条目构成了搜索结果的正文内容。每个条目就是一个结果集,包括标题(title)、摘要信息(abstract)及源文件所链接 URL。这是我们需要分析处理的内容部分,而且每个结果页面上都会有若干个结果条目,一般为 10 个,这由不同的源搜索引擎而定。

(3)分页显示链接 URL:由于现在多数搜索引擎发展得都很成熟,其数据库也很丰富,因此对于每个搜索需求一般都会有大量的返回结果,而且都采取分页显示的方式,即每页显示一定数目的返回结果条目。这样,每个结果页面就包含有指向分页显示的链接(属于内部链接),这些指向分页显示的内部链接也是我们要分析获取的。

3.2.3.2 基于源搜索引擎 Spider 设计

这里设计的 Spider 利用了队列法和多线程技术,并针对源搜索引擎的特点进行了一些特别的设计。

(1)多线程:由于各个源搜索引擎的搜索过程是独立的,我们可以利用多线程技术给每个源搜索引擎开设一个 Spider 线程,让它们并行搜索、处理结果,这样就加快了搜索和处理的速度。

(2)队列:本系统只需要获取一类内部链接——分页显示链接;而且分页显示链接结构简单,每个源搜索引擎都有固定的格式,一般情况下都是有效的,所以我们只需要一个队列——待处理队列。

(3)生成包装器抽取搜索结果内容:这里的包装器主要用来识别 HTML 标签并抽取相应的内容,我们针对每个 SSE(源搜索引擎)返回结果页面的 HTML 标签的特点分别设计相应的抽取规则(正则表达式),用来识别出相应信息所在的 HTML 标签并抽取出返回结果条目内容及分页显示链接 URL,存入系统数据库的相应位置,备搜索结果处理模块使用。

3.3 结果处理模型

3.3.1 四级结果集思想

• 元搜索引擎要处理的是非常庞大的结果集(多个源搜索引擎 SSE 返回的结果集)

• 我们需要精度更高的算法来对这个庞大的结果集进行排序(一般按相关度进行排序),但这会耗费较多的时间,从而造成搜索速度较慢。

• 设置多级结果集,在庞大的较低结果集中用低级快速的算法挑选出一部分好的结果,构成较高级的结果集;在较高级的结果集中应用较高级(精度高)的算法,形成更高级的结果集。

• 目标是提高排名在前的少数结果的准确度(与用户查询相关度较高,提高用户查询的满意度)。

• 依据是用户最终可能浏览的网页少于 100 个。

3.3.2 四级结果集的实现

(1)符号约定

• S_i 表示元搜索引擎调用的第 i 个源搜索引擎,假设共调用 M 个源搜索引擎;

• R_i 表示第 i 个源搜索引擎返回的结果的结果集;

• W_i 为 S_i 的权重;

• N_i 表示从 R_i 中根据 W_i 提取的结果的数量。

(2)结果提取

各个源搜索引擎返回结果的集合,称为一级结果集。

① 对每个独立搜索引擎 S_i 赋以权重 W_i(注意此处源搜索引擎赋给的权值越小,代表此源搜索引擎的权威度越高);

② 将每个 R_i 中第 N_i 个结果后面的记录全部删除,并按照综合相关度排序中的方法计算出每个 R_i 中前 N_i 个结果的排序权值,此时各个源搜索引擎返回结果的集合称为二级结果集;

其中:

$$N_i = 0.1 * |R_i| * w_i / \sum_{i=1}^{M} w_i$$

$|R_i|$ 表示集合 R_i 的基数。

(3)结果排序

按照二级结果集中每个返回结果的排序权值的升序进行排序(排序权值越小,代表返回结果的相关度越大),并取出前 n 个结果形成三级结果集。其中:

$$n = c * \sum_{i=1}^{M} N_i$$

n 的大小由常数 $C(<1)$ 决定,我们取 0.8。

(4)去重、消除死链接

去除三级结果集中的死链接与重复记录,最终形成四级结果集(最终以一定的显示方式返回给用户)。

具体的操作为:先按照排序权值从大到小依次测试当前三级结果集中的记录。如果链接无效,直接从三级结果集中删除,直到测试到第一个链接有效的记录,再把该记录保存到四级结果集中。同时,在三级结果集中删除该条记录,然后依次取出当前三级结果集中权值较大的有效链接记录(无效链接直接从三级结果集中删除),与四级结果集中的每条记录进行比较。如果有相同 URL 的记录,视为重复结果,只保留排序权值较高的记录。

3.4 综合相关度排序

一般来说,成熟的源搜索引擎都有自己的搜索结果排序方法,比如 Google 的 PageRank 算法、百度的超链接算法。对于元搜索引擎来说,如何对各源搜索引擎返回的结果进行综合排序,关系到该元搜索引擎的搜索效率、查准率等指标。

本元搜索引擎的排序采用下列方法:首先评估每个源搜索引擎的权威度,综合该 SSE 的服务质量、搜索的查准率和查全率、用户的口碑等指标,给其一个评分权值(注意此处源搜索引擎赋给的权值越小,代表此源搜索引擎的权威度越高),记第 j 个 SSE 的权值为 W_j;然后记录每个 SSE 的返回结果在本 SSE 中的排序位置,记每个 SEE 结果集中第 i 返回结果的排序位置为 L_i(其值从该返回结果的相应字段取得);最后,用 $S_i = W_j \times L_i$(S_i 代表第 j 个 SSE 中的第 i 个返回结果的排序权值)所得的 S_i 作为该返回结果的最终排序权值。

计算最终排序权值的流程图如图 3 所示。其中,

SSEResultsTable:为一级结果集数据表;

LocationNum:为 SSEResultsTable 数据表(一级结果集数据表)中的一个字段,表示返回结果中的一条记录在其所属 SSE 中的排序位置;

authorityWeight:源搜索引擎列表 SSETable 数据表中的一个字段,代表 SSE 的权威度,比如 Google 为"1";

SSEID:SSE 的标识 ID,表示每条记录所属的 SSE,根据此字段来关联 SSEResultsTable 与 SSETable 两个数据表,进

(下转第 130 页)

```
postprune ()
}
```

结论 针对目前的知识发现过程模型在实际应用中存在挖掘周期长、对大型数据库的知识发现支持不够的问题，提出了基于数据抽取器的知识发现模型。该模型利用标准的SQL语言构造数据抽取器，为不同的学习算法准备数据，数据挖掘算法可以通过数据抽取器从DBMS中简单地得到所需要的统计值，数据抽取器可以针对不同的算法抽取特定的数据，减少数据挖掘算法对数据库直接调用的次数，避免了直接对大型数据库的数据进行调用，使得对大型数据库进行快速数据挖掘成为可能。可以加快知识发现过程，提高数据挖掘效率，实现对于大型数据库的知识发现。

为数据抽取器定义了统一的接口。任何一个使用该接口的学习算法都可以用不同的数据抽取器进行数据抽取，不同的学习算法也可用同一数据抽取器进行知识提取，从而实现多目标学习。最终用户可以通过定义数据抽取器的方法将学习任务所涉及到的可能的数据范围设定，这样数据集中就包含了最终用户的一些背景知识和经验。这样一方面最大程度地利用最终用户的经验知识；另一方面，由于给出了数据范围，学习算法处理的数据量减少，有利于提高学习速度。

最后设计了SQL- C4.5算法，该算法实现了利用数据抽取器为决策树算法C4. 5抽取必要的统计数据，实现了C4.5决策树的构建。

参考文献

1 Quinlan J R. Expert systems Discovering rules by induction from large collections ofin the Microelectronic Age. Edinburgh University Press,1979
2 Berson A,Smith S J. Data warehousing, Data Mining,&OLAP. McGraw-Hill Co.,1997. 113～150,333～514
3 Scotney B,McClean S. Efficient knowledge discovery through the integration ofheterogeneous data. Information and Software Technology,1999,41:569～578
4 Quinlan J R. C4. 5: Programs for Machine Learning. Morgan Kaufmann, 1993
5 Pei J,Han J,Mortazavi-Asl B,Zhu H. Mining Access Pattern efficiently fromWeb logs. In: Proc. 2000 Pacific Asia Conf. on Knowledge Discovery and Data Mining (PAKDD'00), Kyoto, Japan, April 2000
6 William J. Long, Medical informatics: reasoning methods. Artificial Intelli- gencein Medicine,2001,23:71～87
7 Ziarko W,Yao Y Y,Rough Sets and Current Trends in Computing (RSCTC 2000). Berlin: Springer-verlag, 2001
8 Wang Jue,Wang Ju. Reduction algorithms based on discernbility matrix: the ordered attributes method. Journal of Computer Science & Technology, 2001, 16(6):489～504
9 Wang Jue,Wang Ju. Reduction algorithms based on discernbility matrix: the ordered attributes method. Journal of Computer Science & Technology, 2001, 16(6):489～504
10 Lin T. Y,Yao Y. Y,Zadeh L A. Data Mining, Rough Sets and Granular Computing. New York: Physica-Verlag, 2002
11 Ziarko W,Yao Y Y. Rough Sets and Current Trends in Computing (RSCTC 2000). Berlin : Springer-verlag, 2001
12 Skowron A. Rough sets and Boolean reasoning. In: W. Pedryczed, ed. Granular Computing: An Emerging Paradigm. New York: Physical Verlag,2001. 95～124
13 Polkowski L,Tsumoto S,Lin T Y. Rough Set Methods and Applications: New Developments in Knowledge Discovery in Information Systems. New York: Physica -Verlag, 2000

(上接第109页)

而获取该记录所属SEE的权威度(authorityWeight)。

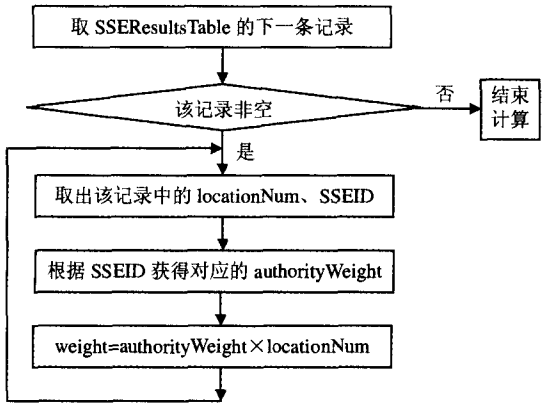

图3 计算最终排序权值的流程图

4 系统实例

随着网上教学信息资源的膨胀发展，一般资源搜索引擎在教学资源方面的查全率和查准率很难满足广大教学资源检索者的要求。要想获得一个比较全面、准确的结果，教学资源检索者就必须反复调用多个搜索引擎，然后在各个搜索引擎的结果中综合出最适合自己的内容。为了减轻教学资源检索者学习检索技巧、选择资源搜索引擎以及综合多个搜索结果的负担，我们以计算机学科领域资源为例，建立了一个面向计算机教学资源的元搜索引擎原型系统。

该试验在Windows 2000 Professional、Internet Information Server 5.0服务器、Access数据库、VB. NET软件开发环境下试验成功，试验结果如图4所示。本原型系统能够较好地满足用户进行计算机学科领域资源检索的要求。

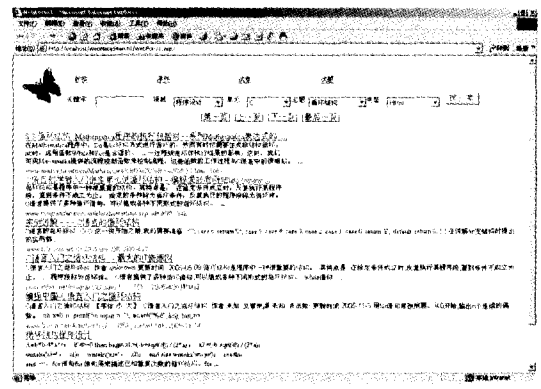

图4 搜索结果示例

结束语 由于元搜索引擎扩大了搜索的覆盖面，解决了搜索的可扩展性，方便获取来自多个搜索引擎的信息，而且提高了信息检索的查全率和查准率。把元搜索引擎技术引入到领域资源中，进而建立领域资源本体库[7]，并结合现代语义网技术，进一步提高本搜索引擎的智能化水平，对于推动领域资源的发展和共享必将起到十分积极的作用。

参考文献

1 Selberg E,Etzioni O. Multi-Engine Search And Comparison Using The MetaCrawler. In:Proceedings of the Fouth World Wide Web Conference '95, Boston USA, Dec. 1995
2 李广建,等. 元搜索引擎及其主要技术. 情报科学 2002(2)
3 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):376～378
4 张弓强. 一种元搜索引擎的查询结果处理模型. 北京工业大学
5 董荣胜,古天龙. 计算机科学与技术方法论. 北京:人民邮电出版社
6 Heaton J. 网络机器人Java编程指南. 童兆丰,李纯,刘润杰,译. 北京:电子工业出版社
7 蔡铭. 面向网络化制造基于语义的资源获取、智能检索和服务匹配原理与技术研究