文章编号: 1672 - 9315 (2007) 03 - 0448 - 04

一种元搜索引擎框架模型的设计*

郭 晔,李建廷,王浩鸣
(西安财经学院 计算机科学系,陕西 西安 710061)

摘 要:元搜索引擎是搜索引擎之上的搜索引擎。用户递交检索请求,元搜索引擎接收该请求后,把它提交给多个预先选定的搜索引擎成员,集中所有的查询结果并以统一的格式返回给用户。首先概述了元搜索引擎的原理和现状,分析了当前研究元搜索引擎的难点所在,并提出改进方案。在此基础上,设计了元搜索引擎的总体框架,提出了查询代理、搜索代理、运算代理三大功能模块,并阐述各代理的功能流程。

关键词:搜索引擎;元搜索引擎;信息检索
中图分类号: TP 311. 5 **文献标识码**: A

A framework model design for meta-search engine

GUO Ye, LI Jian-ting, WANG Hao-ming
(Dept. of Computer Science, Xi'an University of Finance and Economics, Xi'an 710061, China)

Abstract: Meta-search engine is the set of commercial search engines. It takes the request from the user and sends it to the member search engines. And retrieved results got from different search engines are integrated into one unified format and presented to the users. The principles and actuality of the meta-search engine are discussed. The difficulties of making the meta-search engines more power are pointed out. Based on this discussion, we draw up some methods for improving the performance of it. The framework is designed, which includes three modules: enquiry agent, search agent and operation agent.

Key words: search engine; meta-search engine; information retrieval

元搜索引擎是一个搜索其他引擎的搜索引擎。一个元搜索引擎以用户输入的查询关键字作为输入,然后将该关键字同时提交给多个成员搜索引擎,并将这些搜索引擎的返回结果按照一定的算法重新排序并反馈给用户。元搜索引擎是用来提高单个搜索引擎的查准率 (Precision)和查全率 (Recall)而出现的一种新的搜索模式。

1 搜索引擎的实现难点

由于元搜索引擎环境所特有的一些特征,给创建高效的元搜索引擎带来一定困难。

1.1 成员搜索引擎的自治性

元搜索引擎的成员搜索引擎通常都是独立建造的,每个搜索引擎自己决定该索引哪类文本集为用户提供服务,自己决定文本如何表示、索引及更新索引的时间,自己决定相似函数的计算,而文本与给定查

* 收稿日期: 2007 - 04 - 20
基金项目:陕西省自然科学基金项目 (2005F08),陕西省教育厅专项基金项目 (06JK300, SVJYB06278)
作者简介:郭 晔 (1961 -),女,陕西泾阳人,副教授,主要从事海量数据环境中的信息检索、数据挖掘的研究.

询的相似度是通过相似函数计算得到的[1]。通常,很多商业搜索引擎认为自己的相似函数及其它一些信息是保密的,不愿对公众提供足够的关于其引擎的设计和统计信息,目前尚没有足够有效的方法能够独立地找到这些商业搜索引擎的有关信息[2]。

1.2 成员搜索引擎之间的相异性

各基本成员搜索引擎之间存在诸多的不同,具体而言,不同的搜索引擎采用不同的方法来确定标识一篇文本的关键字,具有不同的索引方式;采用不同的关键字权值的确定方法;以不同的相似函数来确定查询和文本之间的相似度[1]。同时,它们可能索引了页面的不同版本,有可能同一个页面在不同的搜索结果中差异性很大。

1.3 元搜索引擎的全局结构与成员搜索引擎间的相异性

全局结构与成员搜索引擎之间存在相异性[3]。具体表现在:元搜索引擎的全局接口使用自己的特定的相似函数 (全局相似函数)来计算文本的全局相似度,而各成员搜索引擎的局部相似函数很可能与之不同;另外,元搜索引擎全局接口计算关键字的方法可能与成员搜索引擎上的计算方法也不同。

2 元搜索引擎的改进

结合目前元搜索引擎的研究现状,在充分考虑元搜索引擎实现难点的基础上,可以采用下面的方法对元搜索引擎设计进行改进:

1) 一般元搜索引擎得到的返回结果很多,降低了查准率。可以考虑在用户输入界面上询问用户所需要查询的范围,如工业、文学、新闻 ……,确定用户的兴趣所在,对于每一个范围都对应有若干个在该领域较出色的成员搜索引擎,使查全率和查准率有所提高。

2) 元搜索引擎在收到用户的查询请求后,最简单的方法是向所有的成员搜索引擎广播该请求,但这实际上是不一定可行的,应该采用一定的搜索引擎调度机制来选择对于用户查询有潜在用处的成员搜索引擎。

3) 针对元搜索引擎收集的成员搜索引擎过少,增加新的引擎困难,可将基本成员搜索引擎的查询语法、请求格式放入知识库,当要发送查询请求时,从知识库读取信息产生的查询请求。这样如果引擎的语法发生变化或者要新增引擎时,只需更改或增加知识库中的信息即可。

4) 有些成员搜索引擎不能或不能很好地支持布尔操作,对于它们将分别发送独立关键字,然后对返回的结果在数据库中进行布尔操作,得到所需结果。

5) 对于从多个成员搜索引擎返回的结果,采用一定的措施和算法,去除不必要的、重复的链接与信息,尽可能将有效的结果反馈给用户。

6) 每个用户对信息资源有各自不同的兴趣,所以利用元搜索引擎为各个用户提供适合其特点的个性化信息服务,不仅能减轻用户的搜索负担,而且能增强软件对用户的亲和感,提高用户的查询效率。

3 元搜索引擎总体框架设计

元搜索引擎的总体设计框架如图 1所示,主要由查询代理、运算代理、搜索代理三部分组成,其支持的后台数据库则由个性数据库、关键词数据库、搜索规则数据库、网页数据库、索引数据库等组成,有关各数据库的功能描述 (图 1)。

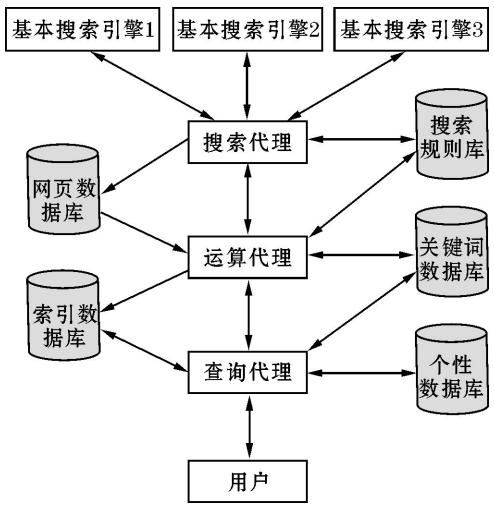

图 1　元搜索引擎的总体设计框架

Fig 1　Framework of meta-search engine

网页数据库:收集基本搜索引擎的查询结果,包括最新查询时间、关键词、下载状态等基本信息。

索引数据库:记录各关键词的查询结果,包括网页标题、网页简要、网页内容、网页地址、连接状态、网

页相关度等信息。

搜索规则库:记录各基本搜索引擎对关键词的查询响应情况,包括返回结果数量、响应时间、用户评价值等。

关键词数据库:记录查询关键词的查询频率。

用户个性数据:记录用户查询关键词的历史清单;记录用户对关键词的点击情况;记录用户查询癖好信息;记录用户信息定制需求情况。

3.1 搜索代理

搜索代理根据运算代理的要求,通过调用搜索规则库,向成员搜索引擎发出提取网页的指令,接收成员搜索引擎的查询结果,并将搜索的网页结果保存到网页数据库中。搜索代理能自动识别各成员搜索引擎的响应情况,如果在限定时间内没有得到成员搜索引擎的响应,则降低该成员搜索引擎的查询响应评价值。

3.1.1 元搜索引擎数据源的选择

搜索代理作为元搜索引擎与成员搜索引擎的唯一查询接口,实现向成员搜索引擎发出查询指令,以及提取和解析它们的查询结果。目前,各种元搜索引擎在选择数据源上的方法不尽相同,首先确定搜索引擎的重复度问题,即如果搜索引擎 A的搜索结果与搜索引擎 B的搜索结果极为相似,则优先选择响应时间最短的搜索引擎。

3.1.2 元搜索引擎搜索规则库的确定

为了能够实现“自动为用户选择召回率最高、响应速度最快的搜索引擎数据源”,则要求建立搜索规则库,在该规则库中设定“提取比重”和最近“平均搜索时间”参数,以便为用户提供最佳搜索结果。搜索规则库的模型如表 1所示。

假设经过用户选定共有 n个搜索引擎 $S_1, S_2, S_3, \cdots, S_n$,它们各自的提取比重为 $W_1, W_2, W_3, \cdots, W_n$,最近提取 10个结果的“平均搜索时间”为 $T_1, T_2, T_3, \cdots, T_n$,则需要考虑如何根据召回率和最新平均查询响应时间来确定向各基本搜索引擎提取的链接数量。

表 1 元搜索引擎搜索规则库建立模型

Tab.1 Rules of searching by meta-search engine

序号	基本搜索引擎	提取比重	最近“平均搜索时间”	设定提取记录数量	设定提取网页时间
1	搜索引擎 S_1	W_1	T_1	N_1	$ST_1 = T_1 * N_1$
2	搜索引擎 S_2	W_2	T_2	N_2	$ST_2 = T_2 * N_2$
3	搜索引擎 S_3	W_3	T_3	N_3	$ST_3 = T_3 * N_3$
…	…	…	…	…	…
n	搜索引擎 S_n	W_n	T_n	N_n	$ST_n = T_n * N_n$

设每个成员搜索引擎提供 m 个搜索结果,故各基本搜索引擎提供的总结果记录数为 $m \times n$个记录,因此,提取的总记录数

$$\text{Record} = m \times n = \sum_{i=1}^{n} N_i = N_1 + N_2 + N_3 + \cdots + N_n$$

第 i个基本搜索引擎提取的记录数为

$$N_i = \frac{W_t}{W_1 + W_2 + \cdots + W_n} \times \text{Record}$$

由此可以得到第 i个基本搜索引擎提取记录所需时间为:$ST_i = N_i \times T_i$,则元搜索引擎系统的最大提取时间为

$$ST_{\max} = \text{MAX}\{ST_1, ST_2, \cdots, ST_i, \cdots, ST_n\}$$

将这个时间设定为当前元搜索引擎系统的最大限定时间,当某成员搜索引擎在搜索过程中,超出了该最大提取时间还没有输出搜索结果,则取消该搜索进程[4,5]。

3.2　运算代理

运算代理负责的运算功能主要包括有:根据用户查询代理提供的关键词清单,调用关键词数据库,计算各查询关键词的权重,并向搜索代理发出搜索要求;负责解析网页结果,通过对检索结果的评价与处理,将所得到的记录重新排序,以获得更高的准确率;将生成关键词与网页清单保存到索引数据库中。

3.2.1　解析基本搜索引擎搜索结果网页

运算代理解析成员搜索引擎搜索结果网页的过程为:当搜索代理将成员搜索引擎的网页搜索结果保存到数据库后,运算代理将该网页结果从数据库提取出来,进行解析,生成"关键词 - 成员搜索引擎 - 搜索结果记录"的清单。

3.2.2　检索结果的评价处理

搜索引擎在生成结果的时候可以通过两种方式来提高用户对搜索结果的满意程度,一是通过将用户真正需要的搜索结果记录返回给用户;二是给用户提供二次检索的功能。元搜索引擎在满足用户要求方面与普通的搜索引擎一样,都是将搜索到的网页按照一定的优先关系显示给用户,因此结果排序是元搜索引擎的一个非常重要的研究领域,结果排序的好坏直接影响着整个元搜索引擎系统的实用性。

3.3　查询代理

查询代理负责解析用户提交的关键词,包括多关键词查询,提取核心关键词。查询代理作为元搜索引擎系统与用户的接口,其界面的友善程度关系重大,优秀的搜索引擎必须有独特的功能才能引起用户的青睐。本元搜索引擎系统也与其它搜索引擎一样,提供个性化设置如搜索引擎数据源的选择,具有一定的个性信息定制功能;能学习反馈用户的选择情况,为进一步提高元搜索引擎准确率创造条件。

查询代理首先检查索引数据库的历史情况,检查索引数据库的对应查询关键词记录情况,若存在该关键词对应的搜索结果,则直接向用户输出查询结果。这样能够充分利用以前结果,提高搜索响应速度。

搜索引擎的相关度运算排序结果是否满足用户需求,是每个搜索引擎必须考虑的问题,而充分利用用户选择情况,对此进行学习反馈,不断改进系统,设计一个具有学习能力的搜索引擎是元搜索引擎追求的目标之一。

不同用户有不同的查询需求,查询代理应能充分利用个性数据库,记录、存储用户的查询癖好,包括中文搜索引擎的数据源选择、显示记录数量、排序显示顺序等信息。

元搜索引擎系统首先根据结果排序算法对检索结果进行排序并提交结果给用户,在用户进行浏览查看的同时记录并保存浏览顺序,从而调整结果记录对应基本搜索引擎的分值,进而可以作为排序算法的一个考虑因素,用于在下次检索时调整结果排序的顺序。搜索结果与用户选择的分值调整的构思(表 2)。

表 2　搜索结果与用户选择的分值调整表

Tab. 2　Adjust between searching results and selection of user

搜索结果排序	结果记录页面的初始分值	用户选择顺序对网页分值的变化										
		第 1次	第 2次	第 3次	第 4次	第 5次	第 6次	第 7次	第 8次	第 9次	第 10次	无点击
No. 1	100	0	- 5	- 10	- 15	- 20	- 25	- 30	- 35	- 40	- 45	- 100
No. 2	95	+5	0	- 5	- 10	- 15	- 20	- 25	- 30	- 35	- 40	- 95
No. 3	90	+10	+5	0	- 5	- 10	- 15	- 20	- 25	- 30	- 35	- 90
No. 4	85	+15	+10	+5	0	- 5	- 10	- 15	- 20	- 25	- 30	- 85
No. 5	80	+20	+15	+10	+5	0	- 5	- 10	- 15	- 20	- 25	- 80
No. 6	75	+25	+20	+15	+10	+5	0	- 5	- 10	- 15	- 20	- 75
No. 7	70	+30	+25	+20	+15	+10	+5	0	- 5	- 10	- 15	- 70
No. 8	65	+35	+30	+25	+20	+15	+10	+5	0	- 5	- 10	- 65
No. 9	60	+40	+35	+30	+25	+20	+15	+10	+5	0	- 5	- 60
No. 10	55	+45	+40	+35	+30	+25	+20	+15	+10	+5	0	- 55

(下转第 456页)

2005, (1 - 3): 1 788 - 1 790.
[5] 唐 炬,宋胜利. 局部放电信号在变压器绕组中传播特性研究 [J]. 中国电机工程学报, 2002, 22(10): 91 - 96.
[6] 张嘉祥. 变压器线圈波过程 [M]. 北京:水利电力出版社, 1982.
[7] 李六零,胡攀峰,邱毓昌. 不同变压器绕组模型对计算快速暂态过电压的影响分析 [J]. 西安交通大学学报, 2005, 39(10): 1 160 - 1 164.
[8] Miki A. A calculation method for impulse voltage distribution and transferred voltage in transformer windings[J]. IEEE Transactions on Power Apparatus and Systems 1978, PAS - 97(3): 930 - 938.
[9] 杨学昌,戚庆成,崔益彬. 暂态分析中变压器线圈等值参数的确定 [J]. 清华大学学报, 1996, 36(9): 100 - 105.
[10] 胡永红,陈汉雄. PSpice在电力系统静态稳定性的应用 [J]. 四川电力技术, 2003, (4): 53 - 54.

(上接第 451页)

对于给定关键词,当元搜索引擎输出结果后,初步设定其分值从结果第 1位至第 10位分别为 100分、95分、……、60分、55分,如果用户最先选择 No. 1结果,则 No. 1的分值不变,如果是第二次才点击 No. 1结果,则 No. 1的分值减 5分,即从 100降为 95分;反之,如果用户最先选择 No. 10结果,则 No. 10的分值增加 45,即 No. 10的分值从 55增加到 100分,成为分值最高的 100分;若用户对以上结果一直没有点击,则它们的网页分值都置 0;系统不断跟踪用户的选择情况后,多数用户的选择就会保证元搜索引擎排序结果的准确性。查询代理充分利用用户的点击情况,及时调整索引数据库的结果排序,不断学习反馈,达到不断提高查询准确率,满足用户查询要求目的。

4 结 论

本文在对目前元搜索引擎总结、分析的基础上,提出元搜索引擎的设计思路,主要介绍元搜索引擎的总体设计模型以及各主要部分的功能,同时对各个模块实现时的设计方法进行了描述。从本文所提出的元搜索引擎框架来看,如何利用用户提交的检索关键字和点击查看的历史记录将检索结果的相关度、用户的个性化需求、检索结果的排序三者有机的结合起来是一个难点。具体讲,就是能否通过充分考虑用户检索关键字的历史记录和用户点击查看的历史记录去影响基本搜索引擎的选择,同时影响检索结果的排序。这将是未来研究的一个主要方面。

参考文献:

[1] Cutler M, Shih Y, Meng W. Using the structures of HTML documents to improve retrieval[C]. //NSITS'97. USENIX Symposium on Internet Technologies and System. Monterey, California. 1997: 241 - 245.
[2] Meng W. Yu C, Liu K. Building effective and efficient metasearch engines[Z]. Submitted to ACM Computing Surveys(under revision). 1999.
[3] 张俭恭,陈定权,吴振新. 关于搜索引擎与元搜索引擎的讨论 [J]. 信息检索技术, 2002, (2): 36 - 38.
[4] Wray L Buntine, Karl Aberer, Ivana Podnar, et al. Opportunities from open source search[J]. - (WI2005), 2005, (9): 19 - 22.
[5] Koraljka Golub, Anders Ardo. Importance of html structural elements and metadata in automated subject classification. - (ECDL 2005)[C]. //Vienna, Austria. LNCS, 2005, 3652: 368 - 378.
[6] 王 霞,杨炳儒. Web搜索结果挖掘的研究与应用 [J]. 计算机工程与应用, 2003, 39(14): 187 - 189, 207.
[7] 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, 29(6): 118 - 121.
[8] 陈俊杰,薛 云,宋翰涛,等. 基于 Agent的元搜索引擎的研究与设计 [J]. 计算机工程与应用, 2003, 39(10): 33 - 36.
[9] 张廷华. Web元搜索引擎的改进 [J]. 计算机应用, 2002,. 22(2): 105 - 107.