元搜索引擎在油田领域信息分布式搜索的应用①

文必龙，彭成晖

(东北石油大学 计算机与信息技术学院，大庆 163318)

摘 要：当前大型集中式企业搜索引擎面临规模扩展、数据更新快速和用户需求专业化、多样化等一系列挑战。因此，在整个油田内建立一个高效、高质量的分布式搜索引擎体系架构，并保证不破坏各公司原有企业搜索引擎的基础之上，实现数据共享变得越来越重要。通过分析油田内部公司搜索引擎现状，结合元搜索引擎技术，提出构建一个合理的油田内部分布式搜索引擎体系架构的解决方案。主要从接口规范，查询调度，排序策略三个方面入手，解决分布式搜索引擎在油田应用查询中存在的问题。

关键词：元搜索引擎；调度；排序；分布式；信息

Application of Meta-Search Engine to Oil-field Information Distributed Search

WEN Bi-Long, PENG Cheng-Hui

(School of Computer and Information Technology, Northeast Petroleum University, Daqing 163318, China)

Abstract: The current large-scale centralized enterprise search engines face with a series of challenges. Such as the scale expansion, fast data update and user needs specialization and diversification. Thus, to establish an efficient, high-quality distributed search engine architecture in the area of oil-field and to realize data sharing based on ensuring does not destroy their existing enterprise search engine is becoming more and more important. In this paper, through analyzing the current situation of the search engines in oil-fields enterprise, and combining with meta-search engine technology, we proposed a reasonable solution about building distributed search engine architecture in the area of oil-field. To solve the existing problems of distributed search engine in application queries of the oil-field, this article mainly uses three technologies, such as interface specifications, query scheduling and ordering strategy.

Key words: meta-search engine; schedule; sort; distributed; information

在互联网上搜索引擎服务的热潮一浪高过一浪的同时，也注意到另外一个现象，企业信息化的风起云涌，提升了工作效率，也创造了更多的价值，但与此同时企业累积了成千上万的海量数据。这些数据都是企业的宝贵财富，如果不能有效地加以利用，结果就只会形成占用 IT 资源的信息垃圾。但是，这些信息往往分散在企业的各个角落，找到它们如同大海捞针，对信息的有效利用更是难上加难。随着企业信息化建设的推进，搜索引擎是必不可少的企业内部数据的快速增长促进了企业搜索引擎的发展[1]。

在油田范围内，许多下属企业都拥有了自己的企业搜索引擎，并且现有的企业搜索引擎基本可以满足当前的需要。

随着企业的高速发展和信息的快速膨胀，当前大型集中式企业搜索引擎面临规模扩展、数据更新速度和用户专业化、多样化需求等一系列挑战，在整个油田内建立一个高效高质量的分布式搜索引擎体系架构越来越重要。

目前油田内部使用的搜索引擎包括：Microsoft 的 Search Server 系列产品、IBM 的 OmniFind 搜索套件、Oracle 公司的 Secure Enterprise Search(SES)软件和国产的 TRS。本文借鉴元搜索引擎的思想，制定一套搜

① 基金项目:国家科技重大专项(2008ZX05023-05-05)
收稿时间:2011-08-22;收到修改稿时间:2011-09-23

索排序策略，用来解决异构分布式企业搜索引擎查询结果的重排序和融合，并根据企业中不同数据源类型，制定出一种基于元搜索引擎的统一的结果处理机制。

1 现状分析

1.1 元搜索引擎

元搜索引擎[2]将现有的搜索引擎看成一个逻辑意义上的整体，为用户提供统一的查询界面，并利用成员搜索引擎的查询功能，得到所需的搜索结果。元搜索引擎主要有以下三部分构成：首先是检索请求预处理部分，用来实现用户的个性化检索设置的要求、成员搜索引擎的调度策略、检索时间、返回结果集的限制等；然后是检索接口代理部分，用来实现将用户的个性化查询请求转化为可被成员搜索引擎识别的固定格式；最后是检索结果处理部分，用来实现把调用的成员搜索引擎检索到的结果去重、合并、排序和按一定的格式返回给用户[3]。

现有的元搜索引擎主要有：搜魅网（someta），集合了百度、google、搜狗、雅虎多家主流搜索引擎的结果，提供网页、资讯、网址导航等聚合查询，另外，搜魅网突破了元搜索引擎没有自己的蜘蛛的瓶颈，提供了网站查询的功能；比比猫（Bbmao），独创国际领先的聚类和去重技术。搜索结果汇集各大搜索引擎结果，搜索结果智能分类整理，去掉重复搜索结果，并拥有直接搜寻文档和强大网络收藏夹等多元功能。对于记者、教授、高管等知性、高端且惜时如金的人群比较适用，支持中英文搜索。

1.2 问题的提出

元搜索引擎的关键技术主要包括用户请求的识别转化、成员搜索引擎的调度[4,5]以及搜索结果的合成。在现有技术的基础上，元搜索引擎技术应用在现有的企业搜索引擎中，建立分布式搜索引擎体系，要解决如下问题：

现有的元搜索引擎主要针对 WEB 的应用，在油田中，每个品牌的企业搜索引擎也只是实现了自己品牌的分布式检索，也可以实现多结点结果融合，但是油田企业中每个分公司的自主权比较大，可能所采用索引不同机制的企业搜索引擎，以现有的手段和方法，如果要在各异构搜索引擎之上使用元搜索引擎，那么这些异构的企业搜索引擎就不能实现结果的融合，所以针对不同品牌的企业搜索引擎系统组成的元数据搜索引擎，如何建立一种适于于油田的、方便集成的油田信息搜索引擎体系？

元搜索引擎引用的大多数十个以内的成员搜索引擎。但如果像油田企业，随着企业内部数据量的增加，成员搜索引擎的数目也会相应增加，可能达到成百上千个成员搜索引擎，如果每次搜索都需要对这几百个搜索引擎发送请求，势必会影响搜索效率，所以有必需建立一种适用于该情况的查询调度策略。

元搜索引擎在企业搜索应用中主要针对员工。除了要考虑利用原有搜索引擎排序策略，并且根据员工所处的工作岗位，分析员工在企业内部所扮演的角色，把员工所关心的搜索内容排在前面，所以需要结合员工角色，提出一套基于用户的排序策略。

2 解决方案

2.1 接口规范

异构搜索引擎系统是相关的多个搜索引擎系统的集合，可以实现数据共享和透明访问，每个搜索引擎系统在加入异构搜索引擎系统之前本身就已经存在，拥有自己的 DBMS。异构搜索引擎的各个组成部分具有自身的自治性，实现数据共享的同时，每个数据库系统仍保有自己的应用特性、完整性和安全性控制。

油田内部各公司搜索引擎并入搜索体系之前，必须要提供正在使用的企业搜索引擎的相关服务接口，对这些接口进行封装，实现本文所提出的标准接口。统一查询接口格式：

resultQuery：查询接口。用来获取结果集，并返回 ResultElement 对象。public List<ResultElement> resultQuery(string query, int[] datasourceids, int startindex, int totalnum)，其中 query 是用户输入的关键词，datasourceids 是设定查询范围，startindex 是定位从结果集中多少条开始取，totalnum 是一次取多少条记录。

ResultElement 对象：ResultElement 对象中包括 Title，Snippet，Url，SorFactor 属性，分别表示结果的标题，摘要，数据类型，结果指向的链接，排序因子。结果通过排序因子排列。ResultElement 对象如表 1。

本文以 Oracle 公司开发的企业搜索引擎 SES（Secure Enterprise Search）为例，通过阅读 SES 的开发手册，得知可以使用 SES 所提供的 WebService 服务接口来获取搜索结果。首先需要创建搜索服务类：OracleSearchService stub = new OracleSearchService();

其中提供了 setSoapURL 方法，设定提供搜索服务的地址：stub. setSoapURL(string queryUrl),搜索服务 Url 的格式为 http://<host>:<port>/search/query/OracleSearch。通过 doOracleSearch 方法传入用户输入的关键字，返回 OracleSearchResult 类型，OracleSearchResult 中提供了获取记录标题、摘要、Url 地址等信息的方法。

表 1　数据元组成成分

属性名	属性类型	中文名称
Title	String	结果的标题
Snippet	String	结果的摘要
Url	String	结果的 Url 链接
SorFactor	String	结果的排序因子

通过以上实例，可以将油田内部的企业搜索引擎所提供的服务接口进行再次封装，形成统一查询接口。

2.2 调度策略

元搜索引擎没有自己的索引数据库，要依靠调用成员搜索引擎搜索各自的索引数据库来实现搜索，各个成员搜索引擎的搜索返回结果构成了元搜索引擎的搜索返回结果。由于各成员搜索引擎在不同信息搜索主题下会表现出不同的搜索性能，差异较大，因此对成员搜索引擎的选择十分重要。再加上各个成员搜索引擎的性能都是在不断地发生变化，所以对成员搜索引擎的选择并不能是静态不变的。

为了解决上述问题，将用户与查询调度结合起来，通过分析油田用户所在的子公司以及用户的历史访问次数，对相关的成员搜索引擎进行访问；如果用户知道所要查找的内容属于哪个公司，还可以对成员搜索引擎直接进行选择，用人工干预的方式，帮忙搜索引擎提高查准率，通过以上两种方法相结合，可以油田信息搜索引擎的搜索效率大大提升，从而形成不同于传统元搜索引擎的查询调度策略。

调度策略引入元数据的思想，元数据可以描述信息资源或数据对象，其目的在于使用户能够发现资源、识别资源、评价资源，对相关的信息资源进行选择、定位和调用，追踪资源在使用过程中的变化，实现信息资源的整合[6]。

元数据可以对用户信息以及用户的查询结果进行描述，通过分析，最终实现将元数据与搜索引擎结果排序结合起来。每一个油田用户都有一个唯一标识自己身份的账号(DN, distinguished name)[7]。用户往往只关心所在单位的查询结果。用户使用账号登录时，通过 LDAP 协议获取用户在油田 LDAP 服务器中的完整信息并存入元数据库中。在成员搜索引擎加入到本文提出的分布式体系之前，管理员需要建立对用户行为进行描述的信息表（表 2），以及存储成员服务器地址的信息表（表 3）。

表 2　用户访问信息

字段名	属性类型	中文名称
UserName	String	用户名称
UnitName	String	单位名称
VisitedUrl	String	已访问 Url 链接
VisitedTimes	Int	访问次数

表 3　服务器信息

字段名	属性类型	中文名称
UnitName	String	单位名称
ServiceUrl	String	服务器 IP 地址

一个完整 DN 账号（唯一标识名）含有 DC(Domain Component)，OU (Organizational Unit)，CN(Common Name)三个元素。例如：用户通过 ptr/test 登录，系统可以获取用户的在 LDAP 服务器中的唯一标识名：CN= test,OU= 勘探开发研究院 ,DC= ptr,DC= petrochina。通过分析得知用户所在单位，在用户没有人工勾选指定的访问范围时，元搜索引擎将会访问该用户所在单位的成员搜索引擎，得到返回结果。

在图 1 中，如果搜索用户的组织单位 OU=勘探开发研究院，在 A、B、C、D、E 五个成员搜索引擎中，A 和 C 是属于勘探开发研究院的成员服务器（服务器信息可从表 3 中获取），那么根据调度策略，当搜索用户输入关键词后，主服务器访问成员搜索引擎 A 和 C，并结合用户访问的历史记录，获取成员搜索引擎返回的结果。如果用户想要获取所有成员搜索引擎的结果，也可根据需要，手动选择要访问的成员搜索引擎。

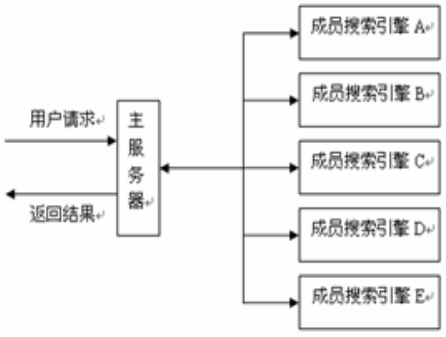


图 1　调度示意图

2.3 排序策略

由于各子公司的搜索引擎只对该公司的数据进行采集并建立索引，所以各成员搜索引擎中结果重复率会比 WEB 中元搜索引擎结果低得多。针对这种情况，可以引入轮询排序算法[8]。轮询法最初虽然按照搜索引擎的性能给予排序，但没有充分考虑到成员搜索引擎之间的差异以及用户角色融入的问题。

为了解决轮询法在油田信息搜索结果排序应用中存问的问题，通过分析用户所在单位和部门以及记录用户查询过的结果，通过计算得出权重，并且与轮询算法结合到一起，从而提高用户的查准率。

当用户以匿名身份登录时，采用简单的轮询法进行排序。轮询算法：s=1+count(rs)(n-1)，其中 s(Sor-Factor)：排序因子，表示在记录总结果集中的位置，count(rs)结果集个数(同时也可以表示线程数)，n：记录在子结果集中的位置。

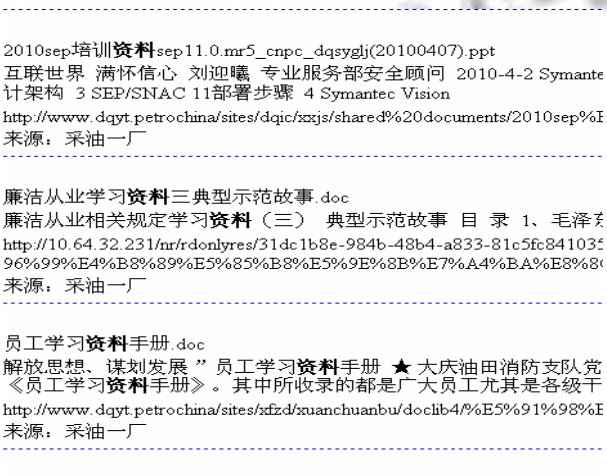
当用户登录时，在元数据库中记录用户对某一地址的访问次数 v，可以改进上述算法为

$$s=\begin{cases}1+count(rs)(n-1) & (v=0)\\ [1+count(rs)(n-1)](h*v) & (v>0)\end{cases}$$

其中 h 为常数，可以根据实际需要设定大小，h 值越大，访问次数 v 对结果排序的影响越大。最终把结果展示给用户。

3 实验结果

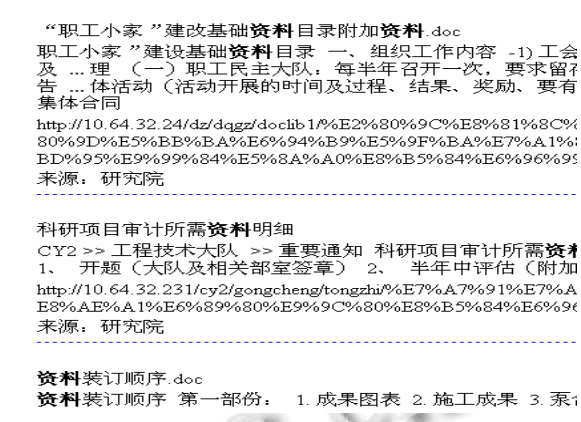
用户以不同身份搜索关键词“资料”，得到的结果来源不同(如图 2，图 3)。


2010sep培训**资料**sep11.0.mr5_cnpc_dqsyglj(20100407).ppt
互联世界 满怀信心 刘迎曦 专业服务部安全顾问 2010-4-2 Symante
计架构 3 SEP/SNAC 11部署步骤 4 Symantec Vision
http://www.dqyt.petrochina/sites/dqjc/zxjs/shared%20documents/2010sep%E
来源：采油一厂

廉洁从业学习**资料**三典型示范故事.doc
廉洁从业相关规定学习**资料**（三） 典型示范故事 目 录 1、毛泽
http://10.64.32.231/nr/rdonlyres/31dc1b8e-984b-48b4-a833-81c5fc84103
96%99%E4%B8%89%E5%85%B8%E5%9E%8B%E7%A4%BA%E8%8
来源：采油一厂

员工学习**资料**手册.doc
解放思想、谋划发展”员工学习**资料**手册 ★大庆油田消防支队党
《员工学习**资料**手册》。其中所收录的都是广大员工尤其是各级干
http://www.dqyt.petrochina/sites/xfzd/xuanchuanbu/doclib4/%E5%91%98%E
来源：采油一厂

图 2 用户以采油一厂员工身份搜索的结果


“职工小家”建改基础**资料**目录附加**资料**.doc
职工小家”建设基础**资料**目录 一、组织工作内容 -1) 工会
及 ... 理 （一）职工民主大队：每半年召开一次，要求留
告 ... 体活动（活动开展的时间及过程、结果、奖励、要有
集体合同
http://10.64.32.24/dz/dqgz/doclib1/%E2%80%9C%E8%81%8C%
80%9D%E5%BB%BA%E6%94%B9%E5%9F%BA%E7%A1%
BD%95%E9%99%84%E5%8A%A0%E8%B5%84%E6%96%9
来源：研究院

科研项目审计所需**资料**明细
CY2 >> 工程技术大队 >> 重要通知 科研项目审计所需**资**
1、 开题（大队及相关部室签章） 2、 半年中评估（附加
http://10.64.32.231/cy2/gongcheng/tongzhi/%E7%A7%91%E7%A
E8%AE%A1%E6%89%80%E9%9C%80%E8%B5%84%E6%9
来源：研究院

资料装订顺序.doc
资料装订顺序 第一部份： 1. 成果图表 2. 施工成果 3. 泵

图 3 用户以研究院员工身份搜索的结果

4 结语

本文主要实现把元搜索引擎思想引入油田信息搜索引擎中，针对油田内部异构成员搜索引擎，结合油田内部用户身份标识的特性，建立标准的查询接口规范，提出了一种主服务器访问成员搜索引擎的查询调度策略，本调度策略在设计过程中，统计用户访问历史，合理选择成员搜索引擎，获取查询结果，解决了由于成员搜索引擎过多而造成查询效率偏低的问题，使主服务器的性能得到较好的发挥。通过改进的轮询算法，提高了油田用户使用元搜索引擎的查准率。

参考文献

1 文必龙,李添,李娜,高鹏.企业搜索引擎安全搜索的研究.齐齐哈尔大学学报,2010,5:1–3.

2 吴小兰,汪琪.元搜索引擎研究综述.图书馆学理论研究,2009,53(9):46–49.

3 王金栋.元搜索引擎调度策略及结果排序算法的研究[硕士学位论文].燕山大学,2010.

4 李村合,孟文杰.基于分类评价的元搜索引擎调度策略.计算机工程与设计,2008,29(5):1065–1066.

5 Meng WY, Wu ZH, Yu C, Li ZG. A highly scalable and effective method for metasearch. ACM Trans. on Information Systems, 2001,19(3):310–335.

6 文必龙,李智新,王英艳.基于元数据的企业搜索引擎研究.郑州轻工业学院学报(自然科学版),2008,(6):5–6.

7 王征.对目录服务中 LDAP 技术的分析.科技创新导报,2011,(9):5–6.

8 曹林,韩立新,吴胜利.元搜索引擎排序技术综述.计算机应用研究,2009,(2):412–413.