文章编号: 1000-2243(2008)S0-S122-03

一种智能型元搜索引擎功能模型

颜晓玉

(福州大学图书馆, 福建 福州 350108)

摘要: 分析了现有元搜索引擎的功能, 指出其不足并提出一种智能型的元搜索引擎模型. 通过建立本地数据库, 自动提供用户需求的扩展、分发和对搜索结果的过滤、排序, 提高网络信息搜索的查全率和查准率.

关键词: 元搜索引擎; 查全率; 查准率; 数据库
中图分类号: TP391.3 文献标识码: A

A functional model of intelligent meta-search engines

YAN Xiao-yu
(Library of Fuzhou University, Fuzhou, Fujian 350108, China)

Abstract This paper analyzes the existing meta-search engine function, pointing out its deficiencies and made an intelligent meta-search engine model. By establishing a local database, to provide the user needs to expand and distribute, to filter and sort the search results automatically, thereby enhancing recall ratio and pertinency ratio of network information search.

Keywords meta-search engines; recall ratio; pertinency ratio; database

如何从海量的网络信息资源中快速、有效地提取信息正成为当前搜索引擎的主要发展方向[1]. 目前, 互联网上可用的搜索引擎很多, 典型代表有 Google和百度, 搜索内容可以覆盖互联网上绝大多数网页内容. 在 Google中输入“网络信息技术”, 完成搜索过程用时 0.15 s, 得到 200多万条相关的答案; 在“百度”, 用时 0.001 s, 得到信息 2.05×10^6 条. 虽然反应速度迅速, 并返回大量的搜索结果, 但检索效果并不理想. 谷歌搜索结果排序为:

- 西英拓网络信息技术有限公司——广西最大的 IDC业务提供商 域名注册 ...
- [DOC]农村信息化示范单位初选名单公示公告
- [DOC]信息产业部公布的信息技术专业课程

……

显然, 所得结果不是用户期待的内容. 用户难以对海量的信息进行分析以获取所需的有效信息. 此外, 对于同义关键词的查询也会有许多不同的结果, 表 1是在“百度”对“网络”一词搜索的结果.

表 1 搜索结果

Tab.1 The results of search

关键词	t/s	信息数
网络信息技术	0.001	2 050 000
互联网信息技术	0.082	1 400 000
Internet信息技术	0.115	189 000
Web信息技术	0.084	143 000

在大多数搜索引擎不能满足个性化查询请求的情况下, 各种个性化搜索引擎应运而生, 有的可限定搜索类别, 有的提供相关信息或将查找结果分类提供, 在一定程度上弥补了普通搜索引擎的不足. 本文针对非营利性的信息服务单位, 如数字图书馆, 提出一种智能型的元搜索引擎: 通过在用户接口对关键词的提交和搜索结果进行自动处理, 以提高网络搜索信息的查全率和查准率.

收稿日期: 2008-06-13
作者简介: 颜晓玉(1976-), 女, 馆员.

1 元搜索引擎工作原理

元搜索引擎是一个“顶层”的搜索引擎, 一般情况下它没有数据库, 更不用网络蜘蛛, 而是一个调用其他现成搜索引擎(SE)的搜索引擎, 如图 1所示. 这个元搜索引擎将用户的查询字串进行处理提交给 n 个搜索引擎进行查询, 然后把返回的结果进行适当处理, 合并成一个统一的结果返回给用户. 如: 预先设置一些典型类型: 论文、图片、音乐, 甚至语言或地区等等, 用户查询时, 针对性地选择搜索类型进行搜索, 从而提高查询质量; 对搜索结果进行处理, 如: 限制从各个独立搜索引擎的结果中提取的数量、利用权重使从各个搜索引擎的结果分占一定比例或“信息获取的训练集策略”[2]等, 一定程度上减少结果的重复性, 从而减少用户浏览负担. 另外, 还可以把搜索结果自动分类再提供给用户选择浏览, 以提高搜索查准率. 可见, 元搜索引擎可以做成一种很有个性的、很有效率的搜索工具. 目前, 有些“底层”搜索引擎也具备了其中某些功能, 这给元搜索引擎设计带来更多的方便.

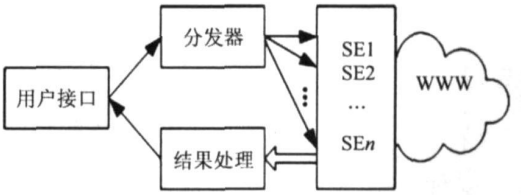

图 1　元搜索引擎的结构

Fig 1　The structure of meta- search engine

2 一种智能型元搜索引擎功能模型

元搜索引擎无需人工干预、不必维护庞大的数据库而且查全率高, 但是, 它的处理速度比较慢, 并且, 这个查全率是通过多个搜索引擎搜索与所提供的关键词相关内容所得结果的总和, 由于各种搜索引擎互相竞争, 追求“查全率”, 导致信息量非常庞大. 另外, 由于没有涉及与关键词同义的内容, 从语义上来说, 搜索结果其实很不完全.

本文提出的元搜索引擎(图 2)与通常的元搜索引擎不同, 它在本地服务器上有一个小规模的数据库, 用来存储用户需求扩展的词汇和词频等动静态信息. 使用时, 从需求处理到结果处理都从该数据库中提取积累的“经验”, 一次性地自动搜索出与用户所提供关键词同义的“全部”相关内容, 并且, 通过词汇的扩展, 有针对性地分配给下层相关的搜索引擎, 实现真正的“查全”.

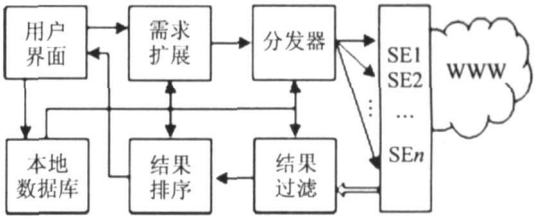

图 2　智能型元搜索引擎功能框图

Fig 2　Function block diagram of intelligent meta- search engine

用户需求扩展的方式主要有 2种: 基于语义和基于统计. 基于语义的方法是分析用户需求的含义, 将语义上最相似的信息扩展进来. 但自然语言理解是一个尚未解决的课题, 真正的语义分析并不能自动进行, 往往需要用户的参与. 基于统计的用户需求扩展方法主要是利用统计分析技术计算出跟已知用户需求最相似的信息, 然后把它们扩展到用户需求中来. 比如, 用户当前浏览的内容说明用户当前兴趣, 通过分析用户当前浏览的内容, 就可以自动选择那些跟用户需求关键词相似度最大的词进行扩展. 这类方法由于可以实现自动的扩展, 因而应用最为广泛. 但是, 如前所说, 用户浏览时不可能同时进行多个同义词的搜索, 所以, 通过分析用户的浏览内容进行扩展, 有可能“误导”, 以致慢慢偏离搜索目标.

2 1 用户需求扩展和分发

对用户需求扩展从语义上进行扩展. 例如用户需要查找“网络资源”, 实际上也需要“互联网资源”、“Internet资源”、“Web资源”等等, 然后, 进行词组的自动延伸: “…的采集”、“…的利用”等. 通常, 延伸的关键词用户是会提供的. 最后, 分发给下层搜索引擎对应的类别进行搜索.

为实现基于语义扩展, 要在本地服务器建立一个数据库, 通过对词频等信息积累、分析, 动态地管理这些词汇.

2 2 用户需求扩展数据库的建立

由于真正的语义分析不能自动进行, 目前只能由人工建立. 先积累用户使用的关键词, 通过词频统

计, 在达到一定规模后就可以从热门词汇开始建立同义词库. 另外, 通过词频统计自动调整其权重, 在搜索及结果排序中取得优先.

2 3 搜索结果的处理

搜索结果的处理主要有搜索结果的过滤和搜索结果的排序[3]. 搜索结果的过滤是指从各个独立搜索引擎的搜索结果中进行提取和剔除. 由于有本地数据库, 就可以简单地用数据库中的不同同义词的词频统计数来分配其权重; 同时, 采用词组优先排队来提高"查准率".

3 结语

通过建立本地数据库, 对用户需求的自动"扩展"和"分发"提高了查全率, 对搜索结果的"过滤"和"排序"保证了高的查准率. 另外, 有了本地数据库, 随时可以处理历史数据, 使得自动分析和处理搜索结果更加容易.

本地数据库的维护问题在一定程度上削弱了元搜索引擎原来的优点, 也是提高"双率"的代价. 不过, 本模型针对某单位特定人群, 是在一个大学的数字图书馆中提供专业性和知识性强的信息搜索服务, 而将其它的内容留给普通搜索引擎, 不作为超级搜索门户那样企求一揽子解决问题, 数据库规模可以小一些, 也可以自动剔除词频小的记录. 这样突出本单位的特色, 讲究小范围的优质服务, 显然可以提高搜索效率.

参考文献:

[1] 孙 莹. 融合人工智能技术 搜索引擎工具即将发生巨变 [J/OL]. 2006- 08- 21. CNET科技资讯网, http: //www. cnet-news. com. cn/.

[2] 张强弓, 喻国宝, 廖湖声, 等. 一种元搜索引擎的查询结果处理模型 [J]. 通讯和计算机: 中文版, 2005, 2(2): 19- 20.

[3] 杨道玲. Web资源采集与保存研究 [D]. 武汉: 武汉大学, 2005.

(责任编辑: 杨 青)