个性化高效元搜索引擎的设计与实现*)

胡 亮 许永诚 高 文 胡利平
(中国农业大学信电学院 北京 100083)

摘 要 本文介绍了一个高效的元搜索引擎系统SMS(Smart Meta Searcher),采用检索实例知识库对用户的检索意图进行推理,同时给出一套独特的星级排行评价策略,通过用户行为分析技术为用户提供个性化信息检索服务,以及其在未来搜索引擎个性化、智能化、专业化和多媒体搜索的发展方向所作的探索工作。

关键词 元搜索引擎,搜索引擎,个性化信息服务,信息检索,用户行为分析

Design and Realization of A Personalized and Efficient Meta-search Engine

HU Liang XU Yong-Cheng GAO Wen HU Li-Ping
(College of Information and Electrical Engineering, China Agricultural University, Beijing100083)

Abstract This paper introduces an effective meat-search engine SMS(Smart Meta Searcher). SMS adopts the knowledge base of query case to ratiocinate the query intentions of users. SMS proposes a particular strategy of star rank. It could adjust itself to user's interests and provide personalized information retrieval services with the technology based on user behavior analysis. It also gives a brief introduction about what we have done for explorering the developmental direction of personalization, intelligentization, specialization & multimedia search of the future search engine.

Keywords Meta-search engine, Search engine, Personalized information service, Information retrieval, User behavior analysis

1 引言

目前Intertnet上用于信息检索的技术主要是搜索引擎技术,搜索引擎收集Internet上的网页数据,建立庞大的索引数据库,为用户提供关键词检索功能。当用户查找某个关键词的时候,所有在页面内容中包含了该关键词的网页都将作为搜索结果被搜出来。在经过复杂的算法进行排序后,这些结果将按照与搜索关键词的相关度高低,依次排列。搜索引擎的最大优点是:信息的覆盖面较大,信息新颖,而且对搜索结果的相关性排列上,搜索引擎将其认为相关性高的检索结果排列在前。但由于搜索引擎使用的信息检索技术智能水平的限制,以及对自然语言理解的制约,因此其也存在着不足之处,主要是数据冗余度 、检索方式、数据覆盖率、准确率等几个方面。

由于不同的搜索引擎的索引数据库不相同,而且采用了不同的索引结构与检索、排序技术,因此检索效果是不一样的,在一个搜索引擎上找不到的网页可能在另一个搜索引擎上找到。通常单个搜索引擎能找到的相关信息不超过所有相关信息的一半,用户一般通过多个搜索引擎进行检索才能比较全面地检索所需的内容,同时需要在不同搜索引擎的检索结果中挑选所需要的内容,因而对于用户来说很不方便。

为了解决以上传统搜索引擎存在的问题,提供智能化、个性化、专门化服务的元搜索引擎是一个比较好的可行方案。元搜索引擎相当于多个搜索引擎集成为的一个门户,提供统一的检索界面,用户提交信息检索请求,由元搜索引擎进行加工,转换成多个独立搜索引擎的检索模式,然后模拟用户同时在其他多个引擎上进行搜索,并将所有的结果处理后以统一的方式返回给用户。智能化的元搜索引擎采用多种优化策略对搜索结果进行整理,删除搜索结果中大量的冗余数据与失效链接,同时以更合理的排序提供给用户。

2 元搜索引擎的相关工作

2.1 国外的主要元搜索引擎

MetaCrawler是一个并行式元搜索引擎,具有优秀的清晰性和详细的组织性,可以同时调用6个独立Web引擎;提供全面的用户接口与丰富的逻辑检索功能;排序是基于评分策略的,同时有效地消除了大量的重复结果,保证了高质量的搜索结果。

Dogpile是目前性能较好的并行式元搜索引擎之一,它可以同时调用25个独立引擎,其中Web搜

*)该课题得到国家高技术研究发展计划(编号:2002AA243031)资助。胡 亮 硕士研究生,主要研究方向为智能信息检索系统。许永诚 副教授,硕士生导师,研究方向为网络安全、信息检索。

索引擎14个;采用独特的并行和串行相结合的查询方式,支持布尔算符和模糊查询,可设置最大查询时间;支持搜索方式比较全面,包括WWW、FTP、News(新闻论坛)等。

ProFusion是并行式元搜索引擎,在智能化的搜索技术、对查询的实用提示和个人化搜索服务方面做得比较优秀,可同时调用9个独立Web搜索引擎;采用C/S结构,用户接口使用JavaApplet开发,对于每个搜索引擎都有一个相对应的类描述;其数据融合是基于返回记录的排序位置。

国外的元搜索引擎都属于英文搜索引擎,不支持中文。

2.2 国内的主要元搜索引擎

3721疯狂搜索是国内做得比较好的商业化元搜索引擎,其搜索范围极为广泛,采取联合搜索的模式,可以同时检索数十个中文搜索引擎;融合结果是基于URL的唯一性与标题/简介与检索词的相关程度以及标题/简介的文字长度;排序主要考虑检索结果的标题/简介与检索词的相关程度、结果来自多少个独立的搜索引擎、搜索结果在这些引擎中的排列位置、搜索引擎的权威性等;相关性、优先级和权值越高的结果排序越靠前。

万纬中文元搜索是并行的中文元搜索引擎,调用9个支持中文检索的Web搜索引擎;可以选择最大等待结果时间;搜索结果可按相关度、时间、域名和引擎分类。

国内大多数元搜索引擎都是实验室中开发,而且未实用化,一般用于学术研究,不提供公众网访问方式。

3 SMS元搜索引擎系统设计

本文介绍了一个元搜索引擎系统——SMS(Smart Meta Searcher),其关键工作主要集中在如下几个方面:搜索引擎选择、结果融合算法、用户行为分析与星级排行评价因子。它采用用户行为分析技术,对检索的关键词进行提取加工处理,选择最优的搜索引擎组,按各搜索引擎的特点将其转换为更精确的检索条件,大大提高了检索的准确率。同时利用高效的融合算法对返回结果分析处理,用星级排行因子标注每条记索结果来描述检索返回的文档与用户所要找的目标文档的差距,提高了检索的查全率、查准率与较好的排序。

SMS(http://202.205.91.20)是本文作者设计实现且实用化的一个元搜索引擎,其系统结构设计图如图1所示。主要包括:用户接口、检索记录模块、行为分析模块、搜索引擎选择模块、检索条件转换模块与融合模块。SMS的主要工作流程如下:用户通过用户接口提交检索请求,SMS提取关键词、用户IP(本系统采用IP地址作为用户标识),由检索记录模块进行分类后再存储到数据库户。然后系统根据IP调用以往的搜索记录,通过相关性算法寻找相关检索高频词汇,精确用户的检索范围。最后,系统按加权评估算法选择适当的搜索引擎,将用户的请求提交给各搜索引擎从而得到检索结果,再利用融合算法处理检索结果以统一的形式返回给用户。如果用户认可此次检索是达到要求的,则由检索记录模块记录本次检索相关数据,标志其为一次成功的检索,同时将本次检索的关键词添加到用户IP地址对应的词频数据库中,用于下次检索用户行为分析。

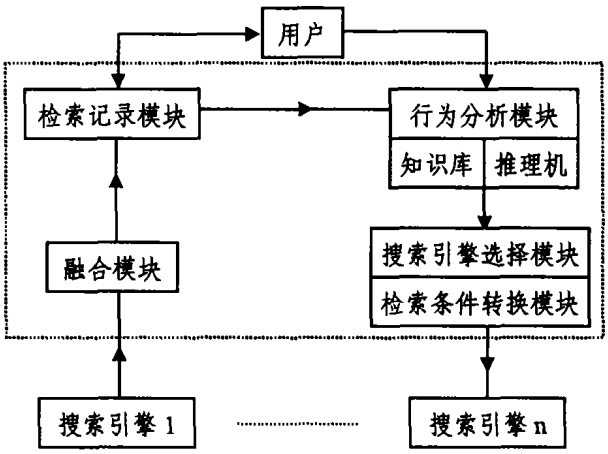

图1 SMS元搜索引擎

4 SMS元搜索引擎实现的关键技术

4.1 原始搜索引擎XML配置文件

由于各个原始搜索引擎的访问方式不相同,我们为每个原始搜索引擎建立一个XML的配置文件。每个对应的XML配置文件记录了原始搜索引擎的检索提交URL地址以及与检索相关的参数。如果原始搜索引擎更新或地址更改了,我们只需要修改对应的XML配置文件。SMS根据XML配置文件将用户提交的检索请求转换为该搜索引擎处理的URL提交地址,然后提取返回检索结果。这样,通过XML文件描述原始搜索引擎的方式使得原始搜索引擎在改动时,不需要更改主程序,便于程序的模块化。

4.2 用户行为分析与星级排行因子

用户行为分析技术是搜索引擎个性化的核心技术。本系统的主要思路是考虑同类或同专业用户的检索行为与检索意图在很大程度上是基本相似的,通过与同类用户以往的检索行为的对比来预测用户的检索意图。

SMS首先根据用户检索历史记录中的关键词组对用户进行归类,当用户提交检索请求时,SMS对提取检索词,通过同义词字典对检索词进行同义转换,如电脑、计算机属于同义词,提炼出更精确的

检索词再提交给各个独立搜索引擎，存储返回的检索结果，然后采用基于距离度量的实例推理模式，从知识库中匹配与返回结果相似的以往同类用户的检索实例，计算其星级排行因子，用于描述检索返回结果中的文档与用户所要找的目标文档的差距。以下是用于用户行为分析的主要策略与定义：

同类用户集：SMS 系统利用 IP 地址作为用户的标识，对于每次来自同一个 IP 的检索关键词，如果用户表示满意则认为本次检索成功，同时将本次的用户提交的关键词记录到以 IP 地址为主键的数据库中。这样，对于每个用户 j 来说，都对应一个关键词的集合 $K_j=(k_1,k_2,\cdots,k_i)$，其中 $k_y(1\leqslant y\leqslant j)$ 为用户 j 检索过的关键词。如果用户 m 与用户 n 的关键词集 K_m 与 K_n 的相似度 X_{mn} 在阈值 f 以上，则表示这两个用户属于一个同类用户集。

设 $K_1=(k_1,k_2,\cdots,k_m)$ 与 $K_2=(k_1,k_2,\cdots,k_n)$ 是两个线性序列，$L=K_m\cap K_n=(l_1,l_2,\cdots,l_k)$，其中 $1\leqslant k\leqslant\min(m,n)$ 且 $l_1,l_2,\cdots,l_k$ 排列次序同其在 $K1$ 和 $K2$ 中出现的次序一致。定义线性序列 L 的长度 $Len(L)=k$，称 $\mathrm{Max}(k)$ 为 K_1 和 K_2 的相似度，记为 X_{mn}，计算公式如下：

$$X_{mn}=\mathrm{Max}((Len(L))=\mathrm{Max}(k)$$

检索实例集：用户 n 提交检索请求，这里用文档的元信息来标识一个检索结果，文档用 $T_j=(t_{j1},t_{j2},\cdots,t_{jm})$ 形式的 m 维矢量表示，如果用户 n 选择满意则表示该文档为一次成功的检索实例文档，记为 $P_v=(p_{v1},p_{v2},\cdots,p_{vm})$。对于检索词序列 $K=(k_1,k_2,\cdots,k_x)$，存在一个检索实例集 $T(K)$，其中 $T(K)$ 表示所有同类用户的所有检索实例 P_v 的集合，如下：

$$T(K)=(P_1,P_2,\cdots,P_v)$$

同义检索词：对于检索词序列 K_1、K_2，如果 $T(K_1)\subseteq T(K_2)$，且 $T(K_1)\supseteq T(K_2)$，则定义 K_1 与 K_2 为同义检索词序列，同时在实际数据库中所有同义检索词序列的检索实例集数据都用同一条记录存储表示。

星级排行因子：计算检索词序列 $K=(k_1,k_2,\cdots,k_n)$ 检索返回结果中的任意记录 j 的星级排行因子 $R_g(j)\in\{0,1,2,3,4,5\}$，如果用文档的元信息来标识一个检索结果，那么文档可以用 $T_j=(t_{j1},t_{j2},\cdots,t_{jm})$ 形式的 m 维矢量表示，这样，检索结果 T_j 与检索词序列 K 的检索实例集 $T(K)$ 中任一元素 $P_v=(p_{v1},p_{v2},\cdots,p_{vm})$ 的相关度就可以用向量的余弦来度量，计算公式如下：

$$SIM(T_j,P_v)=\cos(\vec{T_j},\vec{P_v})=\frac{\sum_k t_{jk}p_{vk}}{|\vec{T_j}||\vec{P_v}|}$$

$$=\frac{t_{j1}\times p_{v1}+t_{j2}\times p_{v2}+\cdots\cdots+t_{jk}\times p_{vk}}{\sqrt{t_{j1}^2+t_{j2}^2+\cdots+t_{jk}^2}\times\sqrt{p_{v1}^2+p_{v2}^2+\cdots+p_{vk}^2}}$$

其中，$T_j=(t_{j1},t_{j2},\cdots,t_{jm})$ 为检索结果 j 的文档元数据（如标题、摘要、作者、大小）矢量，$P_v=(p_{v1},p_{v2},\cdots,p_{vm})$ 为对应检索词序列 K 的检索实例集 $T(K)$ 中任一矢量，则有

$$Rg(j)=5\times\sin(\prod_k SIM(T_j,P_k))$$

该因子用于描述检索返回结果中的文档与用户所要找的目标文档的差距，在页面输出中以五角星的数量表示。

4.3 结果融合处理

结果融合处理是元搜索系统的最关键技术之一。一个元搜索系统的检索质量很大程度上取决于其采用的融合算法。SMS 处理过程如下：

1）将从各个独立搜索引擎的返回的检索结果保存在虚拟磁盘上的一个文件中。

2）对相同的检索结果进行处理，主要依据 URL 是否相同、文件名是否相同、文档的元信息（标题、摘要、作者、大小）是否相同等策略。如果上述策略不能执行则比较文档的全文，相似度在阈值以上的则认为是同一篇文档。

3）通过计算每条记录的相关度对检索结果进行排序。SMS 主要考虑以下几个因素：检索结果 j 在各个独立搜索引擎的顺序 R_j、包含检索结果 j 的独立搜索引擎的个数 N_j、页面的更新时间 T_j，将这三个因子线性组合得到一个基于权值的排序方法，对于检索结果 j 其相关度计算公式定义如下：

$$W_j=k_1\times R_j+k_2\times N_j+k_3\times T_j$$

其中，k_1、k_2、k_3 为常量，且 $k_1+k_2+k_3=1$。k_1、k_2、k_3 根据不同的需要赋值，比如要提高检索结果在各个原始搜索引擎的顺序 R_j 因子的作用，则可以设定较大的 k_1 值增加其在排序中的影响力度。通过测试，SMS 取 $k_1=0.5$，$k_2=0.4$，$k_3=0.1$，返回结果排序是相关度越大位置越靠前。

4）计算返回结果每条记录的星级排行因子，该因子用于描述检索返回的文档与用户所要找的目标文档的差距，在页面输出中以五角星的数量表示。

5）从检索历史数据库中找出同类相关检索词列表，按检索词被检索的次数排序输出给用户。

4.4 原始搜索引擎选择方法

作为信息来源的原始搜索引擎的选择关系到系统检索的速度与质量是否优越。因此，SMS 在选择原始搜索引擎的算法上主要从速度与质量上考虑，速度方面主要考虑网站连接速度、检索速度，质量方面主要考虑原始搜索引擎的索引页面数量、更新速度以及预定的加分权值进行评估。

4.5 虚拟磁盘

为了提高系统运行的效率，SMS 系统采用了虚拟磁盘技术，使用内存来虚拟硬盘。系统利用 2.5G 的内存虚拟了一个磁盘，将主机整个运行相关的文

件拷贝到其中，在该虚拟磁盘上运行 SMS(包括其他相关的支撑系统，如 Apache 等)。这样，因为内存的速度远远高于硬盘，所有数据交换都是在内存中进行，因此 SMS 的支持环境不存在传统的 IO 瓶颈问题，也提高了主机的运行速度。通过测试，证明该方法有效地减低了普通方式 Web Server 访问磁盘的时间，提高了整个 Web 系统性能的 15%左右。同时，该技术使得 SMS 系统的运行速度比没采用虚拟磁盘时提高了 2 个百分点左右。

5 SMS 元搜索引擎系统的评价指标与性能测试分析

SMS 系统不但吸收目前成功的元搜索系统的大部分优点，而且在某些方面的设计具有自己的特色。同时，在传统检索系统的评价指标基础上突出了未来搜索引擎具有的个性化、智能化、专业化与多媒体搜索的特性。通过以下各个不同方面的测试，SMS 系统在大部分基本评价指标方面超过传统搜索引擎的平均水平，某些方面明显优于其他元搜索引擎。

1)覆盖率(Coverage)：本系统不仅集成了 51 个独立搜索引擎，其中包括 11 个 Web 中文搜索引擎、30 个 Web 英文搜索引擎、10 个 FTP 搜索引擎，而且还集成了 10 个元搜索引擎，从信息的覆盖面与支持的独立搜索引擎与元搜索引擎数目来说，是国内最多的。

2)查全率(Recall)：由于支持了当前 Internet 上几乎所有的最优秀门户型独立搜索引擎，在查全率方面，尤其是中文搜索(包括 GB 与 BIG5)，具有优秀的表现；在随机选择关键词检索的条件下，我们比较了对 SMS 与 Google、Altavista、Yahoo、百度、天网、慧聪等 10 中英文独立搜索引擎检索结果的数量，SMS 比单个的搜索引擎检索到的结果数量平均高出 12.52%，其中中文关键词比单个中文搜索引擎平均高出 26.29%。

3)查准率(Precision)：SMS 系统采用用户行为分析技术、相关词与星级评价机制，准确地把握用户检索意图，自动匹配关键词组合，同时利用 IP 地址为用户标识存储的历史检索记录，对用户分类，从而达到很好的检索效果；在测试中，我们以一次检索结果前 20 条内能找到目标信息的次数占总检索次数的比例为评价标准，通过随机提取 1000 个英文关键词对 MetaCrawler、Dogpile 等 10 个元搜索引擎进行检索，SMS 的查准率为 82.16%，优于同类搜索引擎平均 2 个百分点左右。

4)响应时间(Response Time)：通过采用虚拟磁盘与多线程技术，同时对返回的检索结果进行过滤，SMS 在检索速度上不低于单个搜索引擎。

5)用户负担(User Effort)：SMS 的用户界面简洁明了，属于纯净搜索引擎；搜索结果网页文字描述采用索引带关键词的部分，关键词高亮显示，同时显示网页地址与网页历史记录；系统布局与界面设计从用户使用简便出发，充分考虑用户负担，使其对用户检索的影响达到最低。

6)个性化(Personalization)：SMS 通过跟踪分析用户的搜索行为，采用群体行为分析与实例推理技术，提高了用户的搜索效率；同时，用户可以自己订制搜索引擎，用户决定结果中是否出现网站描述，还可选是否高亮显示关键词；每页显示多少条结果可以从 10 到 100 条之间自由设置；各条目各颜色也可任意设置；选择哪些独立搜索引擎参与。

7)智能化(Intelligentization)：本系统在检索中引进专家系统的设计思路，建立动态的用户成功检索知识库，对用户的查询计划、意图、兴趣方向进行推理、预测并为用户提供有效的检索关键词；在测试中，我们以用户表示满意为检索成功，以检索成功次数占总检索次数的百分比为评价标准，建立了一个 25000 条成功检索记录的知识库，然后让 35 个不同行业的用户进行随意检索，SMS 系统在推理用户检索意图方面的准确率达到了 42.15%。

进一步的工作与结论 在测试结果中，SMS 元搜索引擎在覆盖率、查全率与查准率都优于单个搜索引擎；在设计思想与实现方法上力求突破技术上的限制，充分吸收其他元搜索引擎的优点；同时在对探索下一代搜索引擎的个性化、智能化、专业化、多媒体搜索发展方向上做了一些有意义的工作。该系统的各模块算法在选择上都通过了严格的性能测试分析，同时利用了一些现有的成果；引进专家系统中的知识库与推理机技术使得系统的检索质量得到了很大的提高。

综上，在各类信息检索系统中，SMS 作为一个追求高效的检索质量与全面的个性服务的元搜索引擎，其所采用的实现技术与设计思路，使得其不失为一个具有实际应用意义的信息检索系统。

参 考 文 献

1 Lawrence S, Giles C L. Accessibility of Information on the Web [J]. Nature, 1999, 400(8): 107～109

2 Aslam J, Montague M. Models for metasearch[A]. In: the Proc. of the 24th ACM SIGIR conf. on Research and Developement in Information Retrieval[C], Sept. 2001. 276～284

3 Montague M, Aslam J. Relevance score normalization for metasearch[A]. In: the Proc. of the ACM Tenth Intl. Conf. on Information and Knowledge Management (CIKM) [C], Nov. 2001. 427～433

4 阳小华，刘振宇，谭敏生，等. 元搜索引擎系统合成算法的约束条件[J]. 软件学报，2002，13(7)：1264～1267

5 皮鹏，张国印. 智能元搜索引擎的研究[J]. 应用科技，2001，28(8)：24～26