个性化元搜索结果整合算法的研究

胡文江[1],李 磊[1],姜文涛[2],高永兵[1]
(1.内蒙古科技大学 信息工程学院,内蒙古 包头 014010;
2.西安应用光学研究所,陕西 西安 710065)

摘 要:针对当前元搜索引擎存在的问题,提出一种个性化元搜索结果整合算法。首先对成员搜索引擎根据相应条件设定权值,对各成员搜索引擎得到的搜索结果按贡献量加权分块排序,根据用户检索词条与兴趣库和元搜索结果的文本相关度对块内搜索结果进行整合排序。实验结果表明,该算法能够满足不同用户的个性化需求,在保证了搜索结果查全率的同时提高了查准率,很大程度上改善了用户检索效果和效率。

关键词:个性化元搜索;多重排序;权重;排序整合;相关度
中图分类号:TP391 **文献标识码**:A **文章编号**:1674-7720(2013)15-0054-04

Personalized metasearch results integration algorithm

Hu Wenjiang[1], Li Lei[1], Jiang Wentao[2], Gao Yongbing[1]
(1.School of Information Engineering, Inner Mongolia University of Science and Technology, Baotou 014010, China;
2.Xi'an Institute of Applied Optics, Xi'an 710065, China)

Abstract: In connection with the existing problems in search engines, this paper presents a personalized metasearch engine result integration algorithm. Firstly, this algorithm merges the search engine results into personal fusion sort with user's interests, and then it combines the traditional multi-level matching algorithm with text relevance of search results sort integration. It shows that the algorithm is able to meet the individual needs of different users through the experimental results. This algorithm can improve the precision greatly while ensure the search results recall rate; and it improves the effectiveness and efficiency of users' searching.

Key words: personalized metasearch; multiple sort; weights; sort integration; relevancy

随着 Internet 的迅速普及,网络上的信息量成指数增长。由于网络上的信息是海量和无组织的,具有分布散、动态变化、结构复杂等特点,人们在互联网上检索信息主要依靠搜索引擎。单个搜索引擎检索机制、范围、算法等的不同,导致同样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]。LAWRENCE S[2]等研究表明,任何一个搜索引擎索引的 Web 页面都不超过页面总数的 1/3。因此,要想获得一个全面、准确、符合用户需求的搜索结果,就必须反复调用多个搜索引擎进行搜索。

如何在无比庞大的网络信息库中更快速、更准确地找到能够满足用户所需的信息,已经成为 Web 搜索领域研究的热点问题。搜索引擎的优劣、成功与否完全由用户对其搜索结果的满意度决定。目前存在的搜索引擎实用性不断加强,在一定程度上满足了人们检索信息的需求,但其自身在查全率和查准率方面与生俱来的局限性,无法为用户提供更全面、更精准的检索需求。

元搜索引擎是一种基于搜索引擎的搜索引擎[3],其搜索过程是首先将用户提交的搜索请求分发给多个成员搜索引擎,再将各个成员搜索引擎的搜索结果进行整合反馈给用户。元搜索引擎没有独立的数据库,而是依赖于其他成员搜索引擎,因此元搜索引擎在进行搜索时会遇到这样的问题:成员搜索引擎的搜索算法不同、源数据库和数据采集技术不同、各个成员搜索引擎返回文档没有统一的全局相似度等问题,造成搜索结果各不相同,纷繁复杂而不能满足用户的搜索需求。元搜索引擎最终要为用户提供个性化搜索服务[4]。针对这种情况,设计一种元搜索引擎搜索结果排序整合算法,对各个搜

索引擎返回的文档进行分析、排序，尽可能返回更加贴切的结果给用户，具有很大的可行性。本文针对元搜索引擎中搜索结果整合问题，提出了一种基于用户兴趣的结果整合排序方法，在检索过程中考虑了用户兴趣，实现了元搜索引擎的个性化，既保证了搜索结果的查全率，又提高了查准率。

1 个性化元搜索引擎结果整合算法的实现

结果合成是元搜索引擎的一个核心问题，是将多个搜索引擎的检索结果整合到一起的过程。以往元搜索引擎中没有根据成员搜索引擎权值的结果贡献量概念，也没有根据用户使用情况动态进行调整，在合成算法中，如 Comb Sum、CombMNZ、Round-Robin 等[5-6]没能结合用户兴趣和成员搜索引擎的优先级问题，使得得到的结果合成效果不是很理想。为此，本文提出一种算法，结合了以往各种优秀算法的长处，加入了用户兴趣、成员引擎结果贡献量及多维排序，为用户提供更加全面、准确、高效、个性化的检索服务。算法思想是：首先对成员搜索引擎根据其排名信息和用户偏爱度设定权值，得到各成员搜索引擎的结果贡献量，根据多维排序算法对结果分块，对各个搜索引擎以及用户所关注内容设置权重，根据用户反馈动态更改相应权重，根据用户查询和用户兴趣库与搜索结果文档相关度权值整合排序，将最终最符合用户个性化需求的结果返回给用户，既保证了搜索结果的查全率，又提高了查准率。

1.1 算法的实现过程

系统初次使用成员搜索引擎的排名信息作为搜索结果贡献量选择依据。假设系统要从 N 个成员搜索引擎中检索出 n 篇文档，则排名第 i 位的搜索引擎需要返回的文档数量为 n_i，计算方法如式(1)所示。

$$n_i=n\times\frac{2(1+N-i)}{N(N+1)} \tag{1}$$

考虑元搜索引擎效率，N 取值为 $(1,10]$，再根据用户对相应成员搜索引擎结果的点击反馈及用户的偏爱程度生成式(2)，其中 c_i 为用户在使用过程中对成员搜索引擎 i 的结果点击反馈总数，成员搜索引擎根据 c_i 得分动态学习，α_i 为用户对成员搜索引擎 i 的偏爱程度，取值为 $[0,0.5]$，当用户未选择相应成员搜索引擎时，设为 -1，成员搜索引擎 i 的最大结果数收敛于 $n/2$，以保证元搜索引擎的查全率及效率。计算方法为：

$$n_i=n\left(\frac{N-3}{2(N+1)}\times\frac{c_i}{\sum_{k=1}^{Y}c_k}+\frac{2(1+N-i)}{N(N+1)}(1+\alpha_i)\right) \tag{2}$$

其中，Y 为用户的历史总点击次数。

对各成员搜索引擎的搜索结果按先后次序进行排序。首先为其赋初值 $W(r_i)$，初值为：

$$W(r_i)=\left(\frac{N-3}{2(N+1)}\times\frac{c_i}{\sum_{k=1}^{Y}c_k}+\frac{2(1+N-i)}{N(N+1)}(1+\alpha_i)\right) \tag{3}$$

其中，r_i 为搜索结果。

对搜索引擎结果中的重复进行处理，同一结果若被多个成员搜索引擎检索到，则其局部相关度相对较高，进行去重处理，更新结果权值 $W(x,r_i)$，如式(4)所示。

$$W(x,r_i)=\left(\frac{N-3}{2(N+1)}\times\frac{c_i}{\sum_{k=1}^{Y}c_k}+\frac{2(1+N-i)}{N(N+1)}(1+\alpha_i)\right)+(x/10) \tag{4}$$

其中，x 为该结果 r_i 重复次数，用户对该结果的点击反馈使相应成员搜索引擎点击总数为 c_i+1。

对成员搜索引擎得到的结果根据多级匹配算法进行分块，考虑效率和各成员搜索引擎特点，将结果分为 5 块，每个成员搜索引擎的 $[1,n_i/5]$ 个结果组成元搜索引擎第一组搜索结果，$[n_i/5+1,2n_i/5]$ 组成元搜索引擎第二组搜索结果，依次分块，$[4n_i/5+1,n_i]$ 组成元搜索最后一组搜索结果。

搜索结果的标题和摘要集中了文档的主要信息，用它们来代替文档计算全局相关度能很好地得到近似值[7]，能有更好的检索时效。

定义 1 设查询串 q 有 X 个词条，每个搜索结果 r_i 的摘要中包含有这 X 个词条的 N 个 $(N\leqslant X)$，则 r_i 与查询串 q 的词条匹配等级为 N。

对每个结果 r_i：

(1)仿照摘要排序法，计算结果 r_i 与查询串 q 的普通相关度 $Rq(q,r_i)$。

先计算每个词条与结果 r_i 的相关度，计算方法为：

$$Rl(l_j,r_i)=\sum_{k=1}^{\mathrm{Occ}}\ln(\mathrm{Len}(r_i)/\mathrm{Loc}(l_j,k,r_i)) \tag{5}$$

其中，$\mathrm{Len}(r_i)$ 表示搜索结果 r_i 摘要的长度；$\mathrm{Occ}(l_j,r_i)$ 表示词条 l_j 在 r_i 摘要中出现的次数；$\mathrm{Loc}(l_j,k,r_i)$ 表示词条 l_j 在 r_i 摘要中第 k 次出现的位置。

再计算查询串与结果 r_i 的相关度 $Rq(q,r_i)$，计算方法为：

$$Rq(q,r_i)=\sum_{j=1}^{X}Rl(l_j,r_i) \tag{6}$$

(2)计算结果 r_i 与查询串 q 的词条匹配等级 $MG(q,r_i)$。

(3)计算查询串 q 与文档 r_i 的最终相关度 $R(q,r_i)$，相关度计算方法为：

$$R(q,r_i)=MG(q,r_i)\times Rq(q,r_i) \tag{7}$$

1.2 结合用户兴趣库计算用户查询与搜索结果相关度

对于查询结果 r_i，元搜索引擎解析出其中包含的所有特征词组组成的特征词集 $TS(r_i)$。定义查询结果对于 CS 的特征词集：

$$\mathrm{TCS}(r_i)=\{t\mid t\in T(c)\wedge c\in CS\wedge t\in TS(r_i)\}$$

表示为 $(t_1,t_2,\cdots,t_k)$，k 为 $\mathrm{TCS}(r_i)$ 中特征词的个数。

定义 2 对于用户兴趣类别 $c(c\in CS)$，查询结果 r_i 对于 $\mathrm{TCS}(r_i)$ 的权重 $u(r_i,c)$，表示为 $(w_1,w_2,\cdots,w_k)$。其中 $w_i(1\leqslant i\leqslant K)$ 是对应特征词 $t_i(t_i\in\mathrm{TCS}(r_i))$ 在用户兴趣类别 c 中的权重，表示为：

$$W_i=\begin{cases}1, t_i\in \mathrm{TCS}(r_i)\cap T(c)\\ 0, t_i\notin \mathrm{TCS}(r_i)\cap T(c)\end{cases} \quad (8)$$

定义 3 $\mathrm{TCS}(r_i)$ 中术语的加权向量 $x(r_i)$，表示为 $(x_1,x_2,\cdots,x_k)$，其中 $x_i(1\leqslant i\leqslant k)$ 为特征词 $t_i(t_i\in \mathrm{TCS}(r))$ 对查询结果 r_i 的重要性，定义为：

$$x_i=\begin{cases}a, t_i \text{ 是 } r_i \text{ 标题中的特征词}\\ b, t_i \text{ 是 } r_i \text{ 摘要中的特征词}\end{cases} \text{且} \sum a+\sum b=1 \quad (9)$$

根据上述定义，计算查询结果 r_i 与用户兴趣类别 $c(c\in CS)$ 的相似度为：

$$\sin(r_i,c)=\frac{\sum_{i=1}^{k}(w_i\times x_i)}{\sqrt{\sum_{i=1}^{k}w_i^2\times\sum_{i=1}^{k}x_i^2}} \quad (10)$$

根据用户查询 q 和 q 对应的用户兴趣类别 c 的相似度 $\sin(q,c)$，计算查询结果 r 与用户查询 q 在兴趣类别 c 上的相似度 $\sin(r_i,q,c)=\sin(r_i,c)\sin(q,c)$。计算查询结果 r_i 与用户查询 q 相似度为：

$$\sin(r_i,q)=\frac{1}{|CS|}\sum_{c\in CS}\sin(r_i,q,c) \quad (11)$$

其中，$|CS|$ 为集合中兴趣类别个数。

1.3 计算用户查询与搜索结果的最终相关度

最终相关度 $\mathrm{Final}(r_i)$ 计算如下：

$$\mathrm{Final}(r_i)=A\times R(q,r_i)+B\times\sin(r_i,q)+C\times\left(\frac{N-3}{2(N+1)}\times\frac{c_i}{\sum_{k=1}^{Y}c_k}+\frac{2(1+N-i)}{N(N+1)}\right)(1+\alpha_i)+(x/10)$$

其中，A、B、C 为常数，$A+B+C=1$。A、B、C 表示不同得分对于最终相关度的重要性，对每块内结果根据最终相关度得分进行排序。

2 实验结果与分析

2.1 实验环境

本次实验主要使用个性化元搜索引擎与常用的搜索引擎（Google、百度、yahoo、sogou）进行搜索性能对比，其中使用成员搜索引擎（Google、百度、yahoo、sogou、bing、youdao、360），通过实验的验证，个性化搜索引擎有较好的性能。

验证实验所使用的试验硬件平台为 Intel(R)Core(TM)2 Duo CPUE7400，主频为 2.8 GHz，2 GB 内存，320 GB 硬盘。操作系统为 Linux，所有算法及实验系统用 C、PHP 实现。

2.2 评价标准

在信息检索中，衡量系统的基本指标主要是：查全率（Recall）、查准率（Precision）及响应时间等。

（1）查全率

$$\text{查全率}=\frac{\text{检索出的相关文档数}}{\text{文档集中所有的相关文档数}}\times 100\%$$

本实验系统由于多个成员搜索引擎的同时使用，所以能得到较高查全率。即

$$U(S_1,S_2,\cdots,S_n)\geqslant \mathrm{Any}(S_1,S_2,\cdots,S_n)$$

（2）查准率

$$\text{查准率}=\frac{\text{检索出的相关文档数}}{\text{检索出的文档总数}}\times 100\%$$

2.3 实验结果

查准率是衡量一个搜索引擎的重要指标之一，表 1 是各个搜索引擎的查准率比较。从表 1 可以看出，个性化搜索较其他几种常用搜索在查准率性能方面要高。

表 1 各个搜索引擎的查准率比较

序号	搜索引擎	查准率/%
1	百度	68
2	Google	73
3	sogou	63
4	雅虎	65
5	个性化元搜索	86

表 2 所示为各个搜索引擎的平均响应时间对比，可以看出，百度和 sogou 的平均响应时间明显优于其他搜索引擎，主要是因为其服务器在国内，所以响应时间较短；而谷歌和雅虎则由于服务器的地域性，响应时间较长；个性化元搜索由于结合了用户的兴趣和多个搜索算法，因此响应时间最长。

表 2 搜索引擎平均响应时间对比

序号	搜索引擎	平均响应时间/s
1	百度	0.34
2	Google	1.61
3	sogou	0.49
4	雅虎	1.10
5	个性化元搜索	2.67

本文提出了一种个性化元搜索引擎结果整合算法，在考虑了成员搜索引擎自身特点与用户兴趣及用户查询的相关度等因素后，引入成员搜索引擎结果贡献量及结果分块，加入了词条匹配等级的概念，更好地体现了用户的个性需求。经实验验证，此算法能够使用户在庞大繁杂的元搜索结果中更快找到自己感兴趣的结果，较其他元搜索引擎以及其他整合排序算法大大提高了查询的效率和效果。

参考文献

[1] 梁美玉，杜军平，高田.基于领域知识的个性化智能语义检索系统 [J]. 中南大学学报（自然科学版），2011(42)：866.

[2] LAWRENCE S, GILES C L. Searching the World Wide Web [J]. Science, 1998, 280(5360): 98-100.

[3] 李广建，黄崑.元搜索引擎及其主要技术[J].情报科学，2002，2(2)：22-27.

[4] 徐娟，王群.3G 融合计费解决方案探讨[J].电信快报：网

络与通信,2008(9):13-17.

[5] NAIK S K, MURTHY C A. Hue-preserving color image enhancement without gamut problem[J]. IEEE Transactions on Image Processing, 2003, 12(12): 1591-1598.

[6] HUANG K, WANG Q, WU Z. Color image enhancement and evaluation algorithm based on human visual system[C]. 2004. Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing, 2004, 3: iii-721-4 vol. 3.

[7] MONTAGUE M, ASLAM J A. Relevance score normalization for metasearch[C]. Procee dings of 10th International Conference on Information and Knowledge Management.Atlanta, USA, 2001:427-433.

[8] WHITE R W, KAPOOR A, DUMAIS S T. Modeling long-term search engine usage[M]. User Modeling, Adaptation, and Personalization, Springer Berlin Heidelberg, 2010.

[9] SI L, CALLAN J. Using sampled data and regression to merge search engine results[C]. Proceedings of the 25th annual international ACM SIGIR conference on Research and development in information retrieval. ACM, 2002: 19-26.

[10] 张卫丰,徐宝文,周晓宇,等.元搜索引擎结果生成技术研究[J].小型微型计算机系统,2003,24(1):34-37.

[11] KEYHANIPOUR A H, MOSHIRI B, PIROOZMAND M, et al. Web fusion: fundamentals and principals of a novel Meta search engine[C]. Neural Networks, 2006. IJCNN'06. International Joint Conference on IEEE, 2006: 4126-4131.

[12] BINGRU L Y C X Y. Research on Web mining-based intelligent search engine[J]. Computer Engineering and Applications, 2002(4): 11.

(收稿日期:2013-04-12)

作者简介:

胡文江,男,1959年生,教授,主要研究方向:数据库和计算机网络。

李磊,男,1981年生,硕士研究生,主要研究方向:数据库和计算机网络。

姜文涛,男,1984年生,硕士,主要研究方向:计算机网络与信息处理。