第35卷 第10期 计 算 机 工 程 2009年5月
Vol.35 No.10 Computer Engineering May 2009
·人工智能及识别技术· 文章编号：1000—3428(2009)10—0198—03 文献标识码：A 中图分类号：TP354

基于查询扩展和数据融合的检索过程优化

王 非
(广东外语外贸大学信息科学技术学院，广州 510006)

摘 要：介绍典型的检索过程优化方法——数据融合和基于相关度反馈的查询扩展，前者通过集成多个检索结果提高检索性能，后者执行多次查询，依据前次结果修改/扩展用户查询，以求更好地反映用户信息需求，并在此基础上提出一种新的检索过程优化方法——HQD方法，由相关度反馈结果生成多个替代查询，在检索这些替代查询后，采用求和余弦法生成最终检索结果。仿真实验结果表明，该方法是有效的。

关键词：相关度反馈；数据融合；检索过程优化

Optimization of Retrieval Procedure Based on Query Expansion and Data Fusion

WANG Fei
(School of Information Science & Technology, Guangdong Foreign Study University, Guangzhou 510006)

【Abstract】Typical optimized retrieval procedure methods such as data fusion and relevance feedback-based query expansion are introduced. While data fusion improves retrieval performance by merging multiple retrieval results, relevance feedback revises user query according to previous retrieval result and runs the new query to improve retrieval performance. On the basis of this, a novel optimized retrieval procedure method called HQD is presented, which selects top-ranked documents from the relevance feedback result, runs these documents as surrogate queries, and merges the retrieval results using a sum cosine measure. Experimental results show this method is effective.

【Key words】 relevance feedback; data fusion; optimization of retrieval procedure

1 概述

现有数据融合方法可分为多检索机制单一查询[1]与多查询单一检索机制2大类[2]。研究发现：对于同一信息检索任务，尽管不同检索者会使用不同查询，其检索结果却会包含类似的相关文献集，而包含的不相关文献集却不尽相同，因此，一个最简单的投票机制都会使这些共有的相关文献脱颖而出。数据融合会提高相关文献的检出几率，降低非相关文献的检出几率，从而提高检索性能。

在合并多个检索结果集时，已有研究多采用Ad Hoc方法，如有些研究者使用求和，有些则使用正规化得出最终相关度取值。Ad Hoc方法无法保证有效性，同一种方法在不同文献集和系统上可能表现出较大的性能差异。

相关度反馈方法通过扩展查询重新估计用户信息需求，可分为：实反馈，即人工反馈，用户被要求决定初始结果中哪些文献是相关的；自动反馈，也称为伪反馈或盲反馈，初始结果中排在前面的几篇文献被自动认为是相关的，无需用户参与；如果在自动反馈的基础上，初始结果中排在最后的几篇文献被自动认为是无关的，并且在反馈机制中也使用它们参与查询扩展，则称这种反馈为完整自动反馈。

现代相关度反馈方法基本沿袭Rocchio的想法。Rocchio使用一种自适应词条加权方法来修改用户初始查询：$q' = aq_0 + b\sum_{d_i \in R} d_i - c\sum_{d_i \in I} d_i$，该方法的最大困难是确定系数$a$, b, c。由于无法从理论上确定哪组取值会得到最佳检索性能，因此研究者通常采用试错方法确定它们。

本文提出一种检索过程优化方法——HQD方法，采用相关度反馈技术生成多个替代查询，对替代查询的检索结果使用数据融合，并生成最终检索结果。

2 HQD方法

HRD方法基本过程分为3个步骤：

(1)用初始查询 q_0 进行一次向量空间模型机制下的常规检索，即计算 q_0 和文献集 D 中每篇文献 $d_i(i=1,2,\cdots,N)$的余弦夹角 $c_{0i}=\cos(q_0,d_i)$。降序排列 c_{0i} 对应文献生成新文献集 D'。

(2)选取(手动、自动皆可)D'中前 k-1 篇文献作为 q_0 的替代查询。k 的取值可以人工赋予，也可采用启发式选择规则自动生成，这 k-1 篇文献连同 q_0 记为 K。对 K 中每个查询，重复第(1)步，得到 $c_{ij}=\cos(d_i,d_j)$，其中，$d_i \in K$, $d_j \in D$ 及 k 个排序列表：

$<c_{0,1},c_{0,2},\cdots,c_{0,N}>$
$<c_{1,1},c_{1,2},\cdots,c_{1,N}>$
…
$<c_{k-1,1},c_{k-1,2},\cdots,c_{k-1,N}>$

(3)采用求和方式合并这 k 个列表生成最终检索结果。HQD方法与数据融合及相关度反馈之间的联系很明显。第(2)步的基本理论是相关度反馈，检索结果中排在前面的文献(理想状态下，理论上最相关的文献)能反映相关文献使用查询词的状态，因此，可被视为用户初始查询的“某种修正”。

基金项目：国家自然科学基金资助项目(70473066)
作者简介：王 非(1976－)，男，博士，主研方向：信息检索，电子商务，知识管理
收稿日期：2008-11-10 **E-mail：**gdufs.wangfei@gmail.com

第(3)步的理论依据是数据融合，通过使用多重查询然后合并这些查询返回的结果可以提高检索性能。由于在第(2)步中没有采用自适应学习方法，各个查询被视为相互独立，因此HQD方法避开了容易出错的系数选择问题，而第(3)步所采用的求和方法同其他Ad Hoc方法相比也更容易调整。

在确定HQD方法中，第(2)步所需的k篇文献是HQD方法的一个要点。在理想状态下，这k-1篇文献都应该是相关的，但是只有采用人工查询扩展时才会如此，如果第(2)步使用自动查询扩展，这一点就无法保证了。下面讨论HQD方法如何处理这一问题。

3 集合K的确定

在HQD方法第(2)步中生成的集合K非常关键，如果K中出现了非相关文献，则会使替代查询严重偏离用户的信息需求，进而影响第(3)步数据融合的效果，降低HQD方法的性能。

K集合元素有人工选择和自动选择2种确定方法。如果采用人工选择，HQD方法可以根据需要将k设为任意值(如2,3,4等)，借此调整集合K的规模，进而影响检索性能，但是这需要用户的额外参与。本文倾向于自动选择，故不讨论如何在HQD中实现人工查询扩展。

HQD方法实现自动选择的基本思路是：使用若干启发式规则生成一个临界值，然后将第(1)步初始检索结果中所有满足临界值要求的文献自动选择进集合K。

文献[3]研究了检索结果排序得分的分布问题，发现：

(1)在向量空间模型返回结果中，文献和查询间的余弦取值依据排序位置呈指数下降，如图1所示。

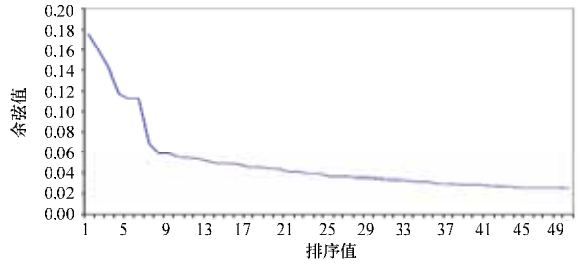

图1 余弦值分布

(2)排序位置靠前文献的余弦取值之间的差异$\cos(q,d_i)$-$\cos(q,d_{i-1})$比排在后面的文献之间的差异大，如图2所示。

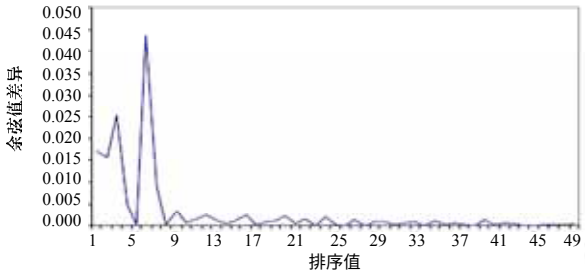

图2 余弦值差异分布

这种现象是由于排在尾部的不相关文献的余弦取值基本上都逼近于0，因此差异较小，而排在前面的相关文献分别从不同角度不同程度的捕捉到用户需求，同时相关文献的数量较少，因此，其余弦取值的差异会有较大变化。

基于该检索结果分布理论，HQD方法采用3条自动选择规则：

规则1 文献d对应的余弦取值不得少于列表中最高余弦取值的50%。

规则2 如果对结果列表中任意5篇连续文献序列S求其得分的标准差，那么d所在S的标准差不得少于列表中出现的最高标准差的20%。

这里的临界值50%和20%是个经验值，可以调整以控制集合K的规模。实验表明，当取50%和20%时，规则1返回的平均文献数为9，即集合K的规模平均为10(q_0自动成为K的一员)。

在查询结果列表中，相关文献集和不相关文献集在向量空间模型机制下会表现出图1和图2所示的分布形态。如果要从该图上分开两者，分界点就应该位于曲线最陡处。据此，HQD方法采用的第3条自动选择规则为：

规则3 在初始结果列表中，以相邻2篇文献的余弦值差异最大处为分界，排序位置在该分界点之前的文献全部可以入选集合K。

在规则3中，分界点的出现位置是不定的，受具体查询和具体文献集的制约。在图2中，分界点出现在第5位并不是普遍现象。

通过规则1~规则3所设立的临界值，HQD方法自动的从第(1)步返回的结果列表中生成集合K。

4 HQD方法性能评估

将向量空间模型和相关度反馈模型作为测评基准，使用TREC测试集作为评估数据，并采用201号~300号查询。

HQD方法默认使用向量空间模型机制，故采用文献[4]关键词加权机制对201号~300号查询以及测试集文献进行了向量化。

由于规则3比较严格，因此规则3和规则1、规则2被分开使用，总共测评6套机制：

(1)VS机制，使用向量空间模型的检索系统，加权机制采用Buckley方法。

(2)RF10，基本相关度反馈机制，采用目前常用的伪反馈，自动认定VS机制返回结果的前10位为相关文献，查询修正采用Rocchio方法。

(3)RFK12，增强型相关度反馈机制，使用规则1和规则2自动从VS机制返回结果中生成集合K作为替代查询。

(4)HQD12，第(2)步采用规则1和规则2生成集合K的HQD方法。

(5)RFK3，增强型相关度反馈机制，使用规则3自动从VS机制返回结果中生成集合K作为替代查询。

(6)HQD3，第(2)步采用规则3生成集合K的HQD方法。

已有研究表明，在TREC测试集上运行向量空间模型返回的前10篇结果中，平均只有3篇相关文献，这意味着如果在TREC上直接采用RF10机制会有较强的噪声词条干扰[4]。由于相关度反馈机制使用Rocchio方法，如果b值大过a值(即认为检中文献比q_0更好地反映用户信息需求)就会加强这种干扰，测试参数<a=3, b=7>, <a=4, b=8>, <a=8, b=18>后，证实了这种情况，因此加强了q_0在相关度反馈中的重要程度，以求抵消干扰，经过一些实验，选择了<a=8, b=3>作为相关度反馈机制的Rocchio参数。

表1为经过综合后的主要性能数据，可以看出，HQD方法(HQD12, HQD3)和相关度反馈系统(RF10, RFK12, RFK3)的检索性能要明显优于向量空间系统，HQD方法和相关度反馈系统相比，性能提高不明显。

表 1　基于 TREC 数据集的 HQD 方法测评结果

测试机制 性能数据	VS	RFK12	HQD12	RFK3	HQD3	RF10
检中相关文献数	13 802	13 802	13 802	13 802	13 802	13 802
相关文献总数	7 114	7 488	7 239	7 358	7 282	7 512
最小平均查准率	0.000 3	0.000 0	0.000 0	0.000 0	0.000 0	0.000 0
最大平均查准率	0.783 4	0.760 0	0.770 0	0.740 0	0.770 0	0.750 0
平均查准率	0.156 8	0.180 0	0.170 0	0.170 0	0.170 0	0.170 0
R 查准率	0.231	0.240	0.240	0.240	0.240	0.240
与 VS 相比平均查准增强率	—	0.135	0.118	0.095	0.087	0.111
与 RF 相比平均查准增强率	—	—	0.013 3	—	0.011 2	—
T 检验 H0：HQD 和 RF 平均查准率相同	—	—	0.03	—	0.02	—
T 检验 H0：HQD 和 VS 平均查准率相同	—	—	0.36	—	0.23	—

HQD 方法是对向量空间模型的改进，采用 2 个前提假设：

(1)求和余弦方法的有效性。求和余弦基本按照文献同前 k 位文献中心点的距离远近排序文献。由于前 k 位文献一般情况下是不同作者独立创建的，因此替代查询时，可以视为用户信息需求的独立表达，其中心点就可视为用户信息需求的一个近似估计。从这一点出发，这 k 篇文献(也就是集合 K)的中心点 s 越逼近用户信息需求，求和余弦方法就越充分有效，那么 HQD 方法对性能的优化就越明显。

(2)集合 K 比 q_0 更好地反映用户信息需求。众多研究已经证明此点，表 1 中 RFK12 和 RFK3 的性能数据也证实了这点。假设用户信息需求是固定的，其对文献相关性的判定是一致的，且文献集包含的相关文献一定可以满足其信息需求，那么用户信息需求可视为文献集中所有相关文献的均值 u，这与 Rijsbergen 提出的聚类假设(相关文献会聚集在文献集形成的文献空间中的某个中心点周围)相符合。

可以看出，HQD 方法的性能直接受集合 K 的影响，其性能反比于集合 K 中心点 s 和 u 之间的距离。由于 HQD 以向量空间模型为基础，因此 s 和 u 之间的距离又反比于集合 K 中真实相关文献的比例。

HQD 方法和相关度反馈方法都采用假设(2)，使用替代查询集合 K 来表示 u，因此，两者性能表现要比使用 q_0 表示 u 的 VS 方法好。

假设(1)指出求和余弦方法的性能反比于 s 和 u 之间的距离，受 TREC 数据集所限，HQD 方法和相关度反馈方法所得集合 K 中真实相关文献的比重并不会太高，因此，s 并不会很靠近 u，HQD 方法的第(3)步求和余弦并没有发挥多大作用，而只有前 2 步的 HQD 方法和相关度反馈方法很相似，因此，在测评中两者表现出相近的性能。

鉴于此，通过采用其他自动选择规则，如关键词、聚类等，提高集合 K 中真实相关文献的比例，HQD 方法的性能还有进一步上升的可能。

5　结束语

实验结果表明，HQD 方法能有效提高检索性能。理论分析表明 HQD 方法还可以采用更多的自动选择规则生成更好的集合 K，其实验性能还能进一步得到提升。

参考文献

[1] Kantor P. Predicting the Effectiveness of Naive Data Fusion on the Basis of System Characteristics[J]. Journal of American Society for Information Science, 2000, 51(13): 1177-1189.

[2] Lee J. Combining the Evidence of Different Relevance Feedback Methods for Information Retrieval[J]. Information Processing & Management, 1998, 34(6): 681-691.

[3] Manmatha R, Rath T. Using Models of Score Distributions in Information Retrieval[C]//Proc. of the 24th ACM Conf. on Research and Development in Information Retrieval. New York, USA: [s. n.], 2001.

[4] Buckley C, Salton G, Allan J. The Effect of Adding Relevance Information in a Relevance Feedback Environment[C]//Proc. of the 17th Annual Int'l ACM Conf. on Research and Development in Information Retrieval. New York, USA: [s. n.], 1994.

编辑　陈　文

(上接第 95 页)

参考文献

[1] 柳赛男, 柯映林, 李江雄. 基于调度策略的自动化仓库系统优化问题研究[J]. 计算机集成制造系统, 2006, 12(9): 1438-1443.

[2] 冯辉宗, 陈　勇, 刘　飞. 基于遗传算法的配送车辆优化调度[J]. 计算机集成制造系统, 2004, 10(10): 81-84.

[3] 田国会, 刘长有, 林家恒. 自动化立体仓库若干优化调度问题及其研究进展[J]. 山东工业大学学报, 2001, 31 (1): 12-17.

[4] 张　攀, 田国会, 贾　磊. 旋转货架拣选作业优化问题的新型混合遗传算法求解[J]. 机械工程学报, 2004, 40(6): 34-38.

编辑　陈　文

(上接第 197 页)

[2] Hart E, Timmis J. Application Areas of AIS: The Past, Present and the Future[J]. Journal of Applied Soft Computing, 2008, 8(1): 191-201.

[3] Castro L N, Timmis J. An Artificial Immune Network for Multimodal Function Optimization[C]//Proc. of the IEEE CEC'02. Honolulu, Hawaii, USA: [s. n.], 2002.

[4] Attux R R F, Loiola M B, Suyama R, et al. Blind Search for Optimal Wiener Solutions Using an Artificial Immune Network Model[C]//Proc. of the IEEE Int'l Conf. on Genetic and Evolutionary Computation for Signal Processing and Image Analysis. [S. l.]: IEEE Press, 2003.

编辑　陈　文