一种基于相关反馈的信息检索模型

金光赫[1,2] 王兴伟[1] 曲大鹏[1,3] 蒋定德[1]
(东北大学信息科学与工程学院 沈阳 110819)[1] (金策工业综合大学应用程序学院 平壤朝鲜)[2]
(辽宁大学信息学院 沈阳 110036)[3]

摘 要 针对现有信息检索系统难以按查询需求处理检索文档的问题,提出了一种基于相关反馈的信息检索模型,分析了查询词分解,推导了相关反馈机制和正规化过程,并进一步阐述了文档提取方法。提出的模型通过相关反馈和查询词扩展,克服了传统方法无法计算文档与查询词之间的相似度问题,并能有效地处理检索文档。仿真结果证明了该模型的有效性和可行性。

关键词 信息检索,相关反馈,查询分解,布尔检索,查询扩展
中图法分类号 TP393 **文献标识码** A

Information Retrieval Model Based on Relative Feedback

KIM Kwang-hyok[1,2] WANG Xing-wei[1] QU Da-peng[1,3] JIANG Ding-de[1]
(College of Information Science and Engineering, Northeastern University, Shenyang 110819, China)[1]
(School of Application Program, Kim Chaek University of Technology, Pyongyang, DPR Korea)[2]
(School of Information, Liaoning University, Shenyang 110036, China)[3]

Abstract In view of the existing information retrieval systems which are difficult to process document retrieval problem according to query demand, this paper proposed one retrieval model based on relative feedback, analyzed the decomposition of the query words, derived the correlation feedback mechanism and normalization process, and further expounded the document extraction method. Through relative feedback and query expansion, this model can overcome the problem that traditional methods can not calculate the similarity between document and query word, and can effectively deal with document retrieval. The simulation results proved the validity and feasibility of the model.

Keywords Information retrieval, Relative feedback, Query splitting, Boolean retrieval, Query expansion

1 引言

经典信息检索理论认为信息需求决定信息检索的效率。网络环境下,信息需求的提出及表达由用户决定。清晰的用户表达是提高检索效率的关键。但是,精确表达出用户查询非常困难[1],因为绝大多数检索系统的标记和检索过程不透明,用户很可能不熟悉检索语言或检索式的表达;此外,用户需求与查询表达的对应转换也可能存在不一致的地方。

纯粹的布尔检索模型无法计算文档和查询词之间的相似值,所以不能将文档按满足查询的程度进行排列。Fuzzy集合模型通过利用索引词在文档内的重要性来计算文档值[2,3],但产生的文档值其准确度往往较差。为了克服Fuzzy集合模型的缺点,开发了MMM、Paice、P-norm模型[4-6]。MMM模型虽然检索速度较快,但效果较差;Paice和P-norm模型检索速度较慢。

为了将用户需求转换成相应的查询表达,研究者们提出了查询扩展的方法,即修正检索表达式以满足信息需求。相关反馈[7]是其中备受关注的一种自动扩展查询方法,其主要思想是:检索系统在初始查询到一组样本文档的基础上,根据用户在样本文档中的相关性选择,构造出改进的查询表达式,再次进行查询,通过调整检索策略来得到更准确的相关文献。相关反馈技术按照用户是否参与,可以分为自动相关反馈(automatic relative feedback)和用户相关反馈(user relative feedback)[8-10]。前者也称为伪相关反馈方法(pseudo relative feedback),是完全自动进行的,通过假定检索结果列表的前 n 篇文献作为相关文献来进行反馈,不需用户做出相关性判断。后者也称为交互式相关反馈(interactive relative feedback),其融入用户参与因素,用户除了对检索出来的文献进行相关性判断外,还可以控制和修改查询[11-15]。

本文提出了一种基于相关反馈的信息检索模型,该模型能将用户输入的自然查询词自动转换为查询词(即关键词),再进行相关检索,从而得到所需的结果;同时,将该模型与MMM、Paice、P-norm等模型进行了性能比较。

2 布尔逻辑方法

2.1 布尔逻辑检索

目前使用的大部分文档检索系统都基于布尔逻辑检索。

到稿日期:2011-09-29 返修日期:2011-11-28 本文受国家自然科学基金项目(61070162,71071028,70931001),高等学校博士学科点专项科研基金课题(20100042110025),中央高校基本科研业务费专项资金(N090504003,N090504006)资助。
金光赫(1978—),男,博士生;王兴伟(1968—),男,教授,博士生导师。

在布尔逻辑检索中，查询语句是由表达概念的查询词和表现它们之间逻辑关系的布尔运算符号所组成的。通过检索中包含的查询词的准确组合，就可以检索到与查询词完全一致的文档集合。但是，布尔逻辑检索在索引过程及信息检索中存在一些明显的缺点：检索结果过多或没有输出结果，无法准确地控制文档数量；不能表现索引词和查询词的相对重要性；只能找出标记与用户输入的查询词完全一致的文档；无法准确地计算出文档与用户需求之间的匹配度，因此难以对结果进行合理的排序；检索最重要的问题是难以清晰地表达出用户的需求，难以进行反馈更新。

2.2 正规化

为了更准确地计算出文档值，扩展的布尔检索模型使用索引词的加权值。

索引词的加权值由逆文档频度和索引词的出现频度得到。在文档 i 中，索引词 k 的加权值 W_{ik} 是索引词 k 的逆文档频度 IDF_k 与它在文档 i 中的出现频度 TF_{ik} 的乘积：

$$W_{ik} = IDF_k * TF_{ik} \tag{1}$$

式中，$IDF_k = \log_2(N/n_k)$，N 是文档集中文档的数量，n_k 表示有索引词 k 出现的文档的数量。

为了使索引词的加权值处于[0,1]，做正规化处理：

$$W_{ik} = \frac{TF_{ik}}{\max TFindocument_i} \cdot \frac{IDF_k}{\max IDFindocument_i} \tag{2}$$

3 相关反馈

使用文档的题目和内容作为系统的输入，利用朝鲜语词类标记，将索引词抽取出来构成向量。

目前，为向量做加权值赋值的方法主要有二进制向量(TF_{bin})和以文档内单词的出现频度构造向量(TF)的方法。

为了表示更精确，使用正规化构造法(TF_{norm})。

现在，以为索引词 W_i 在文档 S_j 中组成的向量 $S_j = (W_{1j}, W_{2j}, W_{3j}, \cdots, W_{nj})$ 做加权值赋值 $s_j = (w_{1j}, w_{2j}, \cdots, w_{nj})$ 为例：

$$TF_{bin}法：w_{i,j} = \begin{cases} 1, & \text{if word } i \text{ is in sentence } j \\ 0, & \text{otherwise} \end{cases} \tag{3}$$

$$TF法：w_{i,j} = freq_{ij} \tag{4}$$

式中，$freq_{ij}$ 表示索引词 i 在文档 j 中的出现频度。

$$TF_{norm}法：w_{ij} = \frac{freq_{ij}}{\max freq_j} \tag{5}$$

式中，$\max freq_j$ 表示在句子 j 中出现的索引词的最高频度。

3.1 相关反馈

相关反馈在信息检索系统中的机制见图 1。用户初始提交的需求通常比较模糊，系统自动对返回的结果进行相关排序；然后由用户做进一步相关性判断(事实上，用户通常只查看返回结果列表中的前 10～20 篇文档)，指出真正相关的结果文档；系统重新构建查询，并对结果进行相关排序，再由用户做相关性判断，该过程循环进行，直至得到满意的结果。

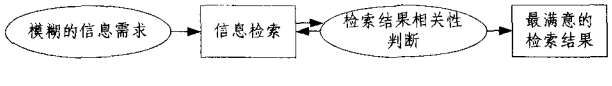

图 1 信息检索系统中的相关反馈机制

3.2 一般相关反馈

在信息检索系统中，一个重要的环节是为文档构造出高效的索引词。在具体检索过程中，如果用户认为得到的结果比较匹配，但不是最终结果，那么系统会自动开始扩展查询。

为使用户得到准确的结果，一般相关反馈是增加相关文档中的索引词的权值，减少不相关文档中的索引词的权值，然后重新生成新的查询词。这样可以增加相关文档被检索的概率，从而修正查询词，如图 2 所示。

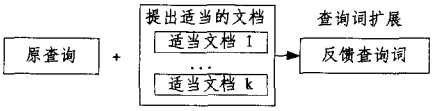

图 2 一般相关反馈

4 查询词分解相关关系

4.1 查询词分解相关反馈

利用适当文档使用相关正反馈(Relative Positive Feedback)扩展初级查询词。把 k 个文档合并起来用于初始查询词的扩展，所以扩展之后查询词仍只有一个，这样返回的结果中容易出现不适当的文档；而且，即使是使用适当文档扩展查询词，也容易使查询词范围过宽。

为了解决这个问题，对 k 个适当文档进行单独的查询词扩展，产生 k 个对应的反馈查询词，从而提高信息检索的效率，如图 3 所示。

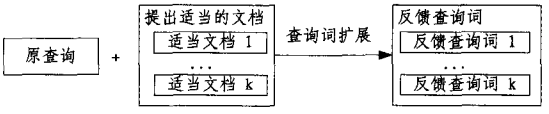

图 3 查询词分解相关反馈

通过相关反馈，查询词扩展的过程可以表示如下：

$$Q^{new} = \alpha Q^{old} + \beta \sum_{D_i \in R} D_i - \gamma \sum_{D_i \in \bar{R}} (D_i) \tag{6}$$

式中，Q^{new} 是表示扩展后的反馈查询词向量，Q^{old} 表示扩展前的查询词向量，D_i 是代表文档向量，R 是相关文档的集合，$\bar{R}$ 是不相关文档的集合。α、β 以及 γ 是用于调整的 3 个常数，β 部分用于正反馈，γ 部分用于负反馈，α、β、γ 的取值比率决定了在调整检索词的权重时，原先的查询、相关文档、不相关文档之间的相对重要性。几种常用取值方法：$\alpha=\beta=\gamma=1$ 或 $\alpha=\beta=1, \gamma=0$。当 $\gamma=0$ 时，即不使用不相关文档集合的信息，只用相关文档集合的信息，也叫做正反馈。正反馈经常用于伪相关反馈。

4.2 适当文档的提取

初始查询词和各文档之间的相似度计算使用余弦相似度计算公式：

$$\text{sim}(S_j, Q^0) = \frac{S_j \cdot Q^0}{|S_j| \times |Q^0|} = \frac{\sum_{i=1}^{t} w_{ij} \times w_{iq}}{\sqrt{\sum_{i=1}^{t} w_{ij}^2} \times \sqrt{\sum_{i=1}^{t} w_{iq}^2}} \tag{7}$$

式中，S_j 是表示文档 j 中的索引词加权值组成的向量，Q^0 表示初始查询词的索引词加权值组成的向量，$|S_j|$ 表示向量 S_j 的模，w_{ij} 和 W_{iq} 是索引词 i 在各个文档和初始查询中对应的加权值，t 是构造各文档向量和初始查询向量的索引词的总个数。式(6)中，相似度值最高的 k 个文档定性为适当文档。

5 信息检索模型

图 4 是本文提出的自然语查询句的信息检索模型。

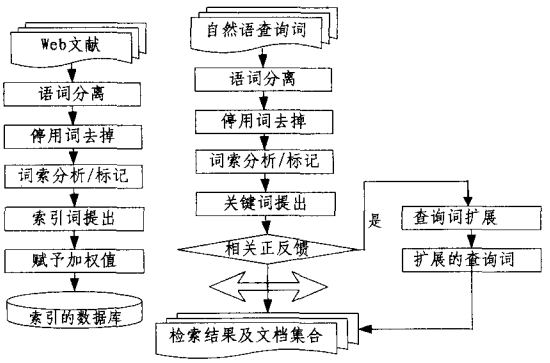

图 4　自然语查询句的信息检索模型

在提出查询词阶段，用户通过词素分解和词类区分法(Part-of-Speech Tagging)提取出名词、形容词和动词等作为自然语句中关键的查询词。自然语形态的查询词提取方法与前阶段的索引词提取的方法一样。过程如下：

1)首先将空格、特殊文字等区分出来，然后提取语词；

2)在语词内通过区分英文字母、数字、特殊文字等来索引，并按照提取的英文字处理 Stemming；

3)将特殊文字和数字等难以选出的语词利用停用词词典来去除；

4)取出的语词用词类区分法对分解名词、收取名词等实际词素进行提取；

5)在 3)中取出的实际词素是名词或动词索引；

6)在 3)中取出的实际词素有 3 个音节，并且在不发生名词分解时对前两个音节进行索引；

7)抽取的语词用名词提取机抽选名词，当提出的名词在 3)中取出的实际词素里不存在长度为 6 字(12bite)以上的词时就附加为索引词；

8)如果事先提出的查询词在数据库里没有，那么添加新的索引词；

9)在被提取的查询词上给予加权值方法来生成查询词目录，查询词不仅可以提取单词，也可以提取复合词；

10)在被提取的查询词上进行第一次检索，根据适当文档的顺序排列文件；

11)对于 K 个适当文档生成 K 个相关反馈查询词来扩展查询词，根据被扩展的查询词进行第二次检索。

6　实验及结果

本文使用朝鲜中央科技通报社的部分资料作为参考资料，开展性能评价。这些资料由 500 个文档与 21 个索引词构成。

评价信息检索系统的效果是召回率(recall rate)与准确率(precision rate)。召回率是检索出的适当文档数和文档库中所有的适当文档数的比率，准确率是检索出的适当文档数与检索出的文档总数的比率。

准确率(P)、召回率(R)的测试值公式如下：

$$P=A/(A+C) \tag{8}$$

$$R=A/(A+B) \tag{9}$$

式中，A 是系统检索出来的适当文档的数量，B 是系统检索到的不适当文档的数量，C 是适当但系统没有检索到的文档的数量。

表 1 列出了 MMM、Paice、P-norm 和 RF 的检索准确率。

表 1　检索结果的比较

模型	检索准确率
MMM	0.327
Paice	0.318
P-norm	0.362
RF	0.602

在本文中为了评价检索效果，对查询词进行了平均检索准确率测试。把对每个查询词的检索准确率固定成召回率为 0.25、0.5、0.75，再计算检索准确率的平均值。

在表中得到的检索效果是各自检索效果的最佳值。RF、P-norm 模型能够提供比 MMM、Paice 模型更为准确的检索效果。MMM 模型又比 Paice 模型呈现出更好的检索效率。

从表 2 可以看出，索引词分解的相关反馈(Index Term Splitting Relative Feedback，ITS-RF)的检索结果比 RF 的结果在准确率上提高了 5.48%。

表 2　检索准确率的比较

模型	检索准确率
RF	0.602
ITS-RF	0.635

当限定文档数量为 10 个或 20 个时，各种信息检索模型的性能比较如表 3 所列。文档数量限制为 10 个时，RF 的召回率比 P-norm 的召回率提高了 14%，在检索准确率上提高了 21%；而 ITS-RF 的召回率又比 RF 的召回率提高了 3%，检索准确率也提高了 3%。

表 3　P-norm、RF、ITS-RF 的准确率与召回率的比较

检索文档数	P-norm		RF		ITS-RF	
	准确率	召回率	准确率	召回率	准确率	召回率
文档≤10	0.43	0.24	0.64	0.38	0.67	0.41
文档≤20	0.397	0.48	0.578	0.68	0.625	0.70

文档数量限制为 20 个时，RF 的召回率比 P-norm 的召回率提高了 20%，在检索准确率上提高了 18.1%；ITS-RF 比 RF 在召回率和检索准确率上分别提高了 2%和 4.7%。

图 5 显示出，本文研究的索引词分解的相关反馈法的准确率和召回率比其他的方法更高。

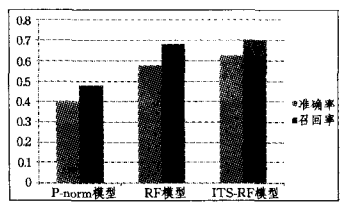

图 5　准确率和召回率的比较

图 6 表示查询文件数为 20 时，随着召回率的变化，各方法准确率的变化情况。如图 6 所示，召回率的提高使得更多的适当文档被检索出来，不适当的文档也随着被检索出来，准确率反而下降。

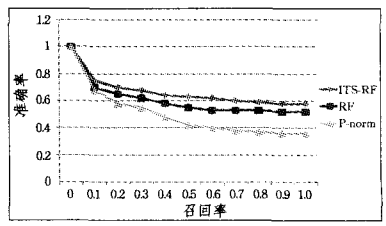

图 6　P-norm、RF、ITS-RF 的实验结果

结束语 本文提出了一种基于相关反馈的信息检索模型，分析了查询词分解，推导了相关反馈机制和正规化过程，并进一步阐述了文档提取方法。

提出的模型通过相关反馈和查询词扩展，能克服传统方法无法计算文档与查询词之间的相似度的缺点。通过对比分析和性能评价结果表明，本文所提方法是切实有效的。

参 考 文 献

[1] Zhang S D, Chen Y. Research on domain ontology-based intelligent information retrieval system [J]. Key Engineering Materials, 2011, 460-461: 300-304

[2] Hong W-S, Chen S-J, Wang Li-hui, et al. A new approach for fuzzy information retrieval based on weighted power-mean averaging operators[J]. Computers and Mathema-tics with Applications, 2007, 53(12): 1800-1819

[3] Chen S-J, Chen S-M. Fuzzy information retrieval based on geometric-mean averaging operators[J]. Computers and Mathematics with Applications, 2005, 49(7/8): 1213-1231

[4] Tombros A, Sanderson M. Advantages of Query Biased Summaries in Information Retrieval [C] // Proceeding of ACM-SIGIR98. 1998: 2-10

[5] Lee J H, Kim M H, Lee Y J. Information Retrieval Based on Conceptual Distance in Is-a Hierarchies [J]. Journal of Documentation, 1993, 49(2): 188-207

[6] 迟呈英，战学刚，姚天顺. 基于 p 范式模型的检索[J]. 中文信息学报，2000，14(4)：35-41

[7] He D Q. A study of self-organizing map in interactive relevance feedback [C] // Proceedings of the 3rd International Conference on Information Technology: New Generations. Las Vegas: IEEE, 2006: 394-401

[8] Paredes R, Deselaers T, Vidal E. A probabilistic model for user relevance feedback on image retrieval [C] // Proceedings of 5th International Workshop on Machine Learning for Multimodal Interaction. Utrecht: Springer, 2008: 260-271

[9] Xu Y, Jones G, Wang B. Query dependent pseudo-relevance feedback based on wikipedia [C] // Proceedings of the 32nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. Boston: ACM, 2009: 59-66

[10] Borji A, Jahromi M Z. Evolving weighting functions for query expansion based on relevance feedback [C] // Proceedings of the 10th Asia-Pacific Web Conference on Progress in WWW Research and Development. Shenyang: Springer, 2008: 233-238

[11] Vitsentiy V. A user interface of relevance feedback for interactive information retrieval systems [C] // IEEE Workshop on Intelligent Data Acquisition and Advanced Computing Systems: Technology and Applications. Dortmund: IEEE, 2007: 449-453

[12] Chandramouli K, Kliegr T, Nemrava J, et al. Query refinement and user relevance feedback for contextualized image retrieval [C] // Proceedings of 5th International Conference on Visual Information Engineering. Xi'an: IEEE, 2008: 453-458

[13] Lv Y H, Zhai C X. Positional relevance model for pseudo-relevance feedback [C] // Proceedings of the 33rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. Geneva: ACM, 2010: 579-586

[14] 黄名选，严小卫，张师超. 基于矩阵加权关联规则挖掘的伪相关反馈查询扩展[J]. 软件学报，2009，20(7)：1854-1865

[15] Pu Q, He D Q. Pseudo relevance feedback using semantic clustering in relevance language model [C] // Proceedings of the 18th ACM Conference on Information and Knowledge Management. HongKong: ACM, 2009: 1931-1934

(上接第 126 页)

深入研究。从应用实际出发对模型变化描述进行全面分析和归纳，全面研究模型演化所应遵循的约束，从而提高模型变化的语义表述能力、特性保持和一致性判定能力，结合模型驱动开发工具，使其能更加准确、有效地反映并支持工程化的模型驱动开发。

参 考 文 献

[1] 钟林辉. 构件化软件开发中演化信息的获取和应用技术研究[D]. 北京：北京大学，2007

[2] 刘辉，麻志毅，邵维忠. 模型转换中特性保持的描述与验证[J]. 软件学报，2007，18(10)：2369-2379

[3] Yuan W, Ying S. Modeling requirements evolution with π-calculus[C] // 2nd Conference on Power Electronics and Intelligent Transportation System (PEITS). Shenzhen, China, IEEE Computer Society, 2009: 355-358

[4] Lormans M. Monitoring requirements evolution using views[C] // 11th European Conference on Software Maintenance and Reengineering (CSMR). Amsterdam, Netherlands: IEEE Computer Society, 2007: 349-352

[5] 刘辉，麻志毅，邵维忠，等. 元建模技术研究进展[J]. 软件学报，2008，19(6)：11

[6] Sprinkle J, Karsai G G. A domain-specific visual language for domain model evolution[J]. Journal of Visual Languages and Computing, 2004, 15(3/4): 291-307

[7] Alanen M, Porres I. Difference and Union of Models[C] // Lecture Notes in Computer Science. Springer, 2003

[8] Altmanninger K. Models in Conflict-Towards a Semantically Enhanced Version Control System for Models [C] // Lecture Notes in Computer Science. 2008: 293-304

[9] Chen K, Sztipanovits J, Abdelwalhed S, et al. Semantic anchoring with model transformations [C] // Proceedings of the 3rd European conference on Model driven architecture-foundations and applications (ECMDA-FA05). Springer, 2005, 3748: 115-129

[10] Scheidgen M, Fischer J. Human comprehensible and machine processable specifications of operational semantics [C] // Proceedings of the 3rd European conference on Model driven architecture-foundations and applications (ECMDA-FA07). Springer, 2007, 4530: 157-171

[11] 李雷，吴从. 集值分析[M]. 北京：科学出版社，2003

[12] Herrmann C, Krahn H, Rumpe B, et al. An algebraic view on the semantics of model composition[C] // Lecture Notes in Computer Science. Springer, 2007, 4530: 99-113