基于最大边缘相关的伪相关反馈方法

闫 蓉 高光来

(内蒙古大学计算机学院 呼和浩特 010021)

摘 要 反馈文档的质量是制约伪相关反馈方法性能的主要因素。为了提高反馈文档的鲁棒性,提出一种基于最大边缘相关的伪相关反馈方法 RMMR(Reorder Maximal Marginal Relevance)。该方法通过对查询初检结果进行重调序,使得排序后的前 k 个文档中,文档间的相似度最小且与查询相关的数目最大。最后,利用查询纯度将影响性能的候选扩展词剔除后进行二次查询。实验结果表明,该方法可以有效地提高反馈文档的鲁棒性。

关键词 查询扩展,伪相关反馈,最大边缘相关,查询清晰度

中图法分类号 TP391.3 **文献标识码** A **DOI** 10.11896/j.issn.1002-137X.2015.6.057

Pseudo Relevance Feedback Based on Maximal Marginal Relevance

YAN Rong GAO Guang-lai

(College of Computer Science, Inner Mongolia University, Hohhot 010021, China)

Abstract The performance of PRF(Pseudo-relevance feedback) is heavily dependent upon the quality of 'pseudo-relevant' documents. In order to improve PRF robustness, this paper proposed a novel approach named RMMR(Reorder Maximal Marginal Relevance). Its aim is to make the minimum similarity between the two documents and the maximum number of relevant with the query for the top-k ranked documents by means of reordering the first-pass retrieval result. At last, query clarity was used to filter the set of expanded queries for the second-pass. Evaluation of this proposal shows important improvements in terms of PRF robustness.

Keywords Query expansion(QE), Pseudo relevance feedback(PRF), Maximal marginal relevance(MMR), Query clarity

1 引言

文本检索的主要目标就是在海量信息中帮助用户找到与其信息需求(Information Need)相关的文档。然而,对于目前主流的基于关键词的搜索方式,用户必须通过构造有限的查询词来表达信息需求。通常,用户本身的信息需求就是模糊的或是不确切的,或是具有多面性的。另外,查询本身的表述方式又具有多样性。这些都使得查询返回结果中有很多都是与用户查询需求不相关的,用户需要花费时间阅读大量相关文档才能获得自己确实所需的信息。这就要求搜索引擎提供的查询结果不仅要保证与查询内容的相关性和有用性,还要让用户收获意料之外的发现。如何使得检索更有效,成为信息检索领域研究的重点和难点。

查询扩展通过对用户原始查询添加扩展词进行查询重构,弥补用户初始查询信息不足的缺陷,可以有效解决用户表达问题。伪相关反馈作为一种有效的自动的查询扩展方法[1-3],其基本思想就是假设初检结果排名靠前的 k 个文档都与初始查询相关,并通过从中抽取扩展词,自动重构新的查询来完成文本检索。这种查询扩展方式使得在初检结果中没有检索到的相关信息也可以被检索到,从而提高查询性能。但其关注的重点是用户查询词的形式体现,而不是用户的实际信息需求。在这种假设下,得到的相关文档中有很多是非常相似的,冗余信息多,质量不高,不能很好地体现用户不同层面的查询需求[4]。特别是当查询词出现歧义时,往往反馈的文档集会偏向于某一个主题,而该主题往往并不是用户潜在的查询意图[5]。用户必须阅读大量相关文档才能获得自己确实所需的信息,用户体验得不到满足。进一步,文献[6,7]表明用户的初始查询中至少16%都是带有歧义的;有超过75%的查询具有更复杂的信息需求。并且,伪相关反馈的性能主要受制于反馈文档的数量和质量[4,8,9]。那么,在这种模糊的信息需求前提下,如何考虑用户需求的多面性来描述查询词的多角度含义,提高反馈文档的质量,选取高质量的更能反映用户需求的扩展词,成为提高查询性能的关键。文献[10]的研究表明,提高用户体验的较好的办法就是给用户提供尽可能多且不同的信息,而这些信息中至少会有一个是与用户需求相关的。

综上所述,为了提高反馈文档质量,提出一种新的伪相关反馈方法 RMMR。具体描述为:首先利用最大边缘相关算法,将初检查询结果中的冗余信息去掉,同时提高靠后的其它不相似且相关文档的排名,让更多的反映用户需求不同层面的内容都尽量在重排后的结果中出现;接着,利用查询的真相关 Judgments,进一步调整最大边缘的相关排序结果,进行二

到稿日期:2014-06-23 返修日期:2014-09-19 本文受国家自然科学基金(61263037),内蒙古自然科学基金重大项目(2011ZD11)资助。

闫 蓉(1979-),女,博士生,讲师,主要研究方向为自然语言处理、信息检索,E-mail:csyanr@imu.edu.cn;高光来(1964-),男,硕士,教授,主要研究方向为智能信息处理、数据挖掘等。

次重排序，让与查询相关的文档排名尽量靠前，减少查询主题偏移；最后在重排后的查询结果中，寻找更有价值的反馈信息，以能够很好地契合用户查询需求的信息作为最终扩展词，再进行伪反馈。

2 最大边缘相关 MMR

2.1 最大边缘相关算法

最大边缘相关算法是由 Carbonell[11] 等提出的，是一种实现文本摘要的方法，追求的目标是查询结果文档的不相似。其目的主要是解决查询结果的多冗余性，同时增加查询结果的新颖性，提高查询结果质量，增加用户体验。其主要思想是对结果文档进行重排序，依据的是文档的相关性和新颖性来对文档结果进行排序，将与查询相关并且和先前被选择的文档间保持较小的相似性的文档的排名提前，以达到重排序的目的。假设用户潜在查询意图是多类别的信息需求，查询结果一方面要保证满足各种类别的信息需求，即覆盖面的最大化，另一方面还要使得文档冗余度最小，即文档间的相似度最小化。

设 C 表示一个文档集合，$d_i, d_j \in C$，Q 表示一个用户查询，$Rank0$ 表示已得到的一个以相关度为基础的初检文档集合，S 表示 $Rank0$ 中已经被选取的文档集合，$d_j \in S$，$Y = Rank0 - S$ 表示未被选取的文档集合，$d_i \in Y$。$\arg\max^k[*]$ 表示给出集合 k 个最大元素的索引。

文档的最大边缘相关定义如下：

$$MMR(Q,C,Rank0) = \arg\max^k_{d_i \in Y}(\lambda Sim_1 d_i, Q) - (1-\lambda)\max_{d_j \in S} Sim_2(d_i, d_j)) \quad (1)$$

MMR 算法描述如下：

输入：以相关度为基础的初检结果文档集，记为 Rank0。

输出：重排序的文档集，记为 Rank1。

Step1 从已得到的一个以相关度为基础的初检文档集中取前 K 个文档，记为 $D_r = IR(C,Q,K)$；

Step2 选取 $\max Sim_1(d_i \in D_r, Q)$ 作为第一个，即 $Rank1 = \{d_i\}$；

Step3 从 D_r 中首先去掉文档 d_i，记为：$D_r = D_r \backslash \{d_i\}$；

Step4 若 D_r 不为空，do：

(1)找到使得 $\max MMR(Q, D_r, Rank1)$ 的文档 d_i；

(2)给 Rank1 中添加文档 d_i，记为：$Rank1 = Rank1 \cdot \{d_i\}$；

(3)$D_r = D_r \backslash \{d_i\}$。

其中，$Sim_1(d_i, Q)$ 计算的是文档 d_i 与查询 Q 之间的相似度，$Sim_2(d_i, d_j)$ 计算文档 d_i 和 d_j 间的相似度，$\lambda \in [0,1]$ 是调节系数。

2.2 相似度计算

MMR 算法的核心是两个相似度算法的选取。关于文档与查询间的相似度 Sim_1 的计算，本文直接利用经典的 BM25 检索方法的结果。对于文档间的相似度 Sim_2 的计算，本文采用了基于 TF-IDF 的向量空间模型方法。基于 TF-IDF 的向量空间模型文档相似度的实现思路是：用词在文本中出现的频率及在文档集中出现该词的频率来表征词的权重，通过计算向量之间的 cosine 系数或 Jaccard 系数来确定文档间相似度。

其中，每一个词项 w_i 的 TF-IDF 值采用式(2)计算：

$$TF\text{-}IDF(w_i) = tf(w_i) \times idf(w_i) = tf_j(w_i) \times \log(N/df(w_i)) \quad (2)$$

式中，$tf_j(w_i)$ 表示词项 w_i 在文档 j 中出现的频率，N 表示文档集中所有文档的总数，$df(w_i)$ 表示文档集中出现词项的文档数。

文档间相似度的计算公式如下：

$$Sim_{\text{cosine}}(T,T') = \frac{\sum_{i=1}^{n} T_i * T_i'}{\sqrt{\sum_{i=1}^{n} T_i^2 * \sum_{i=1}^{n} T_i'^2}} \quad (3)$$

$$Sim_{\text{Jaccard}}(T,T') = \frac{\sum_{i=1}^{n} T_i * T_i'}{\sum_{i=1}^{n} T_i^2 + \sum_{i=1}^{n} T_i'^2 - \sum_{i=1}^{n} T_i * T_i'} \quad (4)$$

其中，$T = \{w_1, w_2, \cdots, w_n\}$ 和 $T' = \{w'_1, w_2', \cdots, w_n'\}$ 表示两个不同的文档向量。

3 基于最大边缘相关的伪相关反馈方法的设计思路及实现方案

3.1 设计思路

伪相关反馈假设初检结果集中排名前 k 的文档与原查询是相关的。对初检结果文档集的处理将直接影响伪相关反馈的性能。若初检结果文档中包含过多的冗余文档，或是真正与用户实际需求相关的文档排名靠后，则对检索性能影响较大。最极端的情况是，与用户需求相关的文档不在前 k 个文档之列，这时二次查询后得到的结果文档集中并没有用户想要的信息，用户将不得不重构初始查询，用户体验得不到满足。本文认为，用户的查询需求并不是单一的，而是多层面的，有时甚至是不确定的，要实现自动的查询扩展，就要求初检文档集中的内容既要保证与用户原查询相关，又要保证其与用户多层面需求的一一映射关系，获取最有用的信息来进行伪相关反馈。

MMR 调序后的 $Rank1$ 中的文档并不都与查询 Q 是相关的。通常，$Rank1$ 中排名靠前的 k 个文档与查询 Q 较相关，并且相关文档数越多，伪反馈效果就越好。那么，要保证查询结果与用户多层面需求的一一映射关系，就必须保证出现在 $Rank1$ 中的各文档最大的不相似性，还要保证其中的文档与原查询相关的数目最大，这就要求必须对 $Rank1$ 中的文档进行重调序。通过调整各文档在 $Rank1$ 中的排名，提高相关文档的排名，同时使得不相关文档尽量靠后，保证相关文档尽量出现在初检结果排名前 k 个文档中。图 1 所示为基于最大边缘相关的伪相关反馈方法实现流程。

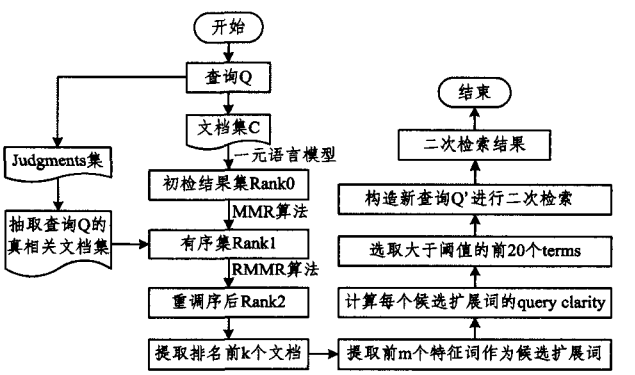

图 1 基于最大边缘相关的伪相关反馈方法

3.2 RMMR 算法实现

输入：MMR 调序后结果 $Rank1$ 集合。

输出：RMMR 重调序后结果 $Rank2$ 集合。

RMMR 算法描述如图 2 所示。

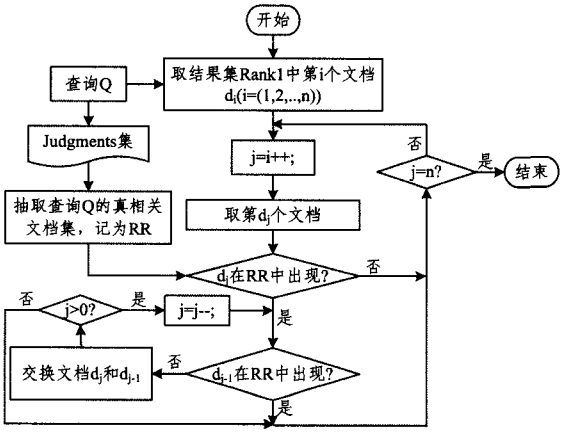

图 2　RMMR 算法流程

3.3　候选扩展词的筛选

由 Cronen-Townsend[12] 等人提出的查询纯度(Query clarity)是查询性能预测的一种方法。查询纯度实际上是查询语言模型和整个文档集语言模型的相对熵或 KL(Kullback-Leibler)距离，它表征的是查询在多大程度上精确地刻画了用户的信息需求，并将相关文档与不相关文档区分开来的能力，即查询的特异性(specificity)。不同于 Shah[13] 等的研究，本文不是用 Query clarity 来调整扩展查询词的权重，而是要利用它过滤掉 clarity 值较低的候选扩展词，保证伪反馈的第二次检索过程的查询结果与各查询是相关的，从而提高查询性能。

4　实验结果及分析

4.1　实验数据集及相关参数设置

实验数据集采用的是简体中文 xinhua2002－2005 年共 308845 个文档。查询集是简体中文 ACLIA2-CS-0001－ACLIA2-CS-0100，共 100 个查询。利用 Indri(http://www.lemurproject.org)工具建立索引和进行查询。初检的相关度排序方法选用的是典型的一元语言模型 LM(Language Model)方法。实验中统一采用 Dirichlet 平滑方法，并在实验过程中统一设置平滑参数为固定值 1000。关于 MMR 算法中参数 λ 的选取，实验中采用贪心策略，当 λ 值取 0.7 时效果最好。文档间相似度采用 cosine 系数计算。对于词项的词频统计信息，主要利用 SOGOU 提供的 2006 年 10 月，涉及互联网语料规模 1 亿页面，约为 15 万条高频词的词频信息。Baseline 选取标准的 BM25 伪反馈方法。实验中初检查询文档个数设定为 $K=50$，并设置统一从排名前 $k=10$ 个文档中，筛取前 20 个 *terms* 进行伪反馈。

4.2　预处理

由于本文处理对象为中文数据，因此在进行检索和查询前，首先对数据集和查询集均进行了预处理，主要包括：分词(采用的是中科院计算所的 ICTCLAS)和去除停用词(采用的是哈工大的中文停用词表 Stoplist)。另外，在进行文档相似度计算时，为了提高算法执行效率，对数据集中各文档又进行了去除虚词处理。

4.3　实验结果及分析

本文工作的核心是在伪反馈之前，尽可能地保证在排名前 k 个文档中与查询相关的文档数最大，并且文档间的相似度最小，即提高伪反馈文档的质量，从而整体上提高查询性能。实验结果验证了提出方法的科学性。表 1 和图 3 分别示出利用 Lemur(http://www.lemurproject.org)工具得到的 RMMR 方法与 Baseline 方法的比较结果和 Precision-Recall(精度-召回)对比分析曲线图。

表 1　RMMR 与 Baseline 方法实验结果比较

评价方法	Baseline	RMMR
MAP	0.2690	0.2912(+8.3%)
nDCG	0.3997	0.4224(+5.8%)
R-prep	0.3022	0.3278(+8.5%)
bpref	0.2829	0.3034(+7.2%)
P@5	0.5239	0.5389(+2.9%)
P@10	0.5056	0.4917(−2.7%)

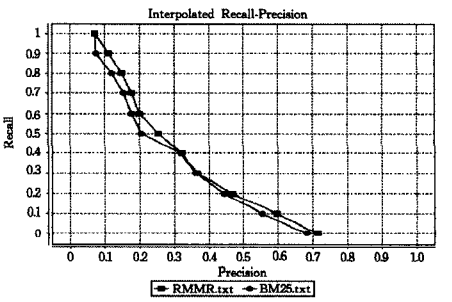

图 3　BM25 和 RMMR 的 Precision-Recall 曲线图

从表 1 和图 3 分析来看，RMMR 方法比 Baseline 方法在检索效果上有了比较明显的提高，在 MAP、nDCG、R-prep、bpref 和 P@5 各指标上都分别有不同程度的提高，说明这种重调序方法对于提高检索效果是有效的。但其在 P@10 这个指标上却是有所下降的。分析其原因在于，RMMR 方法对初检结果进行二次重调序，其目标是使得重排序后的结果文档间的相似度最小，使与用户初始查询语义相关的所有文档尽量出现在排名靠前的 k 个文档中。若某个查询 Q 本身所具有的语义特性较多，那么就会出现排名靠前的文档中与真实 Judgments 内容不同的数目会增加，使得 P@n 值下降。如文献[14]所述，对于信息检索系统的查询结果，实际中往往要做的是查询结果的相关性和查询结果类别多样性的折中。

进一步，本文针对各查询分别在 Baseline 和 *Rank*2 中排名前 10 个文档中出现的相关文档数目进行了比较，结果如图 4 所示。

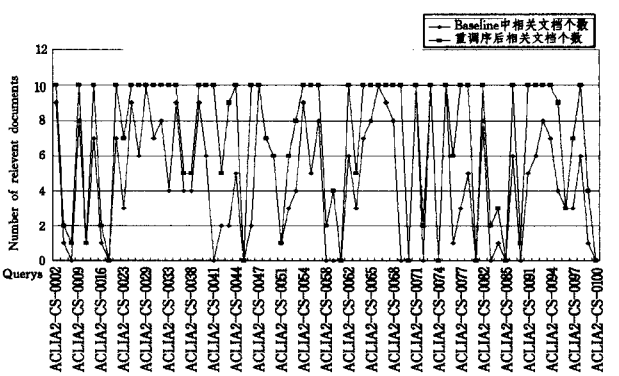

图 4　Top-10 中各查询在 Baseline 和 *Rank*2 中出现相关文档的数目比较

从图 4 中可以看出，进行重调序后的 *Rank*2，在排名前 10 个文档中出现的相关文档数明显更高。在前 10 个文档中，*Rank*2 和 Baseline 中平均出现相关文档个数分别约为 6.89

(下转第 287 页)

feature discretization and selection [J]. Pattern Recognition, 2012,45(9):3048-3060

[10] Lin T,Granular Y. Computing on binary relations I:Data mining and neighborhood systems[C]// Skoworn A,Polkowshi L,eds. Proc. of the Rough Sets in Knowledge Discovery. Physica-Verlag,1998:107-121

[11] Lin Tsau-young. Encyclopedia on complexity of systems science [M]. Berlin:Springer,2009:4339-4355

[12] Yang Xi-bei, Li Xin-zhe, Lin Tsau-young. First GrC model neighborhood systems: the most general rough set models [C]// 2009 IEEE International Conference on Granular Computing. Beijing,China:IEEE,2009:691-695

[13] Yao Y Y. Relational interpretation of neighborhood operators and rough set approximation operators[J]. Information Sciences, 1998,111(198):239-259

[14] Wu W Z,Zhang W X. Neighborhood operator systems and approximations[J]. Information Sciences,2002,144(1-4):201-217

[15] 胡清华,于达仁,谢宗霞. 基于邻域粒化和粗糙集逼近的数值属性约简[J]. 软件学报,2008,19(3):640-649
Hu Qing-hua,Yu Da-ren,Xie Zong-xia. Numerical Attribute Reduction Based on Neighborhood Granulation and Rough Approximation [J]. Journal of Software,2008,19(3):640-649

[16] Hu Q H, Yu D R, Liu J F, et al. Neighborhood rough set based heterogeneous feature subset selection[J]. Information Science, 2008,178(18):3577-3594

[17] 王丽娟,吴陈,杨习贝,等. 邻域系统粗糙集和覆盖粗糙集[J]. 计算机科学,2013,40(1):221-224
Wang Li-juan,Wu Chen,Yang Xi-bei,et al. Neighborhood System Based Rough Set and Covering Based Rough Set[J]. Computer Science,2013,40(1):221-224

[18] 张颖淳,苏伯洪,曹娟. 基于粗糙集的属性约简在数据挖掘中的应用研究[J]. 计算机科学,2013,40(8):223-226
Zhang Ying-chun,Su Bo-hong,Cao Juan. Study on Application of Attributive Reduction Based on Rough Sets in Data Mining [J]. Compute Science,2013,40(8):223-226

[19] 李智玲,胡彧. 改进的属性约简算法在数据挖掘中的应用研究[J]. 计算机技术与发展,2012,22(10):47-50
Li Zhi-ling,Hu Yu. Application Research of Improved Attribute Reduction Algorithm in Data Mining[J]. Computer Technology and Development,2012,22(10):47-50

(上接第 278 页)

和 4.45,*Rank2* 比 Baseline 平均出现相关文档个数高出近 55%,明显提高了反馈文档的质量。

结束语 为了尽可能地将满足用户信息需求的文档查询出来,减少查询语义缺失,提出了一种基于最大边缘相关的伪相关反馈方法。通过对初检结果进行两次重定序,既保证了与查询相关文档的优先返回,又可以有效避免伪相关反馈中的查询主题偏移问题。本文的相似度算法仅选取了基于向量空间的 cosine 系数来计算文档间相似度,进一步的工作是选取更合适的相似度计算方法,提高算法运行速度。

参 考 文 献

[1] Collins-Thompson K. Reducing the risk of query expansion via robust constrained optimization[C]// Proceedings of the 18th ACM conference on Information and Knowledge Management. ACM,2009:837-846

[2] Raman K,Udupa R,Bhattacharya P,et al. On improving pseudo-relevance feedback using pseudo-irrelevant documents[M]// Advances in Information Retrieval. Springer Berlin Heidelberg, 2010:573-576

[3] Zhai C, Lafferty J. Model-based feedback in the language modeling approach to information retrieval[C]// Proceedings of the 10th international conference on Information and Knowledge Management. ACM,2001:403-410

[4] He B, Ounis I. Studying query expansion effectiveness[M]// Advances in Information Retrieval. Springer Berlin Heidelberg, 2009:611-619

[5] Vargas S,Santos R L T,Macdonald C,et al. Selecting effective expansion terms for diversity[C]// Proceedings of the 10th Conference on Open Research Areas in Information Retrieval. 2013:69-76

[6] Clarke C L A,Kolla M,Cormack G V,et al. Novelty and diversity in information retrieval evaluation[C]// Proceedings of the 31st annual international ACM SIGIR conference on Research and Development in Information Retrieval. ACM,2008:659-666

[7] Song R,Luo Z,Nie J Y,et al. Identification of ambiguous queries in web search[J]. Information Processing and Management, 2009,45:216-229

[8] Amati G,Carpineto C,Romano G. Query difficulty,robustness, and selective application of query expansion[M]// Advances in information retrieval. Springer Berlin Heidelberg,2004:127-137

[9] Parapar J, Presedo-Quindimil M A, Barreiro á. Score distributions for Pseudo Relevance Feedback[J]. Information Sciences, 2014,273:171-181

[10] Teevan J,Dumais S T,Horvitz E. Characterizing the value of personalizing search[C]// Proceedings of the 30th annual international ACM SIGIR conference on Research and Development in Information Retrieval. New York:ACM,2007:757-758

[11] Carbonell J, Goldstein J. The use of MMR, diversity-based reranking for reordering documents and producing summarics [C]// Proceedings of the 21st annual international ACM SIGIR conference on Research and Development in Information Retrieval. New York:ACM,1998:335-336

[12] Cronen-Townsend S,Zhou Y,Croft W B. Predicting query performance[C]// Proceedings of the 25th annual international ACM SIGIR conference on Research and Development in Information Retrieval. New York:ACM,2002:299-306

[13] Shah C, Croft W B. Evaluating high accuracy retrieval techniques[C]// Proceedings of the 27th annual international ACM SIGIR conference on Research and Development in Information Retrieval. New York:ACM,2004:2-9

[14] Karimzadehgan M, Zhai C. A learning approach to optimizing exploration-exploitation tradeoff in relevance feedback[J]. Information retrieval,2013,16(3):307-330