一种基于索引模型融合的面向问答的信息检索方法*

郭稷[1]　骆卫华[2,3]

1. 北京大学软件与微电子学院，北京 102600

2. 中国科学院计算技术研究所智能信息处理重点实验室，北京　100190

3. 中国科学院研究生院，北京 100049

guojpku@126.com, luoweihua@ict.ac.cn

摘　要：IR4QA (Information Retrieval for Question Answering)是日本国立情报局组织举办的第 7 届国际跨语言检索评测(NTCIR-7)提出的新任务，其目的是研究信息检索和问答技术融合的有效方法，寻找带来最好的问答系统性能的信息检索策略。因此，IR4QA 比传统信息检索任务更具挑战性。在 NTCIR-7 中，我们探索了不同索引模型和检索模型的性能，提出了一种基于不同索引模型的融合方法，并使用贪心方法调节各索引模型在组合模型中的权重。实验结果表明，对于传统的信息检索和 IR4QA 任务，我们的组合模型都具有很好的性能。

关键词：IR4QA，信息检索，模型融合

A Method for IR4QA Based on Index Model Combination

Ji Guo[1], Weihua Luo[2,3]

1. School of Software and Microelectronics, Peking University, Beijing 102600, China

2. Key Laboratory of Intelligent Information Processing, Institute of Computing Technology, Chinese Academy of Sciences, Beijing 100190, China

3. Graduate University of Chinese Academy of Sciences, Beijing 100049

guojpku@126.com, luoweihua@ict.ac.cn

Abstract: Information Retrieval for Question Answering (IR4QA) is a new task in NTCIR-7. It is intended to find out the best IR strategy that brings about the best end-to-end QA performance and which IR and QA techniques work together well. It's more challenging than traditional ad hoc IR. In NTCIR-7, we investigate the performances of different index models and retrieval models, train a combination model based on different index models and propose a greedy search algorithm to find the optimal parameter setting for each index model in the interpolation. Experimental results show that our combination model achieves a high recall and good precision on both ah doc IR and IR4QA tasks.

Keyword: IR4QA, information retrieval, model combination

1. 引言

面向问答的信息检索(Information Retrieval for Question Answering，简写为 IR4QA)是日本国立情报局（NII）组织举办的第 7 届国际跨语言检索评测(NTCIR-7)提出的新任务，其目的是研究信息检索和问答技术融合的有效方法，寻找带来最好的问答系统性能的信息检索策略。传统的信息检索评价方法鼓励系统返回更多的相关文档，所以召回率被认为是衡量系统性能的更重要因素。然而，IR4QA 更趋向于返回不同的、与答案更相关的文档，即使该文档与查询词相关性不高。例如，将问题“Who is George Bush?”提交给 QA 系统，则描述第一夫人 Laura Welch 的文档，也被认为是构成答案的好的候选文档。因此，IR4QA 任务的相关性评价标准已不同于传统的 ad hoc 检索任务，IR4QA 更具挑战性。在 NTCIR-7 中，我们探索了传统信息检索方法在 IR4QA 任务中的性能，提出了一个基于不同索引模型的融合模型，用于检索主题的相关文

*本文工作得到了国家“八六三”高技术研究发展计划基金项目（项目编号：2007AA01Z438）的支持。本文实验所用数据集采用了日本NII发布的NTCIR数据集。

档，并使用贪心算法来调节参数，使融合模型的性能达到最佳。在不用数据集上的多组实验表明，对于 ad hoc 检索和 IR4QA 任务，我们的组合模型都能取得高召回率和较好的准确率。

2. 相关工作

中文信息检索目前主要使用两种索引策略：一种是以中文词语作为基本索引单元。另一种是N-gram。其中，Unigram（单个字）和Bigram（连续两个字）是最常使用的索引单元。在中文信息检索中，两种索引策略都取得了很好的效果。Nie使用不同的方法来切分文档和查询词[1]，通过大量的实验表明，最大正向匹配分词和中文汉字一起作为索引单元，系统达到最佳性能。[2]使用统计语言建模信息检索模型，探索了Unigram、Bigram以及它们的组合在中文信息检索中的性能。[3]的研究显示，被大多数检索系统共同认可的结果通常也与人工选择的结果相一致，因此，如果一个系统对于所有的查询都能给出与多数系统相一致的结果，则这个系统就有较好的检索性能。

3. 索引模型融合

3.1 动机

目前主要的信息检索模型的性能通常都不稳定，即某个模型对于某些查询性能较好，而对于另一组查询则很差。对于同一查询，不同的检索模型往往会给出不同的有序文档列表。通过选择适当的模型，其结果可以提供很好的互补性，从而给出比单系统更好的检索结果。因此，在统一的框架内把这些模型的结果融合起来，应该是改进检索系统性能的一种可行的思路。模型融合需要解决两个关键问题：一是选择哪些模型，二是这些模型的结果以何种形式组合起来。

在考虑模型融合时，检索的模型可以从两类范畴中挑选：检索模型和索引模型。检索模型包括布尔模型、空间向量模型、概率模型、统计语言模型等[4,5,6,7]。索引模型通常包括Unigram、Bigram和词索引模型(Word)等。不同模型在不同检索任务和测试集上的性能有所不同。

由于IR4QA是一种间接评测，很难直接用IR或QA的评价标准直接去优化IR4QA参数，因此我们利用不同模型的性能，建立多个模型的线性组合模型，期望获得一个稳定、健壮的模型。受统计机器翻译中系统融合思想的启发[8]，我们认为，在不同的模型下文档和查询词的相关程度不同。这个相关程度可以表示为不同模型的加权和，即：

$$Score(d,q)=\sum_{j}\alpha_{j}Score_{j}(d,q) \qquad (1)$$

其中，$Score_j$(d, q)表示文档在每个模型下的得分，α_j 是各个模型在线性组合中的权重。

3.2 我们的方法

不同的检索模型从不同的角度表示了文档、查询词以及它们的相关性。所以，不同检索模型产生的相关文档列表往往存在较大差异，因此很难把这些结果集成到同一个有序文档列表中。与此相反，在基于索引模型的融合方法中，使用相同的模型检索模型，而仅仅是索引单元不同，因此产生的排序文档列表的差异性较小，从而使得结果的组合更容易实现，同时融合之后的结果与各个系统之间的一致性更好。因此我们最终采用基于索引的模型融合。

我们的组合模型是线性模型，各模型的权重依赖于具体的检索任务。对于 IR4QA，需在开发集上调节各模型的权重，使融合模型的性能达到最优。为调节参数，我们定义了目标函数：

$$\delta(\bigcup_{i=1}^{n}Model_i)=MAP(\bigcup_{i=1}^{n}Model_i) \qquad (2)$$

输入：模型集合 $M = \{M_1, M_2, \ldots M_n\}$
输出：权值集合 $\Lambda = \{\alpha_1, \alpha_2, \ldots, \alpha_n\}$

设D是在线性组合模型中确定了权重的模型集合

步骤 1：初始化 $D = \varnothing$
步骤 2：
(1) 对于任意模型组合对 $< M_i, M_j >$，其中 $M_i \in M \cap M_j \in M \cap i \neq j$
调节参数 $< \alpha_i, \alpha_j >$ 获得其最优的MAP值$MAP_{i,j}$
(2) 确定获取MAP值最高的特征对 $< M_f, M_s >$的权重 $< \alpha_f, \alpha_s >$
从M中移除M_f，M_s，并将它们添加到集合D。将α_f，α_s添加到集合 Λ
步骤 3：
while $M \neq \varnothing$
(1) *for* 每个模型$M_k \in M$
连同已确定权值的模型一起，调节模型M_k的权值，取得最高的MAP值MAP_k
(2) 确定获得最高MAP值的模型M_l及其权重α_l。从M中移除M_l，并将其添加到D中。将α_l添加到Λ中。

图 1　调节模型权重的贪心算法

这里$\bigcup_{i=1}^{n} Model_i$代表不同模型的集合，MAP 是指在检索任务中融合模型的平均准确率。参数调节的目标就是寻找使融合模型的 MAP 达到最优时的各个模型在线性加权组合模型中的权重。我们使用了一个贪心算法来调参，如图 1 所示。

4. Ad hoc 检索实验

我们在 ad hoc 检索和 IRQA 任务上进行了大量的实验，目的是为了测试组合模型在不同检索任务上的性能，探索 ad hoc 检索和 IR4QA 检索任务相关性评价的差异。我们使用 Lemur Toolkit 作为我们的基线检索系统。NTCIR-5 C-C 检索任务包含 50 个主题，文档集合包括 901,466 个文档。我们的实验包括检索模型选择和索引模型融合。

4.1 检索模型选择

Lemur 实现了多个信息检索模型。各个模型的检索性能不同。表 1 显示了部分检索模型和索引模型在 NTCIR-5 C-C 任务上的性能：tfidf（简单 tfidf），fb_tfidf（带反馈的 tfidf），okapi（简单 okapi），fb_okapi（带反馈的 okapi），kl_dir（使用 KL 距离的语言建模检索模型，Dirrichlet 平滑），mixfb_kl_dir（使用 KL 距离的语言建模检索模型，Dirrichlet 平滑，带伪相关反馈）。我们进行了三组实验：第一组使用主题的 title 部分(T run)。第二组使用主题的 description 部分(D run)。第三组使用 narrative 部分(N run)。所有模型的参数使用系统缺省参数。

表 1 的实验结果显示，KL 距离的统计语言建模检索模型，并使用 Dirrichlet 平滑和伪相关反馈，取得最优的检索结果。因此我们选择该模型作为信息检索任务的主模型。同时，我们调节了伪相关反馈的反馈文档数和反馈词数目，使得不同索引模型在模型融合前达到了最佳性能。

4.2 索引模型融合

通过调节参数，在索引模型融合前，Unigram 模型、Bigram 模型和 Word 模型分别达到了最优的性能。索引模型融合就是将三者的检索结果进行线性加权，以确定最终的检索结果。

使用带反馈的统计语言模型进行索引模型融合面临两个问题。一个是不同索引模型下，每个文档的得分不能直接比较，因此不能将直接进行线性加权求和。我们用结果中得分最高文档的得分（即有序列表中的第一个）来归一化各个文档的得分。归一化后的得分都落在 0 到 1 之间，这样不同索引模型下文档的得分就具备了可比性。另一个是索引模型融合的时机。存在三种融合策略：（1）在

表 1 不同检索模型和索引模型在 NTCIR-5 C-C 任务上的性能

IR model	Mean Average Precision								
	T run			D run			N run		
	Unigram	Bigram	Word	Unigram	Bigram	Word	Unigram	Bigram	Word
tf_idf	0.2497	0.2664	0.2515	0.2164	0.2465	0.2052	0.2716	0.3032	0.2729
fb_tf_idf	0.2724	**0.3162**	0.2950	0.2569	0.3088	**0.2608**	0.2844	0.3362	0.2837
okapi	0.2609	0.2646	0.2582	0.2065	0.2393	0.1769	0.2835	0.3172	0.2737
fb_okapi	0.2931	0.3005	**0.3277**	0.1946	**0.3130**	0.2345	0.1678	**0.3390**	0.2773
kl_dir	0.3031	0.2615	0.2706	0.2588	0.2270	0.2013	0.3154	0.3019	0.2790
mixfb_kl_dir	**0.3443**	0.2860	0.3057	**0.2933**	0.2983	0.2513	**0.3343**	0.3378	**0.2992**

每个索引模型各自完成初始检索后，在反馈前进行模型融合；（2）每个模型进行初始检索，各自完成反馈后进行融合；（3）采用两阶段融合，即在每个索引模型各自完成初始检索后，在反馈前进行模型融合；使用第一阶段的融合结果，在各索引模型完成反馈后，融合各返回结果得到最终检索结果。在我们的实验中，第二种策略的性能最佳。图 2 给出了索引组合模型在 NTCIR-5 C-C 检索任务上的性能。从图 2 中可以看出，使用 0.8U+0.2W+0.15B 融合方案，组合模型的性能达到最优，MAP 值为 0.3669。这表明，对于 ad hoc 检索，索引模型融合提高了系统的性能。

5 IR4QA 检索实验

IR4QA 检索任务在评价指标上不同于传统的 ad hoc 检索，因此，在 NTCIR-5 数据集上调节的参数将不适用于新的检索任务。我们需要根据具体的数据集、主题和相关性评价标准，调节各索引模型的参数以及它们在模型融合中的权重。表 2 简要介绍了 NTCIR-7 中 EC-CS 跨语言 IR4QA 检索任务和 CT-CT 任务的测试集数据情况。

5.1 开发集构造

为了根据IR4QA任务的特点和相关性评价标准调节模型参数，我们手工构造了开发数据集。我们从EPAN[1]网站下载了开发集的问题和参考答案。对于每个问题，参考答案所在的文档被视为相关文档。我们分别为EN-CS和CT-CT两个任务构造了开发集数据。EN-CS开发数据集包含 78 个主题，CT-CT数据集包含 71 个主题。

5.2 EN-CS 检索

本文中我们不讨论翻译模块的实现。读者可以参阅文献[9]。在我们的实现中，翻译模块和检索模块并非完全独立。在 EN-CS 检索中，翻译质量对于组合模型的参数设置有着重要的影响。我们的检索模型是 KL 距离的语言建模检索模型，并使用 Dirrichlet 平滑和伪相关反馈。开发集实验只使用主题的 question 部分。使用 Unigram 模型，MAP 值是 0.1847；使用 Bigram 模型，MAP 值是 0.2018；使用词模型的 MAP 值是 0.1955。

我们使用图 1 所示的贪心算法进行了索引模型融合。图 3 显示了索引组合模型的实验结果：

- Unigram+Bigram：当融合方案是 0.45U+0.55B 时性能达到最优，MAP 值是 0.2131。
- Unigram+Word：当融合方案是 0.5U+0.5W 时性能达到最优，MAP 值是 0.2146。
- Word+Bigram：当融合方案是 0.35W+0.65B 时性能达到最优，MAP 值是 0.2079。

由于 Unigram 和词索引融合后的效果最佳，我们确定他们在组合模型中的权重，然后调节 Bigram 索

[1] http://aclia.lti.cs.cmu.edu:8080/epan/index.jsp

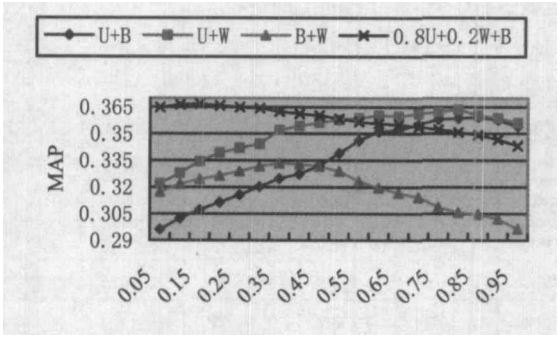

图 2 索引组合模型在 NTCIR-5 C-C 检索任务上的性能

表 2 NTCIR-7 数据集情况

任务	数据集	主题数	文档总数
EN-CS	Xinhua	97	545,162
	Lianhe Zaobao		
CT-CT	CIRB020	95	1,150,954
	CIRB040		

引模型的权重。当组合方案是 0.5U+0.5W+0.15B 时系统取得最佳效果，MAP 值达到 0.2168。

5.3 CT-CT 检索

由于缺少繁体中文分词工具，在 CT-CT 检索实验中，我们将繁体中文检索任务的文档集和主题转化成 GB 编码。CT-CT 检索实验的过程和 EN-CS 相同。使用 Unigram 索引模型，MAP 值是 0.2722；使用 Bigram 索引模型，MAP 值是 0.3077；使用词索引模型的 MAP 值是 0.2057。可以看到词索引模型的性能很差，原因在于，简体文档和繁体文档在描述同一事物时的用词存在差异。比如 Bush 在中国大陆通常被翻译成"布什"，但港台则翻译为 "布希"。这种差异导致词索引模型中存在很多切分错误，这必然会影响系统的性能。图 4 是索引组合模型在 CT-CT 检索任务上的实验结果。

受I2R工作的启发[10]，我们试图加入重排序技术，使得包含答案的文档排名更靠前。但实验结果表明，基于关键词扩展的重排序技术对于我们的模型没有帮助。探索更适合IR4QA任务的重排序技术将留给未来的工作。

6 官方测试集结果

在NTCIR-7 正式评测中，我们提交了 5 个系统参加EN-CS跨语言检索任务，提交了 4 个系统参加CT-CT检索任务。各个系统采用不同的模型[9]。表 3 显示了NTCIR-7 IR4QA子任务各系统的性能，我们的系统在两个子任务中取得了很好的性能[11]。在EN-CS子任务中，系统MITEL-EN-CS-03-T取得了与单语检索可比的性能，在所有跨语言参赛系统中排名第一。在CT-CT子任务中，我们的系统包揽了所有参评系统的前四名，其中系统MITEL-CT-CT-02-T性能最优。

在官方报告中，我们的系统返回了最多的相关文档，所以召回率最高。由于当前检索评测广泛使用 Pooling 技术获取相关文档集，召回率高的系统更容易取得高分。然而，开发集上的实验让我们对自己的方法有更深入的了解。在开发集实验中，系统取得高召回率，但准确率在很多问题上却不令人满意。这是因为 ad hoc 检索和 IR4QA 检索任务的相关性评价标准不同。例如，对于问题"Who is Hu Jintao?"，用户想知道关于中国国家主席的详细信息，因此介绍胡锦涛生平事迹的文档被认为是构成答案的最佳文档。即使在该文档中"胡锦涛"仅出现一次，该文档也应该在检索结果中排名靠前。相反，介绍胡锦涛访问日本的文档，即使关键字"胡锦涛"出现多次，也应该在返回结果中排名较低。但不幸的是，使用传统检索的相关性评价标准，后者将会在检索结果中排名很高。因此，传统的信息检索评价标准已不适用于 IR4QA 任务。但由于目前还没有直接针对 IR4QA 任务优化的目标函数，所以我们最终使用了模型融合，通过在现有评价标准上的风险最小化策略，在保持准确率不降低的情况下提高系统的高召回率，从而优化系统的整体性能。

表 3 NTCIR-7 IR4QA 子任务官方结果

Task	Run id	Mean AP	Mean Q	Mean nDCG
EN-CS	MITEL-EN-CS-01-T	0.5849	0.6005	0.7949
	MITEL-EN-CS-02-T	0.5693	0.5858	0.7847
	MITEL-EN-CS-03-T	**0.5959**	**0.6124**	0.7947
	MITEL-EN-CS-04-D	0.5789	0.5950	0.7907
	MITEL-EN-CS-05-TD	0.5898	0.6058	**0.8003**
CT-CT	MITEL-CT-CT-01-T	0.5791	0.5963	0.7835
	MITEL-CT-CT-02-T	**0.5839**	**0.6018**	**0.7873**
	MITEL-CT-CT-03-D	**0.5839**	0.6013	0.7869
	MITEL-CT-CT-04-T	0.5645	0.5783	0.7648

7. 结论

NTICR-7 评测是我们参加的第一次跨语言检索评测。我们测试了各种索引模型的性能，并使用索引模型融合提升了检索性能。通过一种参数调节的贪心算法来调节各索引模型在线性组合模型中的权重。NTCIR-7 评测结果表明，我们提出的方法能够获取高召回率和好的准确率。然而 ad hoc 检索和 IR4QA 检索的评价标准存在很大差异，我们的方法在提高准确率上还有很大提升空间，因此对融合结果重排序将是一种有效的策略，这也是我们下一步要探索的重点。

参 考 文 献

[1] J.Y. Nie., J. Gao, J. Zhang, M. Zhou. On the use of words and n-grams for Chinese information retrieval. In the 5th International Workshop on Information Retrieval with Asian Languages. IRAL-2000, Hong Kong, 2000. pp. 141–148

[2] L.X. Shi, J.Y. Nie. Using Unigram and Bigram Language Models for Monolingual and Cross-Language IR. In the Proceedings of NTCIR-6 Workshop Meeting, Tokyo, Japan, 2007, pp.20-25

[3] Tetsuya Sakai, Noriko kando. Are Popular Documents More Likely To Be Relevant? A Dive into the ACLIA IR4QA Pools. In the proceedings of The Second International Workshop on Evaluating Information Access, Tokyo, 2008. pp.8-9

[4] J. Lafferty and C.X. Zhai. Document Language Models, Query Models, and Risk Minimization for Information Retrieval. In the Proceedings of ACM SIGIR, New Orleans, LA USA, 2001, pp. 111 - 119

[5] J. Ponte and W.B. Croft. A language modeling approach to information retrieval. In the Proceedings of ACM SIGIR, Melbourne, Australia, 1998, pp. 275 - 281

[6] C.X. Zhai & J. Lafferty. A Study of Smoothing Methods for Language models Applied to Ad Hoc Information Retrieval. In the Proceedings of ACM SIGIR, New Orleans, LA USA, 2001, pp. 334 - 342

[7] C.X. Zhai & J. Lafferty. Model-based Feedback in the Language Modeling Approach to Information Retrieval. In the Proceedings of 10th international conference on Information and knowledge management, Atlanta, GR, 2001, pp. 403 - 410

[8] A.I. Rosti, N.F. Ayan et al. Combing Outputs from Multiple Machine Translation Systems. In the Proceedings of NAACL HLT 2007, Rochester, NY, 2007, pp. 228 - 235

[9] Weihua Luo, Tian Xia, Ji Guo, Qun Liu. ICT-Crossn: The System of Cross-lingual Information Retrieval of ICT in NTCIR-7. In Proceedings of NTCIR-7. Tokyo, 2008, pp. 132-139

[10] L.P. Yang and D.H. Ji. I2R at NTCIR5. In the Proceedings of NTCIR-5 Workshop Meeting, 2005, Tokyo, Japan. Available at http://research.nii.ac.jp/ntcir/publication1-en.html

[11] Teruko Mitamura, Eric Nyberg, Hideki Shima, Tsuneaki Kato, et al. Overview of the NTCIR-7 ACLIA: Advanced Cross-Lingual Information Access. In the proceedings of NTCIR-7, Tokyo, 2008, pp.77-114