IR4QA 系统中基于维基百科的查询扩展*

周斌，刘茂福，陈建勋

武汉科技大学计算机科学与技术学院 武汉 430065

E-mail: zb_zhoubin@163.com, liumaofu@wust.edu.cn, cjxwh@wust.edu.cn

摘要：针对信息检索中文档与查询之间的词不匹配问题，研究者们提出了许多有效的解决方法，其中，查询扩展是一种非常重要的技术手段。本文提出了一种基于维基百科的查询扩展方法。通过对此方法在 IR4QA 系统中的表现分析，表明对于某些特定类型的问题，该方法可以使查准率有一定的提高。

关键词：IR4QA，维基百科，查询扩展

Query Expansion in IR4QA System Based on Wikipedia Article Content

Zhou Bin, Liu Maofu, Chen Jianxun

College of Computer Science and Technology, Wuhan University of Science and Technology, Wuhan 430065

E-mail: zb_zhoubin@163.com, liumaofu@wust.edu.cn, cjxwh@wust.edu.cn

Abstract: Researchers have proposed many effective solutions to resolve term mismatch between query and documents in information retrieval system. Query expansion is one of the important technical approaches. In this paper, we propose a query expansion method based on Wikipedia articles. According to the performance of this method in IR4QA, we conclude that this approach can improve the precision in some specific types of questions.

Keywords: IR4QA, Wikipedia article, query expansion

1. 引言

问答系统(Question Answering，QA)，是指根据用户以自然语言提出的问题找出一个明确的答案，是信息检索与自然语言处理相结合的研究领域[1]。例如，对问题“2008年奥运会在哪举办？”，QA系统将直接回答“北京”。IR4QA即根据用户提出的问题，在目标文档集中使用信息检索技术将所有包含答案的文档检索出来提供给用户，它不生成问题的最终答案。

在问答系统中，有许多重要的因素影响着信息检索的表现。其中一个因素就是问题关键词的获取[2]，对于同一个问题，采用不同的分词方法，往往会得到不同的关键词。另外一个重要因素就是问题中的关键词和检索文档中的词不匹配[3]，由这些关键词直接生成的 Query 往往不能有效检索出包含相关信息的文档[4]。这是因为自然语言本身具有语义模糊性和歧义性，也就是一词多义现象，使得简单地使用这些关键字，可能导致返回一批不相关的文档。例如，使用“Windows”关键字检索 Windows 操作系统的信息，可能会得到一些关于房屋窗户的信息。另外，自然语言还具有多词一义现象，即对于同一个内容能有不同的表达方式，例如，“这部电影非常好看。”和“这部电影很棒。”表达的都是一个意思。这两个方面因素的影响会造成查询中的信息丢失和信息的冗余，从而导致文档检索的准确度和召回率不佳[5]。

* 本文承湖北省自然科学基金（项目号 2009CDB311）的资助

查询扩展是一种有效的提高文档检索准确度和召回率的信息检索方法。查询扩展的基本思想是分析原有的查询项，根据一定的原则选择与原查询项高度相关的扩展查询项，加入原有查询形成新查询。它的作用在于可以解决信息检索领域长期困扰的词不匹配问题，弥补用户查询信息不足的缺陷以及消除多义词的歧义。目前查询扩展技术已成为公认的能够改善信息检索查全率和查准率的关键技术之一。

本文在 IR4QA 系统中使用了基于维基百科的查询扩展方法。维基百科是一个多语种、基于网络的、自由的百科全书的内容项目，它里面的每一个条目都有相应的页面进行详细介绍，目前共有 30 万多篇条目以中文撰写。本文使用中文维基百科的内容，以与问题相关的页面作为基础，在页面中定位与问题相关性最高的段落，然后进行查询扩展。

文中的实验基于我们在 NTCIR-8 会议中的 IR4QA 任务所做的相关工作。NTCIR 是由日本情报信息研究所(National Institute of Informatics)主办的多语言处理国际评测会议，主要关注中、日、韩等亚洲语种的相关信息处理。

本文的第二部分简单的介绍了我们本次的 IR4QA 系统，第三部分详细描述了对于维基百科的处理方法，第四部分是实验结果，在第五部分给出了结论。

2. IR4QA 系统简介

本系统的设计思想是：首先通过对问题的分析，将得出的初始查询在维基百科中进行检索，提取相关的维基百科文档作为查询扩展的基础文档。在这个文档中找到和问题相关性最高的相应的段落，进行查询扩展，得到扩展后的新查询，再用新查询在目标文档集中进行检索，得到最终的查询结果。

根据以上设计思想，将系统分为四个部分：建立索引模块，问题分析模块，文档检索模块以及查询扩展模块。系统的详细结构如图 1 所示。

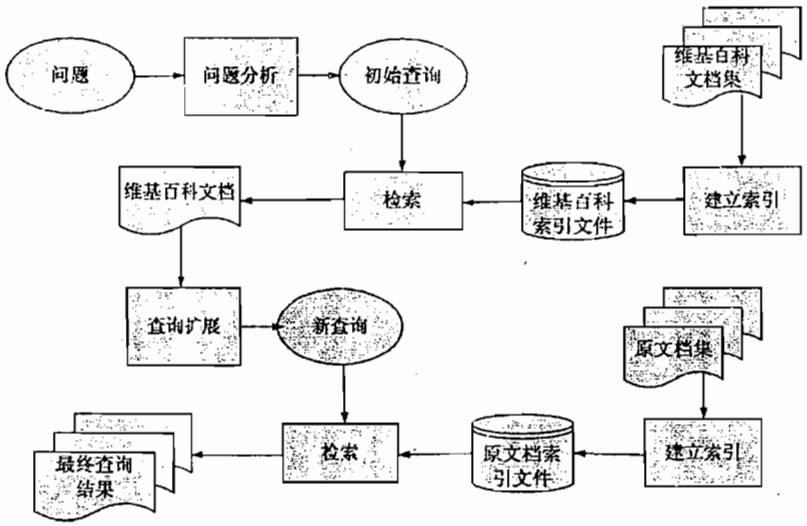

图 1. 系统结构图

在建立索引模块里，我们采用基于词典的最长匹配算法对目标文档集和维基百科的内容进行分词，并去掉其中的停用词。在问题分析部分，根据给定的问题，将其分为定义、传记、事件、

关系、原因等不同的类型。对于不同类型的问题，使用不同的模板去提取命名实体，并采用了与建立索引时同样的分词方法来提取出关键词，这些关键词组成了初始的查询内容。系统使用向量空间模型（VSM）的方法对全部文档进行检索。在对维基百科进行检索的时候，有时候会出现多个条目对应一个问题的情况，而且有些问题可能包含多个命名实体，这个时候就需要综合不同的条目内的段落进行查询扩展。

3. 对维基百科的处理

对维基百科的处理可以具体分为段落定位和查询扩展两步。段落定位找出和初始查询相似度最高的 n 个段落，查询扩展提取这些段落中的扩展词，加入初始查询中形成新查询。

3.1 段落定位

大多数维基百科的条目内容都有很长的说明性的文字，其中有很多内容与问题并不相关，因此，我们只需要对里面的一个或几个与问题相关的段落进行查询扩展。根据不同类型的问题，我们采用不同的定位方式，对于人物传记、定义类型的问题，我们直接使用第一段进行查询扩展。例如“高仓健是谁？”属于人物传记类型，于是直接在维基百科的高仓健条目中取第一段内容进行查询扩展。对于其他类型的问题，我们采用如下的方法：

首先根据如下的向量空间模型计算公式来计算初始查询和维基百科段落的相似度：

$$SimVsm(D_1, D_2) = \frac{\sum_{k=1}^{n} W_{1k} \cdot W_{2k}}{\sqrt{\sum_{k=1}^{n} W_{1k}^2 \cdot \sum_{k=1}^{n} W_{2k}^2}} \qquad ①$$

其中 $$W_{ik} = \frac{tf_{ik} \cdot \log(N / n_k)}{\sqrt{\sum_{k=1}^{n} tf_{ik}^2 \cdot [\log(c + N / n_k)]^2}} \qquad ②$$

tf_{ik} 表示关键词 T_k 在文档 D_i 中的文档内频数，N表示全部文档的总数量，即维基百科条目中的段落总数，n_k 表示包含关键词 T_k 的文档数量，c为常数，取0.01，是为了防止出现分母为0的情况。

例如问题“郭晶晶和吴敏霞的关系是？”，从这个问题中我们用模板得到两个命名实体——“郭晶晶”和“吴敏霞”，它们在维基百科中都各对应了一个条目，将这两个条目下的段落组成一个文档集，每个段落作为一个文档，然后用上面的公式计算初始查询和这些段落的相似度。使用①式计算出相似度后，按降序排列，取其中前n个段落用来进行查询扩展。

3.2 查询扩展

为了使查询扩展更有效，我们选取的扩展词必须与初始查询所表征的主题或概念具有很好的相关性。在本文中采用基于局部共现的查询扩展方法，该方法属于基于局部文档集分析的方法[6]。相对而言，局部分析的方法能有效地将那些与查询表示的主题相关的词加入初始查询中，因而其扩展的效果要更好一些。

所谓共现，指的是两个词项在一定的文本窗口范围内共同出现。本文将文本窗口定义为维基百科中一个段落的范围。首先对上面得到的 n 个段落分词，对于其中的某个词项 w 和问题中的某个查询词 q，定义 w 和 q 在文档 d（本文为维基百科的一个段落）中的共现频度为 $coof(w,q|d)=tf(w|d)\times tf(q|d)$，其中 $tf(\bullet|d)$ 表示一个词项在 d 中的出现次数。由于自然语言的数据稀疏性，文档 d 中不同词项的出现次数之间可能存在较大的差异，在减少这种差异对共现信息的负面影响的同时，还需要根据共现信息来区分不同词项的重要性。定义词项 w 和查询词 q 在局部文档集 S（即为上面得到的 n 个段落）中的共现度 $cood(w,q|S)$ 为 w 和 q 在 S 中的所有文档内的平均共现频度。本文采用如下的式子来计算共现频度[7]：

$$cood(w,q|S)=\frac{\sum_{d\in S}[\log(tf(w|d)+1.0)\times\log(tf(q|d)+1.0)]}{n} \quad ③$$

令 $cohd(w,Q|S)$ 表示词项 w 与查询 Q 在局部文档集 S 中的关联度，定义为：

$$cohd(w,Q|S)=\prod_{q\in Q}(cood(w,q|S)+1.0) \quad ④$$

仅用上面这个式子来选择扩展词会出现一个问题，那就是会有一些出现频率很高但是区别率很低的词，比如停用词，显然我们不能将它们作为扩展词。因此，我们使用的是如下的函数：

$$f(w,Q|C,S)=\sum_{q\in Q}idf(q|C)idf(w|C)\log(cood(w,q|S)+1.0) \quad ⑤$$

其中 C 为问题中的所有命名实体对应的维基百科条目内的段落总和。

使用上式计算出每个词项与查询 Q 在局部文档集 S 中的关联度后，选择其中的前 m 个作为最终的扩展词。还是以问题“郭晶晶和吴敏霞的关系是？”为例，先对 3.1 中得到的 n 个段落分词，然后使用③式分别计算其中的某个词项（如河北）与初始查询中的查询词（郭晶晶、吴敏霞）在这 n 个段落中的平均共现频度，接着根据计算出来的平均共现频度使用⑤式计算此词项（河北）与初始查询的关联度，其它词项也采用同样的方法计算与初始查询的关联度。最后根据关联度的大小，选择了“女子”、“跳水”、“3 米板”、“三米板”等扩展词。

很明显，不同的扩展词具有不同的重要性，因此，查询扩展要考虑的另外一个问题是如何对扩展后的查询进行词项权重的分配。本文中，我们采用如下的权重分配方法。

$$w(q|Q_{new})=p\cdot w(q|Q)+k\cdot avg(boost)\cdot\frac{score(q)}{MaxScore}w(q|d) \quad ⑥$$

其中 $w(q|Q)$ 为查询词 q 在初始查询 Q 中的权重，$w(q|d)$ 为查询词 q 在文档 d 中的权重，p 和 k 为两个大于 0 的可调参数，在实验中我们将 p 设为 1.0，k 设为 0.9。

经过以上两步的处理，就可以将从维基百科段落中提取的扩展词加入到初始查询中去了。

4. 实验结果

本次实验的文档集使用的是 2002 年至 2005 年的新华日报。问题集为 NTCIR 会议提供的 100

个不同类型的问题。根据我们对问题的分析，各类问题的数量情况如下表所示。

表 1 全部问题分类情况表

问题类型	人名/机构组织	关系	时间/地点	人物传记	定义	事件	原因	列举
问题数量	9	17	10	11	10	18	22	3

从分类情况看来，关系、事件和原因这三种类型的问题在问题集中占的比重比较大，总和达到了 57%，而其他类型的问题数量较少。

下表统计了各问题在维基百科中检索的情况，分为一个问题对应多个页面、一个问题对应一个页面和找不到相关页面的情况。另外，对找到页面的问题也统计了其中定位不到具体段落的数量。

表 2 在维基百科中检索的统计结果

类型	对应多个页面	对应一个页面	找不到页面	找不到段落
数量	36	60	4	28

从上表可以看到虽然大部分问题都可以在维基百科中找到相应页面，但是仍然有一部分定位不到段落，也就是说里面的内容回答不了所提的问题，加上找不到页面的问题，共有 32 个相关问题在维基百科中找不到答案。根据统计，这 32 个问题中关系和原因类型的问题数占的比重比较大，其中原因类型的问题数量达到了 17 个占总数量的 53%。详细情况如下表：

表 3 问题分类情况表

问题类型	人名/机构组织	关系	定义	事件	原因	列举
问题数量	3	6	1	4	17	1

本文认为，对于关系、原因类问题在维基百科中检索的效果并不理想，是和维基百科的内容形式有关的。我们知道，维基百科是一部在线的电子百科全书，它里面的内容绝大多数都是说明、描述类型的，因此像定义、人物传记之类的问题更符合维基百科的内容组织形式。例如问题“高仓健是谁？”可以根据高仓健找到维基百科中高仓健的页面，然后定位到第一段；问题“俄罗斯“K-159”号核潜艇为什么会沉没？”则找不到相关的维基百科页面；问题“巴厘岛爆炸和本·拉丹的关系？”虽然能找到相关的维基百科页面，但是却找不到可以回答问题的段落。

从返回的结果评测来看，对于像定义、人物传记、机构组织、时间地点这些类别的问题，使用维基百科扩展后的结果有了一定提高，但是对于其他类型的问题，则表现不是很好。我们使用 P@N 评测指标对部分结果进行了评测。

对某个特定的查询，P@N 指的是该查询的检测结果中前 N 篇文档的准确率[8]，对多个查询构成的查询集进行评测时，P@N 表示该查询集合中所有查询对应的 P@N 的算术平均值。在本文中，把 P@10、P@20、P@30 作为实验的评测指标。下表是定义、人物传记这类问题的结果对比情况。

表 4 P@N 评测结果

P@N	初始结果	使用维基百科后结果	提高百分比(%)
P@10	0.447	0.493	10.3
P@20	0.326	0.344	5.5
P@30	0.304	0.317	4.3

从表中我们看到，在使用了维基百科后准确率的确有了提高，最高达到了 10.3%。但是随着 N 的加大，准确率是有一个下降的趋势的。

在本次任务中，我们对每个问题都提交一个文档数不大于 1000 的排列，官方返回的中文检索评价结果如下表所示：

表 5 测评结果

测评指数	Mean AP	Mean Q	Mean nDCG
测评结果	0.2694	0.293	0.4881

从评价结果来看，整体的表现并不理想。我们分析除了上面提到的维基百科内容组织形式的原因外，在查询扩展的方法上可能还有些问题。本文采用的是基于维基百科的查询扩展方法，但是最终检索的目标是会议提供的文档集，因此在查询扩展中加入的扩展词可能在目标文档集中并不存在，或者目标文档集对同一个问题的描述与维基百科并不相同，这就使得查询扩展后的新查询并没有起到预期的效果。

5. 结论

在信息检索中，查询扩展是解决词不匹配问题的一种有效技术手段。本文探讨了一种基于维基百科的查询扩展的方式，首先通过初始查询，对维基百科进行检索，找到与问题相关度最高的段落，进行查询扩展，形成新的查询，然后使用该查询对目标文档集进行检索，得到最终结果。从实验结果来看，对于特定的问题类型，如定义、人物传记等类型，准确率有了一定的提高。但是此方法还存在着一些问题，扩展词和目标文档联系性不强，需要在接下来的研究中继续探讨。

参考文献

[1] VOORHEES E M, TICE D M. Building a question answering test collection[C]. Proceedings of the 23rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. New York: ACM Press, 2000:200-207

[2] Pantel P. and Lin D. A Statistical Corpus-Based Term Extractor[C]. Proceedings of AI 2001. Ottawa, Canada: Springer-Verlag, 2001:36-46.

[3] Salton G. and McGill MJ. Introduction to Modern Information Retrieval[M], New York, NY: McGraw-Hill. 1983.

[4] Huang T S, Mehrotra S, Ramchandran K. Multimedia Analysis and Retrieval System (MARS) Project[C]. In Proceedings of the 33rd Annual Clinic on Library Application of Data Processing-Digital Image Access and Retrieval, March 1996.

[5] Rijsbergen van. A new theoretical framework for information retrieval[C]. In Proceedings of 1986 ACM Conference on Research and Development in Information Retrieval, 1986:194-200.

[6] Xu J X. and Croft W B. Improving the Effectiveness of Information Retrieval with Local Context Analysis[J].ACM Transactions on Information Systems, 2000, 18(1):79-112.

[7] 丁国栋, 白硕, 王斌. 一种基于局部共现的查询扩展方法[J].中文信息学报, 2006,20(3):84-91.

[8] Baeza-Yates R, Ribeiro-Neto B. Modern Information Retrieval[M]. New York: Addison-Wesley-Longman, 1999.