第34卷 第17期 计 算 机 工 程 2008年9月
Vol.34 No.17 Computer Engineering September 2008
·人工智能及识别技术· 文章编号：1000—3428(2008)17—0186—02 文献标识码：A 中图分类号：TP18

问答式信息检索中模式优化及性能评价

杜永萍，叶乃文
(北京工业大学计算机学院，北京 100124)

摘 要：问答式信息检索是新一代搜索引擎，集成自然语言处理和信息检索科学的研究成果，提高信息检索效率。该文介绍问答式信息检索中的模式优化及其应用，并进行客观评价。模式在问答式信息检索中有两个重要应用——查询扩展和答案抽取。实验结果表明，在TREC标准测试集上，采用模式匹配策略实现答案抽取，能有效地提高问答式信息检索系统的准确率。

关键词：信息检索；模式优化；性能评价

Pattern Generalization and Performance Evaluation in Question Answering Information Retrieval

DU Yong-ping, YE Nai-wen
(Institute of Computer Science, Beijing University of Technology, Beijing 100124)

【Abstract】 Open domain Question Answering(QA) represents a challenge of natural language processing, aiming at returning exact answers in response to natural language questions. Pattern matching is an effective strategy for QA. This paper demonstrates the technique of pattern generalization and its applications in QA. There are two important applications of pattern in QA. Pattern-based query expansion retrieves more exact snippets including the correct answer and the performance is improved. Another application is the answer extraction which applies the pattern matching approach. Experimental result on TREC data indicates that the pattern matching method is effective.

【Key words】 information retrieval; pattern generalization; performance evaluation

1 概述

目前，为了有效查找、利用计算机可读文本信息，信息检索与信息抽取技术日益重要。问答式(Question Answering, QA)信息检索集成了自然语言处理和信息检索科学的研究成果，把大量原来需要用户来完成的操作，如查询关键词的生成和答案的搜索，都交给计算机自动完成，减轻了用户的负担，提高了信息检索和利用的效率。这个任务看似简单，但对于计算机具有很大的挑战性。

TREC(Text REtrieval Conference)会议于1999年设立问题回答项目，促进该领域的技术交流。TREC是文本检索领域最权威的国际评测会议，由美国国家标准技术局(NIST)和国防部高级研究计划局(DARPA)组织，为各种方法和不同系统提供了一个公平竞争的舞台。

问答式信息检索区别于通常意义下文本检索的关键在于答案抽取模块，它实现了精确答案的抽取，该模块可以采用不同策略实现，如：逻辑推理[1]，模式匹配[2-3]等。

由于答案在语料中的表达形式是灵活的，给准确识别带来很大难度，采取模式匹配是一种有效的方法。建立一个完善的模式知识库。总体上有两种途径进行该知识库的构建：人工构建与自动学习。在TREC10参加评测的单位中，InsightSoft[3]使用了从大规模语料中提取出的简单模式来匹配抽取答案，人工构建了庞大的模式库。目前，也有其他研究机构实现了自动学习简单模式并用于问题回答。ISI[2]所获取到的模式还有一定局限性，包含词形但没有任何外部知识(如语义知识信息等)。一个良好的模式可以融入语义信息，使模式有更高的可扩展性与可靠性。

2 模式构建与优化

2.1 模式构建

针对不同问题类型，获取其相应的答案抽取模式。答案抽取模式，也即答案出现的上下文，描述了问题的答案句可能表现的不同形式，在问答式信息检索系统中实现答案抽取。

根据问题类型分别对其答案模式进行评价。评价采用数据挖掘中的指标：可信率(confidence)。可信率越高的模式抽取到正确答案的可靠性越高，而低可信率的模式则不能保证其正确性。有关答案抽取模式的获取算法与评价算法详见文献[4]。

示例问题类型及其答案抽取模式(含可信率值)见表1。其中，问题类型/问题为：When did Q_LCN Q_DoVerb Q_BNP ?
Sample question:When did Hawaii become a state?

表1 答案模式示例

答案模式	可信率
Q_LCN Q_DoVerb Q_BNP in <A>.	0.91
Q_DoVerb Q_BNP in <A>, Q_LCN	0.82
in <A>, Q_LCN Q_DoVerb Q_BNP	0.73

<A>标记表示问题的答案，其他标记表示问题中的不同成分，其含义如下：

Q_LCN= “Hawaii”

Q_DoVerb= “become”

基金项目：北京工业大学博士科研启动基金资助项目(52007012200701)

作者简介：杜永萍(1977－)，女，讲师、博士，主研方向：自然语言处理；叶乃文，副教授

收稿日期：2007-11-20 **E-mail：**ypdu@bjut.edu.cn

Q_BNP= “a state”

在利用答案模式进行匹配抽取答案的过程中[4]，出现在<A>标记位置的信息将被选作问题的侯选答案。

2.2 模式优化

对于学习到的答案模式，在实际应用的过程中，发现部分特殊类型的模式存在的问题，降低模式的可适用性，示例如下：

示例

问题类型：When be Q_PRN Q_DoVerb? (Q_DoVerb= “born”)

问题： When was Abraham Lincoln born?

片段：Abraham Lincoln (1809-1865), the sixteenth president of the United States.

答案：1809

答案模式：Q_PRN (<A> - 1865)

(注：Q_PRN= “Abraham Lincoln”)

新问题：When was Thomas Jefferson born?

对于这样一个具有同样问题类型的新问题“ When was Thomas Jefferson born? ”，答案模式 Q_PRN (<A> - 1865)由于存在特殊的词“ 1865 ”而无法适用。

对于如上示例中的答案模式，用作回答同样问题类型的新问题，适用性极差，主要原因在于，这些答案模式包含与和原问题紧密相关的信息，如“ 1865 ”。而这些信息对于其他问题基本上毫无意义。为了解决这一缺陷，需要采取策略对这种形式的模式进行泛化操作，以消除原来存在的特殊文本信息。

在经过对这类特殊答案模式的观察，可以发现，存在的特殊文本信息通常为一些专有名词如：日期，地名，机构名等。笔者利用了实体名识别工具对模式进行泛化，用规定的实体名标记集代替相应的文本信息，解决模式适用性弱的问题，使其具有较好的鲁棒性，达到模式优化的目的。标记集如下：

日期- DATE；人名-PRN；地名- LCN；机构名-ORG；数字-NUM

上述示例的答案模式在经过泛化之后为

示例答案模式：Q_PRN (<A> - DATE)

对模式进行如上的扩展，并特别将如下问题类型在 TREC13 的测试集上做了分析，如表 2 所示：

When be Q_PRN Q_DoVerb? (Q_DoVerb=born)

Where be Q_PRN Q_DoVerb? Q_DoVerb=born)

表 2 模式泛化前后实验结果比较

Experiment	*precision*/(%)	*MRR*
1	40	0.50
2	42	0.52
3	36	0.36
4	40	0.46

Experiment1 代表系统回答问题类型“ When be Q_PRN Q_DoVerb? ”时，没有进行模式优化的实验结果。

Experiment2 代表系统回答问题类型“ When be Q_PRN Q_DoVerb? ”时，进行模式优化后的实验结果。

Experiment3 代表系统回答问题类型“ Where be Q_PRN Q_DoVerb ? ”时，没有进行模式优化的实验结果。

Experiment4 代表系统回答问题类型“ Where be Q_PRN Q_DoVerb? ”时，进行模式优化后的实验结果。

3 模式应用

模式在问答式信息检索系统中有两个重要的应用：查询扩展和答案抽取。

3.1 查询扩展应用

答案模式描述了问题答案可能出现的不同表达形式，这是一个非常好的资源，可以充分利用起来，实现查询扩展，使检索模块得到更好的性能。

并非所有的答案模式都被用作查询扩展，而是选取可信率高的答案模式，对它们进行实例化，生成查询。如：

问题：Where is Mount Olympus?

(Q_LCN= “Mount Olympus”)

答案模式：Q_LCN is located in <A>.(Confidence=0.85)

答案模式实例化：Mount Olympus is located in <A>

扩展查询：“Mount Olympus is located in”

如上查询提交给Google后，会返回很多包含正确答案的片段，如“ …Mount Olympus is located in Northern Greece… ”，这样，可以更加准确地定位正确答案。

搜索引擎性能的评价，通常采用的指标为精度(precision)和召回率(recall)，具体含义为

$$\text{精度} = \frac{\text{检索到的相关文献数目}}{\text{检索到的文档总数}} \quad (1)$$

$$\text{召回率} = \frac{\text{检索到的相关文档数目}}{\text{文档集中所有相关文档总数}} \quad (2)$$

借助 Google 实现检索，并基于 Web 信息，笔者无法判定计算召回率所需要的“ 文档集中所有相关文档总数 ”，在这里，主要通过精度来评价所生成查询的质量。

在 TREC13 的测试问题集上，可以选择疑问词为 Where 类型的问题，评价基于模式的查询扩展性能，如表 3 所示。

表 3 检索模块查询性能评价

精度	基本查询性能/(%)	扩展查询性能/(%)
P@5	42.8	57.1
P@10	50.0	78.5
P@20	82.1	82.1
P@30	82.1	82.1

其中，P@*n* 代表搜索引擎对每个提交的查询返回前 *n* 个结果时，系统的精度。

如表 3 所示，扩展查询在 P@5 与 P@10 的性能明显高于基本查询的性能，而在 P@20 与 P@30，基本查询与扩展查询的性能没有差异。实验结果说明，基于模式的查询扩展可以将包含答案的精确片段排序在返回结果列表的前端，有效地提高了检索效率。

3.2 答案抽取应用

获取答案模式就是为了在问答式信息检索系统中实现答案抽取。具体抽取方法见文献[4]。利用 TREC 提供的问题集与答案集，笔者进行几组实验来测试模式匹配方法的性能。

系统按照不同疑问词类型问题的性能分布，其实验结果如表 4 所示。

表 4 不同疑问词类型问题性能分布 (%)

TREC*n*	When	Who	What	Where	How
TREC8	32.5	28.3	24.8	35.2	30.2
TREC9	40.0	28.0	39.0	38.0	20.0
TREC10	31.4	30.4	22.6	37.6	30.8
TREC11	32.4	32.6	25.3	41.2	32.4

对于不同疑问词类型的问题，where 类型问题性能最稳定，在不同测试集上，准确率始终最高；而 what 类型问题则准确率始终最低。

4 结束语

问答式信息检索中实现问题回答，模式匹配是一种有效的方法。本文介绍了问答式信息检索中的模式优化及其应用。

（下转第 190 页）