计算机研究与发展
Journal of Computer Research and Development
DOI:10.7544/issn1000-1239.2018.20170232
55(5):958-967,2018

基于粗糙集知识发现的开放领域中文问答检索

韩 朝[1,2,3] 苗夺谦[1,2] 任福继[3] 张红云[1,2]
[1](同济大学电子与信息工程学院 上海 201804)
[2](嵌入式系统与服务计算教育部重点实验室(同济大学) 上海 201804)
[3](德岛大学工学部 日本德岛 7708506)
(1990hanzhao@tongji.edu.cn)

Rough Set Knowledge Discovery Based Open Domain Chinese Question Answering Retrieval

Han Zhao[1,2,3], Miao Duoqian[1,2], Ren Fuji[3], and Zhang Hongyun[1,2]
[1](*College of Electronic and Information Engineering, Tongji University, Shanghai* 201804)
[2](*Key Laboratory of Embedded System and Service Computing (Tongji University), Ministry of Education, Shanghai* 201804)
[3](*The Faculty of Engineering, Tokushima University, Tokushima, Japan* 7708506)

Abstract In the information retrieval (IR) based open domain question answering system (QA system), the main principle is that first use the semantic tools and knowledgebase to get the semantic and knowledge information, then calculate the matching value of both semantic and knowledge. However, in some practical applications of Chinese question answering, because of the uncertainty of both the Chinese language representation and the Chinese knowledge representation, the current methods are not very effective. To solve this problem, a rough set knowledge discovery based Chinese question answering method is proposed in this paper. It uses the method of rough set equivalence partitioning to represent the rough set knowledge of the QA pairs, then uses the idea of attribute reduction to mine out the upper approximation representations of all the knowledge items. Based on the rough set QA knowledgebase, the knowledge match value of a QA pair can be calculated as a kind of knowledge item similarity. After all the knowledge similarities of one question and its answer candidates are given, the final matching values which combines rough set knowledge similarity with traditional sentence similarity can be used to rank the answer candidates. The experiment shows that the proposed method can improve the *MAP* and *MRR* compared with the baseline information retrieval methods.

Key words question answering (QA) system; information retrieval (IR); rough set; knowledge discovery; text mining

---

收稿日期:2017-04-01;**修回日期**:2017-05-23
基金项目:国家自然科学基金项目(61673301,61273304,61573255);高等学校博士学科点专项基金项目(20130072130004);安徽省高校优秀青年人才基金项目(gxyq2017056)
This work was supported by the National Natural Science Foundation of China (61673301, 61273304, 61573255), the Specialized Research Fund for the Doctoral Program of Higher Education of China(20130072130004), and the Universities Outstanding Young Talents Foundation of Anhui Province (gxyq2017056).
通信作者:苗夺谦(dqmiao@tongji.edu.cn)

摘　要　基于信息检索的开放领域问答系统,其主要原理是先使用语义分析工具和知识库获得确定性的语义和知识等信息,然后再进行问答句匹配度计算.但在实际的中文问答系统应用中,由于中文语言表达的不确定性和中文知识表达的不确定性大量存在,现有的匹配度计算方法不适合大量不确定性存在的应用场景.针对这一问题,提出了一种基于粗糙集知识发现的中文问答检索方法,利用粗糙集的属性约简方法和上近似概念从已标注的问答语料库中发现并表示知识,再结合传统的句子相似度方法对问句和候选句进行匹配度计算.实验结果表明:相对传统的问答检索方法,该方法在 MAP 和 MRR 两个评测指标上均有提升.

关键词　问答系统;信息检索;粗糙集;知识发现;文本挖掘

中图法分类号　TP391

问答系统是当前自然语言处理研究领域的热点.现有的问答系统按照答案的产生方式主要分为2种:1)基于信息检索的问答系统,即在已经给定了候选答句的情况下,针对输入的问句返回最匹配问句的一个或多个候选答句;2)基于自然语言生成的问答系统,即不给定候选答句,利用自然语言生成的相关技术直接构造答句并返回.由于现有的自然语言生成技术尚未成熟,在现实应用中基于信息检索的问答系统仍然是主流[1-2].

基于信息检索的问答系统主要是通过计算问句和若干个候选答句的匹配度来获取最能匹配问句的答句,其中,匹配重心主要集中在问句和答句的话题关键点.例如,在问句"腾讯公司的老板是谁?"中,【"腾讯公司","老板"】即为该问句的话题关键点,而在候选答句"腾讯公司的老板是马化腾."和"腾讯公司的总部在中国深圳."中,前者的话题关键点是【"腾讯公司","老板"】,后者是【"腾讯公司","总部"】,因此前者与问句有更高的匹配度.

相比于英文问答系统,中文问答系统存在如下问题:首先,由于中文的语言特性,中文问答系统的问答匹配度计算主要先通过自然语言处理工具对问句和答句进行分词和词性标注等预处理;然后再对分词后的问句和答句进行句子相似度计算.尽管中文分词技术已经相对成熟,但由于中文的语言灵活特性,中文的语句表达存在大量不确定性,中文语言处理工具得到的预处理结果有时不能完全满足后续分析过程的需要.例如在给定上下文情况下,"苹果"和"苹果公司"都是指代"苹果公司",但若候选句中"苹果公司"被略写为"苹果",而分词工具将"苹果公司"作为命名实体切分,"苹果"作为名词切分,且同义词词库中缺乏"苹果"和"苹果公司"的同义关联的话,在后续的处理过程这2个词将会被视作2个不同的对象,进而影响后续的匹配度计算结果.

其次,中文的问答知识的表达方式也存在大量的不确定性.例如,给定问句"黎明来自哪个国家?"、"黎明的国籍是什么?"、"黎明是在哪个国家出生的?"、"黎明是中国人还是韩国人?",这些问句都可以选择"黎明出生在中国."作为最匹配答句,但这些问句的话题关键点可以表达成【"黎明","国籍"】、【"黎明","出生地"】、【"黎明","出生地点"】等多种方式,因而增加了问句和候选答句的话题相似度的计算难度.

以上2种情况可以总结为中文语言表达的不确定性和中文知识表达的不确定性.在实际应用中,由于这2种不确定性的大量存在,现有的利用中文语义分析工具和中文知识库获得的确定性的信息的匹配度计算方法不适合大量不确定性存在的应用场景.

针对上述不确定性问题,本文从粗糙集理论的角度,提出了一种基于粗糙集知识发现的中文问答检索方法,利用粗糙集的属性约简方法和上近似概念从已标注的问答语料库中发现并表示知识,利用获得的粗糙集问答知识结合传统的句子相似度方法对问句和候选句进行匹配度计算.

1　相关工作

在不确定性信息处理方面,现有的处理不确定性信息的理论主要有模糊集(fuzzy set)[3]、粗糙集(rough set)[4]和商空间(quotient space)[5]等.其中粗糙集理论是由 Pawlak[4]首先提出,并在实际的理论研究和应用研究过程中扩展出了模糊粗糙集[6]、邻域粗糙集[7]、变精度粗糙集[8]等多种模型.粗糙集模型的关键在于不同的等价关系下的上近似、下近似和边界域的确立,并依靠不同的等价划分对知识进行不同程度的粒化,从而得到不同的概念或范畴[9-10].

粗糙集理论已经在文本情感分析[11]、知识约简[12]和数据挖掘[13]等多个领域得到了广泛应用.

在文本信息检索方面,文本检索技术可以分为2个部分.首先是文本的语义或话题的向量表示.常见的向量化模型有经典向量空间模型(vector space model, VSM)[14]、TF·IDF向量空间模型[15],以及近年来应用比较广泛的LDA模型(Latent dirichlet allocation)[16]和LSI模型(Latent semantic indexing)[17],此外还有近年来备受关注的深度学习领域的Word2Vec词向量模型[18],以及基于Word2Vec发展而来的Doc2Vec模型[19].其次,对文本向量化后,通常使用余弦相似度来表示2个文本向量之间的相似程度[20].除此之外,Jaccard指数(Jaccard coefficient)[21]、Ochiai指数(Ochiai coefficient)[22]等方法也可以用来计算2个对象之间的关联程度.

问答系统和传统的信息检索的区别在于,传统信息检索中直接由用户输入关键词,而在问答系统中,不论是基于知识库的问答系统还是基于检索的问答系统,用户输入的都是自然语言表达的问句而非关键词串,因而问答系统首先要解决问句和答句的关键词抽取[23],之后才是根据候选句或文档,或根据知识库,返回匹配问句的答句的传统信息检索过程.在问句关键词抽取方面,现有的方法多是利用自然语言处理工具分析得到初步的词汇、POS(part of speech)标记、语法成分等信息后,挖掘问句和答句之间关联规则或分类特征,例如,文献[23]给出的基于中文句法的中文问答方法、文献[24]给出的基于篇章语义的中文问答方法、文献[25]给出的基于POS标记特征和规则挖掘的英文问答方法.

问答系统的研发过程往往基于不同的应用背景,而不同背景下问答系统的预期功能不同,所用语料、知识库以及问答系统的指标要求也不尽相同,因而不同的文献中提到的问答系统的评测语料和评测指标也不相同[26-27].目前国际上英文问答系统的相关评测有TREC QA Track[28],NTCIR QALab[29].而在中文问答系统方面,国内的NLPCC自2015年开始举办开放领域中文问答系统评测比赛[30].

2 基本概念

2.1 粗糙集的基本概念

定义1[31]. 给定知识库 $K=\{U,S\}$, 其中 U 为论域,S 表示论域上的等价关系簇,则 $\forall X\subseteq U$ 和论域 U 上的一个等价关系 $R\in IND(K)$,其子集 X(又称概念或信息粒)关于 R 的下近似 $\underline{R}(X)$、上近似 $\overline{R}(X)$、边界域 $bn_R(X)$ 和负域 $neg_R(X)$ 分别为

$$\underline{R}(X)=\{x|(\forall x\in U)\wedge([x]_R\subseteq X)\}, \quad (1)$$

$$\overline{R}(X)=\{x|(\forall x\in U)\wedge([x]_R\cap X\neq\varnothing)\}, \quad (2)$$

$$bn_R(X)=\overline{R}(X)-\underline{R}(X), \quad (3)$$

$$neg_R(X)=U-\overline{R}(X), \quad (4)$$

其中,下近似 $\underline{R}(X)$ 表示根据等价关系 R 判定肯定属于 X 的元素集合,上近似 $\overline{R}(X)$ 表示根据等价关系 R 判定肯定或可能属于 X 的元素集合,边界域 $bn_R(X)$ 表示根据等价关系 R 暂时无法判定是否肯定属于 X 的元素集合,负域 $neg_R(X)$ 表示根据等价关系 R 判定肯定不属于 X 的元素集合.

2.2 向量的归一化和余弦相似度

定义2[32]. 给定 n 维向量 $\boldsymbol{A}=(a_1,a_2,\cdots,a_n)$,其归一化后的向量 $\boldsymbol{A}'$ 为

$$\boldsymbol{A}'=\frac{\boldsymbol{A}}{\|\boldsymbol{A}\|}=\frac{\boldsymbol{A}}{\sqrt{\sum_{i=1}^{n}a_i}}. \quad (5)$$

定义3[32]. 给定2个 n 维向量 $\boldsymbol{A}=(a_1,a_2,\cdots,a_n)$ 和 $\boldsymbol{B}=(b_1,b_2,\cdots,b_n)$,其余弦相似度为

$$Sim=\frac{\boldsymbol{A}\cdot\boldsymbol{B}}{\|\boldsymbol{A}\|\times\|\boldsymbol{B}\|}=\frac{\sum_{i=1}^{n}a_i\times b_i}{\sqrt{\sum_{i=1}^{n}a_i^2}\times\sqrt{\sum_{i=1}^{n}b_i^2}}. \quad (6)$$

3 基于粗糙集的问答系统知识发现和表达

给定一个问句 *ques* 和一个由若干个候选答句构成的集合,候选答句集合可以划分为2个部分:1)跟问句的匹配度较高的答句集合,在此称为正匹配句集合,记作 Set_p;2)跟问句的匹配度较低的集合,称为负匹配句集合,记作 Set_n.首先将 *ques* 和 Set_p,Set_n 中的每个句子都做了分词处理,即每个句子都视作一个词的集合.对于每一个词,可以根据其在问句和正、负匹配集合中出现的情况总共分为7类,用【*ques*|||Set_p|||Set_n】的方式来分别标记该词在句子、正匹配集合和负匹配集合中的出现情况,如表1所示.

当给定1个问句和若干个候选答句时,被选入正匹配集合的候选答句满足2个条件:

1) 这类候选答句和问句满足相对最细粒度下的话题相似.例如,候选答句①“腾讯公司的老板是

马化腾."和候选答句②"腾讯公司的总部在中国深圳.",在第1层次粒度下都是"腾讯公司"相关话题,但在进一步粒化分析后,前者的话题变为【"腾讯公司","老板"】,后者的话题变为【"腾讯公司","总部位置"】,因此若问句为"腾讯公司的老板是谁?',因问句的较细粒度的话题为【"腾讯公司","老板"】,因此候选答句①相比候选答句②有更高的匹配度.

2) 候选句之所以能成为问句的答案,是因为其含有同等问句粒度下问句所缺失的信息.例如上述例子中的候选答句①之所以能成为问句的答案,是因为其除了含有【"腾讯公司","老板"】这一与问句相同粒度的话题信息外,还含有"马化腾"这一问句所缺失的答案信息.

表1中的7类词汇在一定程度上反映了问答句的不同粒度下的话题信息和答案信息.例如,"腾讯公司"为在问句和正、负匹配集合中均出现的词(标记为【1‖1‖1】),即表示所有问答句都是"腾讯公司"相关;"谁"可能为只在问句中出现的词(标记为【1‖0‖0】);"老板"为只在问句和正匹配句中出现的词(标记为【1‖1‖0】).

Tabel 1　Word Tag and Meaning

表1　词标记和含义

Number	Tag	Meaning
1	【0‖0‖1】	Exist only in Set_n
2	【0‖1‖0】	Exist only in Set_p
3	【0‖1‖1】	Exist in both Set_p and Set_n, but not in ***ques***
4	【1‖0‖0】	Exist only in ***ques***
5	【1‖0‖1】	Exist in both ***ques*** and Set_n, but not in Set_p
6	【1‖1‖0】	Exist in both Set_p and ***ques***, but not in Set_n
7	【1‖1‖1】	Exist in ***ques***, Set_p and Set_n at the same time

上述过程是利用不同的句子集合判定词的标记的训练过程.我们可以把词看做划分规则,不同标记的词则为该词可以对句子划分入问句、正匹配和负匹配的划分程度,则训练过程是通过训练文本获得划分规则的过程,而检索(测试)过程则为利用划分规则和问句将候选句划分入正、负匹配句集的过程.根据粗糙集理论,当给定问句和正、负匹配句集后,问句的话题和对应的答案信息所构成的问答知识的下近似词汇更有可能在标记为【1‖1‖0】,【1‖0‖0】,【0‖1‖0】的词汇集合中,而标记为【0‖0‖1】的词汇即为该问答知识的负域.但在实际应用的场景中,由于语言表达的不确定性存在,可能会存在【1‖1‖0】词汇缺失等情况,反而在【1‖0‖1】词汇集合中可能找到问句的话题信息.例如,给定问句"《线性代数》这本书的内容有哪些?'及其2个候选答句,候选答句①"第一章 行列式"和候选答句②"《线性代数》的出版年是2009年.",候选答句①为更匹配的答句,但①中完全不存在和问句的相同词汇,反而是问句和候选答句②的相同词汇之一"线性代数"在粗粒度下反映了该问句的话题范围.因此,在通过训练用的问句和正、负匹配集分词后得到类别后,我们仅去掉"是"、"的"等停用词,常用标点,以及该话题类别的负域词汇(即标记为【0‖0‖1】的词汇),用剩余的词汇集合和相应的标记表示一个【问句-答案】范畴的上近似.

有2种特殊的训练情况:1)给定问句和候选答句,候选答句全部为正匹配句,则负匹配句集合为空集∅,此时训练得到的粗糙集问答知识则不包含【0‖1‖1】,【1‖0‖1】和【1‖1‖1】标记的词;2)给定问句和候选答句,候选答句全部为负匹配句,但由于我们的训练目标是挖掘出问句和对应答句的话题和答案信息,因此这类训练样本需要在训练前剔除.

表2给出了一个问句及其候选句集的示例,问答句集选自NLPCC-ICCPOL2016[30]国际会议上基于文档的开放领域问答系统评测比赛的公开训练数据集.所有的句子都经过了分词预处理,用"\\"标记切分位置.本例中负匹配句只列出前2条,其他负匹配句省略.带有上标的词为最终在问答范畴知识中出现的词,用右上标标注了序号.表3给出了表2示例的粗糙知识表达,其中,标记为【0‖0‖1】的词为负域词,需要被约简掉,因此未在表3中列出.

Tabel 2　Question and Its Positive/Negative Items

表2　问句和正、负匹配句示例

Category	Word Segments and Labels
ques	"黄山"(9)\\烟(5)\\打破(6)\\了\\原本(3)\\哪(1)\\两(11)\\个(4)\\地方(2)\\高档(14)\\烟(5)\\称霸(7)\\的\\局面(8)\\?
Set_p	一举\\打破\\了\\"\\沪(10)\\、\\云(15)\\"\\高档(14)\\烟(5)\\一统天下(13)\\的\\局面(8)\\.
Set_n	① 黄山(9)\\,\\是\\香烟\\的\\一\\个(4)\\品牌\\. ② "\\黄山(9)\\"\\烟(5)\\是\\安徽中烟工业公司\\蚌埠卷烟厂\\1958\\年\\开发\\的\\.

Tabel 3 Rough Set QA Knowledge Based on Table 2

表 3 由表 2 得到的粗糙集问答知识

Number	Word	Tag
1	哪	【1‖0‖0】
2	地方	【1‖0‖0】
3	原本	【1‖0‖0】
4	个	【1‖0‖1】
5	烟	【1‖1‖1】
6	打破	【1‖1‖0】
7	称霸	【1‖0‖0】
8	局面	【1‖1‖0】
9	黄山	【1‖0‖1】
10	沪	【0‖1‖0】
11	两	【1‖0‖0】
12	一举	【0‖1‖0】
13	一统天下	【0‖1‖0】
14	高档	【1‖1‖0】
15	云	【0‖1‖0】

4 基于粗糙集问答知识的问答检索

在训练得到一系列粗糙集问答知识后，当问答系统获取到新的问句和候选答句后，其问句和答句的匹配度 QAM：

$$QAM=\alpha\times SSim+\beta\times KMatch, \tag{7}$$

其中，$SSim$ 为问句和答句的语句形式相似度，可以用传统的向量化模型得到句子向量后用余弦相似度计算；$KMatch$ 为对粗糙集问答知识库中的问答知识的最高匹配程度；α 和 β 分别为形式相似度和知识匹配度的权重系数. 计算 $KMatch$ 的过程如算法 1 和算法 2 所示：

算法 1.

输入：问句、候选答句；

输出：所有候选答句的假定范畴最大相似度.

1) 对问句和所有的候选答句分词.

2) 从候选答句中选择一个句子，先假定其为正匹配句，其他句子为负匹配句. 按照粗糙集问答知识的训练过程，将全部的词汇进行假定知识标记，去除标记为【0‖0‖1】的词汇. 此时得到一个假定正句下的假定问答知识范畴.

3) 利用算法 2 中的计算过程计算假定问答范畴和问答知识库中的最大相似度.

4) 重复步骤 2 和步骤 3，直至遍历得到所有候选答句的假定范畴最大相似度.

算法 2.

输入：知识库、算法 1 步骤 2 得到的假定问答知识范畴；

输出：候选答句相对假定平均范畴的最大相似度.

1) 从问答知识库中选择一个粗糙集问答知识.

2) 比对假定范畴知识和该粗糙集问答知识的全部词库，若词条和标记均相同，则对应标记数目加 1. 按照(【0‖1‖0】，【0‖1‖1】，【1‖0‖0】，【1‖0‖1】，【1‖1‖0】，【1‖1‖1】)的标记顺序得到一个维度为 6 的计数向量 $\boldsymbol{A}$.

3) 判定 $\boldsymbol{A}$ 中计数总数. 若小于阈值 C，则返回 $KMatch=0$ 并执行步骤 6，否则执行步骤 4.

4) 判定所得元素中是否只包含问句相关元素或只包含答案相关元素，即先计算标记【$ques$ ‖ Set_p ‖ Set_n】的 $ques$，Set_p 各位的总计数(对应记为 X 和 Y)，若 X 和 Y 当中有一个为 0，则返回 0 并执行步骤 6，否则执行步骤 5.

5) 将计数向量 $\boldsymbol{A}$ 和假定平均知识范畴向量 $\boldsymbol{K}$ 归一化后用余弦相似度公式计算两者相似度并返回，并执行步骤 6(归一化公式和余弦相似度公式采用 2.2 节的式(5)和式(6)).

6) 重复步骤 1～5，遍历知识库，最后返回该候选答句相对假定平均范畴 $\boldsymbol{K}$ 的最大相似度.

由算法 1 和算法 2 可知，QAM 的计算过程中，过滤阈值 C，假定平均知识范畴向量 $\boldsymbol{K}$，形式相似度权重系数 α 和知识匹配度权重系数 β 均会影响最终计算出的 QAM 分数. 其中，过滤阈值 C 的主要作用是初步滤掉与候选句匹配可能性极低的问答知识，以提升系统对问答知识库的遍历速度. C 应至少为 1，即候选句与被匹配的问答知识应至少有一个元素相同.

$SSim$ 是由传统文本模型得到的向量余弦相似度. $KMatch$ 是归一化后的知识向量和假定平均范畴 $\boldsymbol{K}$ 的余弦相似度，是不同标记词语的分布相似度，归一化后得到的余弦相似度与 $SSim$ 属于同一个数量级，因此可以加权相加. 但由于不同的问答系统的应用背景不同，对应的语料库的特点也不同，因此 $SSim$ 的取值权重 α 和 $KMatch$ 的取值权重 β 也应该随着语料特点而改变. 本文实验中的开放领域问答语料涉及到的知识领域较广且多为书面语，文本形式相似度和知识相似度都是重要的元素，因此 α 和 β 暂定为 1，后续我们会研究如何根据不同的语料库特征设置更合适的 α 和 β.

5　实验和结果

实验采用 NLPCC-ICCPOL2016 国际会议上基于文档的开放领域中文问答系统评测比赛的公开数据集和评测工具.该评测的公开数据集包含训练集和测试集 2 个部分,其中训练集包含 8 772 条问句,对应问句总共有 181 882 条候选答句;测试集总共有 5 997 条问句,对应问句总共有 122 531 条候选答句.实验采用与该评测比赛相同的 MRR(mean reciprocal rank)和 MAP(mean average precision) 2 个评测指标.MRR 的计算公式为

$$MRR = \frac{1}{Q}\sum_{i=1}^{|Q|}\frac{1}{rank_i}, \tag{8}$$

其中,$|Q|$为问句总数,$rank_i$ 表示第 i 个问句对应的候选答句中第 1 个正确答句的排名.若没有正确答案则令$\frac{1}{rank_i}=0$.

MAP 的计算公式为

$$MAP = \frac{1}{Q}\sum_{i=1}^{|Q|}AveP(C_i,A_i)\ , \tag{9}$$

$$AveP(C_i,A_i) = \frac{\sum_{k=1}^{n}(P(k)\times rel(k))}{\min(m,n)}, \tag{10}$$

其中,m 是正确的正匹配数目,n 是系统给出的正匹配句数目,若 $\min(m,n)=0$,则令 $AveP(C_i,A_i)=0$.如果系统给出的排名为 k 的候选句是正确的正匹配句,则 $rel(k)=1$,否则 $rel(k)=0$.$P(k)$为系统给出的前 k 个候选句中正确的正匹配句所占的比例.

实验中,用传统的 LSI 模型向量化后的余弦相似度检索方法作为对比实验中的 baseline1(记为 LSICosine),用传统的 LDA 模型向量化后的余弦相似度检索方法作为对比实验中的 baseline2(记为 LDACosine),用 Doc2Vec 模型向量化后的余弦相似度检索方法作为对比实验中的 baseline3(记为 D2VCosine).中文的分词预处理使用了 ICTCLAS(NLPIR)工具包[33],所有 baseline 实验均使用 Gensim 工具包实现[34].通过 baseline 实验确定,获取最高 MAP,MRR 值时的 LSI,LDA 维度为 1 400 左右,Doc2Vec 维度为 1 000 左右,3 种模型对应的 MAP,MRR 见表 4.由于 LDA 模型在本语料上的表现暂时最好,因此将 LDA 模型得到的分数用于本文提出方法中的 QAM 的 $SSim$ 部分的分数.

首先将设置 QAM 的过滤阈值 $C=2$,即只要满足有一个问句相关和一个答案相关的元素则进入后续计算步骤;α 和 β 均设置为 1.QAM 中最关键的步骤在于假定平均知识范畴向量 $\boldsymbol{K}$ 的取值.一个直觉上的取值倾向是,令问句和正匹配句中的相同词汇尽可能多,其次是问句词和答案词,最后是其他类别的词,即:标记为【1|||1|||0】的词应该尽可能地占最大比例,其次是【1|||0|||0】和【0|||1|||0】,最后是其他标记的词.实验过程中,当 $\boldsymbol{K}=(2,1,2,1,4,1)$(归一化后为(0.384 9, 0.192 5, 0.384 9, 0.192 5, 0.769 8, 0.192 5))时,本文提出的方法所得到的(记为 RKMethod)的 MAP,MRR 值如表 4 所示:

Tabel 4　Experiment Result when $K=(2,1,2,1,4,1)$, $C=2$

表 4　$K=(2,1,2,1,4,1)$,$C=2$ 时实验结果

Method	*MAP*	*MRR*
LSICosine	0.537 2	0.537 6
LDACosine	0.638 6	0.639 2
D2VCosine	0.329 0	0.330 0
RKMethod	**0.644 9**	**0.645 7**

实验结果证明在 MAP 和 MRR 两个评测指标上,基于粗糙集知识的方法比 3 个 baseline 方法均有所提高,从而证明了该方法的有效性.在本文实验所用语料上,3 个 baseline 方法中 Doc2Vec 并未取得预期理想的效果,其主要原因是 LDA 模型体现的是文本的话题信息,Doc2Vec 模型体现的是词及其所在的上下文信息,而在本文的问答系统和所用的问答语料中,话题信息对后续分析问答话题的帮助更大.

若将 $\boldsymbol{K}$ 先固定为(1,1,1,1,1,1),而后以 0.5 为步长逐步增加每个位置上的元素的权重(例如,(1,1,1,1,1,1),(1.5,1,1,1,1,1),…,(6,1,1,1,1,1),其归一化后各个元素的权重变化见图 7),依次测试各个位置上权重增加后对最终实验结果的影响,其 MAP 和 MRR 的变化分别如图 1~6 所示.

由图 1~6 可以看出,【1|||0|||0】标记词和【1|||1|||0】标记词在相对权重增加后,MAP 和 MRR 值均有所提升,而其他的词汇总体上对结果是下降的影响.其物理含义是:疑问词和【问句-正匹配句】中的共同话题词越多,越能够确定所需要的问答知识.当标记为【1|||1|||0】的词汇在向量归一化后的权值超过 0.7 时,MAP 和 MRR 值可以超过 0.65(如图 5 和图 7 所示).可见该实验结果符合认知.

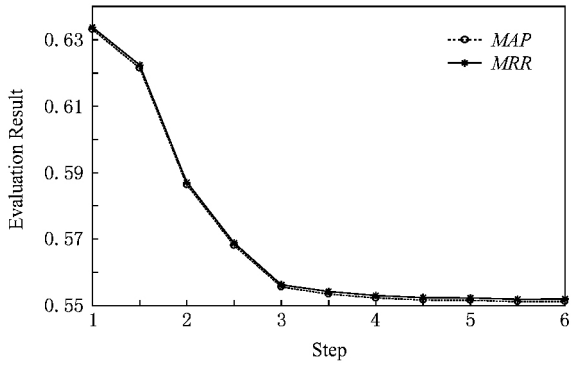

Fig. 1 Result of increasing【0|||1|||0】weight

图1 【0|||1|||0】词汇增加权重后结果

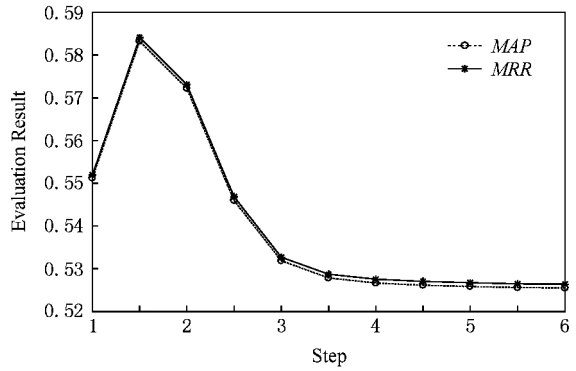

Fig. 2 Result of increasing【0|||1|||1】weight

图2 【0|||1|||1】词汇增加权重后结果

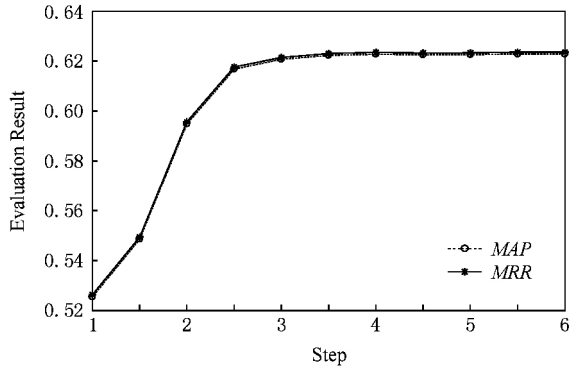

Fig. 3 Result of increasing【1|||0|||0】weight

图3 【1|||0|||0】词汇增加权重后结果

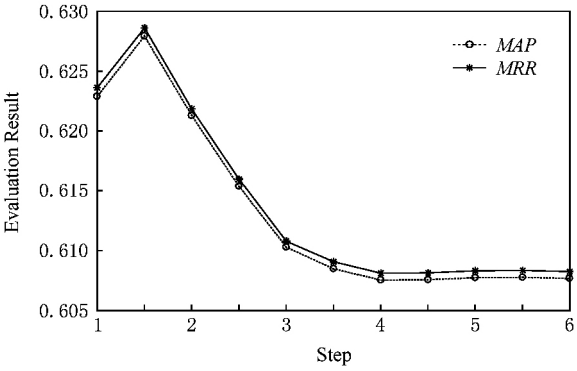

Fig. 4 Result of increasing【1|||0|||1】weight

图4 【1|||0|||1】词汇增加权重后结果

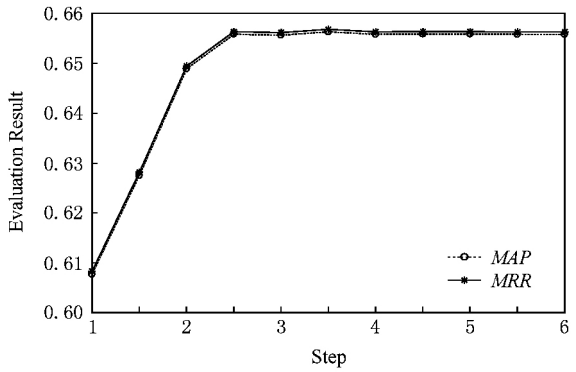

Fig. 5 Result of increasing【1|||1|||0】weight

图5 【1|||1|||0】词汇增加权重后结果

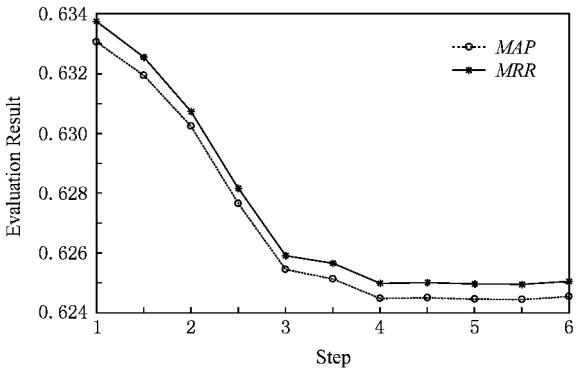

Fig. 6 Result of increasing【1|||1|||1】weight

图6 【1|||1|||1】词汇增加权重后结果

当各个标记的词的权重增加到一定程度后，MAP 和 MRR 值逐渐趋于稳定. 这是由于训练得到的粗糙集问答知识中的词汇标记分布是在一定范围内的，并不是无限数量，因而计数向量 A 的各个位置上的值也是落在一定范围内，所以后续若继续增加假定平均知识范畴向量 K 中的单独某个元素的权重，A 和 K 的余弦夹角仍不会发生大幅度变化.

若固定 $K=(2,1,2,1,4,1)$，令过滤阈值 C 从1开始以1为步长逐步增加，MAP 和 MRR 以及遍历1次测试数据集所需的时间如图8所示.

由图8可知，当过滤阈值为1和2时，实验得到的 MAP 和 MRR 值为最高，但相对耗时也比较长，遍历一次测试数据集所需时间为1 400～1 600 s. 但当阈值超过3后，耗时大大减少，仅需700 s左右，约为阈值为1和2时耗时的一半. 随着阈值的增加，所用耗时不再发生大幅度变化，当阈值超过8以后

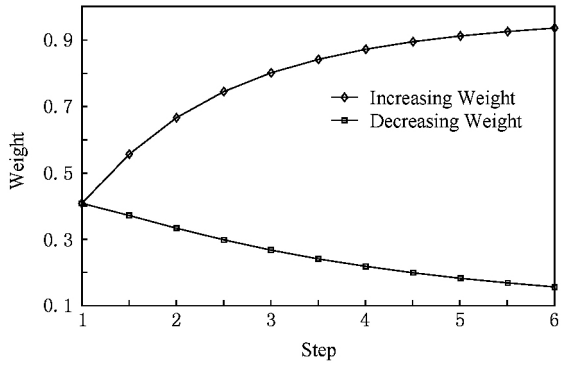

Fig. 7　Weight changing with increasing/decreasing weight

图 7　随步长增加的元素权重变化

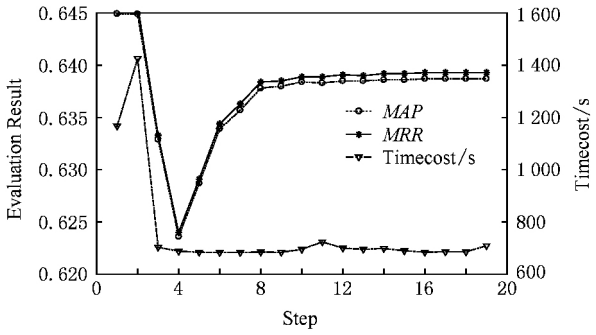

Fig. 8　Result of increasing C

图 8　逐步增加 C 的实验结果

MAP 和 *MRR* 值也不再大幅波动，*MAP* 值保持在[0.637 8，0.638 7]区间内，*MRR* 的值保持在[0.638 4，0.639 3]区间内，2 个区间内的最高值仅仅比 3 个 baseline 实验所得的最高 *MAP* 和 *MRR* 值高 0.0001，但因为增加知识匹配过程因而耗时要高于 baseline 实验. 这是因为过高的过滤阈值使得候选句匹配知识库中条目的概率大大减少，最终使得本文的方法退化为单独的向量余弦相似度方法. 因此，从耗时和效果 2 个方面综合考虑后，过滤阈值选 1 或 2 最为合适.

6　总　　结

本文提出了一种基于粗糙集知识发现的中文问答检索方法，利用粗糙集的属性约简方法和上近似概念从已标注的问答语料库中发现并表示知识，利用获得的粗糙集问答知识结合传统的句子相似度方法对问句和候选句进行匹配度计算. 基于粗糙集问答知识的方法的优势在于，其上近似的知识表达方式既可以存储【问句-答案】知识，也可以存储语言表达知识，还可以从多个正、负匹配句中挖掘出潜在的问答句语言表达信息. 实验结果表明：相对传统的问答检索方法，该方法在 *MAP* 和 *MRR* 两个评测指标上均有提升.

在理论研究方面，该方法还有 4 个方面的提升空间：

1）本文实验中的假定平均知识范畴向量 K 为人工设置，而能否从训练集中挖掘出最有效，适应最广的假定平均知识范畴向量 K，或者能否根据实际情况动态调整最合适的假定平均知识范畴向量 K 是本文方法的一个待解决问题.

2）在知识匹配计算过程中，本文方法是词语和标记同时命中时计数一次，而在某些情况下训练集中的标记分布并不理想，会导致粗糙集问答知识表达的偏差，例如标记【1 ||| 1 ||| 1】和标记【1 ||| 0 ||| 1】，【1 ||| 1 ||| 0】之间的偏差，因而如何确定有效的计数方式也是后续的研究工作之一.

3）在不同应用背景下的问答系统中如何确定最优的形式相似度权重系数 α 和知识匹配度权重系数 β 也是后续的研究方向之一.

4）本实验中的数据集为单个问句、正匹配句集、负匹配句集的形式，因而挖掘出的粗糙集问答知识仅能从答案中发掘出潜在的答句表达. 后续研究中可以尝试将数据集扩展成同义句问句集、正匹配句集和负匹配句集的形式，通过增加同义句问句以挖掘出更多潜在的问句表达，以应对实际的中文问答系统和问答检索的对语言灵活性的需求.

在实际应用方面，由于在部分应用场景下问答系统需要返回一个或少量数目的候选答句，因此，依据本文方法获得候选答句的匹配度分数后，如何选定一个合适的临界值也是今后的研究项目之一.

参　考　文　献

[1] Yang Ye, Jiang Peilin, Ren Fuji, et al. Classic Chinese automatic question answering system based on pragmatics information [C] //Proc of the 7th Mexican Int Conf on Artificial Intelligence. Los Alamitos, CA: IEEE Computer Society, 2008: 58-64

[2] Hu Haiqing, Ren Fuji, Kuroiwa S, et al. A question answering system on special domain and the implementation of speech interface [J]. Computational Linguistics & Intelligent Text Processing, 2006, 3878(3): 458-469

[3] Zadeh L. Fuzzy sets [J]. Information & Control, 1965, 8(3): 338-353

[4] Pawlak Z. Rough sets [J]. International Journal of Parallel Programming, 1982, 11(5): 341-356

[5] Zhang Ling, Zhang Bo. The quotient space theory of problem solving [C] //Proc of the 9th Int Conf on Rough Sets, Fuzzy Sets, Data Mining, and Granular Computing. Berlin: Springer, 2003: 11-15

[6] Dubois D, Prade H. Rough fuzzy sets & fuzzy rough sets [J]. International Journal of General Systems, 1990, 17(2-3): 191-209

[7] Hu Qinghua, Yu Daren, Liu Jinfu, et al. Neighborhood rough set based heterogeneous feature subset selection [J]. Information Sciences, 2008, 178(18): 3577-3594

[8] Ziarko W. Variable precision rough set model [J]. Journal of Computer & System Sciences, 1993, 46(1): 39-59

[9] Miao Duoqian, Li Daoguo. Rough Set: Theory, Algorithms and Applications [M]. Beijing: Tsinghua University Press, 2008 (in Chinese)
(苗夺谦, 李道国. 粗糙集理论、算法与应用[M]. 北京: 清华大学出版社, 2008)

[10] Wang Guoyin, Yao Yiyu, Yu Hong. A survey on rough set theory and applications [J]. Chinese Journal of Computers, 2009, 32(7): 1229-1246 (in Chinese)
(王国胤, 姚一豫, 于洪. 粗糙集理论与应用研究综述[J]. 计算机学报, 2009, 32(7): 1229-1246)

[11] Zhang Zhifei, Miao Duoqian, Nie Jianyun, et al. Sentiment uncertainty measure and classification of negative sentences [J]. Journal of Computer Research & Development, 2015, 52(8): 1806-1816 (in Chinese)
(张志飞, 苗夺谦, 聂建云, 等. 否定句的情感不确定性度量及分类[J]. 计算机研究与发展, 2015, 52(8): 1806-1816)

[12] Lang Guangming, Miao Duoqian, Yang Tian, et al. Knowledge reduction of dynamic covering decision information systems when varying covering cardinalities [J]. Information Sciences, 2016, 346/347: 236-260

[13] Wang Guoyin, Zhang Qinghua, Hu Jun. An overview of granular computing [J]. CAAI Trans on Intelligent Systems, 2007, 2(6): 8-26 (in Chinese)
(王国胤, 张清华, 胡军. 粒计算研究综述[J]. 智能系统学报, 2007, 2(6): 8-26)

[14] Salton G, Wong A, Yang Chungshu. A vector space model for automatic indexing [J]. Communications of the ACM, 1975, 18(11): 273-280

[15] Aizawa A. An information-theoretic perspective of TF-IDF measures [J]. Information Processing & Management, 2003, 39(1): 45-65

[16] Blei D, Ng A, Jordan M. Latent dirichlet allocation [J]. Journal of Machine Learning Research, 2003, 3(1): 993-1022

[17] Hofmann T. Probabilistic latent semantic indexing [C] // Proc of the 22nd Int ACM SIGIR Conf on Research and Development in Information Retrieval. New York: ACM, 1999: 50-57

[18] Mikolov T, Sutskever I, Chen Kai, et al. Distributed representations of words and phrases and their compositionality [C] //Proc of the 27th Annual Conf on Neural Information Processing Systems. New York: Curran Associates Inc, 2013: 3111-3119

[19] Le Q, Mikolov T. Distributed representations of sentences and documents [J]. Computer Science, 2014, 4: 1188-1196

[20] Aliguliyev R. A new sentence similarity measure and sentence based extractive technique for automatic text summarization [J]. Expert Systems with Applications, 2009, 36(4): 7764-7772

[21] Rice J, Belland R. A simulation study of moss floras using Jaccard's coefficient of similarity [J]. Journal of Biogeography, 1982, 9(5): 411-419

[22] Ghanbari M, Tahery R. Similarity coefficient [J]. Fluid Phase Equilibria, 2011, 310(1/2): 82-89

[23] Sun Ang, Jiang Minghu, He Yifan, et al. Chinese question answering based on syntax analysis and answer classification [J]. Acta Electronica Sinica, 2008, 36(5): 833-839 (in Chinese)
(孙昂, 江铭虎, 贺一帆, 等. 基于句法分析和答案分类的中文问答系统[J]. 电子学报, 2008, 36(5): 833-839)

[24] Wang Zhiqiang, Li Ru, Liang Jiye, et al. Research on question answering for reading comprehension based on Chinese discourse frame semantic parsing [J]. Chinese Journal of Computers, 2016, 39(4): 795-807 (in Chinese)
(王智强, 李茹, 梁吉业, 等. 基于汉语篇章框架语义分析的阅读理解问答研究[J]. 计算机学报, 2016, 39(4): 795-807)

[25] Archana S M, Vahab N, Thankappan R, et al. A rule based question answering system in Malayalam corpus using vibhakthi and POS tag analysis [J]. Procedia Technology, 2016, 24: 1534-1541

[26] Ray S K, Shaalan K. A review and future perspectives of Arabic question answering systems [J]. IEEE Trans on Knowledge and Data Engineering, 2016, 28(12): 3169-3190

[27] Dwivedi S K, Singh V. Research and reviews in question answering system [J]. Procedia Technology, 2013, 10: 417-424

[28] Voorhees E M. The TREC question answering track [J]. Natural Language Engineering, 2001, 7(4): 361-378

[29] Sasaki Y. Question answering as abduction: A feasibility study at NTCIR QAC1 [J]. IEICE Trans on Information & Systems, 2003, 86(9): 1669-1676

[30] Duan Nan. Overview of the NLPCC-ICCPOL 2016 shared task: Open domain Chinese question answering [C] //Proc of the ICCPOL&NLPCC 2016: Natural Language Understanding and Intelligent Applications. Cham, Switzerland: Springer International Publishing, 2016: 942-948

[31] Pawlak Z. Rough Sets: Theoretical Aspects of Reasoning about Data [M]. Dordrecht, Netherlands: Kluwer Academic Publishers, 1991

[32] Nengsih W. A comparative study on cosine similarity algorithm and vector space model algorithm on document searching [J]. Advanced Science Letters, 2015, 21(10): 3321-3323

[33] Zhang Huaping, Yu Hongkui, Xiong Deyi, et al. HHMM-based Chinese lexical analyzer ICTCLAS [C] //Proc of the 2nd SIGHAN Workshop on Chinese Language Processing. Stroudsburg, PA: Association for Computational Linguistics, 2003: 184-187

[34] Rehurek R, Sojka P. Software framework for topic modelling with large corpora [C] //Proc of LREC 2010 Workshop on New Challenges for NLP Frameworks. Valletta, Malta: University of Malta Press, 2010: 45-50

Han Zhao, born in 1990. PhD candidate. Student member of CCF. Her main research interests include natural language processing, affective computing, dialogue system and information retrieval.

Miao Duoqian, born in 1964. Professor, PhD supervisor. Senior member of CCF. His main research interests include rough set, granular computing, data mining.

Ren Fuji, born in 1959. Professor, PhD supervisor. His main research interests include natural language processing, artificial intelligent, humanoid robot and affective computing (ren@is.tokushima-u.ac.jp).

Zhang Hongyun, born in 1972. Associate professor, PhD supervisor. Her main research interests include principal curve, granular computing, data mining (zhanghongyun@tongji.edu.cn).