基于粗糙集理论的中文知识问答的知识谓词分析

韩 朝[1,3] 苗夺谦[1,2] 任福继[2,3]
(同济大学电子与信息工程学院 上海 201804)[1]
(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)[2] (德岛大学工学部 德岛 7708506)[3]

摘 要 在基于知识的问答系统中,问句中的知识谓词信息分析结果将会对知识元组的整体匹配效果产生影响。中文短问句中的知识谓词的信息表达方式存在着不确定性,这些不确定性的表达增加了知识谓词分析的难度。从粗糙集理论的角度,提出了一种问句中的知识谓词的分析方法,对问句中的知识谓词的弱相关表达进行约简,使问句中与知识谓词强相关的表达词能更有效地与知识元组中的知识谓词匹配,进而提高系统对知识谓词的整体分析能力。实验结果验证了新方法的有效性。

关键词 粗糙集,问答系统,知识问答,信息检索,短文本相似度
中图法分类号 TP391 **文献标识码** A **DOI** 10.11896/j.issn.1002-137X.2018.06.032

Rough Set Based Knowledge Predicate Analysis of Chinese Knowledge Based Question Answering

HAN Zhao[1,3] MIAO Duo-qian[1,2] REN Fu-ji[2,3]
(College of Electronics and Information Engineering, Tongji University, Shanghai 201804, China)[1]
(Key Laboratory of Embedded System and Service Computing, Ministry of Education, Tongji University, Shanghai 201804, China)[2]
(Faculty of Engineering, Tokushima University, Tokushima 7708506, Japan)[3]

Abstract In knowledge based question answering system, the performance of knowledge predicate analysis can affect the overall match result of knowledge triple. The knowledge predicate analysis of Chinese short question is difficult because of the uncertainty of Chinese knowledge predicate representation. Based on the rough set theory, a new definition of knowledge predicate analysis of knowledge based question snswering was given, and a new method was proposed to analyze the knowledge predicate of question. It can reduce the words which are weakly related with the knowledge predicate, and then the words which are more related with knowledge predicate representation will be used to match the knowledge triples to improve the overall performance of system. The experiment results verify the validity of the method.

Keywords Rough set, Question answering system, Knowledge based question answering, Information retrieval, Short text similarity

1 引言

基于知识的问答系统是当前自然语言处理领域的研究热点之一,其主要原理是在已有的知识库中匹配与问句关联度最高的知识元组,然后根据自然语言生成技术构建并返回答句。知识库中的知识元组通常采用一种通用描述方法 RDF(Resource Description Framework)[1],例如,[《犯罪学》,出版时间,2008年3月]和[《犯罪学》,价格,38元]这两个知识元组是以[subject,predicate,object]为结构的 RDF 表达[1],与其他相同结构表达的知识元组共同存储在知识库中。当系统接收到"《犯罪学》这本书是什么时候出版的?"这一问句后,系统将分别分析出问句中包含知识主语(knowledge subject)和知识谓词(knowledge predicate)的词,即分别对应"犯罪学"和"什么时候出版";接着,将其与知识库中的知识元组的对应位置匹配,选取匹配度最高的知识元组,即[《犯罪学》,出版时间,2008年3月];再基于知识元组中作为答案的部分返回答句,如直接返回"2008年3月"或构成完整答句"出版时间为2008年3月",返回最终的答句。

问句和知识谓词的匹配是问句分析的重要环节之一。由于中文语言的特性,中文的知识谓词的表达方式存在着大量的不确定性,例如,"生日"这一知识谓词在问句中既可以用"xx/的/生日/是/什么/时候"表达,也可以用"xx/是/什么/时

到稿日期:2017-12-18 返修日期:2018-01-29 本文受国家自然科学基金项目(61273304,61673301,61573255),高校学校博士学科点专项基金项目(20130072130004)资助。

韩 朝(1990-),女,博士生,CCF 会员,主要研究方向为粗糙集、问答系统,E-mail:1990hanzhao@tongji.edu.cn;**苗夺谦**(1964-),男,博士,教授,CCF 会员,主要研究方向为粒计算、粗糙集,E-mail:dqmiao@tongji.edu.cn(通信作者);**任福继**(1959-),男,博士,教授,主要研究方向为情感计算、对话机器人,E-mail:ren@is.tokushima-u.ac.jp。

候/出生/的”“xx/的/出生年月/是/什么”“xx/的/生日/是/哪/年/哪/月”等多种其他方式表达(例句中的“/”号为分词位置)。在不同的表达方式下,问句中的知识谓词表达项和知识库中的知识元组谓词项并不一定是完全相等的词语集合。因此,如何将不同表达的问句准确地匹配到对应的知识谓词是知识元祖匹配过程中的一个难点。

针对该问题,本文从粗糙集理论的角度出发,给出了基于知识问答系统的知识谓词分析问题在粗糙集理论下的定义,并提出了一种基于粗糙集属性约简理论的中文问句的知识谓词分析方法。将中文词看作知识表达时的属性,利用粗糙集的属性约简方法和边界域概念,从给定的已标注的问句-知识元组语料库中挖掘与知识表达强关联的属性词,通过约简弱关联的属性词,强化强关联属性词在后续匹配中的权重,进而改善匹配效果,提升系统性能。

2 相关工作

国内外关于基于知识库的问答系统已经有了大量的研究成果,主要集中在知识库构建、知识表达、问题分析等领域。针对不同的应用背景,问答系统侧重的功能有所不同,面向的语料特点不同,其相应的知识库的构成方式和问题分析的方法也不相同。文献[2-3]以商品或服务咨询问答系统为应用背景,针对不同领域的问题风格定义问句类别,对问句进行分类后,结合句法依存分析结果,采用不同的答案选取方法。文献[4]给出了一种语义模型,在已经获取可能的实体的前提下选取最相关的知识谓词。文献[5]采用了复述问句评分的模式进行问答检索。文献[6]从知识管理的角度对问答系统的知识自动化识别、知识创建等问题进行了研究。文献[7]对问答知识的表示学习方法进行了研究。文献[8]针对交互式问答系统的对话管理进行了研究。

粗糙集理论是一种处理不确定性信息的理论,已经在自然语言处理和知识发现等领域得到了广泛应用。在自然语言处理方面,文献[9]提出了基于粗糙集的短文本的相似度处理方法,利用规则获取方法从文本中挖掘同义词和多义词;文献[10]提出了基于粗糙集的 WEB 搜索方法;文献[11]利用粗糙集属性重要度的概念提出了一种文本分类的优化模型;文献[12]给出了基于粗糙集的文本聚类模型。在知识发现领域,文献[13]提出了基于粗糙集理论的社交网络的描述逻辑概念的挖掘方法;文献[14]提出了一种基于粗糙集理论的客户特征发现模型,帮助第三方支付平台挖掘潜在客户;文献[15]利用粗糙集理论从发动机活塞性能数据库中提取知识并将其用于智能摩托车系统设计;文献[16]利用粗糙集理论挖掘观点知识并将其用于舆情预测分析。

3 基本知识

定义 1[17] 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,$\forall X\subseteq U$ 是 U 上的一个子集(又称概念),R 是 U 上的一个等价目关系。则 x 关于 R 的下近似 $\underline{R}(X)$、上近似 $\overline{R}(X)$和边界域 $bn_R(X)$分别为:

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

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

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

下近似$\underline{R}(X)$表示根据等价关系判定肯定属于 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 $bn_R(X)$表示根据等价关系暂时无法判定是否肯定属于 X 的元素集合。

定义 2[17] 给定一个知识库 $K=\{U,S\}$和知识库上的一族等价关系 $R\subseteq S$,对于任意的 $r\in R$,若 r 满足:

$$IND(R-\{r\})=IND(R)$$

则 $IND(R)$为 R 上的不可分辨关系,称 r 为 R 中不必要的。

Pawlak 粗糙集理论提出,知识是人类或其他生物所具有的分类能力[18]。定义 2 给出了粗糙集理论中知识约简的基本思想,即若消除知识中的某部分内容后不会改变知识的分类结果,则这部分内容是冗余的,可以约简。

定义 3[18] 粗糙集理论中的决策表为如下的四元组:

$$DT=\{U,A=C\bigcup D,V,f\}$$

其中,U 为论域;A 为属性集合,分为条件属性 C 和决策属性 D 两部分;V 为值域;f 为映射函数 $U\times A\rightarrow V$。

与定义 2 中的粗糙集理论的基本约简思想类似,决策表的约简目标是:等价关系族 C 去除了冗余的知识后,得到的等价划分与通过 D 得到的等价划分相同。

4 基于粗糙集约简的分析方法

4.1 预处理

给定一个基于知识的问答系统的训练语料,其中包括若干问句及其对应的知识元组,每一个问句用 *ques* 表示,该问句对应的知识元组用 KT 表示,例如:

ques:“你/知道/城关镇/的/电话/区号/是/什么/吗”,
KT:[S:城关镇,P:电话/区号,O:854]。

在 KT 中,S 代表知识主语,P 代表知识谓词,O 代表知识宾语。对训练语料分词并去除问句中的问号等常用标点符号,来消除问号等出现频率较高的标点符号对概念分析的影响。对于语料库中的每一组问句-知识元组的配对,根据问句 *ques* 及其 KT 的分词集合,以全体词集为等价关系,分别求出每一个知识谓词的边界域词集,过程如下:

$$\underline{S}=S\cap ques$$

$$\underline{P}=P\cap ques$$

$$bn(P)=ques-\underline{S}-\underline{P}$$

此时得到的 $bn(P)$词集既包含了构建知识谓词概念语义的边界域词,又包含了构建知识主语概念语义的边界域词,此外还包含了构建完整句子所需要的句法词汇。当问句所采用的谓词概念的表达词汇与知识元组中的表达词汇重叠较少时,例如问句“xx/是/什么/时候/出生/的/”和知识谓词“生日”并没有词汇上的重叠,则边界域词集的部分词元素构成了与知识谓词同等的概念。如果将与知识谓词的概念表达不相关的词汇约简掉,那么剩下的边界域词将更接近对知识谓词的句式规则表达。

过于精简的约简结果可能会导致无法对多样性的句子表达进行识别。例如，“生日”这一谓词概念的多种表达多次出现在原始训练语料中，而其他谓词概念的训练用例较少，过度约简可能造成在训练语料上仅能将“生日”的某一种表达与其他知识谓词的表达区分出来，但当测试语料中出现“生日”的其他表达时，则会因为表达词被约简而出现判定失败的情况。因此，针对知识谓词规则的决策系统的属性约简原则如下：尽可能保留所有与语义表达区分度较高的词，并约简掉与语义表达弱相关的词。

4.2　知识谓词系统的规则获取

利用每一个问句-知识元组对得到的边界域词集和知识谓词构建一个面向知识谓词表达的决策表 $DT=\{U,A=C\cup D,V,f\}$。根据已有的粗糙集知识和决策表理论，提出以下属性约简定义。

定义 4　用 U 表示经过上述预处理后得到的边界域词-知识谓词的规则库，问句中包含的词为条件属性 C，知识谓词为决策属性 D。属性值域 $V=\{0,1\}$，0 表示该词在句中不存在，1 表示存在。用 U/D 表示决策属性集 D 在 U 上等价划分后得到的集合，$U_i\in U/D$ 是该集合中标记为 i 的一个元素集合。由 U_i 可以得到：

$$u_i^*=\bigcup\{a\mid a\in C,f(a,u\in U_i)=1\}$$

$$U^*=\bigcup\{u_i^*\mid U_i\in U/D\}$$

在由规则库 U 转化得到的规则库 U^* 中，若存在属性词 $a\in C$ 满足：

$$\frac{\sum_{U^*} f(a,u_i^*)}{|U^*|}\geqslant\beta,\ 0\leqslant\beta\leqslant 1$$

则属性词 a 可以约简。

根据定义 4，对 4.1 节预处理后得到的规则系统做以下处理。首先，对训练语料中同等概念的边界域词集做合集运算，然后得到新的决策表。合集运算过程中，更接近知识谓词语义构成的属性词的频度在整体语料中的频度被降低但仍被保留，语义表达弱相关的词的频度排名得到提升。随后，采用传统的频繁项挖掘算法，如 Apriori 算法，对合集处理后的决策表中的频繁-1 项词进行约简。这一步骤的原理是，在已经对强关联词进行了降频处理的前提下，或在没有重复知识谓词决策结果的前提下，仍有某个词 w 的频繁程度过高，则该词对知识谓词的区分能力并不高。

4.3　知识谓词分析

对训练语句进行上述处理后可以得到一个知识谓词的表达规则库。当句子中知识谓词的概念表达与谓语本身无词汇重叠时，这些规则库可以用于识别句子中表达的知识概念。此外，部分知识谓词本身即为其概念表达的组成。因此，将对分词后的测试语句进行两次相似度计算：第一次将测试语句与所有的谓词或谓词词组进行相似度计算，并返回相似度最高的项；第二次将测试语句与训练得到的所有规则词组进行相似度计算，返回相似度最高的项所对应的谓词项。相似度计算方法采用 TF·IDF 向量的余弦相似度方法。两次相似度计算得到的最高值对应的谓词即为分析结果。知识谓词分析的算法步骤如算法 1 所示。

算法 1　知识谓词分析算法

Input：测试语句

Output：测试语句的知识谓词

Step 1　对测试语句进行分词，得到分词序列，并用 TF·IDF 向量空间表示；

Step 2　遍历谓词库，采用余弦相似度方法计算测试语句与每个谓词/谓词词组的相似度，保存其中最大的相似度 sim1 及其对应的谓词词项 kp1；

Step 3　遍历规则库，采用余弦相似度方法计算测试语句与每个规则词组的相似度，保存其中最大的相似度 sim2 及其对应的规则词组所代表的谓词词项 kp2；

Step 4　比较 sim1 和 sim2，若 sim1≥sim2，则返回 kp1，否则返回 kp2。返回项即为所求的知识谓词。

5　实验分析

实验采用了在 NLPCC2016 评测比赛中 Huang 等[19-20]标注处理后的 KBQA 子任务语料[21]，该语料中的训练集包括 14609 条问句及其对应的知识谓词，测试集包括 9870 条问句及其对应的知识谓词。将本文方法与 3 种 baseline 方法进行了对比。方法 1 为问句直接与谓词词库进行相似度匹配；方法 2 为问句与谓词词库和原始训练问句进行匹配，即将问句看作完全未经过约简的谓词表达规则；方法 3 是在方法 2 的匹配库的基础上，对规则库使用 4.1 节提出的基本的预处理方法进行加工，再进行相似度匹配操作；方法 4 是在方法 3 的基础上进一步采用了 4.2 节提出的约简方法进行处理。对所有文本采用 tfidf 的文本向量空间表示，匹配操作采用向量的余弦相似度方法计算得分。采用 Scikit-learn[22] 工具包中的评测接口进行结果评测，实验结果如表 1 所列。

表 1　实验结果

Table 1　Experiment results

评价指标	方法 1	方法 2	方法 3	方法 4
Macro-precision	0.5974	0.5110	0.5889	0.5873
Micro-precision	0.5134	0.4593	0.5464	0.5441
Macro-recall	0.5106	0.4582	0.5444	0.5433
Micro-recall	0.5134	0.4593	0.5464	0.5441
Macro-F1	0.5168	0.4536	0.5336	0.5441
Micro-F1	0.5134	0.4593	0.5464	0.5301
Average Precision	0.61	0.52	0.59	0.60
Average Recall	0.51	0.46	0.55	0.54
Average F1	0.52	0.46	0.54	0.54

由表 1 可知，方法 2 在任意指标上的评价结果均低于方法 1，这是因为将训练问句直接作为句型模板后，与知识谓词表达不相关的词的权重干扰了相关词的权重，影响了其在全部匹配分值中的排名。方法 3 在方法 2 的基础上去掉了负域词，因而相对于方法 2，方法 3 大幅提升了系统的性能。而在进一步约简属性词之后，在其他指标相差不大的情况下，方法 4 相比方法 3 在 Macro-F1，Average Precision 指标上有所提升。可以看到，本文方法（即方法 3 和方法 4），在大部分指标上均优于方法 1 和方法 2，仅在 Average Precision 和 Macro-precision 这两个指标上，其评价指标数值略低于方法 1，这是

因为测试语料中问句的知识谓词的表达偏向直接，即问句和知识元组中的谓词重合部分较高。而在实际应用场合中，本文方法能够处理更多样的语言表达。

结束语 粗糙集理论是解决不确定性问题的重要理论之一。本文针对中文知识问答这一应用背景，从粗糙集理论的角度给出了中文知识问答系统的知识谓词分析问题的定义，提出了基于粗糙集理论的问句知识谓词分析方法。与传统方法相比，本文方法能在保证性能的基础上更好地识别多样化的知识谓词的表达。下一步将结合问题的实体(即知识主语)识别方法，对整个系统的框架进行整合和分析。

参 考 文 献

[1] MILLER E. An Introduction to the Resource Description Framework[J]. Journal of Library Administration, 2001, 34(3/4): 245-255.

[2] DU Z Y, YANG Y, HE L, et al. Question answering system of electric business field based on chinese knowledge map[J]. Computer Applications and Software, 2017, 34(5): 153-159. (in Chinese)
杜泽宇，杨燕，贺樑，等. 基于中文知识图谱的电商领域问答系统[J]. 计算机应用与软件，2017，34(5)：153-159.

[3] ZHANG K L, LI W G, WANG H L, et al. Ontology-based Question Answering System for Aviation Domain[J]. Journal of Chinese Information Processing, 2015, 29(4): 192-198. (in Chinese)
张克亮，李伟刚，王慧兰，等. 基于本体的航空领域问答系统[J]. 中文信息学报，2015，29(4)：192-198.

[4] XIE Z, ZENG Z, ZHOUG, et al. Topic enhanced deep structured semantic models for knowledge base question answering[J]. Science China(Information & Sciences), 2017, 60(11): 110103.

[5] ZHAN C D, LING Z H, DAI L R. Learning Word Embeddings for Paraphrase Scoring in Knowledge Base Based Question Answering[J]. Pattern Recognition and Artificial Intelligence, 2016, 29(9): 825-831. (in Chinese)
詹晨迪，凌震华，戴礼荣. 面向知识库问答中复述问句评分的词向量构建方法[J]. 模式识别与人工智能，2016，29(9)：825-831.

[6] ZENG S, WANG S, YUAN Y, et al. Towards Knowledge Automation: A Survey on Question Answering Systems[J]. Acta Automatica Sinica, 2017, 43(9): 1491-1508. (in Chinese)
曾帅，王帅，袁勇，等. 面向知识自动化的自动问答研究进展[J]. 自动化学报，2017，43(9)：1491-1508.

[7] LIU K, ZHANG Y Z, JI G L, et al. Representation Learning for Question Answering over Knowledge Base: An Overview[J]. Acta Automatica Sinica, 2016, 42(6): 807-818. (in Chinese)
刘康，张元哲，纪国良，等. 基于表示学习的知识库问答研究进展与展望[J]. 自动化学报，2016，42(6)：807-818.

[8] WANG Y, REN F J, QUAN C Q. Review of Dialogue Management Methods in Spoken Dialogue System[J]. Computer Science, 2015, 42(6): 1-7, 27. (in Chinese)
王玉，任福继，全昌勤. 口语对话系统中对话管理方法研究综述[J]. 计算机科学，2015，42(6)：1-7，27.

[9] ZHANG Z Z, MIAO D Q, YUE X D. Similarity measure for short texts using topic models and rough sets[J]. Journal of Computational Information Systems, 2013, 9(16): 6603-6611.

[10] YI G X, HU H P. A Web Search Result Clustering Based on Tolerance Rough Set[J]. Journal of Computer Research and Development, 2006, 43(2): 275-280. (in Chinese)
易高翔，胡和平. 一种基于容错粗糙集的 Web 搜索结果聚类方法[J]. 计算机研究与发展，2006，43(2)：275-280.

[11] LIU H, LIU D Y, PEI Z L, et al. A Feature Weighting Scheme for Text Categorization Based on Feature Importance[J]. Journal of Computer Research and Development, 2009, 46(10): 1693-1703. (in Chinese)
刘赫，刘大有，裴志利，等. 一种基于特征重要度的文本分类特征加权方法[J]. 计算机研究与发展，2009，46(10)：1693-1703.

[12] THANH N C, YAMADA K. Document Representation and Clustering with WordNet Based Similarity Rough Set Model[J]. International Journal of Computer Science Issues, 2011, 8(5): 1-8.

[13] FAN T F, LIAU C J. Rough set-based concept mining from social networks[C]// IEEE International Conference on Fuzzy Systems. IEEE, 2016: 663-670.

[14] CAO L, HUANG G, CHAI W. A knowledge discovery model for third-party payment networks based on rough set theory[J]. Journal of Intelligent & Fuzzy Systems, 2017, 33(1): 1-9.

[15] DAI R, DUAN X. Research on Knowledge Acquisition of Motorcycle Intelligent Design System Based on Rough Set[M]// Computer and Computing Technologies in Agriculture V. Springer Berlin Heidelberg, 2012: 16-27.

[16] CHEN X G, DUAN S, WANG L D. Research on trend prediction and evaluation of network public opinion[J]. Concurrency & Computation Practice & Experience, 2017, 29(4): e4212.

[17] 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.

[18] PAWLAK Z. Rough set approach to knowledge-based decision support[J]. European Journal of Operational Research, 1997, 99(1): 48-57.

[19] HUANG X, WEI B, ZHANG Y. Automatic Question-Answering Based on Wikipedia Data Extraction[C]// International Conference on Intelligent Systems and Knowledge Engineering. IEEE, 2016: 314-317.

[20] https://github.com/huangxiangzhou/NLPCC2016KBQA.

[21] DUAN N. Overview of the NLPCC-ICCPOL 2016 Shared Task: Open Domain Chinese Question Answering[C]// International Conference on Computer Processing of Oriental Languages. Springer International Publishing. 2016: 942-948.

[22] PEDREGOSA F, GRAMFORT A, MICHEL V, et al. Scikit-learn: Machine Learning in Python[J]. Journal of Machine Learning Research, 2011, 12(10): 2825-2830.