

DOI:10.3969/j.issn.1673-4785.
网络出版地址: <http://www.cnki.net/kcms/detail/23.1538.TP.20150326.1014.002.html>

基于基元和知网的问题相关度计算

曹礼园,李卫华

(广东工业大学 计算机学院, 广东 广州 510006)

摘 要:通过对可拓学的基元和复合元与知网的研究,利用 HowNet 中的词语相似度的计算方法改进词语相关度的计算方法,建立基元相关度计算方法,进而建立目标相关度和条件相关度的计算机方法,形成问题相关度的计算。改进了相关度算法,传统方法只通过上下位关系计算词语相关度,将知网描述的其他 15 个关系也考虑进来,并提出了负相关的概念,将相关度的取值范围定在 $[-1, 1]$, 对应了关联函数的值域。通过对基元中特征与特征的对应关系,计算出基元的相关度,进而计算问题的相关度。通过对数据库里的上千个矛盾问题进行测试,证实了这种方法可有效增加策略生成途径,使矛盾问题解决的概率大大提高。

关键词:可拓学;知网;相关度;基元;关联函数

中图分类号:TP311 **文献标志码:**A **文章编号:**1673-4785(2015)02-0234-06

中文引用格式:曹礼园,李卫华. 基于基元和知网的问题相关度计算[J]. 智能系统学报, 2015, 10(2): 234-239.

英文引用格式:CAO Liyuan, LI Weihua. Calculation of correlation problem based on basic element and HowNet[J]. CAAI Transactions on Intelligent Systems, 2015, 10(2): 234-239.

Calculation of correlation problem based on basic element and HowNet

CAO Liyuan, LI Weihua

(College of Computer, Guangdong University of Technology, Guangzhou 510006, China)

Abstract: Based on the basic-element and composite elements of extenics and HowNet research, the words similarity computing method in HowNet is used to improve the calculation method for the relationship of words. The basic-element correlation calculation method is built to establish a method for basic-element target correlation and basic-element conditions correlation. This is used to derive the method to calculate correlation of problem. There is improvement with the correlation algorithm, but the traditional method calculates the correlation of words only by hyponymy. In this paper, an additional fifteen relationships described in HowNet is taken into account. The concept of negative correlation is put forward and the range of correlation is in set $[-1, 1]$, which corresponds to the range of the correlation function. Through the corresponding relationship between the features in the basic element, the correlation between basic elements is calculated and the correlation of problems can be calculated. Based on the test of thousands of contradictory problems in the database, the results showed that this method can effectively increase the amount of strategy generating approaches and the probability of contradictory problems to be solved is greatly promoted.

Keywords: extenics; HowNet; correlation; basic element; dependent function

可拓学^[1]是一门中国原创的新学科,它以形式

化的模型,探讨事物拓展的可能性以及开拓创新的规律与方法,并用于解决矛盾问题。知网^[2](HowNet)是一个以汉语和英语的词语所代表的概念为描述对象,以揭示概念与概念之间以及概念所具有的

属性之间的关系为基本内容的常识知识库。知网的建网方式、知识获取和表达方式、事件概念分类方法和其自行设计的知识数据描述拓学中提出解决矛盾问题的策略生成知识需求。本文将可拓学和知网这个中国原创的理论和应用工具相结合,进一步提高了可拓学智能化水平,拓展了知网的应用。在文献[3]中,蔡文教授首先提出物元模型的概念,文献[4]提出了关联函数的构造方法。知网是我国著名机器翻译专家董振东先生创立的一个知识系统,它包含丰富的语义知识和世界知识。文献[5]通过知识描述结构和上下位关系计算义原相似度,文献[6]和[7]分别提出了一种利用相关度来计算词语相关度的算法。

1 传统的可拓策略生成系统分析问题的不足

传统的策略生成系统是先建立可拓模型,即建立问题 P 的核问题 P_0 的可拓模型为 $P_0 = g_0 \times l_0$, 其中, g_0 为问题的目标, l_0 为问题的条件。再确定问题的关联函数 K , 然后对问题 P_0 进行相容性分析, 若不相容(目标与条件有矛盾), 则对问题的目标或条件进行可拓分析, 得到问题的发散树或相关树, 通过对问题发散树或相关树的“叶”基元进行可拓变换, 生成候选策略集合, 最后对策略集合中的策略进行优度评价, 推荐优度较高的策略供决策者选择。这种分析问题的方法存在着一些不足之处。

首先, 这个解决问题的方法对每个问题都进行整个过程, 当遇到类似的问题时, 系统依然重复上次策略生成的过程, 增加了策略生成的时间。

其次, 这个过程无法对问题进行归类。也找不到问题之间关联性, 因而无法对问题进行归并, 进而找到通用解。同时导致策略的针对性不强。

可见, 在可拓策略生成系统中, 在原问题和核问题建模后, 增加另一个过程——进入矛盾问题库, 用问题相关度判断是否与已有矛盾问题相关(或相同)是非常重要的。如果问题与已在问题库的问题类似, 则可直接进入可拓策略库搜索选取相应的策略, 缩短问题解决的时间, 提高问题解决的概率。

2 词语相关度计算模型

2.1 词语相关概念

定义 1 词语相关度。给定 2 个词语 B_1 和 B_2 , 它

们之间的相关度通过函数 $\text{rele}(B_1, B_2): S \times S \rightarrow [-1, 1]$, 表示集合 S 中 2 个词语 B_1 和 B_2 的相关程度。

相关度函数的性质具有自反性和对称性, 即对于 $B_1 \in S, B_2 \in S$, 形式上有:

1) $\text{rele}(B_1, B_2) \in [-1, 1]$ 。相关度的计算值为 $[-1, 1]$ 中的一个实数。

2) 相关关系具有自反性, $\text{rele}(B_1, B_2) = \text{rele}(B_2, B_1)$ 。

3) $\text{rele}(B_1, B_2) = 1$, 当且仅当 $B_1 = B_2$, 即如果 2 个词汇是词语等价, 则相关度为 1。

4) $\text{rele}(B_1, B_2) = -1$, 当且仅当即存在对义或反义关系, 相关度为 -1 。

5) $\text{rele}(B_1, B_2) = 0$ 。如果 2 个基元没有任何共同特征, 那么其相关度为 0。

定义 2 正相关。给定 2 个基元 B_1 和 B_2 , 如 $\text{rele}(B_1, B_2) > 0$, 则称 B_1, B_2 正相关。

定义 3 负相关。给定 2 个基元 B_1 和 B_2 , 如 $\text{rele}(B_1, B_2) < 0$, 则称 B_1, B_2 负相关。

在知网中, 美丽与漂亮的相关度为 1, 美丽与丑陋的相关度 0.814 815, 在情感分析中, 美丽与丑陋是反义的。美丽和丑陋在句子中是可以替换的, 但这并不符合句子的原意。

负相关是指 2 个词语是对义或者反义, 或者两者在路径中存在对义或反义关系。如: 假设 O_1 与 O_3 是反义(对义)关系(根据《同义、反义以及对义组的形成》获得), 则 O_5, O_9 的相关度就是在路径中存在对义或反义关系, $\text{rele}(O_5, O_9)$ 为负数。

2.2 词语相关度计算

相似的词语相关的可能性也大些, 把词语的相似度和基于直接关系和语义关系挖掘规则获得的关系作为判断是否相关的依据而得出来的基本相关度部分按比例相加, 就得到了词语的相关度。

2.2.1 相似度部分 $\text{Sim}(W_1, W_2)$

2 个词语相似度计算, 直接采用文献[5]中方法计算。

2.2.2 基本相关度部分 $\text{Rele_prim}(W_1, W_2)$

对于 2 个汉语词语 W_1 和 W_2 , 如果 W_1 有 n 个义项(概念): $S_{11}, S_{12}, \dots, S_{1n}$, W_2 有 m 个义项(概念): $S_{21}, S_{22}, \dots, S_{2m}$ 规定, W_1 和 W_2 的相关度是各个概念的相关度绝对值最大的那个值, 也就是说:

$$\text{Rele_prim}(W_1, W_2) = \pm \max_{i=1..n, j=1..m} |\text{Rele_prim}(S_{1i}, S_{2j})|$$

这样,就把 2 个词语之间的相关度问题归结到了 2 个概念之间的相关度问题。

下面来计算 2 个义项 S_1 和 S_2 之间的相关度。

S_1 与 S_2 的知网形式化表示为

$$S_1 = (p_{11}, p_{12}, \cdots)$$
$$S_2 = (p_{21}, p_{22}, \cdots)$$

定义相对相关度 $\text{Re le_prim}_i(S_1, S_2)$, $\text{Re le_prim}_i(S_1, S_2)$ 代表概念相对于概念 S_1 在关系 i 上的相对相关度,定义

$$\text{Re le_prim}_i(S_1, S_2) = \begin{cases} 1, & (p_{1i} = p_{2j}, j = 1, 2, 3, \cdots, n) \\ 0, & \text{其他} \end{cases}$$

它包括直接关系和间接关系。

直接关系包括上下位关系之外、同义关系、反义

关系、对义关系、部件-整体关系、属性-宿主、材料-成品、施事/经验者/关系主体-事件关系、受事/内容/领属物等-事件关系、工具-事件关系、场所-事件、时间-事件关系、值-属性关系、实体-值关系、事件-角色关系、相关关系等 16 种关系。间接关系指根据语义关系挖掘规则获得的关系^[9]。

$$\text{Re le_prim}(S_1, S_2) =$$
$$[\sum w_i \text{Re le_prim}_i(S_1, S_2) +$$
$$\sum w_j \text{Re le_prim}_j(S_2, S_1)] / 2$$

式中: $\sum_1^m w_i = 1, \sum_1^n w_j = 1, i, j = \% , \# , ? , * , @ , \$, \&$, 其中间接关系挖掘规则总结如表 1。

表 1 知网间接关系挖掘规则

Table 1 HowNet indirect relation mining rules

直接关系	关系符	间接关系挖掘规则
部分-整体	%	以下位义原为主要特征的概念与原概念也构成部分与整体关系。
相关关系	#	1) 以相关义原为主要特性或第二特性的概念与原概念相关; 2) 具体相同相关元素的概念是相关的。
材料-成品关系	?	成品的解释义原与材料概念相关。
施事/经验者/ 关系主体-事件 或工作-时间关系	*	1) 以相同事件义原为施事,经验者的概念相互关联; 2) 以事件为角色的概念与原概念相关; 3) 与事件构成时间/地点-事件关系的概念与原概念; 4) 以事件义原为主要特征的概念与原概念。
时间地点- 事件关系	@	1) 与事件存在施事、受事关系的概念与原概念相关; 2) 以事件为角色的概念与原概念相关; 3) 以事件义原为主要特征的概念与原概念相关。
受事/内容/ 领属物-事件	\$	1) 与事件存在施事、受事关系的概念与原概念相关; 2) 以事件为角色的概念与原概念相关; 3) 以事件构成时间/地点-事件关系的概念与原概念相关; 4) 以事件义原为主要特征的概念与原概念相关。
宿主-属性	&	1) 以宿主为主要特征的概念与原概念之间同样存在宿主-属性关系; 2) 以该属性所对应属性值概念与原概念具有属性-值关系。

w_i 是 S_1 对 S_2 对应的关系的权值, w_j 是 S_2 对 S_1 对应的关系的权值,都是可调节的参数。

例如:

国庆:time|时间,day|日,@congratulate|祝贺,#country|国家

烟花炮竹:tool|用具,*whileAway|消闲,*congratulate|祝贺

“国庆”和“烟花炮竹”通过事件“祝贺”相关,计算如下:

在关系@上 S_1 与 S_2 相关,在关系*上 S_2 与 S_1 相关,所以 $\text{Re le_prim}(S_1, S_2) = (w_{@} * 1 + w_{*} * 1) / 2$

2.2.3 词语相关度 $\text{Re le}(W_1, W_2)$ 计算

$$\text{Re le}(W_1, W_2) =$$
$$\beta_1 \text{Sim}(S_1, S_2) + \beta_2 \text{Re le_prim}_i(S_1, S_2)$$

式中: $\beta_1 + \beta_2 = 1$ 。

β_1 和 β_2 是相似度 $\text{Sim}(S_1, S_2)$ 和基本相关度 $\text{Rele_prim}(S_1, S_2)$ 在总体相关度所占的比例。

3 基元及问题相关度计算

3.1 基元概念以及基元与基元的对应关系

3.1.1 基元的概念

定义 4 以物 O_m 为对象, C_m 为特征, O_m 关于 C_m 的量值 V_m 构成的有序三元组:

$$M = (O_m, C_m, V_m)$$

作为描述物的基本元, 称为一维物元, O_m 、 C_m 、 V_m 三者称为物元 M 的三要素, 其中 C_m 和 V_m 构成的二元组 (C_m, V_m) 称为物 O_m 的特征元。

为方便起见, 把物元的全体记为 $\mathfrak{E}(M)$, 物的全体记为 $\mathfrak{E}(O_m)$, 特征的全体记为 $\mathfrak{E}(C_m)$ 。关于特征 C_m 的取值范围记为 $V(C_m)$, 称为 C_m 的量域。

一物具有多个特征, 与一维物元相仿, 可以定义多维物元:

定义 5 物 O_m , n 个特征名 $c_{m1}, c_{m2}, \dots, c_{mn}$ 及 O_m 关于 $c_{mi}(i=1, 2, \dots, n)$ 对应的量值 $v_{mi}(i=1, 2, \dots, n)$ 所构成的阵列:

$$M = \begin{bmatrix} O_m & c_{m1} & v_{m1} \\ & c_{m2} & v_{m2} \\ & \vdots & \vdots \\ & c_{mn} & v_{mn} \end{bmatrix} = [O_m \quad C_m \quad V_m]$$

称为 n 维物元, 其中

$$C_m = \begin{bmatrix} c_{m1} \\ c_{m2} \\ \vdots \\ c_{mn} \end{bmatrix} \quad V_m = \begin{bmatrix} v_{m1} \\ v_{m2} \\ \vdots \\ v_{mn} \end{bmatrix}$$

3.1.2 基元与基元之间的对应关系

在基元中, 最重要是的特征与特征的对应关系。依靠下列方法建立基元的特征与另一基元的特征的对应关系:

- 1) 首先计算 2 个基元的所有特征两两之间的相关度;
- 2) 从所有的相关度值中选择最大的一个, 将这个相关度值对应的 2 个特征对应起来;
- 3) 从所有的相关度值中删去那些已经建立对应关系的特征的相关度值;
- 4) 重复上述 2) 和 3), 直到所有的相关度值都被删除;

5) 没有建立起对应关系的特征与空特征对应。

3.2 基元相关度计算

每一个特征就是一个词语, 特征对应特征, 相当于词语对应词语, 即用词语相关度计算方法计算对应特征相关度。将对应起来的特征分别计算关联度, 最后按比例相加, 就是基元相关, 计算公式为

$$\text{Rele}_B(B_1, B_2) = \sum_1^k R_i \text{Rele}(W_1, W_2)$$

R_i 是一个比例系数, 等于 $1/k$

3.3 问题相关度计算

问题建模是按以下步骤进行:

界定问题 1 的目标 g_1 与条件 l_1 和问题 2 的目标 g_2 与条件 l_2 , 并用基元表示分别建立问题的可拓模型 $P_1 = g_1 \times l_1$ 、 $P_2 = g_2 \times l_2$ 利用基元相关公式分别计算出目标基元和条件基元的相关度以后, 再计算问题相关度 $\text{Rele}_P(P_1, P_2)$, 计算公式为

$$\text{Rele}_P(P_1, P_2) =$$

$$\text{Rele}_{B_g}(B_{g_1}, B_{g_2}) \times \text{Rele}_{B_l}(B_{l_1}, B_{l_2})$$

4 案例分析

利用以下 6 个问题测试问题相关度问题:

P_1 曹冲称象;

P_2 用直尺测量一张纸的厚度;

P_3 把一根长为 2 m、宽为 1.2 m、高为 0.5 m 的桌子抬进高为 2 m、宽为 1 m 的门;

P_4 把一根长为 3 m、半径为 0.2 m 的竹杆抬进高为 2.5 m、宽为 2 m 的城门;

根据文献[1]中的方法, 建立问题的可拓模型 $P = G \times L$, 进而提取核问题:

$$P_0 = g_0 \times l_0$$

$$P_1 = g_1 * l_1 =$$

$$\left[\begin{matrix} \text{大象 } A, & \text{重量}, & x \\ & \text{可分性}, & -1 \end{matrix} \right] *$$

$$[\text{秤 } B, \text{ 称量}, [0, 200]]$$

$$P_2 = g_2 * l_2 =$$

$$\left[\begin{matrix} \text{纸张 } A, & \text{厚度}, & y \\ & \text{可叠加性}, & 1 \end{matrix} \right] *$$

$$[\text{直尺 } B, \text{ 测量}, [1, 20]]$$

$$P_3 = g_3 \times l_3 =$$

$$\left[\begin{matrix} \text{门} & \text{宽} & 1 \\ & \text{高} & 2 \end{matrix} \right] \times \left[\begin{matrix} \text{桌子} & \text{长} & 2 \\ & \text{宽} & 1.2 \\ & \text{高} & 0.5 \end{matrix} \right]$$

$$P_4 = g_4 \times l_4 =$$
$$\begin{bmatrix} \text{城宽} & 2.5 \\ \text{高} & 2 \end{bmatrix} \times \begin{bmatrix} \text{竹杆长} & 3 \\ \text{半径} & 0.2 \end{bmatrix}$$

问题相关度计算

$$\text{Rele_}P(P_1,P_2) = -0.394$$
$$\text{Rele_}P(P_1,P_3) = +0.204$$
$$\text{Rele_}P(P_1,P_4) = +0.075$$
$$\text{Rele_}P(P_2,P_3) = -0.076$$
$$\text{Rele_}P(P_2,P_4) = +0.204$$
$$\text{Rele_}P(P_3,P_4) = +0.654$$

在参数的选择上,由于一般认为%、#、?、*、@、\$、& 所代表的关系权重程度逐渐下降,即表 1 中的关系的重要程度逐步下降。经多次测试,最终设置参数如下: $w_1 = 0.25, w_2 = 0.2, w_3 = 0.2, w_4 = 0.15, w_5 = 0.10, w_6 = 0.05, w_7 = 0.05$, 其中,1~7 分别指关系%、#、?、*、@、\$、&。

对于相关度的太低的相关度值,可以认为它不相关,把相关度大于 0.3 作为相关的阈值。

$$\text{Rele_}P(P_1,P_2) = -0.394$$
$$\text{Rele_}P(P_3,P_4) = +0.654$$

解决策略: $\text{Rele_}P(P_1,P_2) = -0.394$ 说明这个矛盾问题是负相关的。对于 P_1 ,采取复制变换,用测量 N 张纸的厚度 $N \times y \in [1,20]\text{cm}$ 。

通过解决问题 P_1 ,采用可拓变换中的逆变换找出解决问题 P_2 的方法,即利用分解变换(在利用分解变换的之前,先利用置换变换)。

$$\text{Rele_}P(P_3,P_4) = +0.654$$

说明这 2 个矛盾问题是高度正相关的, $P_3、P_4$ 可采取类似的方法解决。对 P_3 ,将桌子旋转,高变为宽,宽变为高,则可抬入门内。对 P_4 也可采取这种方法。

5 实验及结果

根据上述方法,利用 C++,采用 VS 编程环境实现计算问题相关度程序,并对问题库里的 2 413 个矛盾问题的问题相关度计算。实验参数设置如下: $w_1 = 0.25, w_2 = 0.2, w_3 = 0.2, w_4 = 0.15, w_5 = 0.10, w_6 = 0.05, w_7 = 0.05$ 。

其中,1~7 分别指关系%、#、?、*、@、\$、&。阈值为 0.3。

通过计算,其中 1 782 个与其他问题相关。

将相关问题放在一起研究,并运用可拓学策略生成方法再次研究解决矛盾问题,实验结果如下。

表 2 实验结果

Table 2 Result of test

方法	无问题相	基于问题
数量	关度方法	相关度计算
解决的问题数量	1 390	1 539
未解决的问题数量	392	241

在运用基于问题相关度计算方法后,有 73.85% 的问题是相关的,其中有 140 个未解决的问题被解决了,而在已经解决的问题中,有 325 个问题找到了更好的方法改进策略。所以在总体情况下,基于问题相关度方法对策略生成的改进量为 19.27%;其中是将未解决的问题变成已解决 5.80%, 13.46%是将已解决的问题找到更好的策略。

6 结束语

通过计算基元的相关度进而计算问题的相关度,可以使相关度很高的问题归类。正相关度高的问题之间,可通过解决问题 A 而找出类似的方法解决问题 B ,而负相关的问题之间,通过解决问题 A ,而逆向找出解决问题 B 的方法。而将相关问题归类后,可通过归纳总结得出此类问题的一般特征,找出解决此类问题的一般规律,进而改进策略生成方法,使未解决的矛盾问题得到解决,或者使已解决的矛盾问题找到更优的策略。

实践证明,这种方法可有效增加策略生成途径,使矛盾问题解决的概率大大提高。

参考文献:

[1]杨春燕,蔡文. 可拓工程[M]. 北京:科学出版社, 2007: 1-344.

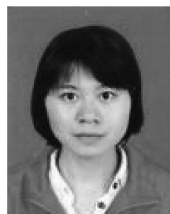
[2]董振东. HowNet [EB/OL]. [2013-05-12]. <http://www.keenage.com>.

[3]蔡文. 可拓集合和不相容问题[J]. 科学探索报, 1983 (1): 83-97
CAI Wen. Extension set and non compatible problems [J]. Science Journal, 1983 (1): 83-97

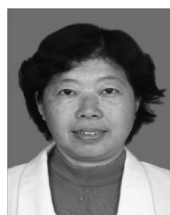
[4]李桥兴,刘思峰. 基于区间距和区间侧距的初等关联函数构造[J]. 哈尔滨工业大学学报, 2006, 38(7) :1097-1100.

- LI Qiaoxing, LIU Sifeng. Elementary dependent function is constructed based on the interval distance and lateral distance[J]. Journal of Harbin Institute of Technology, 2006, 38 (7): 1097-1100.
- [5] 刘群,李素建.基于《知网》的词汇语义相似度计算[J]. 计算语言学及中文信息处理, 2002(7): 59-76.
- LIU Qun, LI Sujian. based on HowNet semantic similarity calculation[J]. Computational Linguistics and Chinese Information Processing, 2002(7): 59-76.
- [6] 赵应秋, 罗军, 张君艳. 基于知网的词语语义相关度计算[J]. 信息技术, 2010 (3): 90-93.
- ZHAO Yingqiu, LUO Jun, ZHANG Junyan. The word semantic relevancy computation based on HowNet[J]. Information Technology, 2010 (3): 90-93.
- [7] 许云, 樊效忠, 张锋. 基于知网的语义相关度计算[J]. 北京理工大学学报, 2005, 25(5): 411-414.
- XU Yun, FAN Xiaozhong, ZHANG Feng. Semantic relevancy computing based on HowNet[J]. Journal of Beijing Institute of Technology, 2005, 25 (5): 411-414.
- [8] 江敏, 肖诗斌, 王弘蔚, 等. 一种改进的基于《知网》的词语语义形似度计算[J]. 中文信息学报, 2008, 22(5): 84-89.
- JIANG Min, XIAO Shibin, WANG Hongwei, et al. An improved word similarity computing method based on HowNet [J]. Journal of Chinese Information Processing, 2008, 22 (5): 84-89.
- [9] 王红玲, 吕强, 徐瑞. 中文语义相关度计算模型研究[J]. 计算机工程与应用, 2009, 45(7): 22-26
- WANG Hongling, LV Qiang, XU Rui. Chinese semantic relativity calculation of model [J] Computer Engineering and Applications, 2009, 45 (7): 22-26.
- [10] 刘宗妹. 本体可拓模型的复合元实现及应用研究[D]. 广州: 广东工业大学, 2010: 1-48.
- LIU Zongmei. Meta ontology extension model research and application of the [D]. Guangzhou: Guangdong University of Technology, 2010: 1-48.
- [11] 李立希, 杨春燕, 李铎汶. 可拓策略生成系统[M]. 北京: 科学出版社, 2006: 1-231.
- [12] 方卓君, 李卫华, 李承晓. 自助游可拓策略生成系统的研究与实现[J]. 广东工业大学学报, 2009, 26(2): 83-89.
- FANG Zhuojun, LI Weihua, LI Chengxiao. The self-help travel extension and implementation[J]. Journal of Guangdong University of Technology, 2009, 26 (2): 83-89.
- [13] 李承晓, 李卫华. 租房可拓策略生成系统[J]. 智能系统学报, 2011, 6(3): 272-278.
- LI Chengxiao, LI Weihua. The extension strategy generating system for rental of intelligent[J]. CAAI Transactions on Intelligent Systems, 2011, 6 (3): 272-278.

作者简介:



曹礼园,女,1987年生,硕士研究生,主要研究方向为智能软件。



李卫华,女,1957年生,教授,主要研究方向为面向 Agent 计算、网络信息系统、智能软件。发表学术论文 40 余篇。