doi:10.3772/j.issn.1002-0470.2023.02.010

基于知识图谱和犹豫模糊理论的复杂产品设计知识检索系统①

张宇飞②* 王宏伟③** 翟 翔**** 牛东晓*** 曹孟媛****

(*浙江工业大学先进制造研究所 杭州 310023)

(**浙江大学伊利诺伊大学厄巴纳香槟校区联合学院 海宁 314400)

(***华北电力大学经济与管理学院 北京 102206)

(****北京电子工程总体研究所复杂产品智能制造系统技术国家重点实验室 北京 100854)

摘 要 针对复杂产品设计知识检索系统的检索精度低、人员使用要求高、匹配知识粒度大等问题,本文开发了一种基于知识图谱(KG)和模糊相似度的设计知识检索系统。结合设计知识需求及历史设计数据,开发知识表示模型以建立设计知识网络;基于知识图谱的三元组存储形式,构建设计知识检索框架;设计基于犹豫模糊理论的句子检索算法,从词和句子层面融合句子的特征属性,并给出不同属性的隶属度计算公式,对比验证了该算法的平均准确率优于其他相关算法。在上述方法和技术的基础上,以柴油发动机设计知识为例开发设计知识检索系统,通过两类设计知识检索实例,验证了所开发系统的有效性和实用性。

关键词 设计知识;知识图谱(KG);犹豫模糊相似度;知识检索

0 引 言

新产品设计研发不断向知识集成方向发展,智能设计逐渐聚焦于设计知识的高效重用,以减少设计活动中的重复性工作。相关研究表明,多达 70% 的设计知识可从历史解决方案中获得,但在整个产品生命周期中,大约只有 28% 的设计知识被重复使用[1]。作为智能设计中知识重用的基础实施平台,检索系统的开发对智能设计过程中知识重用率和产品设计效率等的提升具有重要意义。关于设计知识检索系统的框架和检索方法研究中,Wang 等人[2-3]针对产品设计原理知识开发检索推理框架以提高知识的利用率。设计知识的本体表示从上层抽象为下层实例提供设计知识的分类与组织,涂建伟等人[4]解析设计者的意图,结合“共指关系索引-检索本体-设计知识”映射结构,解决计算机辅助设计系统检索准确率低等问题。这些知识检索系统包括面向本体的夹具设计[5]、多维度语义本体[6]和基于案例推理[7]等,主要由领域人员主导开发,多应用于尺寸、性能指标、型号等参数类知识的检索,故要求使用者了解设计对象的属性及期望的目标知识,但在产品研发前期,设计人员对查询内容难以形成确切的描述[8]。但设计师可在宏观上形成大概的需求描述,如“高强度的材料”虽然并不明确查询的目标是“合金钢”,但却与之高度类似或相关,因此以自然语言查询方式更能满足符合实际查询的目标描述。

在复杂产品设计领域中,知识检索系统多数采用基于设计对象和设计实例的相似性匹配和规则知识库搜索方式,这种检索系统对关键词和数值类进行模板检索多基于向量空间计算[9]或本体语义模型[10-11],都会造成结果过多或过少的情况,但进一步的信息筛选会导致知识的错过或冗余的信息,造

① 国家重点研发计划(2020YFB1707803)资助项目。

② 男,1996 年生,硕士生;研究方向:智能设计与制造,知识图谱;E-mail:zjut_zhang@163.com。

③ 通信作者,E-mail:hongweiwang@zju.edu.cn。

(收稿日期:2021-05-17)

成信息混乱。深度学习技术极大改善了这种情况[12],前提是收集海量的规范数据开展模型训练,这对于设计领域是一个挑战。因此为解决复杂的设计知识在知识匹配中效率低、精确度不足、检索结果冗余等问题,根据上述客观情况,将知识图谱(knowledge graph,KG)技术作为底层技术支撑,并作为知识检索系统的知识源,采用犹豫模糊相似度作为事件类知识的匹配算法,综合提高知识检索的准确度。在此基础上构建原型系统,以柴油发动机设计知识为对象进行实例验证,旨在为设计知识的重用提供系统化工具。

1 知识图谱和犹豫模糊理论

1.1 知识图谱

知识图谱作为知识表示、知识存储和知识重用等多重技术的集成学科,可灵活应用于智能设计中的不同工作场景。复杂产品的生命周期知识中,相比于设备描述[13]等结构化知识,设计知识蕴含多源的异构数据,并且具有复杂重用情景等,这些因素导致设计知识重用的低效率。本文检索的研究对象为复杂产品的设计知识、过程知识组件需求-功能-行为-结构(requirement-function-behavior-structure, RFBS)[14]对基本设计过程知识组织关联、认知过程组件(cognitive process knowledge)建立RFBS知识之间的解释性联系,并关联多源过程知识。这些知识组件以概念和事件的形式构建成复杂的数据空间,有待进一步被存储和挖掘。图形化的存储方式在复杂数据关系结构存储方面取得明显的优势[15-16],对于节点内容及关系的存储具有更强的独立性,其灵活的节点管理与维护为知识的检索查询提供较大的优势。因此知识图谱在产品设计知识检索系统中的应用,将提高知识检索的便利性,并为知识的可视化展示提供基础技术支撑,而这些是传统的设计知识检索系统功能所不具备的。

1.2 犹豫模糊理论

犹豫模糊理论拓展模糊理论,允许元素的隶属度值包含犹豫性[17],即不同集合的隶属度值可由不同数量的隶属度值组成,能够较优地反映实际决策情景中的不同决策偏好[18]。进一步,将犹豫模糊理论应用在检索中,主要考虑检索中的模糊性如提取古籍汉字图像的犹豫模糊特征[19]、对数学表达式的检索[20]等不同的应用中以提高检索效率和准确度。犹豫模糊理论的基础定义如下。

定义1 设 U 为一个非空的论域,当集合 E 满足[21]:

$$E = \{ < x, h(x) > \mid x \in U\} \tag{1}$$

称 E 为犹豫模糊集(hesitant fuzzy sets,HFS),$h(x)$ 是元素 x 对模糊子集 E 的所有可能隶属度值组成的有限集合,并且 $h(x)$ 中的值 $h(x)_i \in [0,1]$,$i = 1, 2, \cdots, n$,n 表示 $h(x)$ 中元素的个数。

对于犹豫模糊决策问题,其决策矩阵为

$$\boldsymbol{D} = (h_{ij})_{m\times n} = \begin{array}{c} \\ Y_1 \\ Y_2 \\ \vdots \\ Y_m \end{array} \begin{array}{c} \begin{array}{cccc} P_1 & P_2 & \cdots & P_n \end{array} \\ \begin{pmatrix} h_{11} & h_{12} & \cdots & h_{1n} \\ h_{21} & h_{22} & \cdots & h_{2n} \\ \vdots & \vdots & \cdots & \vdots \\ h_{m1} & h_{m2} & \cdots & h_{mn} \end{pmatrix} \end{array} \tag{2}$$

其中,$Y = \{Y_1, Y_2, \cdots, Y_m\}$ 表示可能的方案集合,$P = \{P_1, P_2, \cdots, P_n\}$ 表示目标匹配的属性集合。当 $n > 1$ 时,属性的权重 $W = \{w_1, w_2, \cdots, w_n\}^{\mathrm{T}}$,其中 $w_i \in [0,1]$,$\sum_{i=1}^{n} w_i = 1$。$h_{ij}(i = 1,2,\cdots,m; j = 1, 2, \cdots, n)$ 表示当前方案 Y_i 对当前目标属性 P_j 的匹配程度。

产品设计知识的检索中,用户对查询目标(需求)的不确定性以及用户查询对于检索系统的未知性,导致查询检索的模糊性。尤其对于句子的匹配,受限于词向量转化模型的词库以及语义分析工具当前的局限,进一步强化了检索系统的模糊性。深度学习已在句子(短文本)对的相似判断中发挥其优势,但在产品设计等垂直领域中,收集足够的数据用于迭代训练模型参数难以实现。根据上述分析,将犹豫模糊理论应用于中文设计知识检索过程中,主要流程如图1所示。

2 基于知识图谱的知识表示、存储与重用技术

2.1 C-RFBS知识表示模型

知识表示模型是知识网络构建的基础,为对历

史设计案例中的更多的设计知识进行组件捕获、划分、归纳并组织，同时考虑设计师对事件类设计知识的重用需求，在 RFBS 模型的基础上，创新性地引入认知发展理论，建立基于认知发展理论的知识表示模型（C-RFBS），如图 2 所示。

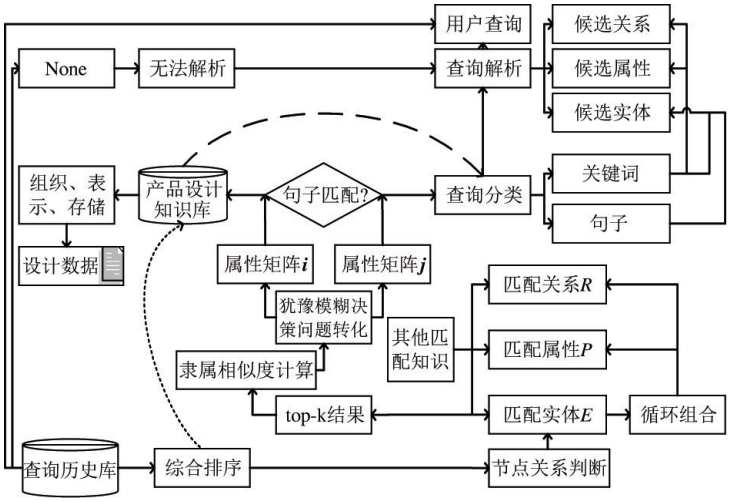


图 1　基于知识图谱和犹豫模糊理论的产品设计知识检索流程

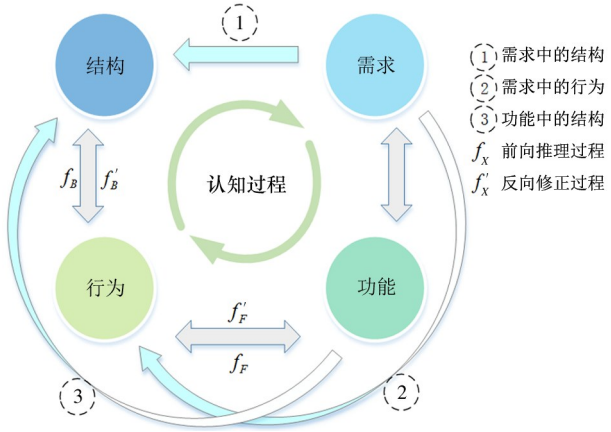


图 2　C-RFBS 知识表示模型

图 2 中，f_X 表示正向的设计推理过程，f_X' 表示设计过程中的反向修正过程，对于 RFBS 模型的公式化描述如下。

（1）需求分析

需求分析作为整个设计阶段的关键，需要设计人员根据自身经验以及市场趋势等外部知识对用户需求进行推理分析，将需求分解为可由功能层实现的最小需求，并获取其中的设计信息：产品的功能、用户期望的结构等作为已知结果。用户的需求 R 可表示为

$$R = \{R_1, R_2, \cdots, R_n\} \tag{3}$$

（2）功能映射

需求分析后，获取期望的产品功能。由于功能与需求之间的映射依赖本次设计活动，因此功能及需求的分解粒度是多态的形式，需求与功能的映射关系包括一对一、一对多和多对一的情况，如式（4）所示。

$$R_i = \{\sum_{j=1}^{m} F_{i,j}\} \tag{4}$$

式中，$\sum_{j=1}^{m} F_{i,j}$ 表示第 i 个需求 R_i 的功能映射结果集合，j 表示功能集合中的第 j 个功能。

（3）功能分解

结合行为映射过程，对不能直接获取其行为映射的功能，进一步分解为子功能集合，确保能够获取实现该功能的行为。同理，针对后续设计过程中的行为及结构有相同的分解过程。

$$F = \{F_1, F_2, \cdots, F_n\} \tag{5}$$

$$Be = \{Be_1, Be_2, \cdots, Be_n\} \tag{6}$$

$$S = \{S_1, S_2, \cdots, S_n\} \tag{7}$$

（4）行为映射

通过需求-功能映射及功能分解获取产品设计知识中的功能信息，对应每个功能单元，获取实现各个功能的预期行为知识 Be。对比产品结构的行为属

性 Bs, 判断设计方案的可行性。

$$F_i = \{\sum_{j=1}^{m} Be_{i,j}\} \tag{8}$$

式中, $\sum_{j=1}^{m} Be_{i,j}$ 表示第 i 个功能 F_i 的期望行为映射结果集合。

(5) 结构求解

获取期望行为后,通过选择或设计合适的结构以实现预期行为 Be, 执行产品的功能,满足用户需求。

$$Be_i = \{\sum_{j=1}^{m} S_{i,j}\} \tag{9}$$

式中, $\sum_{j=1}^{m} S_{i,j}$ 表示第 i 个期望行为 Be_i 的结构映射结果集合。

(6) 认知过程中的同化

对于认知过程的公式化描述为

$$\{R, F, B, S\} + C = \text{New } \{R, F, B, S\} \tag{10}$$

认知过程知识多用于捕获设计过程中的驱动事件知识、行为事件知识和决策事件知识,以构建设计知识组件之间的关联。图 3 表示认知过程知识与 *RFBS* 知识之间的组织形式。

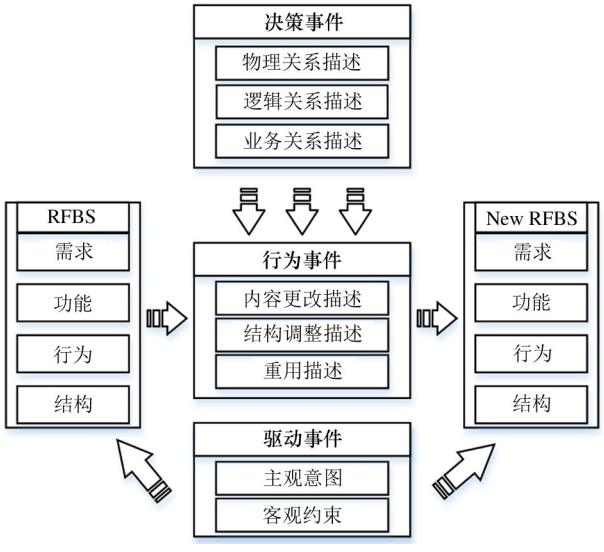


图 3　认知过程知识对 RFBS 知识的组织结构形式

2.2　知识的存储

设计知识的有效重用要求设计知识的存储形式既能定义设计知识组件之间的复杂关系,而且便于知识节点的计算及推理。图形化存储的方式通过节点、属性和关系 3 大组件,将捕获的设计知识按照预定的结构存储。具体而言,设计人员由用户需求、功能需求、结构需求及预期行为等产生知识需求,整理目标知识的可能信息,包括关键词、数值、行为等。其中,关键词包括实体名称、属性名称、属性值、关系名称,甚至事件节点的局部内容等;数值包括确切的值、范围值、模糊值;行为包括关系名称、属性名称、事件节点内容中包含的谓词,这些模糊的知识需求描述主要采用以下 2 种形式检索。

(1) 数据以文件或关系数据库的形式存储,通过基于关键词、模板匹配、本体模型等查询形式,按照给定的查询规则,返回匹配分数最高的 k 个结果。

(2) 知识图谱以三元组 $< h, r, t >$ 的形式对知识进行存储,为知识的检索和推理提供关系路径,其不同于关系数据库构建的关系表,图数据库以双向关系连接实体,因此为查询实体的查询、拓展与推理提供更加灵活的解决方法。本文以三元组的形式存储复杂设计数据作为设计知识检索系统的知识源。

2.3　基于知识图谱的知识检索框架

结合设计知识表示模型及知识检索的特性,设计如图 4 所示的知识检索框架。根据图 4 所示框架,用户查询主要以自然语言的形式展开关键词和句子的查询匹配。

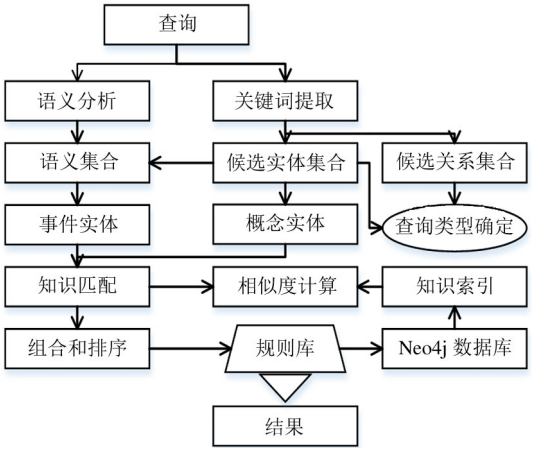


图 4　基于知识图谱的知识检索框架

知识查询与匹配本质上是对查询及知识库数据的匹配,因此针对三元组知识匹配的特征,将知识的

匹配分为表 1 所示的 6 类基本问题,并给出处理原则。查询 $Q = \{e_1, e_2, \cdots, e_n, r_1, r_2, \cdots, r_m\}$ 为查询中匹配到的实体及关系,其中默认属性处理方式同关系、属性值同实体。

表 1 知识图谱中的 6 类查询方式

查询类型	知识图谱中 6 类问题基本处理原则
单实体	判断实体 e 是否为产品名称;是,返回设计摘要;否,返回节点 e 及其属性。
多实体	判断实体 e_i 之间是否存在关系:是,返回被指向实体及其属性,其他实体作为筛选条件;否,返回层级较高的实体及其属性。
单实体 + 单关系	1. 判断 e 和 r 是否属于 $< h, r, t >$ 中的 2 个元素。是,返回空缺的 h 或 t;否,步骤 2。 2. 判断以 r 相连的知识节点 e_s 是否能够与查询实体 e 建立关系。是,返回包含 r、s 的三元组对应知识实体,同时返回包含实体 e 和实体 s 的知识子图。否,删除关系,按照第 1 类问题查询。
多实体 + 单关系	1. 对于每个实体 e_i, 判断 e_i 和 r 之间的关系,处理同第 3 类问题步骤 1; 2. 判断实体 e_i 和 e_j 之间的关系;存在,作为筛选条件。否,放弃。
单实体 + 多关系	同第 3 类查询,对每个关系循环判断。如果关系中包含属性,知识节点排名靠前。
多实体 + 多关系	1. 循环实体 e_i 和关系 r_m, 生成 $< h, r, t >$ 的二元组;带属性知识节点排名靠前。 2. 剔除属性。判断关系是 r_m 和 r_n 之间否存在路径先后关系,是,生成组合固定组合;否,不处理。 3. 步骤 1 处理后获得的属性节点筛选去除无关的实体。以案例局部知识网络对剩余关系和节点进行判断,返回匹配知识节点最多的案例局部网络中包含节点层次最高及层次最低节点的子图。

3 基于犹豫模糊的句子相似度

3.1 句子的模糊相似度

考虑实际检索过程中的模糊性和不确定性,将句子匹配转化为多属性决策过程。将句子 S 作为研究对象,将句子表示为 $S = \{W_1, W_2, \cdots, W_m\}$, 对句子以及 W_i 进行语义分析,获取句子的属性矩阵:

$$\boldsymbol{D} = (p_{ij})_{m\times n} = \begin{array}{c} \\ W_1 \\ W_2 \\ \vdots \\ W_m \end{array} \begin{array}{c} \begin{array}{cccc} P_1 & P_2 & \cdots & P_n \end{array} \\ \begin{pmatrix} p_{11} & p_{12} & \cdots & p_{1n} \\ p_{21} & p_{22} & \cdots & p_{2n} \\ \vdots & \vdots & \vdots & \vdots \\ p_{m1} & p_{m2} & \cdots & p_{mn} \end{pmatrix} \end{array} \tag{11}$$

其中, $s_{ij}(i = 1, 2, \cdots, m; j = 1, 2, \cdots, n)$ 表示词 W_i 对当前匹配属性 P_j 的相似度。

在多属性的决策问题中,一般单维度的决策计算生成二维决策矩阵,而句子作为词的组合排列,导致属性值为多维度矩阵。本文定义句子属性,判断不同句子间的隶属相似度,将句子相似度计算转化为隶属相似度矩阵的匹配。假设查询句子 m 矩阵为 $\boldsymbol{M}$, 其中 $\boldsymbol{M}$ 的维度为 $l_m \times d$, 其中 l_m 为句子 m 的分词数量,d 为句子属性个数。n 为待匹配句子,$\boldsymbol{N}$ 为 n 的属性矩阵,$\boldsymbol{N}$ 的维度为 $l_n \times d$,l_n 为句子 n 的分词数量。

对于知识检索系统来说,句子的相似度匹配是多对一的计算过程,而且句子 $\boldsymbol{N}$ 的长度及内容是未知的,因此对应于 $\boldsymbol{M}$ 中某个位置的评估从数值和数量上具有不确定性。根据式(11)获取句子 m 和句子 n 的属性矩阵 $\boldsymbol{M}$ 和 $\boldsymbol{N}$,将 $\boldsymbol{M}$ 矩阵作为理想矩阵,构建不同属性之间的隶属度计算公式:

$$h_A(X) = \{\langle x, h_A(x)\rangle \mid x \in X\} \tag{12}$$

其中, $h_A(x)$ 表示 x 属于集合 X 的可能程度。分别计算不同属性的隶属度函数,获取如下所示的相似度矩阵。

$$Similarity\ \ Matrix = (s_{ij})_{l\times m} = \begin{array}{c} \\ w_1 \\ w_2 \\ \vdots \\ w_l \end{array} \begin{array}{c} \begin{array}{cccc} w_{m1} & w_{m2} & \cdots & w_{mm} \end{array} \\ \begin{pmatrix} s_{11} & s_{12} & \cdots & s_{1m} \\ s_{21} & s_{22} & \cdots & s_{2m} \\ \vdots & \vdots & \cdots & \vdots \\ s_{l1} & s_{l2} & \cdots & s_{lm} \end{pmatrix} \end{array} \tag{13}$$

其中, P_d 表示第 d 个属性, $s = \sum_{i=1}^{d} w_i h_A(x)$, 因此 s_{ij} 表示当前查询 $\boldsymbol{M}$ 第 i 个词与等待匹配句子 $\boldsymbol{N}$ 的第 j 个词的综合相似度。

在知识检索系统中,待匹配的句子为 N_i, 其中 $i = 1、2、\cdots、n$, n 为知识库中可存储的句子数。因此对于句子的决策维度为 $l \times m \times n$, 如下所示。

$$\overbrace{\textit{Similarity Matrix}_{(l \times m \times n)} = \begin{array}{c} \\ w_{11} \\ w_{12} \\ \vdots \\ w_{1l_1} \end{array} \begin{array}{c} \begin{array}{cccc} w_{m1} & w_{m2} & \cdots & w_{mm} \end{array} \\ \begin{pmatrix} s_{11} & s_{12} & \cdots & s_{1m} \\ s_{21} & s_{22} & \cdots & s_{2m} \\ \vdots & \vdots & \vdots & \vdots \\ s_{l1} & s_{l2} & \cdots & s_{lm} \end{pmatrix} \end{array} \cdots \begin{pmatrix} s_{1m} \\ s_{2m} \\ \vdots \\ s_{lm} \end{pmatrix}}^{n} \tag{14}$$

3.2 句子属性构建

句子作为词的排列与组合,除词本身的信息外,还包括词与词之间及词对句子的影响,考虑词本身的特性(词向量和词性),以及在句子中的特征(句中位置、出现次数(词频)和语法结构信息(第 1 维度))作为句子的特征属性,如表 2 提供"叶片的设计要确保发动机的稳定性"的句子属性特征向量矩阵。

表 2 句子特征属性表

词条	词性	句中位置	词频	语法结构信息
叶片	n	2	1	ATT
设计	v	1	1	SUB
要	v	1	1	ADV
确保	v	1	1	HED
发动机	n	2	1	ATT
稳定性	n	1	1	VOB

文献[22]开发的 LTP 语言技术平台为中文的分词、语义角色分析、词性标注、依存句法分析及语义依存分析提供系统的技术体系。LTP 中定义了 14 种依存句法关系,其中"HED"表示整个句子的核心。句法分析后得到的文本标注信息,剔除 CMP、POB、RAD、LAD 关系中第 2 维度数据(影响句子的含义),保证被匹配的句子最大化包含有用信息。由于每个词都包含检索目标信息,句子相似度计算可以作为句子特征的决策分析过程,首先将查询 S_q 和事件知识(句子/短文本) S_k 表示为词的集合 $S_q = \{w_{q1}, w_{q2}, \cdots, w_{qn}\}$ 和 $S_k = \{w_{k1}, w_{k2}, \cdots, w_{km}\}$, 其中 w_{qn} 和 w_{km} 分别表示 S_q 与 S_k 中的分词。

3.3 基于隶属相似度的隶属度计算函数

根据句子的不同属性,当元素的隶属度为{0, 1},模糊度为 0;元素的隶属度为 0.5,模糊性最大。对于查询中的关键词采用 Word2Vec 模型进行词向量转化。进一步为判断当前词对句子的重要性,采用 Yan 等人[23]获取词对句子的隶属度。设置阈值 $\theta, 0 \leqslant \theta \leqslant 1$, w_i 的不确定函数 I_θ 定义为

$$I_\theta(w_i) = \{w_i\} \cup \{w_j \mid sim(w_i, w_j) \geqslant \theta\} \tag{15}$$

其中, $sim(w_i, w_j)$ 是词 w_i 和 w_j 的相似程度。

$$sim(w_i, w_j) = \cos(w_1, w_2) \tag{16}$$

模糊包含函数 v 定义为

$$v(X, Y) = \frac{|X \cap Y|}{|X|} \tag{17}$$

综上,词对句子的隶属度函数为

$$h(w_j, S_i) = v(I_\theta(w_j), S_i) \tag{18}$$

通过式(15)~(18)能够有效拓展词的语义,根据计算的词对句子的隶属度,衡量词对句子的影响。

针对表 2 中的特征属性,结合文献[24]提出的隶属度计算公式,设置隶属度计算方式如式(19)~(22)。中文词的词性是固定的,但鉴于由词向量的匹配到词可能不具备相同的词性,因此属性词性的隶属度函数 h_{pos} 为

$$h_{\text{pos}}(S_{qn}, Sk_{\text{im}}) = \{x \mid x \in \{0,1\}\} \tag{19}$$

其中, S_{qn} 表示查询输入中的第 n 个词, Sk_{im} 表示第 i 条知识中的第 m 个词,由于中文词的词性在识别模型中是固定,因此,当两词的词性相同时为 1,否则为 0。

查询输入和事件知识不能保证其固定的格式,因此词在句中的位置根据词到句子中心词 HED 的最短距离确定。

$$h_{\text{count}} = \mathrm{e}^{-\left(\frac{count_{Qn} - count_{Kim}}{\sigma}\right)^2} \tag{20}$$

其中, $count_{Qn}$ 表示 w_n 到中心词 HED 的最短距离,并且 $count_{Qn} \in [0.1, 0.4]$。

词的频率决定词对整个句子的影响程度,词频的隶属度函数采用文献[25]中的公式为

$$h_{\text{freq}} = \frac{1}{1 + \mu \mid freq_{Qn} - freq_{Kim} \mid} \tag{21}$$

语法结构分析词在句子中的作用,其隶属度函数为

$$h_{sen} = \{(s, sen(S_{qn}, Sk_{\text{im}}))\} \tag{22}$$

其中，当查询词的结构相同时 $h_{sen}=1$；否则，$h_{sen}=0.5$。

通过上述对隶属度函数的确立，计算式(23)评判 2 个句子的相似程度。

$$Sim(Sq, Sk_i)=1-\sqrt[\lambda]{\left\{\frac{1}{5}\sum\left[\frac{1}{len(Sq)}\sum\mid M_{j\_Sk}-M_{j\_Si}\mid^{\lambda}\right]\right\}} \tag{23}$$

考虑到词对句子的隶属度，导致词对不同句子的影响程度不同，因此修改式(23)的相似度计算公式为

$$Sim(Sq, Sk_i)=1-\sqrt[\lambda]{\left\{\frac{1}{5}\sum\left[\frac{1}{len(Sq)}\sum\mid \alpha M_{j\_Sk}-\beta M_{j\_Si}\mid^{\lambda}\right]\right\}} \tag{24}$$

其中，α 表示词对查询的权重，β 表示对当前知识库中句子的权重。通过词对句子的隶属度函数，可扩展词的语义范围。循环迭代知识库中的每条句子，按照表 3 所示的算法流程，计算句子的相似程度，返回目标查询隶属度较高的句子。

表 3 句子相似度计算流程

句子相似度计算
查询 S_q 和待匹配句子集 *S-set*
设置词向量的 cosine 相似度阈值参数 θ、σ、μ、λ；
S_q 和 *S-set* 中相似度最高的相似度值和句子 $S\text{-}set_i$
步骤 1：整合查询到句子集合：$S_q + S\text{-}set = \{S_i \mid S\text{-}set_i + 1\}$；
步骤 2：根据集合 $\{S_i\}$ 建立单词的论域 U，并对所有的词基于 Word2Vec 进行词向量转化；
步骤 3：剔除表 2 中的非重要词汇，根据式(18)计算 *Sq* 中每个词的隶属度，获取词隶属度集合；
步骤 4：根据步骤 3，同理循环 $S\text{-}set_i$ 中的每个词；
步骤 5：通过词和句子的隶属度作为词的权重，调整相似度计算函数式(23)为式(24)；
步骤 6：根据式(19)～(22)循环 *Sq* 中的每一个词同 $S\text{-}set_i$ 中的每一个词，当 2 个词的语义相似度大于 θ，计算隶属度，生成隶属度矩阵 $\boldsymbol{Mj}\_SK_i$；
步骤 7：根据隶属度矩阵 $\boldsymbol{Mj}\_SK_i$，按式(24)计算 2 个句子的相似度，返回句子相似度向量；
步骤 8：循环 *S-set*，求解与 *Sq* 匹配的最大相似度值并返回。

将输入的句子同样作为一个元素 $\{S_k+1\}$，按照句子属性表生成论域 U。当某个单词不能够匹配句子中的信息时，那么，此时集合为{0, 0, 0, 0}，当完全匹配时，集合为{1, 1, 1, 1}。

3.4 实验分析

(1) 评价指标

目前知识检索的评价指标体系还不完善[26]，因此借鉴文献[27]中采用召回率、准确率和 F_1 测量值作为句子相似度匹配算法的评价指标。召回率是匹配到相关句子的数目与知识库中所有句子数目的比值，衡量知识检索的查全率；准确率是匹配到正确句子的数目与匹配到的所有句子总数的比值，衡量知识检索的查准率；F_1 测量值综合评估召回率和准确率 2 个指标。

$$P=\frac{\text{匹配到正确句子的数目}}{\text{匹配到的所有句子总数}} \tag{25}$$

$$R=\frac{\text{匹配到正确句子的个数}}{\text{数据集中所有相关句子总数}} \tag{26}$$

$$F_1=\frac{2\times P\times R}{P+R} \tag{27}$$

(2) 实验结果及分析

本节实验中，收集 8 M 左右的文本作为训练语料库，人为收集并定义 300 组相似句子对 (S_1, S_2)，并将句长差距过大的 S_1 和 S_2 进行修改或更换后作为实验数据。以基于字/词频(term frequency, TF)、向量空间模型(vector space model, VSM)、语义特征句子相似度和基于模糊句子相似度匹配等 8 种句子相似度计算方法作为对比模型进行多次的实验，取平均结果作为最终比较的数据。由于采用 Word2Vec 训练生成的词向量转化模型会存词表中词汇缺失的问题，本实验采用一维 0 向量作为补充词向量。

通过表 4 可得，自然语言查询更注重词的含义，但在句子的相似度计算实验中，词 TF 的准确率低于字 TF 的方法，通过对比不同的分词工具 jieba、LTP 和 pynlpir 得出，可能由于分词工具对句子分词的准确性未完全覆盖领域中文词汇；采用基于词向量等的语义方法，能够在基于词频的基础上提高一定的准确度，词向量的获取需要预先对语料进行特征训练以获取高质量的词向量模型，当数据量足够

时,采用 Word2Vec 能获得较优的词向量模型,其缺陷是当词不存在时,不能获取该词的向量;在词向量的基础上,添加句子的结构、句长、距离等特征信息,当设置合适的权重时(本实验中优选权重为 0.7、0.1、0.1、0.1),能够进一步改善匹配的准确性。

表 4　不同相似度计算方法对比

检索方法	准确率	召回率	*F*1 值
字 TF	0.630	0.662	0.645
词 TF	0.452	0.490	0.470
词 TFIDF	0.464	0.500	0.482
词共现	0.225	0.375	0.281
基于 Bert 的余弦相似度	0.470	0.475	0.473
基于 Word2Vec 的余弦相似度	0.626	0.640	0.631
特征融合[28]	0.673	0.680	0.677
FRS + 词向量[29]	0.623	0.656	0.634
基于 HFS 的相似度	0.726	0.731	0.733

如图 5 所示,由于在计算过程中考虑词的属性对句子的影响,以及句子赋予词的特性,并通过隶属度函数拓展词的匹配范围,基于 HFS 模型的精确度、召回率和 *F*1 指标在基于特征融合的基础上提升 9% 左右,可满足设计师的应用需求。对于知识检索系统,知识节点的匹配精度会影响系统的有效性,针对设计知识中存在的事件类节点(句子/短文本)知识,利用模糊理论的思想设计相似度计算模型,以提高知识检索系统的查全率和查准率。

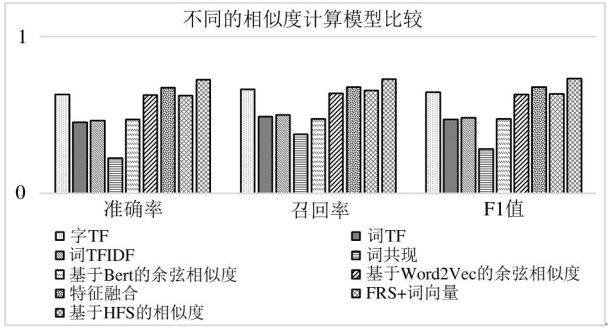


图 5　不同相似度算法的句子匹配效果

4　原型系统

4.1　系统架构

基于上述的知识检索框架及计算方法,设计如图 6 所示的设计知识检索原型系统。采用 Python 处理语言,在 flask 和 react 等开发环境下,以面向 Web 服务的形式搭建系统,用户可在线输入需求描述,系统则通过知识匹配返回最佳的目标知识。进一步地,设计人员可以通过可视化界面拓展期望的关联知识。

4.2　设计知识库

设计知识库以知识图谱为知识源,包括历史设计知识库、查询规则库以及文本属性库等。历史设计知识库以图的形式存储,包括设计过程中的需求、功能、行为、结构、驱动事件、行为事件和决策事件及其拓展知识;查询规则用于被检索知识节点的初步拓展,以知识子图的方式展示被检索的知识节点内容,查询规则一定程度上能够提高知识检索的精度和广度以提供满足实际应用需求的知识关系节点;文本属性库包括设计知识语义属性库和查询语义属性库,其中,设计知识语义属性库是对图谱中存储的事件类知识数据进行语义分析以及词向量的转化,查询语义属性库则是对查询输入的语义分析和词向量转化的记录,该知识库主要用于降低句子解析的次数,提高检索效率。

4.3　查询分类模块

查询分类模块对用户查询进行初步的筛选与分类。识别查询输入中的实体、关系和属性,并对实体进行概念实体和事件实体的区分,从而形成以下 3 类集合有待下一步操作。

(1) 关系、属性集合。首先根据查询输入,基于关键词语义相似度判断查询中可能包含的属性及关系,作为属性集和关系集待组合排序模块中调用。

(2) 概念实体集合。将查询输入进行分词,剔除停用词后,作为关键词集合,进一步判断可能为实体和属性值的词作为实体集合和属性值集合,作为待匹配三元组中的实体元素进行实体的匹配。

(3) 事件实体集合。循环各个关系和属性,对查询中的剩余字段判断是否具有句子特征(主谓宾、主谓、谓宾等),存在,则作为待匹配数据,需要进行句子相似度的计算;不存在时,则仅获取关键词集合。

综合上述待匹配集合,进行实体与关系/属性的

组合,判断是否符合表 1 所示的查询分类。

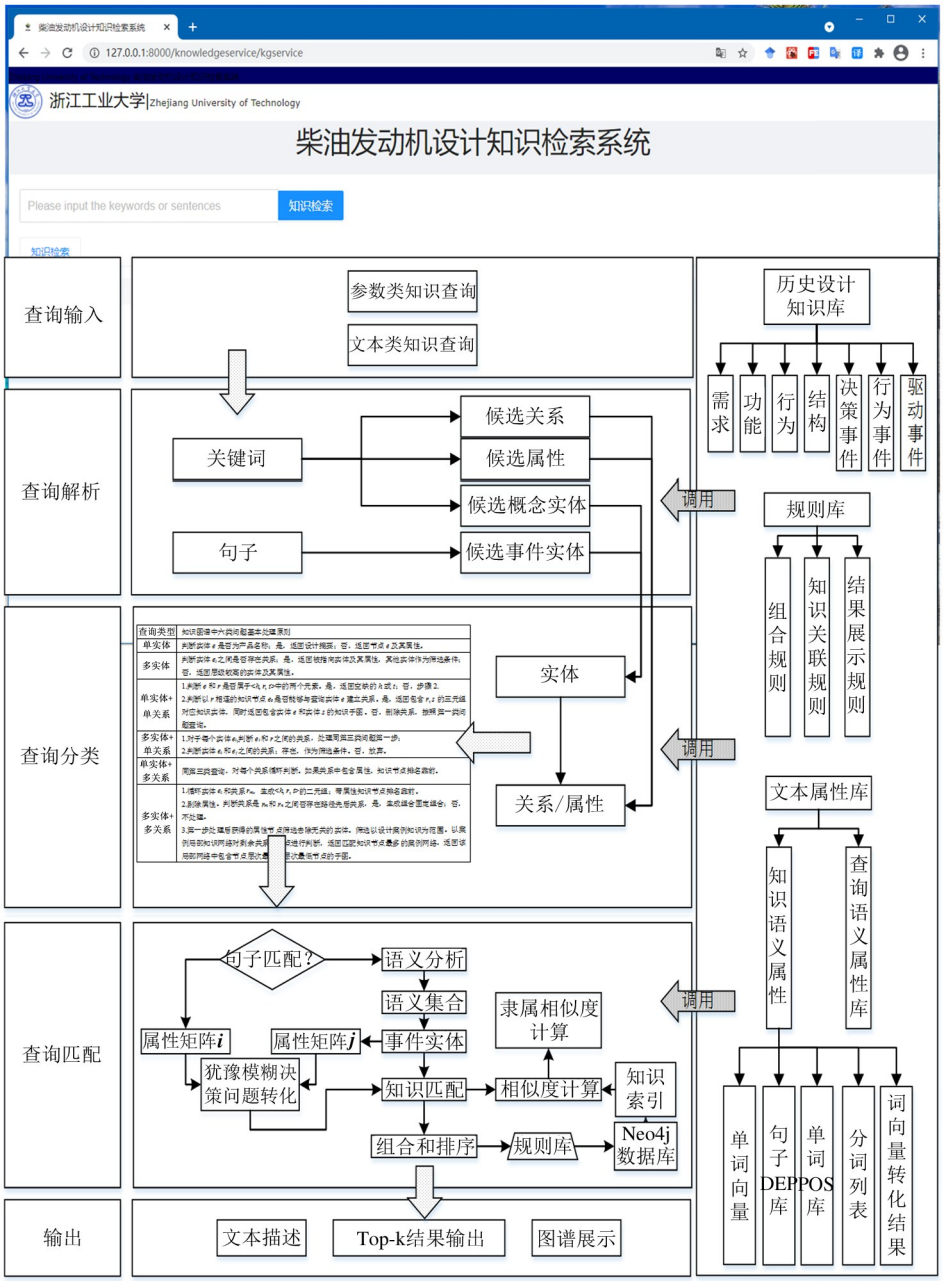


图 6　设计知识检索系统架构

4.4　查询匹配模块

针对关系、属性和概念类设计知识,采用基于 VSM 的相似度计算方式匹配。对于事件类实体,采用模糊相似度匹配后返回。

(1) 关键词匹配。基于 Word2Vec 模型,训练设计领域中的词向量转化模型。采用基于余弦相似度的方式作为关键词相似度的计算方式。设置阈值,当相似度计算结果大于或等于阈值时,将对应的关键词作为候选关键词集合中的元素。

(2) 句子匹配。根据第 3 节提出的句子相似度计算模型,对查询句子进行特征属性矩阵的构建,基于文本属性库,计算查询和事件知识的相似度。若查询不存在查询语义属性库中,则解析该查询并以 json 格式,如本文采用:{"查询":["segment":[None],"vector":[None],"distant":[None],"count":[None],"pos":[None],"dep":

[None]]}的形式进行解析结果的存储,用于下一次的检索。

(3) 路径判断。根据获取的概念类知识节点集合和事件类知识节点集合,循环判断概念知识和事件知识之间的路径距离,当路径距离在预定阈值以内,以知识子图的形式返回并展示;否则,一一返回并展示。

4.5 组合排序模块

关系组合:知识匹配后获取所有可能的目标实体、关系及属性集合,将知识实体分别与关系、属性进行组合,当满足数据库中三元组形式后,返回对应查询结果;否则,仅返回对应的实体内容。组合知识以尽可能充实的方式将知识返回给设计师作设计指导;当目标实体知识为空时,提示设计师重新输入查询。

结果排序:对于查询来说,事件类设计知识蕴含更多的信息。由于查询的模糊性,属性和关系的重要性相对较低,本文设置如表 5 所示的排序规则。

表 5 实体权重排序规则

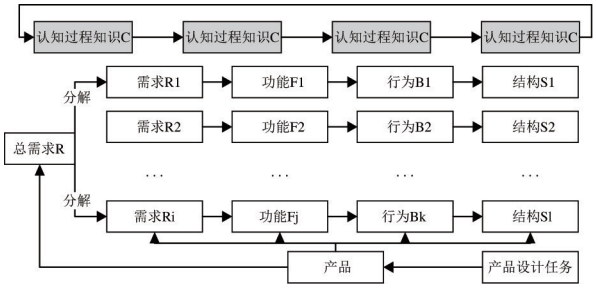
查询定位类别	排序	权重设置
事件实体类	1	0.4
概念实体类	2	0.3
属性	3	0.1
属性值	4	0.1
关系	5	0.1

5 实例验证

柴油发动机作为一种复杂产品,是智能设计落地的前沿行业,其整体的设计过程伴随多学科知识的复杂交互。某企业在长期的柴油发动机设计研制活动中,积累大量的产品研制任务书、设计说明书等等电子版文件,这些文件知识的查询主要通过基于关键词的方式对文档标题进行查找,需要设计师人工进行内容的筛选。目前,由于设计需求的变动和创新发展的需求,期望开发高效的知识检索系统以期提升柴油机设计的整体效率。由于设计产品对象及电子文档的多样性和广泛性,本文仅针对产品研制任务书中的部分知识类型为研究对象作为所开发系统的验证。

5.1 柴油机设计知识表示与存储

基于专家经验、2.1 节所提 C-RFBS 模型及知识的公式化描述式(3) ~ (10),对设计知识 C、R、F、B、S 以及产品、任务名称等知识组件的分类并整合如图 7(a)所示,并以 $< h, r, t >$ 的形式转化为 .csv格式文件用于知识的导入存储。



(a) 知识组织形式

(b) 知识存储形式

图 7 柴油机设计知识的 Neo4j 存储的图形化展示

实验以 Neo4j 图数据库为底层知识源存储工具,对设计知识进行图形化存储,如图 7(b)所示。图形化存储一方面能更高效、低消耗地存储复杂结构数据,另一方面相对于关系型数据库,更便于设计知识检索系统的可视化展示。

5.2 柴油机设计知识检索

基于柴油机设计知识图谱,考虑实际设计活动中常见的设计知识检索方式,展开如下 2 类设计知识检索的案例以验证系统的适用性。

实例 1：文本类信息的查询。在检索页面输入期望知识的描述，如设计师输入“解决柴油机排气管开裂的措施”，通过查询分类模块，转化为实体集合{概念类实体：[柴油机，排气管]，事件类实体：[解决柴油机排气管开裂的措施]}，属性集合{None}和关系集合{None}。因此，该查询为“多实体”类，按照第 2 节中所提的知识检索流程及查询分类原则，表 6 展示为相似度分数大小排名前 2 的匹配结果。

表 6　文本类知识匹配及其相似度

待匹配实体	匹配实体	相似度/%
柴油机	柴油机	100
排气管	排气管	100
	排气歧管	93
解决柴油机排气管开裂的措施	柴油机排气管出现开裂现象	83
	排气歧管出现裂纹	75

通过对上述概念类实体“柴油机”、“排气管”和事件类实体“解决柴油机排气管开裂的措施”的匹配权重分析判断，事件实体具有较高的权重。进一步，判断“排气管”和“柴油机排气管开裂”存在路径关系，并且在预定的路径距离 4 以内，因此，以组合的形式返回知识。如图 7 所示，左侧以本次设计案例的任务名称为标题，并依次显示设计摘要、相关结构、驱动事件、行为事件和决策事件作为本次检索的结果；右侧显示设计案例的图谱化知识用于知识的拓展查询。其中，路径距离为 4 的 CQL(cypher query language)查询判断方式为：CQL _ 4 = " MATCH (m：产品)-[r1]-> (n1)-[r2]-> (n2)-[r3]-> (n3)-[r4]-(n4：行为事件) where m. name = '{0}', n4. name = '{1}' return m. name, n1. name, n2. name, n3. name, n4. name". format(" 柴油机"," 柴油机排气管出现开裂现象")

实例 2：参数类信息的查询。该类查询类似于传统的设计知识检索系统，用于验证该系统可应用于变型设计。查询需定位某参数及参数值的设计案例时，如“排量为 2.5 L，其马力达到 128 hp 的柴油发动机”。如表 7 所示为 10 个柴油机设计案例中的“柴油发动机”实体的部分属性及属性值。

表 7　柴油发动机部分属性及属性值

产品设计案例	排量/L	最大马力/hp	最大扭矩/(N·m)	额定转速/(r/min)
C1	2.29	95	245	3200
C2	2.29	110	280	3200
C3	2.499	116	290	3200
C4	2.499	125	325	3200
C5	2.499	131	290	3600
C6	2.977	136	325	3200
C7	2.977	150	360	3200
C8	2.977	150	400	3200
C9	3.767	150	500	2600
C10	3.767	140	450	2600

设置文本相似度阈值为 95，选择大于或等于预订阈值的产品案例。首先解析查询为{产品：柴油发电机}，{排量：< 2.5；马力：> 128}，将排量与马力进行基于关键词的语义匹配，结果为：{排量：排量，马力：最大马力}。进一步采用改进欧几里得相似度匹配(百分制)，结果如表 8 所示。其中，C4 和 C5 的相似度值大于阈值，因此作为匹配案例返回并综合显示相似度最高的检索结果(如图 9 所示)。

表 8　案例相似度计算结果

案例 C	相似度 S	案例 C	相似度 S
C1	76.67	C7	84.44
C2	87.27	C8	84.44
C3	91.51	C9	84.42
C4	97.88	C10	91.47
C5	97.88	C11	84.44
C6	94.33	C12	80.20

图 8　文本知识检索结果

图 9　参数知识检索结果

上述设计知识检索系统的 2 类查询案例,基本上满足设计师对文本类知识和数值类知识的综合检索需求,可实现历史设计知识辅助再设计活动。所提的方法和原型系统旨在降低用户的专业知识背景限制,简化查询输入,在一定程度上提高知识检索的效率,而且通过模糊语义检索的方式,提高了系统对用户模糊输入的鲁棒性。进一步地,图谱展示的形式能够根据实时的需求,拓展知识节点,提高检索结果的质量,极大降低企业设计活动的成本,提高设计活动效率。

6　结　论

针对设计知识检索中知识组织质量和知识重用效率低等问题,利用知识图谱技术组织、维护和重用设计知识。基于 C-RFBS 模型组织并关联历史产品设计知识,将设计知识作为数据源构建知识图谱并运用到设计知识的检索过程中,针对知识图谱的三元组存储形式将查询归类为 6 类检索问题,在此基础上构建基于知识图谱的检索框架。进一步,采用犹豫模糊相似度计算方法解决检索过程中中文句子匹配精确不足的问题,对比其他 8 种相似度计算方法,验证该算法进一步提升了检索的精确度。在所提模型和方法的基础上,搭建面向 Web 服务的设计知识检索原型系统,以柴油发动机设计知识为数据源,采用两类检索案例验证该系统在实际应用场景下的可行性和有效性。

本文旨在将历史设计知识进行高效率、高质量的重用,为设计师提供快速精准的目标知识,加速智能化设计进程,但对于设计师的意图分析、设计知识的智能化推理计算及系统效果评估等方面,仍有待进一步深入研究。

参考文献

[1] ETTLIE J E, KUBAREK M. Design reuse in manufacturing and services[J]. Journal of Product Innovation Management, 2008, 25(5): 457-472.

[2] WANG H, JOHNSON A, BRACEWELL R. Supporting design rationale retrieval for design knowledge re-use[C]// Proceedings of the 17th International Conference on Engineering Design. Stanford: ICED, 2009:335-346.

[3] WANG H W, JOHNSON A L, BRACEWELL R H. The retrieval of structured design rationale for the re-use of design knowledge with an integrated representation[J]. Advanced Engineering Informatics, 2012, 26(2): 251-266.

[4] 涂建伟, 李彦, 李文强, 等. 一种面向产品创新设计的知识检索模型与实现[J]. 计算机集成制造系统, 2013, 19(2): 300-308.

[5] 张田会, 张发平, 阎艳, 等. 基于本体和知识组件的夹具结构智能设计[J]. 计算机集成制造系统, 2016, 22(5): 1165-1178.

[6] 方伟光, 郭宇, 廖文和, 等. 基于本体的复杂产品设计知识表示和标注方法[J]. 计算机集成制造系统, 2016, 22(9): 2063-2071.

[7] LUO C, WANG X, SU C, et al. A fixture design retrieving method based on constrained maximum common subgraph[J]. IEEE Transactions on Automation Science and Engineering, 2018, 15(2): 692-704.

[8] 余旭, 刘继红, 何苗. 基于领域本体的复杂产品设计知识检索技术[J]. 2011, 17(2): 225-231.

[9] 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.

[10] 尹超, 夏卿, 黎振武. 基于 OWL-S 的云制造服务语义匹配方法[J]. 计算机集成制造系统, 2012, 18(7): 1494-1502.

[11] YANG H H, SUN M Y. Study on application of domain ontology in semantic information retrieval[C]// The 2nd International Conference on Mechatronics and Control Engineering. Dalian: Scientific, 2013: 1662-1665.

[12] 杨德志, 柯显信, 余其超, 等. 基于 RCNN 的问题相似度计算方法[J]. 计算机工程与科学, 2021, 43(6): 1076-1080.

[13] YAN H H, YANG J, WAN J F. KnowIME: a system to construct a knowledge graph for intelligent manufacturing equipment[J]. IEEE Access, 2020, 8: 41805-41813.

[14] CHRISTOPHE F, BERNARD A, COATANEA E. RFBS: a model for knowledge representation of conceptual design

[J]. Cirp Annals-Manufacturing Technology, 2010, 59(1): 155-158.
[15] GONG F M, MA Y H, GONG W J, et al. Neo4j graph database realizes efficient storage performance of oilfield ontology[J]. Plos One, 2018, 13(11): 1-16.
[16] PERCUKU A, MINKOVSKA D, STOYANOVA L. Modeling and processing big data of power transmission grid substation using Neo4j[J]. Procedia Computer Science, 2017, 113: 9-16.
[17] TORRA V. Hesitant fuzzy sets[J]. International Journal of Intelligent Systems, 2010, 25(6): 529-539.
[18] 刘小弟, 朱建军, 张世涛, 等. 考虑属性权重优化的犹豫模糊多属性决策方法[J]. 控制与决策, 2016, 31(2): 297-302.
[19] DU S B, YANG F, TIAN X D. Ancient Chinese character image retrieval based on dual hesitant fuzzy sets[J]. Scientific Programming, 2021, 2021: 1-9.
[20] 徐以聪, 田学东, 李新福, 等. 基于犹豫模糊权重的数学表达式检索[J]. 数据分析与知识发现, 2020, 4(7): 118-126.
[21] VERMA R. Operations on hesitant fuzzy sets: some new results[J]. Journal of Intelligent and Fuzzy Systems, 2015, 29(1): 43-52.
[22] CHE W, LI Z, LIU T. LTP: a Chinese language technology platform[C]// Proceedings of the 23rd International Conference on Computational Linguistics. Beijing: Association for Computational Linguistics, 2010: 23-27.
[23] YAN R, QIU D, JIANG H. Sentence similarity calculation based on probabilistic tolerance rough sets[J]. Mathematical Problems in Engineering, 2021, 2021:1-9.
[24] TIAN X D, WANG J M. Retrieval of scientific documents based on HFS and BERT[J]. IEEE Access, 2021, 9: 8708-8717.
[25] 翟社平, 李兆兆, 段宏宇, 等. 多特征融合的句子语义相似度计算方法[J]. 计算机工程与设计, 2019, 40(10): 2867-2873, 2884.
[26] 刘继明, 于敏敏, 袁野. 基于句向量的文本相似度计算方法[J]. 科学技术与工程, 2020, 20(17): 6950-6955.
[27] 胡雨晴, 纪明宇, 王晨龙. 基于依存句法的句子相似度计算方法[J]. 智能计算机与应用. 2020, 10(4): 113-118,123.
[28] 吴浩, 艾山·吾买尔, 卡哈尔江·阿比的热西提, 等. 融合词性特征的中文句子相似度计算方法[J]. 计算机工程与设计, 2020, 41(1): 150-155.
[29] CHATTERJEE N, YADAV N. Fuzzy rough set-based sentence similarity measure and its application to text summarization[J]. IETE Technical Review, 2019, 36(5): 517-525.

Complex product design knowledge retrieval system based on knowledge graph and hesitant fuzzy theory

ZHANG Yufei[*], WANG Hongwei[**], ZHAI Xiang[****], NIU Dongxiao[***], CAO Mengyuan[****]
([*] Advanced Manufacturing Institute, Zhejiang University of Technology, Hangzhou 310023)
([**] Zhejiang University-University of Illinois Urbana-Champain Institute, Zhejiang University, Haining 314400)
([***] College of Economics and Management, North China Electric Power University, Beijing 102206)
([****] State Key Laboratory of Complex Product Intelligent Manufacturing System Technology, Beijing Institute of Electronic Engineering, Beijing 100854)

Abstract

Aiming at the problems of complex product design knowledge retrieval system such as low precision, high requirements for users, large knowledge granularity, a design knowledge retrieval system based on knowledge graph (KG) and fuzzy similarity is developed. Combining reuse requirements and historical design data, a knowledge representation model is designed to establish a design knowledge network. Based on the three-tuple storage form of knowledge graph, a design knowledge retrieval framework is constructed. In addition, a sentence similarity algorithm based on the hesitant fuzzy theory is developed, which mainly fuses the characteristic attributes from the word and sentence. Further, the membership degree functions of attributes are given, respectively. And the algorithm proposed is compared and verified to be better than other retrieval algorithms. Based on the above methods, a design knowledge retrieval system is developed taking diesel engine design knowledge as an example, and the effectiveness and practicability of the system are verified through two types of design knowledge retrieval cases.

Key words: design knowledge, knowledge graph(KG), hesitation fuzzy similarity, knowledge retrieval