计算机系统应用 ISSN 1003-3254, CODEN CSAOBN
Computer Systems & Applications,2018,27(7):272−277 [doi: 10.15888/j.cnki.csa.006439]

E-mail: csa@iscas.ac.cn
http://www.c-s-a.org.cn
Tel: +86-10-62661041

基于石油领域本体的多源信息融合框架①

宫法明, 崔　佳

(中国石油大学 (华东) 计算机与通信工程学院, 青岛 266580)
通讯作者: 崔　佳, E-mail: s_cuijia@126.com

摘　要: 对多源石油数据的分析是一个很复杂的过程, 容易产生语义和语法上的冲突. 通过利用本体在知识表达和自动推理上的优势, 构建了一个基于本体的石油领域多源信息融合框架, 并在该框架的基础上提出基于本体的元素的相似度算法及融合规则, 经实验分析, 能够提高多源石油数据分析的效率.

关键词: 信息融合; 本体; 石油领域; 多源信息

引用格式: 宫法明,崔佳.基于石油领域本体的多源信息融合框架.计算机系统应用,2018,27(7):272–277. http://www.c-s-a.org.cn/1003-3254/6439.html

Ontology-Based Multi-Source Petroleum Information Fusion Framework

GONG Fa-Ming, CUI Jia

(College of Computer &Communication Engineering, China University of Petroleum, Qingdao 266580, China)

Abstract: Analyzing petroleum data from multi-source is a laborious and complex process, which often results in semantic and syntax conflicts. In this study, we take advantage of the knowledge representation and automatic reasoning of the ontology and propose an ontology-based multi-source petroleum information fusion framework. On the basis of this framework, ontology-based element similarity algorithm and fusion rules are proposed. Experimental results demonstrate that such framework improves the analysis efficiency of petroleum multi-source data.

Key words: information fusion; ontology; petroleum; multi-source

随着全球信息化的推进, 石油领域进入了信息爆炸的时代. 大量来源不同的石油数据缺乏统一的表达方式及语义描述, 给数据分析带来了极大的困难. 实现数据重用和信息共享成为石油行业的巨大挑战[1].

数据融合可以把把不同来源、不同角度的数据结合在一起, 并且为用户提供统一的数据接口[2]. 本体通常用来表示领域知识, 解决数据融合过程中的语义异构问题. 因此, 本文提出了一种新型的数据融合框架, 能够实现不同来源石油数据的融合问题. 为解决上述问题, 本文提出了一种新型的数据融合框架, 能够实现多源数据语义上的融合. 该框架是在一个两层本体结构的基础上实现的. 如图 1 所示, 框架分为 4 层: 源数据层、本体层、融合层和用户层. 源数据层包含来自不同数据源的数据, 本体层是能够实现数据语义融合的两层本体结构, 融合层提供了一些冲突数据的融合规则, 用户层则将融合结果展示给用户.

本文提出了一个四级信息融合框架来解决石油领域的信息融合问题. 文章以下部分的组织结构如下: 第一部分介绍了基于本体的融合模型的研究现状; 第二部分提出了本体的语义模型定义及本体间的映射关系; 第三部分介绍了基于本体的融合规则; 第四部分通过实验证明了该融合框架的可行性; 第五部分对全文进行了总结.

① 基金项目: 科技部创新方法工作专项资助 (2015IM010300); 北京市重点实验室开放课题 (BKBD-2017KF07)
Foundation item: Special Fundation for Innovation Methods of Ministry of Science and Technology (2015IM010300); Open Project of Beijing Key Laboratory (BKBD-2017KF07)
收稿时间: 2017-11-26; 采用时间: 2017-12-15; csa 在线出版时间: 2018-06-27

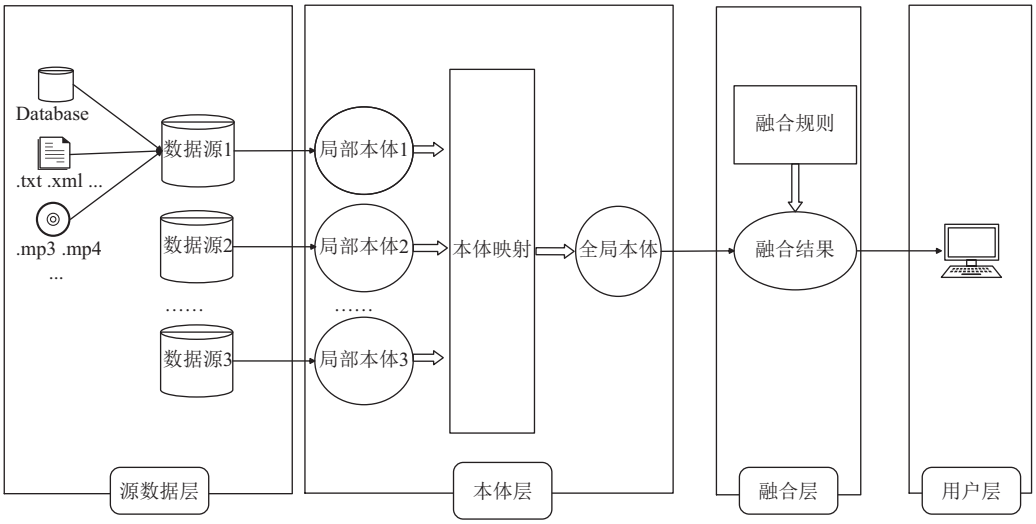

图 1 四层融合框架图

1 研究现状

本体通过领域中的概念及概念之间的关系来表示领域知识[3]. 鉴于本体较强的语义表达和语义推理的能力, 很多研究人员利用本体来解决语义异构问题.

一般来说, 运用本体的方式有三种模式: 单本体模式、多本体模式和混合模式[4]. 单本体模式是通过一个全局本体提供的词汇表表示语义. 这种模式是简单地把所有信息源都和一个全局本体建立起映射关系. 由 Arens 提出的 SIMS 模型[5]就是运用了单本体的方法. 但是单本体模式只适用于所有信息源都是从同一角度描述的情况. 多本体模式突破了这种限制, 每个信息源都有相应的本体与之对应. OBSERVER 模型[6]就是一种基于多本体模式的模型, 通过不同的本体来表示不同数据源的语义, 但是本体之间关系定义困难, 而且在需要添加信息源的时候, 需要添加所有旧本体与新添加的本体之间的映射关系. 为了解决上述两种模式的缺点, Cheng[7]和 Wache 等[8]提出了混合模式, 在多本体的基础上构建了一个全局共享词典, 将不同的本体通过共享词典联系起来. 混合模式的优势在于当需要添加新的信息源的时候, 本体与共享词典之间的映射不需要改动. Visser[9]提出了可以用一个全局本体来代替全局共享词典. 本文提出的模型就是基于混合模式的融合模型.

许多研究人员都在以上三种模式的基础上进行了研究. 赵春江等人[3]提出了一种混合本体结构, 实现了自上而下的融合. 徐赐军等[10]提出了一个基于本体和元数据库的知识融合模型. Boury-Brisset[11]利用本体化方法, 实现了高级别的信息融合, 并将其应用在军事规划领域. 谢能付[12]也在农业领域做了借助本体进行农业信息融合的相关研究, 并且提出了一个针对 Web 信息的只是融合框架[13]. 易善桢等人提出了一种用于数据融合估计的目标地理实体模型和基于图形的本体方法[14]. 王远等人[1]利用全局本体实现了飞机故障数据融合, 对多源数据进行了统一具体的描述. Pai 等人[15]用本体的语义网技术融合军事信息并解决军事中的态势感知问题. 李晓丽等人[16]提出了一个 JDL 模型能够实现一级和二级信息的融合.

本文基于混合模式提出了一种石油领域的融合模型, 能够解决石油信息的语义异构问题, 实现信息融合.

2 两层本体结构

基于混合模式, 我们采用一个两层本体的结构: 全局领域本体和局部本体.

局部本体对应的数据源有不同的存储模式, 例如, 关系数据库, RDF 和结构化、非结构化的数据文件等. 局部本体可以实现信息源内部的语义异构问题, 但是不同的局部本体之间仍然可能存在语义异构. 因此, 需要全局本体来解决上述问题.

全局本体是对领域整体的全局语义定义, 能够为数据融合提供公共的语义描述[17].

2.1 石油本体语义模型

本体是对知识的概念化描述, 包含一系列的领域概念和概念间的关系. 石油领域本体可以定义为 $PO=\{POname, C_s, R_s\}$. 其中 $POname$ 表示本体名, C_s 表示概念集, R_s 表示关系集. 则本体中的对象 e 可以定义为一个四元组 $e=(C, A_s, R_s, I_s)$.

(1) C表示e的概念

(2) A_s是e的属性集, $A_s=\{a_1,a_2,\cdots,a_n\}$

(3) R_s表示关系集, $R_s=\{r_1,r_2,\cdots,r_n\}$. 例如, 是一种、是一部分等.

(4) I_s是e的实例集, $A_s=\{a_1,a_2,\cdots,a_n\}$. 实例集表示在该元素的概念下真实存在的实例集和.

2.2 本体间的映射

局部本体通常是由不同的人员建立的, 因此语义异构很难避免. 建立全局本体与局部本体之间的映射关系就是为了找出本体间的语义联系.

2.2.1 本体间的映射

本体映射的定义如下:

定义 1. 给定两个本体O_1, O_2, 给定两个本体元素e, e'满足$e\in O_1$, $e'\in O_2$. 定义$S(e,e')$为e和e'的相似度, τ为相似度阈值. 那么当$S(e,e')>\tau$的时候就认为e和e'是语义相似的, 并且可以通过某种映射关系M相互转化, $M(e)=e'$. 这个过程就称为本体映射.

在本文中我们只考虑一对一的映射关系. 映射关系如图 2 所示.

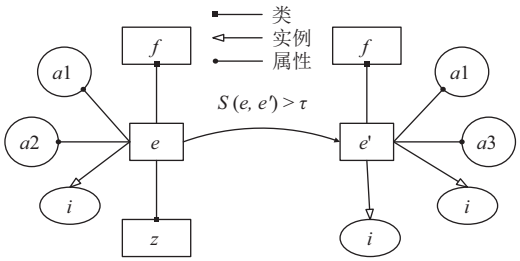

图 2 本体元素映射图

为了计算元素间的相似度关系, 我们依据 Ehrig 等人[18]对相似度度量的定义提出了一个相似度函数$Sim(x,y)$, 定义如下:

定义 2. 元素x, y之间的相似度用$Sim(x,y)$函数计算, 其中$Sim(x,y)\in[0,1]$, 1 表示完全相似, 0 反之.

2.2.2 基于本体元素的相似度算法

本体映射的关键就是不同的局部本体元素间相似度的计算. 本体元素的相似度可以分成四部分: 概念相似度、关系相似度、属性相似度和实例相似度.

令l和l'分别为元素e和e'的概念, 则概念相似度$S_C(e,e')$可以用公式 (1) 来计算.

$$S_C(e,e')=\begin{cases}1.0, & l=l'\\ 0.95, & l'\in syn(l)\\ 0.0, & l'\in ant(l)\\ \dfrac{card(cut(l)\cap cut(l'))}{\max(card(cut(l)),card(cut(l')))}, & \text{其他}\end{cases} \tag{1}$$

其中$cut(l)$表示l的同义词, $ant(l)$表示l的反义词. 同义词和反义词的定义可以借助普林斯顿大学提出的 WordNet[19], 或者重新建立一个石油领域的词表. 函数$card(x)$是用来计算集合x中元素的个数. $cut(l)$通过分词标志将l分成几个基本词. 当$l=l'$时, 相似度为 1; 当l'是l的同义词时, 相似度为 0.95; 当l'是l的反义词时, 相似度为 0; 当l'既不是l的同义词也不是l的反义词时, 相似度用l'和l中基本词的重叠率计算.

例如, 计算$S_C(PlatNO,PlatID)$. $PlatNO$可以被分成 {"$Plat$","NO"}, $PlatID$可以被分成{"$Plat$","ID"}.

假设元素e和e'属于本体O和O', 令P和P'分别为e和e'的父集, 则P和P'的相似度可以用公式 (2) 来表示.

$$S_R^P(e,e')=\begin{cases}0.0, & P=\Phi\text{或}P'=\Phi\\ \dfrac{card(P\cap P')}{\max(card(P),card(P'))}, & \text{其他}\end{cases} \tag{2}$$

如果有且只有一个父集为空, 那么相似度为 0, 其他情况也用重叠率来计算.

类似地, 子集的相似度计算方法与父集相似度一样, 用$S_R^C(e,e')$表示, 则关系相似度的计算方法如公式 (3) 所示.

$$S_C(e,e')=\begin{cases}0.0, & S_R^P=0\text{ and }S_R^C=0\\ S_R^P(e,e'), & S_R^C=0\text{ and }S_R^P\neq 0\\ S_R^C(e,e'), & S_R^P=0\text{ and }S_R^C\neq 0\\ \omega_1S_R^P+\omega_2S_R^C, & \text{其他}\end{cases} \tag{3}$$

如果父集相似度和自己相似度都为 0, 那么关系相似度为 0; 如果只有一个为 0, 那么关系相似度用不为零的那个相似度表示; 如果两个都不为 0, 分别给父集相似度和自己相似度添加了两个权重ω_1和ω_2, 一般来说, $\omega_1+\omega_2=1$且$\omega_1=\omega_2$.

元素属性可以被分为四种: 整数、浮点数、字符和日期. 我们用一个相似度矩阵[20]来计算各种数据类型之间的相似度, 如表 1 所示.

令$a=\{a_1,a_2,\cdots,a_m\}$和$a'=\{a_1',a_2',\cdots,a_n'\}$分别为$e$和$e'$的两个属性集, 那么两个属性元素$a_i$和$a_j'$之间的相似度可以表示为公式 (4).

$$S_a\left(a_i,a_j'\right)=\theta_{a_ia_j'}S_C\left(a_i,a_j'\right) \tag{4}$$

表 1 相似度矩阵

	整数	浮点数	字符	日期
整数	1	0.9	0.1	0.8
浮点数	0.9	1	0.1	0.7
字符	0.1	0.1	1	0.1
日期	0.8	0.7	0.1	1

假设$m>n$, 对于任意的$a_i\in a$都能计算出一个$\max\{S_a(a_i,a_k'),k\in[0,j]\}$, 那么属性相似度$S_A(e,e')$就可以由公式 (5) 来计算.

$$S_A(e,e')=\sum_0^i S_{\max_i}/i \tag{5}$$

令I和I'分别是e和e'的实例集, 那么实例相似度的计算方法如公式 (6) 所示.

$$S_I(e,e')=\begin{cases}0.0, & I=\Phi\text{或}I'=\Phi\\ \dfrac{card(I\cap I')}{\max(card(I),card(I'))}, & \text{其他}\end{cases} \tag{6}$$

最终, 元素e和e'的相似度可以表示为公式 (7).

$$S(e,e')=\theta_cS_C+\theta_rS_R+\theta_aS_A+\theta_iS_I \tag{7}$$

其中θ_c为概念相似度的权值, θ_r为关系相似度的权值, θ_a为属性相似度的权值, θ_i是实例相似度的权值, 并且$\theta_c+\theta_r+\theta_a+\theta_i=1$.

3 基于本体的融合规则

如果不同本体中描述同一个实例的属性值不同, 那么就会反馈给用户不一致的结果. 为了解决这个问题 Motro 等人[21]提出了 5 条解决方法.

(1) 混合结果. 将所有的结果以集合的形势反馈给用户.

(2) 排序结果. 就是在混合结果的基础上, 按照用户的需求进行排序.

(3) 更优结果. 取排序结果中靠前的一个或者几个结果反馈给用户.

(4) 随机结果. 从混合结果集合中随机选取一个.

(5) 融合结果. 将结果集中的所有结果融合成一个.

显然, 融合结果更符合用户的需要. 基于此, 我们提出了一些融合规则, 关键的融合规则定义如下.

定义 3. 多数优先规则. 给定一个结果集$V_d=\{v_1',v_2',\cdots,v_n'\}$, 多数优先规则满足:

$$Majr(V_d)=v_i',\forall v_j\in V_d,v_j\neq v_i',Num\left(v_j\right)\leqslant Num\left(v_j'\right) \tag{8}$$

数优先规则认为出现次数多的那个结果可信度大.

定义 4. 高可信度优先规则. 给定一个结果集$V_d=\{v_1',v_2',\cdots,v_n'\}$, 高可信度优先规则可以表示为:

$$Highcon(V_d)=v_i',\forall v_j\in V_d,v_j\neq v_i',Conf\left(v_j\right)\leqslant Conf\left(v_j'\right) \tag{9}$$

$Conf\left(v_j\right)$表示提供v_j的数据源的可信度, 可信度越高的信息源提供的结果准确度越高.

定义 5. 平均数规则. $V_d=\{v_1',v_2',\cdots,v_n'\}$表示结果集, 平均数规则定义为:

$$Avg(V_d)=\sum_1^n v_i'/n \tag{10}$$

定义 6. 闭区间规则. $V_d=\{v_1',v_2',\cdots,v_n'\}$表示结果集, 闭区间规则定义为:

$$CInt(V_d)=[v_0,v_1],\forall v_i'\in V_d,v_i'\in[v_0,v_1] \tag{11}$$

闭区间规则只适用于属性值为数字类型的情况.

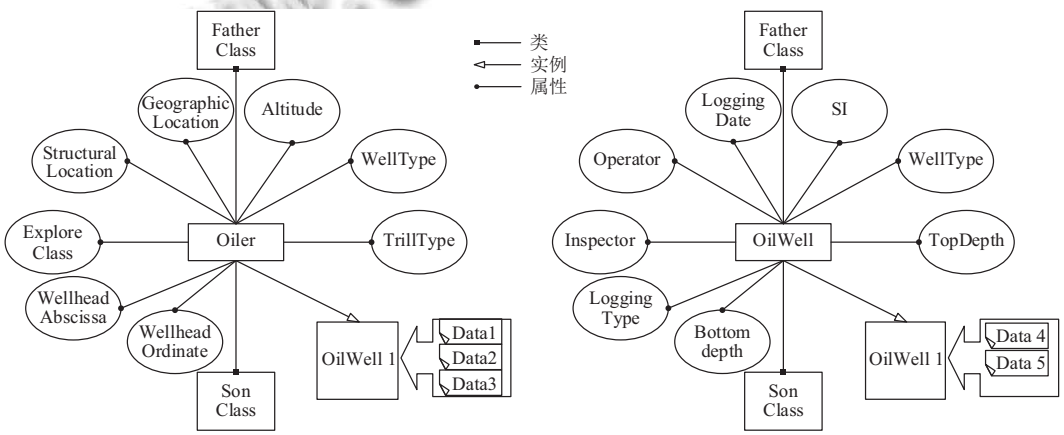

图 3 局部本体中的“油井”元素

4 实验

根据本文提出的框架, 我们开发了一个石油信息融合系统, 并对两个局部本体和一个全局本体进行了实验. 两个局部本体中关于其中关于油井的描述如图 3 所示, 全局本体中的描述如图 4 所示.

由图可见, “Oiler”和“OilWell”都表示“油井”, 但是它们在本体中的呈现形式是不同的, 或者说这两个本体是从不同的角度来描述的“油井”. 运用提出的相似度算法和融合规则, 可以对这两个本体进行融合, 融合之后的结果如图 5 所示.

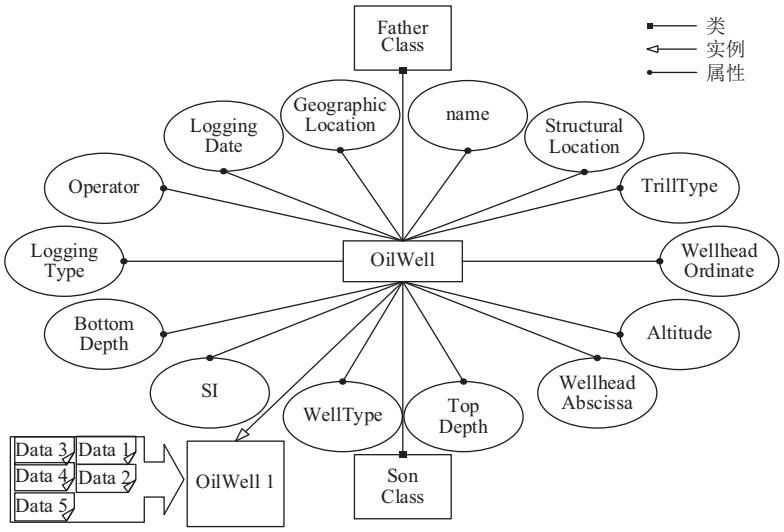

图 4 全局本体中的“油井”元素

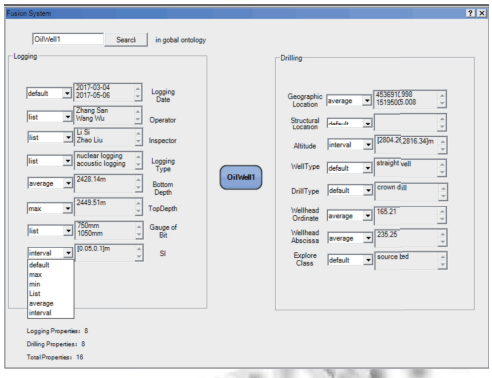

图 5 融合结果图

由实验可以看出, 本文提出的双层本体结构及相似度算法和融合规则能够较精确地实现石油领域的数据融合问题.

5 结论

随着石油行业的发展, 石油领域的数据越来越复杂, 数据融合技术可以更好地分析并使用这些数据. 本文提出了一个基于本体的数据融合框架, 能够解决多源数据的语义异构问题. 本文的融合框架是在石油领域本体的基础上提出的, 但是至今石油领域还没有建立起一个权威的本体. 自动化地构建石油领域的本体将是今后的研究重点.

参考文献

1 Wang Y, Li Q, Sun Y, *et al*. Aviation equipment fault information fusion based on ontology. Proceedings of 2014 International Conference on Computer, Communications and Information Technology. Beijing, China. 2014.
2 Saranya K, Hema MS, Chandramathi S. Data fusion in ontology based data integration. International Conference on Information Communication and Embedded Systems. Chennai, India. 2014.
3 Zhao CJ, Wu HR, Gao RH. Ontology-based multimode information fusion method. 2011 IEEE International Conference on Cloud Computing and Intelligence Systems. Beijing, China. 2011. 55–59.
4 Wache H, Voegele T, Visser U, *et al*. Ontology-based integration of information. Proceedings of ACM Sigmod, 1996, 17(3): 434–437.
5 Arens Y, Hsu CN, Knoblock CA. Query processing in the SIMS information mediator. Tate A. Advanced Planning Technology. Menlo Park, California: AAAI Press, 1996.
6 Mena E, Illarramendi A, Kashyap V, *et al*. OBSERVER: An approach for query processing in global information systems based on interoperation across pre-existing ontologies.

Distributed and Parallel Databases, 2000, 8(2): 223–271. [doi: 10.1023/A:1008741824956]
7 Goh CH, Madnick SE. Representing and reasoning about semantic conflicts in heterogeneous information systems. Cambridge, MA, USA: Massachusetts Institute of Technology, 1997.
8 Wache H, Scholz T, Stieghahn H, *et al*. An integration method for the specification of rule-oriented mediators. Database Applications in Non-Traditional Environments. Kyoto, Japan. 1999.
9 Visser U, Stuckenschmidt H, Wache H, *et al*. Enabling technologies for interoperability. Workshop on the 14th International Symposium of Computer Science for Environmental Protection. Bonn, Germany. 2000. 35–46.
10 徐赐军, 李爱平, 刘雪梅. 基于本体的知识融合框架. 计算机辅助设计与图形学学报, 2010, 22(7): 1230–1236.
11 Boury-Brisset AC. Ontology-based approach for information fusion. Proceedings of the Sixth International Conference of Information Fusion. Cairns, Queensland, Australia. 2003.
12 Xie NF, Cao CG, Guo HY. A knowledge fusion model for Web information. IEEE/WIC/ACM International Conference on Web Intelligence. Compiegne, France. 2005.
13 Xie NF. Research on agricultural ontology and fusion rules based knowledge fusion framework. Agricultural Science & Technology, 2012, (12): 2638–2641.
14 Yi SZ, Shen H, Xiao YF. Ontologies and uncertainty in multi-sources geographical data fusion estimation. Proceedings of the 22nd International Conference on Geoinformatics. Kaohsiung, China. 2014.
15 Pai FP, Yang LJ, Chung YC. Multi-layer ontology based information fusion for situation awareness. Applied Intelligence, 2017, 46(2): 285–307. [doi: 10.1007/s10489-016-0834-7]
16 Li XL, Li WH, Qu YM. Towards ontology–based battlefield information fusion framework for decision support. Proceedings of 2016 International Conference on Computer Science and Electronic Technology. Zhengzhou, China. 2016.
17 Perez AG, Benjamins VR. Overview of knowledge sharing and reuse components: Ontologies and problem-solving methods. Proceedings of the IJCAI-99 workshop on Ontologies and Problem-Solving Methods. Stockholm, Sweden. 1999.
18 Ehrig M, Haase P, Hefke M, *et al*. Similarity for ontologies–a comprehensive framework. European Conference on Information Systems. Regensburg, Germany. 2005.
19 Miller GA. WordNet: A lexical database for English. Communications of the ACM, 1995, 38(11): 39–41. [doi: 10.1145/219717.219748]
20 Cholvy L, Hunter A. Merging requirements from a set of ranked agents. Knowledge-Based Systems, 2003, 16(2): 113–126. [doi: 10.1016/S0950-7051(02)00078-3]
21 Motro A, Anokhin P, Acar AC. Utility-based resolution of data inconsistencies. 2004 International Workshop on Information Quality in Information Systems. Paris, France. 2004.