

电力数据语义冲突检测及消减技术研究

张德刚^{1),2)}, 吴毅¹⁾, 张德海³⁾, 张少泉¹⁾,

¹⁾ (云南电力试验研究院(集团)有限公司电力研究院, 云南昆明 650217)

²⁾ (西北工业大学自动化学院, 陕西西安 710072)

³⁾ (云南大学软件学院, 云南昆明 650091)

摘要: 随着电力企业信息化的加快, 企业内部各种信息系统间的异构数据逐渐增多, 如何有效的消除这些异构数据之间的语义冲突成为一个热点问题。本文根据电力数据语义冲突的具体情况, 提出了一个冲突检测及消减模型, 将电力企业内部的异构数据源进行融合, 有效地解决电力企业内部数据源语义冲突问题。使用结果表明, 该模型可以有效消解异构数据源之间的语义冲突, 为电力企业内部及企业间异构数据的融合和共享提供有力的支持。

关键字: 语义冲突, 冲突检测, 冲突消解

中图分类号: TP319;

文献标识码: A

Detection and resolution of electrical data semantic confliction

ZHANG De-Gang^{1,2}, WU Yi¹, ZHANG De-Hai³, ZHANG Shao-Quan¹

¹⁾(The Electric Power Research Institute, Yunnan Electric Power Test & Research Institute (Group) Co., Ltd. Kunming, Yunnan 650217, China)

²⁾(School of Automation, Northwestern Polytechnical University, Xi'an 710072, China)

³⁾(School of Software, Yunnan University, Kunming, 650091, China)

Abstract: With the acceleration of the construction of the electrical enterprise informatization, heterogeneous data between various information systems gradually increased. In order to effectively eliminate the semantic contradiction between these heterogeneous data, this paper proposes a model for conflicts detection and data mitigation within the enterprise, the application result shows that this model can effectively eliminate semantic conflicts between heterogeneous data sources, and strong support the integration and sharing of heterogeneous data.

Key words: semantic conflict; conflict detection; conflict resolution

1. 引言

大量的信息化管理系统被广泛应用于电力企业的各项业务中, 由于信息化阶段的差异, 不同业务部门的信息系统开发时间和开发单位往往不同, 这就导致这些数据源彼此独立相互封闭, 且相互异构, 使得企业难以综合分析利用内部数据, 影响企业决策。数据异构一般可分为四类: 系统异构、语法异构、结构异构和语义异构^[1,2]。早期常见的解决方案是采用多数据库系统或者联邦数据库系统^[3]较好的解决了数据的系统、语法和结构异构问题, 此后大量出现的基于 XML 的异构数据集成技术研究^[3-5], 进一步有效的解决了语法异构问题, 但是这几类方法都无法解决数据的语义异构问题。

基金项目: 国家自然科学基金(61263043); 云南省教育厅基金(2011Z020);

张德刚(1982—), 男, 云南临沧人, 博士, 研究方向: 企业信息化、大数据分析, E-mail: degangzhang.csg@gmail.com;

吴毅(1964—), 男, 云南昆明人, 高级工程师, 硕士, 研究方向: 企业信息化和企业管理, E-mail: wuyi003@im.yn.csg;

张德海(1977—), 男, 云南临沧人, 副教授, 博士, 研究方向: 知识工程、云计算和大数据分析, E-mail: dhzhang.ynu@gmail.com;

张少泉(1964—), 男, 云南个旧人, 高级工程师, 硕士, 研究方向: 自动控制, E-mail: zsq3733@163.com;

“本体”(Ontology)能够捕获相关领域的知识, 提供对该领域知识的共同理解, 确定该领域内共同认可的词汇, 并从不同层次的形式化模式上给出这些词汇(术语) 和词汇间相互关系的明确定义, 以明确的形式化规范来描述共享概念模型, 通过本体映射在异构数据间建立语义匹配, 可以有效解决语义异构问题, 适合作为数据集成中的通用语义模型来描述异构数据语义^[6-8]。

本文通过分析电力数据语义冲突的具体情况, 利用领域知识构建领域本体, 以领域本体为基础来构建电力异构数据语义冲突检测和消解模型, 有效地解决企业内部数据源语义冲突问题, 为企业决策层综合分析内部数据, 提高决策准确性提供基础。

2. 电力数据语义冲突分析

语义冲突是指不同数据源中采用不同的语义解释数据的意义。形成语义冲突的原因[5]包括: (1) 同一个语义概念使用多种术语表示; (2) 同一名称在不同数据源中表达不同的含义; (3) 相同(或相似)的信息采用不同的结构来表示; (4) 各个分布、自治性数据源中的概念之间存在着各种联系。电力数据中存在的语义冲突包括物理冲突、表冲突、字段冲突、记录冲突等四类, 其中最常见的是字段冲突[8]。

(1) 物理冲突是指数据源的存储格式不同引起的冲突。比如: SCADA 数据是 TXT 格式的非结构化数据, 而电网模型数据是 XML 格式的半结构化数据。

(2) 表冲突是指相同的表名在不同的数据库表示两个不同的实体。例如, 主网和配网数据库都有 Location 表, 但它们只有 61 个共同的字段, 主网特有的字段为 125 个, 而配网特有的字段为 91 个。

(3) 字段冲突包括命名冲突、类型冲突、长度冲突、精度冲突、计量单位冲突、表达方式冲突。例如, asset 表出现在主网和配网数据库中, 但“施工单位”的字段名(英文)在两个表中的含义不同, 而且, 在配网的 asset 表中, 该字段出现了两次, 长度定义一个为 50, 一个为 80, 其英文字段名分别为 YDTF_SGDW, YDPS_SGDW。在不同数据库中, 字段冲突还表现在命名相同, 但表示的信息不同, 例如图 2 所示, 字段 DEPARTMENT_CODE 在数据库中, 代表了创建者 ID、所属部门编码, 施工单位编码, 部门编码和建设单位编码等多种情况。

创建者部门ID	DEPARTMENT_CODE	VARCHAR2(36)	36	创建者部门ID	安全技术交底。
所属部门编码	DEPARTMENT_CODE	VARCHAR2(20)	20	所属部门编码	工种基础信息库。
施工单位编码	DEPARTMENT_CODE	VARCHAR2(20)	20	施工单位编码	施工单位持证情况。
部门编码	DEPARTMENT_CODE	VARCHAR2(20)	20	部门编码	进度计划资源表。
单位编码	DEPARTMENT_CODE	VARCHAR2(36)	36	单位编码	统计及综合评价。
建设单位编码	DEPARTMENT_CODE	VARCHAR2(36)	36	建设单位编码	业主项目部。

图 1. 字段代码相同所表示信息不同

(4) 记录冲突
指描述同一数据的数据记录不同, 因计量单位冲突引起的数值不同。例如, 配网数据库中部分字段名采用汉语拼音的缩写, 但在主网数据库中的对应字段又采用汉语命名。

3 基于本体的语义映射与冲突消解

3.1 语义冲突检测与消解模型

为了消除数据库中的语义冲突, 本文通过领域知识构建局部本体库, 然后在此基础上构建各个本体库之间的语义映射和全局本体库, 消解数据库中的语义冲突, 其处理模型如图 2 所示。

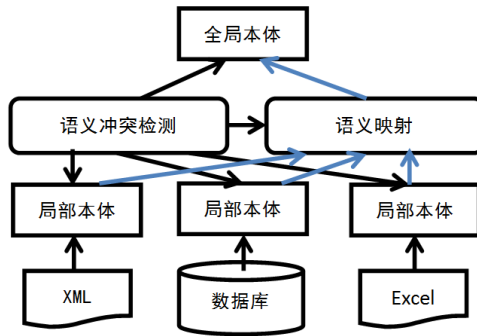


图 2. 基于本体的数据源语义冲突检测与消解模型

在图 2 中，将来自于不同数据源的数据通过领域知识分别构建成多个局部本体，这些局部本体中包含了对应数据源的表、列等信息的映射。根据本体之间的映射关系发现局部本体之间的语义冲突，并通过添加局部本体与全局本体之间的映射来消解该冲突。

3.2 语义冲突的消解

对于表冲突，本文通过局部本体到数据源的映射过程，对实例数据表添加映射规则，在数据库中唯一的标识一个字段，解决同一数据库中不同表之内的字段重复的情况和不同数据库中的不同表重名的情况，抽象出来的新的表名对应的类被加到全局本体中。具体是在原有表结构中增加“所属表格”和“所属数据库”两项，在不同表同名字段和不同数据库同名表中可以据此确定字段所属具体数据库。

对于字段冲突和记录冲突的问题，由于数据字典中的代码和名称字段不是唯一的，因此索引值是相同的数据所表达的意思却不同，如图 3 所示。为了解决这一个问题，本文将不同语义的部分查找出来，进行分类，然后重新命名，以此来消解语义上的冲突，做到唯一查找。

创建人	CREATE_NAME	VARCHAR2(20)	20	创建人	进度计划模板.
创建人	CREATE_NAME	VARCHAR2(20)	20	创建人	里程碑进度明细模板.
创建人	CREATE_NAME	VARCHAR2(20)	20	创建人	里程碑进度模板.
	CREATO	VARCHA			

图 3. 名称和代码都相同的字段示例

以图 3 所示字段为例，首先分别以代码和名称为索引查找出语义冲突并存入文本文档中，然后以代码部分的语义冲突集合为标准，将名称部分的语义冲突集合中与其有交集部分删除，确保语义冲突只被记录一次且包含了所有的语义冲突。然后将重复的字段读出，在数据库中找到与它相同的数据，对这些数据的格式进行调整并分类，语义相同的归为一类，采用相同的标识符表示，并在原有表中增加“解析后名称”和“解析后代码”两个字段，以达到消除语义冲突的目的，修改后的映射表结构如表 1 所示，得到的映射表如图 4 所示。

表 1 实现语义冲突消解的映射表头结构

名称	代码	数据类型	长度	备注	所属表格	解析后名称	解析后代码
解析后中文	原本英文	数据类型	数据长度	备注	所属表格	原本中文	解析后中文
<0>CREATOR_ID	CREATOR_ID	VARCHAR2(20)	20	创建者ID	价格发布信息.	创建者ID	创建者ID<0>
<1>CREATOR_ID	CREATOR_ID	VARCHAR2(100)	100	创建者ID	开工申请明细流水号	创建者ID	创建者ID<1>
<0>CREATOR	CREATOR	VARCHAR2(200)	200	创建者	价格发布信息.	创建者	创建者<0>
<1>MODIFIER_ID	MODIFIER_ID	VARCHAR2(100)	100	修改者ID	工程项目开工申请表.	修改者ID	修改者ID<1>
<1>MODIFIER_ID	MODIFIER_ID	VARCHAR2(100)	100	修改者ID	开工申请明细流水号	修改者ID	修改者ID<1>
<1>MODIFIER_ID	MODIFIER_ID	VARCHAR2(100)	100	修改者ID	开工申请文件明细模板表.	修改者ID	修改者ID<1>
<1>MODIFIER_ID	MODIFIER_ID	VARCHAR2(100)	100	修改者ID	竣工验收申请单.	修改者ID	修改者ID<1>

图 4. 消解冲突后的映射表

经过上述操作可以消解相同域内的数据语义冲突,为了将不同域中消解冲突后的映射表合并,在表上增加所属数据库信息,用以区分字段所属的域,新的映射表的结构如图 5 所示。

A	B	C	D	E	F	G	H	I
解析后英文	原本英文	数据类型	数据长度	备注	所属表格	原本中文	解析后中文	所属数据库结构, 类型
《0》SAFE_APPL\SAFE_APPL	SAFE_APPL	VARCHAR2(36)		安全风险	安全风险评估范围	安全风险	安全风险评估申请流水号	管理表英文映射表
《0》SAFE_APPL\SAFE_APPL	SAFE_APPL	VARCHAR2(36)		安全风险	安全风险评估申请	安全风险	安全风险评估申请流水号	管理表英文映射表
《0》SAFE_APPL\SAFE_APPL	SAFE_APPL	VARCHAR2(36)		安全风险	安全风险评估范围	安全风险	安全风险评估申请流水号	管理表英文映射表
《0》FBS_ID	FBS_ID	VARCHAR2(36)		概算流水	安全风险评估范围	概算流水	概算流水号ID《0》	管理表英文映射表
《0》FBS_ID	FBS_ID	VARCHAR2(36)		概算流水	安全风险评估范围	概算流水	概算流水号ID《0》	管理表英文映射表
《0》FBS_ID	FBS_ID	VARCHAR2(36)		概算流水	安全风险评估范围	概算流水	概算流水号ID《0》	管理表英文映射表

图 5. 整合不同域的总语义冲突映射表示例

在数据表中,由于历史原因导致部分数据名称为空,此类则根据专家知识库分析和判断出其正确的语义,实现语义冲突消解。

4. 数据冲突消解器实现

根据上文所描述的数据语义冲突消解方法,本文实现了电力数据冲突消解器,方便用户在实际的工作中消解新增数据的语义冲突,其界面如图 6 所示。

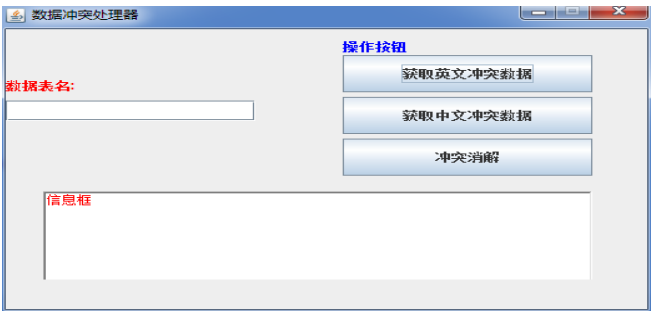


图 6. 数据语义消解器界面

其中,数据表名负责接收想要处理的数据表的表名,信息框负责提示用户所进行的操作是否成功及产生的结果。获取英文冲突按钮实现处理冲突表的英文冲突部分,并生成新的英文处理表。获取中文冲突按钮实现处理冲突表中文冲突部分,并生成新的中文处理表。冲突消解按钮实现将前面生成的两张表合并起来,并消除两者相同的部分,生成综合性的冲突消解表。

点击‘获取英文冲突数据’,获取英文的冲突消解后的数据,信息框提示创建出新的表格,生成的新表如图 7 所示。

yingwengai	yingwen	lexing	changdu	zhushi	suoshubiao	zhongwen	zhongwengai
《0》APNAME	APNAME	UPPER	20	WORKORDER (工单表)	应用程序Appname	应用程序Appname	应用程序Appname《0》
《1》APNAME	APNAME	UPPER	20	WORKTICKET (工作票)	应用程序名称Appname	应用程序名称Appname	应用程序名称Appname《1》
《0》ASSETNUM	ASSETNUM	UPPER	30	ASSET (资产表)	资产号	资产号	资产号《0》
《1》ASSETNUM	ASSETNUM	UPPER	30	WORKORDER (工单表)	确定资产。	确定资产。	确定资产。《1》
《0》ASSETTYPE	ASSETTYPE	UPPER	15	ASSET (资产表)	该资产的预定义类型。	该资产的预定义类型。	该资产的预定义类型。《0》
《1》ASSETTYPE	ASSETTYPE	ALN	200	MONTHLYPLAN (月度计划)	设备类别	设备类别	设备类别《1》
《0》AUTOWOGEN	AUTOWOGEN	YORN	1	ASSET (资产表)	指示当计量器频率已经达到资产值时,是否启动wogen进程的标志。	指示当计量器频率已经达到资产值时	指示当计量器频率已经达到资产值时
《1》AUTOWOGEN	AUTOWOGEN	YORN	1	LOCATIONS (位置表)	指示当计量器频率达到某个位置时,是否启动 wogen 进程的标志。	指示当计量器频率达到某个位置时,	指示当计量器频率达到某个位置时,
《0》CALNUM	CALNUM	UPPER	8	ASSET (资产表)	日历代码	日历代码	日历代码《0》
《1》CALNUM	CALNUM	UPPER	8	LOCATIONS (位置表)	CALNUM	CALNUM	CALNUM《1》
《0》CHANGERY	CHANGERY	UPPER	20	ASSET (资产表)	设备类别	设备类别	设备类别《0》

图 7. 点击‘获取英文冲突数据’后生成的新表

点击‘获取中文冲突数据’,获取中文的冲突消解后的数据,信息框提示创建出新的表格,生成的新表如图 8 所示。

yingwengai	yingwen	leiweng	changdu	zhushi	suoshubiao	zhongwen	zhongwengai
«0» CLASSSTRUCTUREID	CLASSSTRUCTUREID	UPPER	20		LOCSYSTEM (位置系统表)	Classstructureid	Classstructureid «0»
«0» CLASSSTRUCTUREID	CLASSSTRUCTUREID	UPPER	20		SAFETOOLS (安全工具)	Classstructureid	Classstructureid «0»
«0» DESCRIPTION_LONGDESCRIPTION	DESCRIPTION_LONGDESCRIPTION	LONGALN	32000		MONTHLYPLAN (月度计划)	Description_longdescription	Description_longdescription «0»
«0» DESCRIPTION_LONGDESCRIPTION	DESCRIPTION_LONGDESCRIPTION	LONGALN	32000		MONTHLYPROPLAN (生产计划表)	Description_longdescription	Description_longdescription «0»
«0» DESCRIPTION_LONGDESCRIPTION	DESCRIPTION_LONGDESCRIPTION	LONGALN	32000		MONTHLYPLAN (月度计划)	Description_longdescription	Description_longdescription «0»
«0» DESCRIPTION_LONGDESCRIPTION	DESCRIPTION_LONGDESCRIPTION	LONGALN	32000		OPERTICKET (操作票)	Description_longdescription	Description_longdescription «0»
«0» DESCRIPTION_LONGDESCRIPTION	DESCRIPTION_LONGDESCRIPTION	LONGALN	32000		MONTHLYPROPLAN (生产计划表)	Description_longdescription	Description_longdescription «0»
«0» DESCRIPTION_LONGDESCRIPTION	DESCRIPTION_LONGDESCRIPTION	LONGALN	32000		SAFETOOLS (安全工具)	Description_longdescription	Description_longdescription «0»
«0» SENDERSYSID	SENDERSYSID	ALN	50		ASSET (资产表)	ERP集成所用的列 (API)	ERP集成所用的列 (API) «0»
«0» SENDERSYSID	SENDERSYSID	ALN	50		ASSET (资产表)	ERP集成所用的列 (API)	ERP集成所用的列 (API) «0»
#1% SENDERSYSID	SENDERSYSID	ALN	50		LOCATIONS (位置表)	ERP集成所用的列 (API)	ERP集成所用的列 (API) «0»

图 8. 点击‘获取中文冲突数据’后生成的新表

通过语义消解器可以发现和消解数据中的语义冲突，得到异构数据的映射表，可以有力支持电力系统异构数据源的集成、进行分析和利用。

5. 总结

本文通过分析电力数据语义冲突的具体情况，利用领域知识构建领域本体，进而构建电力异构数据语义冲突检测和消解模型，将电力企业内部的异构数据源进行融合，有效地解决企业内部数据源语义冲突问题，并实现了电力数据语义冲突消解器，方便新增数据的语义冲突消解。为电力企业内部及企业间异构数据的融合和共享提供有力的支持，为企业决策层综合分析内部数据，提高决策准确性提供基础。

参考文献

- [1] ZHAN C, PAUL O'B. Domain Ontology Management Environment[R]. In the Proceedings of the 33rd Hawaii International Conference on System Sciences, IEEE, 2000.
- [2] BENEVENTANO, D., ORSINI, M. An Ontology-Based Data Integration System for Data and Multimedia Sources. Semantic Computing, 2009. ICSC '09. IEEE International Conference on, 14-16 Sept. 2009.
- [3] WU X, YAO Y. XML-based heterogeneous database integration system design and implementation [C], Computer Science and Information Technology (ICCSIT), 2010 3rd IEEE International Conference on, 9-11 July 2010.
- [4] TADEUSZ P, Agent and Multi-Agent Systems: Technologies and Applications Third KES International Symposium, KES-AMSTA 2009, Uppsala, Sweden, June 3-5, 2009.
- [5] BUCCELLA A, CECICH A, An Ontology Approach to Data Integration. NR Brisaboa Journal of Computer Science and Technology, 2003.
- [6] QIANG L, TAO H, LIU S. An Ontology-Based Approach for Semantic Conflict Resolution in Database Integration, Journal of Computer Science and Technology, March 2007, Volume 22, Issue 2, pp 218-227
- [7] 龚波, 一种基于本体的数据集成用户接口方法研究[D]: 中国科学技术大学, 2009; [GONG B, The Research on Ontology-Based User Interface Method of Data Integration [D], University of Science and Technology of China. 2009]
- [8] 蔚继承, 杨丽丽. 一种基于本体的异构数据集成方法[J]. 信息化研究, 2013, 01:9-14.
- [Wei Jicheng, Yang Li, A Heterogeneous Data Integration Approach Based on Ontolog, informatization Research, 2013, 01:9-14.]



Degang Zhang received his PhD degree in Control Science and Engineering from Northwestern Polytechnical University of China in 2012. He is a joint training post doctor in School of Automation in Northwestern Polytechnical University and Yunnan Electric Power Test and Research Institute (Group) Co.. His research interests are in the areas of Enterprise informatizaion technology, data mining and big data analysis. He published 19 scientific research articles in international publications and 2 are cited by SCI, 13 are cited by EI. He won patent 2.

张德刚(1982)，男，云南省临沧人，2012 年获西北工业大学博士学位，目前是西北工业大学和云南电力研究院联合培养博士后，主要研究领域为企业信息化、数据挖掘和大数据分析。



Yi Wu is Senior Engineer in Yunnan Electric Power Test and Research Institute (Group) Co. of China. He received the Master degree at Zhejiang University. His main research interests are Enterprise informatizaion technology and business management.

吴毅(1964—)，男，云南昆明人，云南电力研究院高级工程师，硕士，主要研究领域为企业信息化和企业管理。



Dehai Zhang is Associate Professor in School of Software at Yunnan University of China. He received the PhD degree in Systems analysis and integration at Yunnan University. His main research interests are: big data analysis, artificial Intelligence, knowledge engineering and cloud computing. He is member of American Institute of Electrical and Electronics Engineers(IEEE), China Association for Artificial Intelligence(CAAI), International Computer Science and Information Technology Institute(IACSIT).

张德海(1977—)，男，云南临沧人，云南大学副教授，博士，主要研究领域为知识工程、云计算和大数据分析。他是 IEEE、中国人工智能学会(CAAI)，国际计算机科学与技术研究所(IACSIT)会员。



Shaoquan Zhang is Senior Engineer in Yunnan Electric Power Test and Research Institute (Group) Co. of China. He received the BA degree at Chongqing University. His main research interests is automatic control.

张少泉(1964—)，男，云南个旧人，云南电力研究院高级工程师，硕士，主要研究领域为自动控制。