数据空间命名实体的集成方法

王江海 吴扬扬

(华侨大学计算机科学与技术学院 厦门 361021)

摘 要 提出了一种数据空间中的命名实体集成模型(NEIM)及其在异质异构数据源中的集成方法。命名实体模型描述了数据源、实体与实体描述间的关系,能够实现从其中任意一个息查询到其它相关信息。命名实体的集成架构指出了数据空间中命名实体集成要完成的主要任务,包括命名实体的识别、实体的集成映射和实体的统一。集成算法描述了数据空间中异构数据源包含的命名实体及其描述信息的集成方法。针对结构化半结构化数据,它采取构建映射规则,使系统可以在后期持续集成这些数据源中的实体信息,实验验证了集成方法的构建映射规则的有效性。

关键词 数据空间,命名实体,集成

中图法分类号 TP39 **文献标识码** A

Integration Method for Named Entities in Dataspace

WANG Jiang-hai WU Yang-yang

(College of Computer Science and Technology, Huaqiao University, Xiamen 361021, China)

Abstract A named entity integration model(NEIM) was proposed for Dataspace, as well as integration methods for named entities of heterogeneous data sources. Named entity integration model describes the relations among data source, named entity and the descriptions of entity. It supports any inquires from one of them to the other relevant information. The framework of named entity integration points out that the main works of the integration are named entity and its information recognition, entity integration and mapping, and entity resolution. The integrated algorithm represents the integration methods of named entity and its information in heterogeneous data sources. Especially, for the structural and semi-structured data, it constructs mapping rules, makes the system can continuous integration. The experiment validates the mapping rules.

Keywords Dataspace, Named entity, Integration

1 引言

计算机的出现及网络技术的发展使得人们之间可以便捷地分享信息,而信息技术的高速发展使得这种信息分享日渐频繁,信息量不断增长。传统的文件管理已被证明在大量数据的有效管理上难以应付。作为文件管理的替代者,数据库技术已在结构化数据管理方面战功显著,然而对于非结构化和半结构化数据的管理上却是力不从心。2005年,在SIGMOD大会上,数据空间(Dataspace)[1]作为一种新的数据管理方式被提出,拉开了数据空间的研究序幕。数据空间作为一种管理异质异构数据的方法,以一种Pay-as-you-go[2-4]的形式对数据源进行管理。区别于关系数据库系统和集成系统,数据空间不需要预先定义数据源模式,即可自动实现数据源的语义映射,从而完成对数据源的管理。同时,数据空间具有自动演化功能,能不断地进化以提高用户查询结果的完备性。到目前为止,人们从模型[5-11]、索引[12-15]、数据关系发现[16-18]等方面对数据空间进行了研究。

本文从另一角度对数据空间中数据源的关系发现进行了研究。用户数据空间中大量的数据从本质上来看是对现实世界中客观存在的实体的描述信息。即这些信息正是对数据源中所存在的命名实体的描述。命名实体(Named Entity, NE)是指现实世界的人名、地名等专有名称和有意义的时间、日期等数量短语,主要包括3大类(实体类、时间类和数字类)和7小类(人名、机构名、地名、时间、日期、货币和百分比)[19]。用户对数据源的访问,主要是对这些实体信息的查询。因而抽取数据空间中的命名实体及其描述信息加以集成,可以使用户更有效地管理和访问他们的数据。传统的命名实体识别技术只识别待分析汉语字符串中的命名实体名,却并不关注与实体相关的信息。在信息抽取领域中有基于本体的信息抽取技术,它是利用领域本体对领域内的数据信息进行抽取。命名实体作为信息描述的主体对象,其描述信息需要随用户的需求而不同,因而基于本体的方法并不适用。本文将传统命名实体识别技术与信息抽取技术结合起来,根据数据空间对海量数据管理的要求,结合项目组提出的刻面描述模型,提出一个数据空间中命名实体及其相关信息的集成模型。在本文中主要针对数据空间中的人名命名实体及其描述信息的集成

到稿日期:2012-03-28 返修日期:2012-05-29 本文受福建省科技计划重大项目(2011H6016),福建省科技计划重点项目(2011H0028)资助。

王江海(1987—),男,硕士生,主要研究方向为要数据库应用技术,E-mail:hayjeans@hqu.edu.cn;吴扬扬(1957—),女,教授,硕士生导师,主要研究方向为数据库技术、智能数据管理与数据挖掘。

进行介绍。

2 数据空间实体集成模型

2.1 命名实体的集成模型

在现实世界中，同一数据源中一般会出许多异名的命名实体及其描述信息。数据空间作为海量数据管理的新方式，包含了用户大量的数据源。因此数据空间、数据源、命名实体及其描述信息之间的关系可以用图 1 表示。从图 1 可以看出，数据源中包含了不同的命名实体；相同的实体名不一定对应现实世界中的同一对象；不同数据源中会存在同名实体；不同实体间会存在同样的实体描述；实体的同一个描述会出现在不同的数据源。因此提出以下的数据空间命名实体集成模型(Named Entity Integration Model，NEIM)。

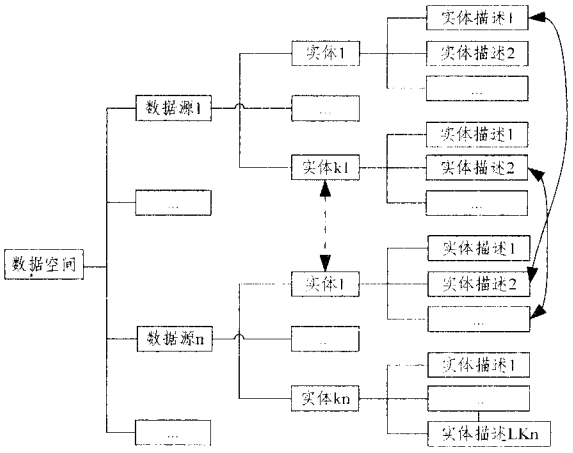

图 1 数据空间、数据源、命名实体及描述信息之间的关系

定义 1 实体指数据空间的数据源中对应于现实世界的命名实体及其相关描述信息集合。其表示为 Entity(id，name，FS，A-VS)，其中，id 表示实体的标识；name 表示实体名称；FS 表示实体的刻面描述集合；A-VS 表示实体描述集合及实体描述间的关系。

定义 2 实体描述是指对数据空间中实体的描述信息。其表示为 Attr(id，name，D)，其中，id 为实体描述 id；name 为实体描述；D 为数据源集合(标识)。

定义 3 数据源指数据空间中存储数据的集合，即对应于数据空间中一般的数据源。其表示为 Datasource(id，E)，id 为数据源标识；E 为实体集合。

在模型中，数据源、实体及实体描述三者的引用关系如图 2 所示。数据源包含若干实体、实体拥有的实体描述、实体描述记录出现的数据源，数据源、实体、实体描述构成一个循环引用。无论是从数据源到实体或实体信息的查询，还是从实体到数据源的查询，都能根据实体描述信息、实体、数据源间的引用完成。

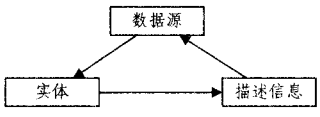

图 2 数据源、实体、实体描述间的引用关系

2.2 人名实体的刻面描述模型

为了实现对人名命名实体的描述信息的集成，本文根据网络中即时聊天工具、BBS、邮件系统对个人信息的收集以及现实生活中人们对个人基本信息的收集，定义了一个基于刻面的人名实体描述模型。人名实体的描述模型以项目组提出的刻面描述模型为基础，将人名实体描述信息划分为 4 个主要刻面：基本信息刻面、联系信息刻面、学历信息刻面和工作信息描述刻面，每个刻面又包含若干属性。人名实体的主要描述模型的主要信息见表 1。

表 1 人名实体的刻面描述信息

基本信息	姓名	直接识别
	性别	根据他，她，与性别描述词(性别，男，女)
	年龄	出生日期，年龄描述词(年龄，X 岁)
	出生日期	
	地点	与人相关的时间地点
	生肖	出生日期，生肖描述词
	血型	血描述词
	星座	出生日期，星座描述词
	身份证号	直接识别
	婚姻	婚姻描述词(＊太太，老公，未婚夫(妻)，爱人)
	政治面貌	直接识别
	家庭	相关家人信息(太太，儿子…)
联系信息	联系地址	直接识别，地点的描述词(省、市、街、路)
	邮箱	XXX@XXX. com
	电话	XXXX－XXXXXXXX
	手机	13 位数字
	主页	
	IM	
	网络昵称	
学历信息	学历	博士，硕士，学士，高中，初中，小学
	学校	＊＊大学、学院
	语言	＊语
	专业	＊系、＊专业
	学习经历	
	论文	判定作者
工作信息	职业	直接识别或专业描述词作分类
	职务	职务描述词
	单位	＊公司，＊学校，＊办公室，…
	时间	
	工作经历	历史工作信息

在表 1 中，部分的人名实体描述信息可以根据其他描述信息确定，如生肖、星座可以根据出生日期确定；有些描述信息由于其唯一性，可被作为区别同名实体是否统一的标准，如身份证号、电话、邮箱等；同时学历信息、工作信息也可以作为计算两个命名实体相似度的因素。

3 数据空间实体的集成框架

本文提出了数据空间命名实体的集成架构(见图 3)。架构主要由 4 部分组成：命名实体的识别、实体描述信息的识别、实体与数据源的映射和实体的统一。其中前两部分可以归结为实体的识别。

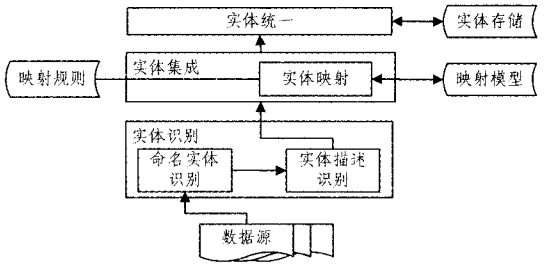

图 3 命名实体的集成框架

实体的识别(包括命名实体与实体描述识别)

在架构中，实体的识别是实现实体集成的先决条件。准

确地识别出数据源中的命名实体，才能保证后续集成的可行性及准确性。只有正确的实体描述识别，才能保证实体信息的真实性与准确性，及实体查询结果的完整性和准确性。实体查询结果的完整性是指所有满足用户查询请求的实体数据都能完整地返回给用户。

实体与数据源映射

实体与数据源映射是建立数据源到实体间的联系的过程。我们研究集成数据源就是为了提升用户查询数据源的效率及查询结果的完整性。实体与数据源映射保证实体从数据源中抽取出来后，能够保持与数据源的联系，保证用户通过实体能够准确地定位数据源。同时，实体到数据源的映射也能通过实体关联数据源的信息找出实体间的关联关系。

实体的统一

在海量数据中，不可避免地会出现同名命名实体。判定这些同名命名实体是否对应于现实世界的同一对象也是必需解决的问题之一。在数据空间的海量异质异构数据中，这一问题更加复杂。由于数据库中数据的结构化，采用数据库中的实体统一(Entity Resolution)技术较容易解决上述问题。在数据空间中，除了结构化数据源，还有半结构化数据和非结构化数据。特别是非结构化的数据，其数据源中的数据表示没有统一的格式，对于文本数据，只是由一组文本串构成。因而在这类数据源中的实体与其它实体进行统一时，不能只采用现有的数据库中的实体统一技术。另一个问题是，在实体统一过程中，实体的合并涉及实体描述信息的来源及合并后实体拆分的难题。

本文主要是对数据空间中命名实体的集成进行论述，因此后文将只涉及到实体集成部分，不再对数据空间中实体的统一进行介绍。

4 数据空间实体集成算法

本节介绍实体集成过程中各类数据源实体集成的主要算法。由于结构化数据源和半结构化数据源的数据结构比较规则，我们提出了先为它们定义数据源到实体的映射规则，再作实体数据集成。为此，我们为结构化和半结构数据源定义了一个数据源映射模型，以方便实现数据源中实体结点到实体模型的映射。映射模型定义如下：

```
〈DataSchema〉
  〈source〉
      〈data name="id"/〉映射 ID
      〈data name="type"/〉数据源类型
      〈data name="sourceId"/〉数据源 ID
      〈data name="key"/〉数据主键列
      〈map〉数据源到实体信息的映射集合
          //实体属性对应的列，元素结点
          〈attr name="名字"〉列，XPath〈/attr〉
          〈attr name=""/〉
      〈/map〉
  〈/source〉
〈/DataSchema〉
```

映射模型通过 XML 定义，最外层 DataSchema 是数据映射模型的根结点，它拥有唯一类型的子结点 source，表示某一数据源到实体的映射模型。映射模型内部结点 data 记录了映射数据源的信息(数据源类型、数据源在数据空间的 Id 等)及 map 结点内实体结点到实体的具体映射规则。

人们存储数据时，通常以具有标示性的标题作为属性的区分标识。算法中使用的实体信息的匹配规则是根据人名实体描述模型中各个属性的特点确定的，需利用结点元素名、属性列名等。因而标题的匹配度计算通过标题进行正则匹配及基于 WordNet 的相似度计算完成。标题的匹配度计算如式(1)所示，其中 sim_w 为标题与预置标题的相似度：

$$\phi_t = \begin{cases} 1, & \text{正则匹配成功} \\ sim_w & \text{正则匹配不成功} \end{cases} \tag{1}$$

内容的匹配度计算，是根据描述信息的特殊格式要求所决定的，如内容的分词标注和正则式匹配等方法确定属性的描述信息，如人名则根据分词工具分词标注识别，电话根据正则式匹配识别等。内容的匹配度计算由三方面构成：分词标识、内含关键字、正则式匹配，其中 ω 表示权重，φ 为 3 个因素的匹配值：

$$\phi_c = \sum \omega\varphi \tag{2}$$

最后数据源中的信息到实体描述的映射规则由标题匹配度和内容匹配度的加权值决定，从达到阈值的所有匹配中取最大匹配度的属性来构建映射规则。数据源中信息到实体描述的匹配度计算为：

$$\phi = \alpha_1 \phi_t + \alpha_2 \phi_c \tag{3}$$

4.1 结构化数据源

在数据空间中，数据库这类结构化数据具有很强的结构性，并且有确定的模式，只要找出命名实体和实体描述，即可直接映射实体。伪代码算法 1 描述了结构化数据源中的实体映射规则算法。

算法 1

```
输入：数据库实体信息匹配规则，数据库信息
输出：数据列到实体属性的映射规则
1. 根据数据库中的信息构造实体信息的虚拟表
  For 每一个实体属性
    For 数据库中的每一列
    //标题计算
    2. 获取数据列的元数据，根据规则计算数据列名的匹配度(列名，
       数据类型)
       //内容计算
    3. 获取一定的数据内容，计算内容的匹配度(内容，分词标注)
       3.1 若有内容的正则式规则，计算内容的正则匹配度
       3.2 若有数据分词标注规则，对数据内容分词，计算内容的分
词标注度
            //计算属性到列的映射
    4. 根据列名匹配度和内容的匹配度，建立数据列到实体信息的映
       射规则
  end
end
return 数据到实体信息的所有映射规则
```

4.2 半结构化数据源

与结构化数据源不同，半结构化的数据源数据结构虽具有一定规则，但其数据内容格式没有完全固定，是一个复杂的树结构，具有多层的数据嵌套。因此在实现映射的过程中，首先要确定数据结点所在的子树层次，然后针对子树来确定映射规则。伪代码算法 2 描述了半结构化数据源中的实体映射规则算法。

算法 2

输入:半结构化数据 XML 实体信息匹配规则,XML 数据源内容

输出:文档元素到实体属性的映射规则

//根据元素结点人名实体所在结点

1. 获取 XML 中的命名实体的人名匹配规则及数据源中文档元素的元数据,根据规则计算文档元素名的匹配度(结点名称)

If 元素结点的名称匹配人名实体的名称规则,且内容分词标注为人名则标识命名实体名识别成功,记录映射

Else

取数据源中所有元素结点的内容进行分词标注,计算内容为人名的标注。

If 内容标注存在为人名的

取匹配度最高的结点为实体名结点,标识命名实体名识别成功,记录映射

Else 内容标注不存在为人名的

Return 空的映射规则,即不映射

End if

End if

2. 计算命名实体相互分隔所在的最小子树(即子树之间拥有共同的父结点)。

//计算实体描述所在对应的元素结点

For 每一个实体的属性

For 最小子树下的每个结点(除命名实体结点外)

根据元素结点名称,计算名称匹配度

根据元素结点内容,计算内容匹配度

End for

End for

3. 根据结点名称匹配度和内容的匹配度,计算数据列到实体信息的映射规则

return XML 中文档元素到实体信息的所有映射规则

4.3 无结构化数据源

关于非结构数据源,这里仅考虑纯文本的无格式化数据。由于非结构数据不像结构化数据或半结构化数据那样可以产生映射规则,同时在处理非结构数据时,要想获取数据源中的所有实体,必需要对数据源中的所有内容进行处理才能完成,因此处理非结构化数据时返回所有实体,而不是返回映射规则。伪代码算法 3 描述了无结构化数据源中的实体映射算法。

算法 3

输入:无结构化数据源内容

输出:命名实体及其描述属性信息

1. 对数据内容是中文分词、标注

2. 根据分词后的人名标注,识别命名实体并生成一个实体对象,记录其出现的位置

定义命名实体出现的上下文(所在句子,段落)

3. 实体属性的映射

For 每一个实体的属性

将上下文中的内容与识别命名实体属性内容做匹配

若匹配成功,则记录到实体的相应属性及属性出现位置

否则,不记录

End for

return 实体的集合

5 实验

本文对从 QQ 娱乐和 Sina 娱乐网站上抽取的明星资料构成的明星数据源进行了数据源到实体的映射。在映射中,我们是通过数据源的部分记录信息来计算映射规则的,因此映射规则生成所用的时间较短,在实际的实体到数据源之间的集成过程中,运行时间则与数据源内容成正比。

本文利用 ICTCLAS 3.0 分词工具进行分词,并利用其标注功能实现人名实体及描述信息的识别。图 4 显示了两个娱乐明星数据库中各个属性对应的人名识别效果。实验结果发现,从中随机选取的 1000 条数据中,在命名实体的识别标注上虽然识别率不高,召回率达到 60%左右,但众多的人名标注所在列与其它列的人名相差较大,表明姓名所在的数据列是人名实体。在数据库、XML 等结构化半结构化数据源中,命名实体常常单独出现。因而,若这些数据源中的某个属性列或节点的分词标注的命名实体标注达到阈值,即可判定该列或该节点就是命名实体列或节点。同时,通常在这些列或节点中未被标记为命名实体的实体可作为未登录词加入到 ICTCLAS 用户词典中,以改善分词效果。

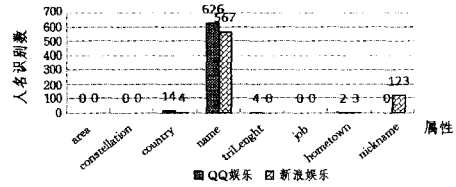

图 4 QQ 娱乐与 Sina 娱乐明星姓名识别效果

在人名实体与数据源映射上,我们对新浪娱乐明星资料数据库的 15 个属性(包括 name, nickname, gender, height, weight, country, area, hometown, college, birthday, bloodType, constellation, company, age, job, 其中 age 和 company 属性值为空)和 QQ 娱乐明星 XML 数据源(根结点为 star,12 个元素结点为 id, name, gender, year, day, area, constellation, bloodType, height, triLength, hobby, job)进行了映射。最后得到的实体与数据源映射见表 2。

表 2 实体与数据源映射结果

新浪娱乐明星(数据库)	〈map〉 〈attr name="名字"〉name〈/attr〉 〈attr name="昵称"〉nickname〈/attr〉 〈attr name="性别"〉gender〈/attr〉 〈attr name="体重"〉height〈/attr〉 〈attr name="身高"〉weight〈/attr〉 〈attr name="生日"〉birthday〈/attr〉 〈attr name="职业"〉job〈/attr〉 〈attr name="血型"〉bloodyType〈/attr〉 〈attr name="年龄"〉age〈/attr〉 〈/map〉
QQ娱乐明星(XML)	〈map〉 〈attr name="名字"〉star/name〈/attr〉 〈attr name="性别"〉star/gender〈/attr〉 〈attr name="身高"〉star/height〈/attr〉 〈attr name="星座"〉star/constellation〈/attr〉 〈attr name="血型"〉star/bloodyType〈/attr〉 〈attr name="职业"〉star/job〈/attr〉 〈attr name="居住地"〉star/area〈/attr〉 〈attr name="家乡"〉star/area〈/attr〉 〈/map〉

从映射结果可知,数据源内的大部分实体描述信息都能被正确识别出来,达到了实体与数据源间的映射要求。在新浪娱乐明星数据库中,包含有姓名和昵称两个属性。姓名一般具有较正式的特点,而昵称则是对人比较随意的称法。从图 4 中可以看出昵称与姓名在分词标注时的区别很大,在人

(下转第 186 页)

[13] 乔玉龙,潘正祥,孙圣和. 一种改进的快速 k-近邻分类算法[J]. 电子学报,2005,33(6):1146-1149
[14] Hart P E. The condensed nearest neighbor rule [J]. IEEE Transactions on Information Theory,1968,14(3):515-516
[15] Wilson D L. Asymptotic properties ofnearestneighbor rules using edited data [J]. IEEE Transactions on Systems, Man and Cybernetics,1972,2(3):408-421
[16] Devijver P,Kittler J. Pattern recognition:A statistical ap-proach [M]. Englewood Cliffs:PrenticeHall,1982
[17] 李荣陆,胡运发. 基于密度的 KNN 文本分类器训练样本裁减方法[J]. 计算机研究与发展,2004,41(4):539-545
[18] 熊忠阳,杨营辉,张玉芳. 基于密度的 kNN 分类器训练样本裁剪方法的改进[J]. 计算机应用,2010,30(3):799-801,817
[19] 刘金岭,王朝,谢少峰. 基于聚类中心初始化的文本分类高效算法[J]. 软件导刊,2010,9(4):47-49
[20] 张孝飞,黄河燕. 一种采用聚类技术改进的 KNN 文本分类方法[J]. 模式识别与人工智能,2009,22(6):936-940
[21] Han Jia-wei,Kamber M. 数据挖掘概念与技术[M]. 范明,孟小峰,译. 北京:机械工业出版社,2007:223-262
[22] 张磊,刘建伟,罗雄麟. 基于 KNN 和 RVM 的分类方法——KNN-RVM 分类器[J]. 模式识别与人工智能,2010,23(3):376-384
[23] 知识发现[M]. 史忠植,译. 北京:清华大学出版社,2002
[24] Youn E,Jeong M K. Class dependent feature scaling method using naive Bayes classifier for textdatamining [J]. Pattern Recognition Letters,2009,30(5):477-485
[25] 罗欣,夏德麟,晏蒲柳. 基于词频差异的特征选取改进的 TF-IDF 公式[J]. 计算机应用,2005,25(9):2031-2033

(上接第 173 页)

名实体识别过程中,具有人名特点但识别率比较高的属性,可以划归为昵称类。QQ 娱乐明星数据源中的 year 和 day 元素结点表示实体的出生年和出生月,但它们之间是分开存放的。由于两个结点间不存在关联信息,无法将它们与日期年月实际联系起来,因此在处理这类相互关联但无实际指定关联信息的数据源属性时,只能采取用户手工映射的方法才能实现。在实际映射过程,也会出现同一属性映射到实体的不同描述信息上去,如 QQ 娱乐明星数据源中的 area(表示明星的所在地域)就映射到了实体的居住地和家乡两个实体描述信息上去。当实体的部分描述信息具有相似的特点或在数据源中的数据包含有实体多方面的描述信息时,这种情况就会出现,因此在映射时部分属性或属性的部分内容可以映射到实体的多个描述信息,但姓名、昵称、血型列,则一般只映射到单个实体描述信息。

结束语 本文提出了一种数据空间命名实体的集成模型及各异质异构数据源中命名实体的集成方法,实现了异构数据源到实体的映射。数据空间命名实体及其描述信息的集成从数据是由实体及其描述信息构成这一特征出发,从数据源中抽取命名实体的主要描述信息,以提高数据源查询能力。数据空间中的实体集成也可以帮助数据空间完成演化,发现数据源间的关系:一方面,根据数据源共同包含的同名实体情况,可以发现拥有相同内容的数据源;另一方面,可以通过数据空间中命名实体在不同数据源的分布情况,发现数据源间的关联。下一步工作是利用数据空间中的实体来发现数据源间的关系,增强数据空间的演化特性。

参 考 文 献

[1] Dong X L,Halevy A. A platform for personal information management and integration[C]//VLDB. 2005
[2] Franklin M,Halevy A,Maier D. From databases to dataspaces:a new abstraction for information management[J]. ACM Sigmod Record,2005,34(4):27-33
[3] Halevy A,Franklin M,Maier D. Principles of dataspace systems [C]//ACM SIGMOD. 2006
[4] Franklin M, Halevy A, Maier D. A first tutorial on dataspaces [J]. Proceedings of the VLDB Endowment,2008,1(2):1516-1517
[5] Dittrich J P,Salles M A V. iDM:a unified and versatile data model for personal dataspace management[C]//VLDB. 2006
[6] Pradhan S. Towards a novel desktop search technique:Database and Expert Systems Applications[C]//Regensburg. Germany, 2007
[7] Ming Z, Mengchi L, Qian C. Modeling heterogeneous data in dataspace[C]//Information Reuse and Integration. IRI 2008. IEEE International Conference,July 2008:404-409
[8] 孟小峰. 数据空间技术研究进展[R]. 北京,2008
[9] Sarma A,Dong X,Halevy A. Data modeling in dataspace support platforms[J]. Conceptual Modeling:Foundations and Applications,LNCS,2009,5600:122-138
[10] Jiang X, Sun X, Zhuge H. A Resource Space Model for Dataspace:Semantics Knowledge and Grid (SKG)[C]//2010 Sixth International Conference. 2010
[11] Yang D,Shen D,Nie T,et al. Layered graph data model for data management of dataspace support platform[J]. Web-Age Information Management,LNCS,2011,6897:353-365
[12] Dong X,Halevy A. Indexing dataspaces[C]//the 2007 ACM SIGMOD International conference on Management of Data. Beijing,China,2007
[13] Sarma A D,Dong X L,Halevy A Y. Uncertainty in Data Integration and Dataspace Support Platforms[M]. Schema Matching and Mapping,part 1,2011:75-108
[14] 刘莉,郭艳艳,吴扬扬. 一种基于基本信息单元的索引[J]. 计算机工程与科学,2011(9):117-122
[15] Chung P W H,Liao Z. Cross-organisation dataspace (COD)-architecture and implementation[C]//International Conference on Computer Science and Software Engineering. Wuhan, China, 2008
[16] 董彦磊,申德荣,寇月,等. 数据空间中数据组织模型以及关联关系发现模型的研究[C]//第 26 届中国数据库学术会议. 南昌,中国,2009
[17] Dong Y,Shen D,Nie T,et al. Discovering Relationships among Data Resources in DataSpace[C]//Web Information Systems and Applications Conference(WISA 2009). 2009
[18] Yang D,Shen D,Nie T,et al. Layered graph data model for data management of dataspace support platform[C]//the 12th International Conference on Web-age Information Management (WAIM '11). Wuhan,China,Springer-Verlag,2011
[19] 孙镇,王惠临. 命名实体识别研究进展综述[J]. 现代图书情报技术,2010(6):42-47