基于数据空间的数据源内容关系发现机制*

曾淑琴,吴扬扬
(华侨大学 计算机科学与技术学院,福建 厦门 361021)

摘 要:数据空间的提出旨在解决模式驱动型的数据管理方式中所遇到的瓶颈,并最终解决数据管理所面临的挑战。而其中数据源内容之间的内部关联性成为数据空间研究的重点。从自然语言处理的角度出发,建立描述数据空间的模式实体,并且综合考虑基本刻面和内容刻面的主要内容,提出基于数据空间的数据源内容的关系发现机制,从而为下一步创建索引、浏览、搜索、查询以及其他服务提供良好的基础。

关键词:数据空间;刻面;自然语言处理
中图分类号:TP391.4 **文献标识码**:A **文章编号**:1674-7720(2012)14-0075-04

Relation discovery mechanism of data source content based on dataspace

Zeng Shuqin, Wu Yangyang
(School of Computer Science and Technology, Huaqiao University of China, Xiamen 361021, China)

Abstract: The proposal of dataspace aims at solving the bottleneck, which the management way of mode-driver encounter, and finally solving the challenge of the data management. The internal correlation between the data source content becomes the focus of the study of dataspace. From the point of view of Natural Language Processing(NLP), this article establishes the model entity of the description of dataspace, comprehensively considering the basic facet and content facet. We propose the relation discovery mechanism of data source content based on dataspace.

Key words: dataspace; NLP; facet

计算机网络的飞速发展以及信息化的推进,使得人类面临着巨大的数据量,而且数据的采集、存储、处理和传播依然与日俱增,数据管理也呈现出海量、共享以及多样化等新的特点。这些新特点使得数据管理技术面临着挑战,DBMS 无法解决这些挑战,如何对数据进行集成并有效的管理成为当前迫在眉睫的研究课题,数据空间就是在这个大背景下应运而生的。

数据空间[1]的概念由 FRANKLIN M、HALEVY A、MAIER D 等人在 2005 年的 SIGMOD 会议上提出,旨在解决数据空间包含的所有与主体(用户)相关的信息,它不是一个信息集成的方法,而是一种信息共存的措施。数据空间淡化模式,凸显数据,支持多种不同的异质异构的数据源,而且具有 pay-as-you-go(演化集成)的特性,强调数据的可关联性和演化性,最终可实现对个人数据的轻量级管理。其中演化集成的思想以及人在数据管理中的主体作用越来越得到关注,对主体人的研究日益成为数据管理技术研究中的重要问题。数据空间强调数据的可关联性,不仅要从用户(主体人)的行为上来获取数据的关联,还要从数据源内容上来获取数据源之间存在的内部关系,对数据源内容关系的发现也成为目前数据空间研究的一个重点难点。本文从自然语言的角度,通过分析数据空间刻面描述模型,对基本刻面和内容刻面进行描述,辅以词语语义相关度的模型,提出一个基于数据空间的数据源内容关系发现机制。

1 相关研究

数据源内容之间的关系发现是数据空间研究的一个重要问题,是创建索引、浏览、搜索、查询以及其他服务的基础。当前的研究前提大多假设已经获得数据之间的关系,但这往往有其局限性,为了解决这个问题,需要提出更加精确的发现数据源之间关系的方法,以便有效

* 基金项目:福建省科技计划重点项目(2008I0021)

地管理数据空间的数据源。参考文献[2]认为采用统一的数据模型来描述数据空间中不同类型的物理数据源是困难的，故而提出一种三层(即物理层、逻辑层、应用数据层)组织结构，文章集中在逻辑数据层，并通过领域本体代表一类数据源资源，从而划分为直接关系和间接关系。参考文献[3]通过关联调整(Reference Reconciliation)来解决数据源复杂信息空间问题，使用基于一个基本框架的算法，通过关联调整传播信息，使用上下文信息、相关实体上的相似性来计算和丰富关联。参考文献[4]提出了使用贝叶斯网络模型来抽取元数据的匹配，通过可能性推理来解决不确定问题，建立数据的关系网，通过元数据匹配来抽取实体之间的关系。参考文献[5]提出了新的分散的语义元数据组织模型 SmartStore，利用元数据的语义来增加相关的文件。参考文献[6]通过制定数据源之间联系，并将每个联系集定义为联系轨迹(Association Trail)，创建来自不同数据源的无联系数据之间的一个增强的关联图。总之，相关方面的研究也都是基于各自对数据空间的描述而进行的，通过本体或推理模型来发现数据之间的关系。本文基于之前所研究的词语相关度模型，综合分析基本刻面和内容刻面，从而确立数据源内容之间的关系发现机制。

2 刻面内容的关系发现机制

2.1 数据空间数据特点

数据空间的数据源是异质异构的，课题组提出了一个 FADSM 模型即基于刻面描述的数据空间模型，通过内容刻面以及基本刻面对数据空间进行描述，并分析刻面的内容来发现数据源本身之间的内部关联性。

数据空间个人数据的特点：(1)多样性和异构型。个人数据均来自不同的数据源，如 Web、Email、文件系统，数据都存储在不同的位置，需要采取统一的方法来制定异构数据源；(2)个性化。缘于不同的知识背景，使用计算机的不同习惯，以及每个人不同的组织数据的方式；(3)复杂结构。RDBMS 都是基于表结构的，但是在 PDS 中，关系都是基于元组级别的，数据源之间可能都是有关系的。

2.2 数据源描述

本文主要从两个方面来讨论数据源，一个是基本刻面，另一个是内容刻面。

将每个数据源作为一个模式实体来描述，每个数据实体都有一个独立的实体标识符。基本刻面是数据源的主体属性，包括文件名、文件类型、访问频率、目录以及大小等。内容刻面是每个数据源的描述性的主体内容，在课题研究组中已经将内容刻面提取出来。将内容刻面的内容进行分词(应用 ICTCLAS 软件)，分词后进行预处理，去除停顿词、虚拟词、语气词等。基于刻面描述的数据空间中数据源实体的表示如图 1 所示。

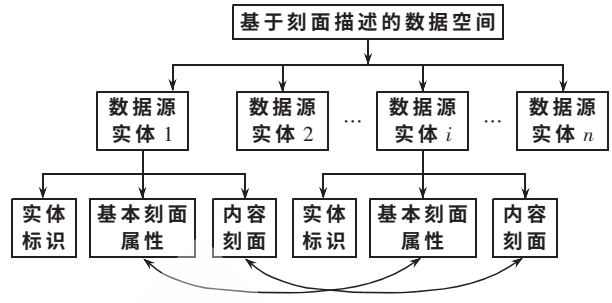

图 1 数据源实体的表示

通过对基本刻面和内容刻面的内容进行分析，对内容刻面进行分词预处理，形成刻面内容主题词集合，即代表了该数据源的实体内容；而基本刻面主要考虑其刻面属性。作为实体的类型集合，综合两者构造数据源实体的语义模式，发现数据源的内部关联性即是发现语义实体模式之间的关联机制。图 2 所示为数据源内容关系发现机制的流程。

图 2 数据源内容关系发现机制的流程图

2.3 语义模式的建立和匹配

本文采用中科院的 ICTCLAS 进行分词。对数据源的刻面内容进行分词预处理，去掉一些修饰词、停用词等，所获得的主题词代表了该数据源的核心内容。而刻面属性需要逐一考虑 4 个属性，对其进行相关的匹配策略。数据源表示为模式实体即[实体标识符，基本刻面，内容刻面]的形式。

语义模式匹配的过程为：(1)考虑基本刻面各刻面属性的匹配程度；(2)过滤掉内容刻面中修饰以及停顿的词语、标点符号、数字、名字等；(3)提取内容刻面的主题词集合以及该数据源所代表的实体类型组成待比较向量；(4)对于基本刻面中[文件类型，目录，访问频率，大小]等条件进行刻面匹配，以此对基本刻面的说明作为基本刻面的匹配策略；对于内容刻面，比较模式向量中的词语与待比较向量中的每个词语的语义相关度(应用相关度模型)。具体的内容刻面匹配策略如下。

2.3.1 基本刻面相关性匹配

基本刻面属性即一个数据实体的主体属性，能够直接获得，如一篇文档的文件名、路径、大小、修改时间、

访问时间等。

(1)文件类型。即实体的属性，如{email，web page，picture，documents}，当然这些类型还可以有二级结构，如picutre：jpg、gif、vsd 等，document：doc、txt、pdf、doxc、wps等。如果文件类型相同，则其匹配值为1，否则为0。

$$TypeMatch(D_p,D_q)=\begin{cases}1, type_{D_p}=type_{D_q}\\0, type_{D_p}\neq type_{D_q}\end{cases} \tag{1}$$

(2)访问频率。由访问时间来决定，频率在[0，1]之间，两个文件的访问时间越接近，则两个数据源有关联的可能性越大。式(2)表示基本刻面属性时间的相关性匹配。

$$BFre(D_p,D_q)=\begin{cases}1, VDate_{D_p}=VDate_{D_q}\\\dfrac{1}{VDate_{D_p}-VDate_{D_q}}, VDate_{D_p}\neq VDate_{D_q}\end{cases} \tag{2}$$

其中，$VDate_{D_p}$为D_p的访问时间，$VDate_{D_q}$为D_q的访问时间，若两个数据源同一天访问，则频率为1，若不同时间访问，则取时间之差的倒数。如果两个文档时间相近，说明这两个文档经常使用，有相关性的可能性也大，在取倒数后，其访问的频率值自然就大了，也就符合经常使用关联性更大的认知。

(3)目录。在同一个子目录下的东西相关性会很大，数据实体在同个文件下的相关性就更大，例如c：\A\D\abc.doc与c:\A\B\cd.doc，这里目录相关性权值为2/3。目录的匹配值即为相同的目录层数/最长的目录级数，如式(3)所示。

$$DirMatch(D_p,D_q)=\frac{SameDir(D_p,D_q)}{BigDir} \tag{3}$$

相同的目录级数越高，说明是相关或相似文档的可能性越大。

(4)大小。一般对document中，如果大小在相同的区间也说明该两个实体相关性更大一些，这里将大小分成几个区间：(0，10 KB]，(10 KB，100 KB]，(100 KB，1 MB]，(1 MB，10 MB]，(10 MB，+∞]，用5个自然数Region={1,2,3,4,5}来代表这5个区间。

$$RegionMatch(D_p,D_q)=\begin{cases}1, Region_{D_p}=Region_{D_q}\\\dfrac{|Region_{D_p}-Region_{D_q}|}{2}, Region_{D_p}\neq Region_{D_q}\end{cases} \tag{4}$$

(5)文件名。文件名是关键字查询搜索，鉴于本课题研究组中有成员正在研究按关键字的查询搜索，本研究是个底层的基础性研究，对此先不作介绍。接下来的工作也可以将这两个方面结合起来，按关键字搜索和按相关度搜索，或者两个相结合起来的综合搜索。

(6)将上述的刻面属性结合式(1)~式(4)，对基本刻面相关的贡献率分别为α、β、γ、η，则其最终的基本刻面相关性的匹配程度为：

$$\begin{aligned}BMatch(D_p,D_q)=&\alpha\times TypeMatch(D_p,D_q)+\\&\beta\times DirMatch(D_p,D_q)+\gamma\times BFre(D_p,D_q)+\\&\eta\times RegionMatch(D_p,D_q)\end{aligned} \tag{5}$$

其中$\alpha+\beta+\gamma+\eta=1$。

2.3.2 内容刻面相关性匹配

将内容刻面的内容进行分词（使用ICTCLAS），分词后进行预处理，去除停顿词、虚拟词、语气词以及人名等。

其匹配策略为：每个数据源内容由一个向量来表示，向量中的元素(即分向量)为词语结点。两个数据源的相关性匹配格式用一个四元组来描述，这个四元组为：<mID，[$First_i$，$FristFrequency_i$]，[$Second_j$，$SecondFrenquency_j$]，RSource>，$i=1,2,\cdots,m$；$j=1,\cdots,n$；其中mID是给定数据源内容的唯一标识符；[$First_i$，$FristFrequency_i$]表示一个向量，它包含一个数据源里所有词语的集合，$First_i$表示该向量的第i个(词语集合的第i个)词语，$FristFrequency_i$则是在向量里该词语出现的次数即词频，m是该数据源集合中的词语个数，同理，[$Second_n$，$SecondFrenquency_j$]表示另一个向量，这个向量也是包含另一个数据源里所有词语的集合，RSource是这两个数据源最后的相关度值。

具体的词语向量描述如下：数据源D_p里内容刻面分词为$[p_1,c_1]\ [p_2,c_2]\cdots[p_i,c_i]\cdots[p_m,c_m]$，其中$p_i$为词语，$c_i$为该词语出现的频数，有$m$个词语；数据源$D_q$里内容刻面分词为$[q_1,f_1]\ [q_2,f_2]\cdots[q_j,f_j]\cdots[q_n,f_n]$，其中$q_j$为词语，$f_j$为该词语出现的频数，有$n$个词语。

空间向量模型[7]是把文本内容的处理简化为向量空间中的向量运算，以空间上的相似度来表达语义上的相似度，而相似性的度量方式是使用余弦距离。此处也运用该原理，以空间上的相关度来表达语义上的相关度，以文档的刻面内容表示为文档空间的向量，通过计算向量之间的相关性来度量文档间的相关性。相关度的度量也使用余弦距离，余弦距离中的向量用词语相关度值来表示，两个数据源的内容刻面的相关度可以使用余弦向量值来表示，即：

$$\cos\theta=\frac{D_p\cdot D_q}{||D_p||\cdot||D_q||} \tag{6}$$

式中D_p、D_q分别表示两个数据源的向量描述形式。

正如上文所述，将式(6)的向量换成两个词语所计算而得的相关度。可以看出式(6)中的分子是两个向量的点积，即将两个向量中分量进行相乘再相加，与内容刻面上对刻面中的词语分量所要进行的相关度计算的思想一致。因此，将文档的向量改为两个文档中的词项的语义相关度，根据两个向量$D_p=[[p_1,c_1]\ [p_2,c_2]\cdots[p_i,c_i]\cdots[p_m,c_m]]$和$D_q=[[q_1,f_1]\ [q_2,f_2]\cdots[q_j,f_j]\cdots[q_n,f_n]]$，点积的定理和计算公式变化为：

$$\begin{aligned}RSource(D_p,D_q)&=D_p\cdot D_q\\&=\sum_m^i\sum_n^j W\_Rele(p_i,q_j)\times\frac{c_i}{f_j},\ i<m;j<n\end{aligned} \tag{7}$$

鉴于此处在分子中，已经对词语项进行了相关度的计算，而在SVM中，对分母的取模是为了保证整个余弦

值的范围在(0,1)之间,而在式(7)中,已经转化为对词语项进行相关度计算了,因此直接使用了点积公式作为计算内容刻面相关度的公式形式,即 *RSource* 就是所求的两个内容刻面的相关度值。

2.3.3 数据源内容关系发现机制

将基本刻面和内容刻面的相关性匹配策略结合起来,本文着重以内容刻面中表达的数据源内容来发现关联关系,因而内容刻面所占的权重会比基本刻面对数据源关系发现的贡献率更大,设基本刻面对数据源内容关系发现的贡献率是 λ,而内容刻面的贡献率是 σ。通过加权值来获得最终数据源内容的关系发现机制,如式(8)所示,其中 $\lambda+\sigma=1$ 且 $\sigma>\lambda$。

$$Match(D_p,D_q)=\Sigma\lambda\times RSource(D_p,D_q)+\sigma\times BMatch(D_p,D_q) \quad (8)$$

2.4 讨论与分析

根据以上对数据空间数据源的刻面模型描述以及对基本刻面和内容刻面的主要内容进行考虑,辅助以词语相关度模型计算,可以从理论上分析出获取数据源内容关系发现机制,并以上述的计算模型来表达其关系程度。但是,这个方案存在一些不足之处:(1)相关度的研究存在一些主观上的误差;(2)分词上出现的误差;(3)考虑内容刻面时,其中的许多主关键字没有考虑到人物名词,人物名词对于发现数据空间中数据源之间的内部关系起到很大的作用,本文主要是考虑数据源的具体内容,而未涉及到具体的人物之间的联系,因此对数据源的关系发现有一定的影响;(4)在基于数据空间对数据源内容的关系发现研究上,存在很多不同的方式,本文作为基础性的研究,因而辅以之前的相关度的研究,从而提出这个数据源内容关系发现机制的方案。

数据空间中的数据源都是异质异构的,且基于数据空间,是数据驱动型的管理手段,这些数据源彼此之间的内部关联性发现是数据空间研究的一个重点难点,国外研究方面,数据源内容关系的发现都是通过制定联系或者是参考协调等方法来完成,而本文研究是以自然语言处理中的词语相关度模型作为突破口,提出一个关系机制来发现数据源之间的关系。

今后的工作将继续完善该策略,特别是在考虑到几个不足之处的影响因素中,尽量减少这些因素所造成的误差,以该策略为基础,实现从相关度上进行数据空间中数据源的检索和查询。

参考文献

[1] 李玉坤,孟小峰,张相於.数据空间技术研究[J].软件学报,2008,19(8):2018-2031.

[2] Dong Yanlei, Shen Derong, Nie Tiezheng, et al. Discovering relationships among data resourcesin DataSpac[C]. IEEE, 2009 Sixth Web Information Systems and Applications Conference, 2009.

[3] Xin Dong. Providing best-effort services in dataspace systems[J]. Doctor of Philosophy University of Washington, 2007(9): 76-81.

[4] Sun Daring, Ma Anxiang, Zhang Bin, et al. Metadata matching based bayesian network in DataSpace[C]. Computer Design and Applications(ICCDA), 2010: 358-362.

[5] Hua Yu, Jiang Hong, Zhu Yifeng, et al. SmartStore: a new metadata organization paradigm with metadata semantic-awareness for next-generation file systems[C]. Proceedings of the Conference on High Performance Computing Networking, Storage and Analysis, Portland, Oregon, USA, 2009.

[6] SALLES M A V, DITTRICH J, BLUNSCHI L. Intensional associations in dataspace[C]. Data Engineering(ICDE), 2010 IEEE 26th International Conference, 2010: 984-987.

[7] Li Yukun, Meng Xiaofeng. Exploring Personal corespace for dataspace management[C]. Fifth International Conference on Semantics, Knowledge and Grid, 2009.

(收稿日期:2012-02-26)

作者简介:

曾淑琴,女,1987年生,硕士研究生,主要研究方向:数据库应用技术。

吴扬扬,女,1957年生,教授,,硕士生导师,主要研究方向:数据库技术和数据挖掘。