基于活动的数据空间数据关系发现

崔 晨,吴扬扬

(华侨大学 计算机科学与技术学院,福建 厦门 361021)

摘 要:基于分析用户日常活动间存在的相关性,利用活动与数据间的紧密联系,提出了一种基于活动的数据空间数据关系发现方法。通过分析用户活动发现数据间的隐含关联。利用日志系统收集用户活动窗口信息,从语义、切换、时间等方面计算活动相关度信息,并从中提取生成数据关系信息。系统还可随用户递增的活动信息及用户对数据关系的反馈,更新数据关系信息,提高系统的服务能力。

关键词:数据空间演化;活动理论;数据关系发现

中图分类号:TP393 **文献标识码**:A **文章编号**:1674-7720(2011)11-0012-04

Activity-based relationship discovery algorithm in dataspace

Cui Chen, Wu Yangyang

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

Abstract: By analyzing the correlation of user's daily activities and the close connection between data and activities, we described an activity-theory based relationship discovery and evolution algorithm in dataspace. Log user's activity window information and calculate activity-related information from the semantic, switching, and time. Then extract data-related information from activity-related information. The system can also evolve more accurate data-related information by using user's feedback and increasingly user's activity information.

Key words: dataspace evolution; activity theory; data relationship discovery

随着数字技术及互联网技术的发展,数据呈现出多样、异质化的特点。传统数据库已不适合对多样异质数据进行有效统一管理,因此,Franklin、Halevy 和 Maier 等人提出了数据空间[1]的概念。数据空间是数据和其关系的集合,系统需要从数据源中发现并抽取出有用的数据关系,以丰富其自身。而且,数据空间演化的目的是为了更好的满足主体需求,数据空间必须能够理解不同类型及来源的数据之间的联系,更好地对相关联数据做出处理,为主体提供更好的服务[2]。

但现阶段,数据空间数据关系的定义、范畴、如何发现关系还未定义。因此,帮助系统自动发现数据关系成为重要研究方向。用户日常活动中隐含了本地数据的独特分类见解。从用户日常活动中发现关系成为数据空间数据关系发现的途径之一。本文参考活动理论[3]系统对用户日常活动的分析方法,设计了自动记录、分析、提取出数据关系的系统。该系统利用日志收集用户活动信息,利用语义、时间、切换相关度等计算活动间的关系。将用户活动信息间的关系转换为用户意识中对数据关系的理解。而且,由于用户的需求和对数据关系的理解会随时间变化,系统会根据用户的使用情况更新数据关系,以更好满足用户需求。

1 相关工作

数据空间成为数据管理领域的一个研究热点。参考文献[4]设计了 Roomba 系统,利用多用户的反馈寻找数据关联。系统需要多用户评判生成候选匹配并生产有益的关系模式。系统依此关系模式进行数据关系的演化。但这种模式是广谱适用的,并不适用某个主体的偏好和习惯,有其局限性。

参考文献[5]认为用户文件的分布和排列中,隐含了用户对数据之间关系的理解。利用对文件的类型、命名规则、结构分析后所得到的关系,自动将相关数据分类并加入数据空间。但其成立初期需要用户手工标注常用

的数据，且缺乏后期对数据关系的调整与优化。

参考文献[6]将用户查询时对数据空间的交互作为数据空间演化的来源。该系统自动为数据资源建立资源摘要，并在初始简单查询时，利用关键字与摘要的匹配，逐渐丰富关键字与资源间的关联。但该系统对用户除了搜索外的其他交互活动利用有限。

参考文献[7]首次提出了在数据空间中进行用户任务挖掘的概念，作者定义其任务为一定数量文件的集合。与本文类似，该方法收集分析用户活动，但该系统仅利用窗口的时序关联，而没有利用语义、切换等发现数据关系。

2 基于活动理论的活动模型

维果斯基提出的“文化-历史心理学”思想是活动理论的源泉。目标导向原则是活动理论的核心。它认为人类活动受到广泛的客体群体影响，既包括自然领域，也包括社会文化领域，因此以维果斯基的三元关系模型为基础，为活动建模：

计算机上，个体代表计算机用户，目标代表所操作的对象，工具代表所使用的软件。此模型代表用户通过计算机达到某种目的而进行活动。本文扩充了此模型，增加了活动发生时间和用户操作，因此表示为：

$$Activity=(user, time, tool, operation, goal)$$

其中 user 为用户，time 代表活动发生的时间，tool 代表处理软件，operation 代表用户操作(如粘贴、复制、切换等)，goal 代表用户的操作目标。

3 活动信息的保存与处理

3.1 活动信息的保存

本文将用户日常活动的窗口标题信息作为活动信息的代表。下面为活动信息记录样例：

2010/7/11 11:29:33 具有自适应邻域探测机制的简化 PSO 算法修改稿 20100711 (用户上次保存的)[兼容模式]-Microsoft Word

2010/7/11 11:29:39 中国图像图形学报 A Journal of Image and Graphics(2010 年 5 期)-万方数据知识服务平台-世界之窗 3.3

3.2 活动信息的处理

活动信息处理是数据关系发现的前提。借鉴了相关工作，本文以活动语义、交互、切换、时间度量活动是否相关。

3.2.1 语义相关度

活动窗口的相关性表现为活动标题内容的相似性。本文采用改进了的向量空间模型 (VSM) 为对象模型。VSM 中，第 i 个对象的矢量模型如下：

$$V_i=\sum_{n=1}^{N} W_{ik}\times T_k$$

其中，W_{ik} 在传统方法中为关键词频度，但本文将 W_{ik} 改为关键词语义相似度之和。计算方法如下：

设 W_{ik} 为关键词 T_k 中对象 $Activity_i$ 中的权重，$D(T_{i1}, T_{i2}, \cdots T_{iN})$ 为 $Activity_i$ 的特征词组。$Seman(T_{i1}, T_k)$ 表示关键词 T_k 与特征词 T_{i1} 的语义相似度。则：

$$W_{ik}=\sum_{n=1}^{N} Seman(T_{in}, T_k)$$

语义相似度利用的中英文 WordNet 的词语相似度计算软件[8]。T_k 表示活动窗口标题中的关键字；V_i 表示第 i 个活动窗口标题的向量空间。通过向量内积计算活动窗口标题相似度，方式如下：

$$Sim(Object_{Activity_i}, Object_{Activity_j})=\frac{V_i\times V_j}{|V_i|\times|V_j|}=\frac{\sum_{k=1}^{N} W_{ik}\times W_{jk}}{\sqrt{\left(\sum_{k=1}^{N} W_{ik}W_{ik}\right)\times\left(\sum_{k=1}^{N} W_{jk}W_{jk}\right)}}$$

由此定义规则 1：

$$RelationV(Activity_i, Activity_j)=Sim(Object_{Activity_i}, Object_{Activity_j})$$

规则 1 表示：若相似度大于某阈值，则语义相关，相关度为 $RelationV(Activity_i, Activity_j)$。

3.2.2 内容交互相关

传统活动分析系统把内容交互作为活动相关的重要特征。随着博客、即时聊天的兴起，用户的活动大多变为信息获取与交流，粘贴复制的代表性作用降低。

因此定义规则 2：

$$IsContentPaste(Activity_i, Activity_j) \Rightarrow RelationC(Object_{Activity_i}, Object_{Activity_j})=0.05$$

规则 2 表示：如果检测两活动窗口有内容交互，则内容交互相关，相关度为 0.05.

3.2.3 切换相关

多任务操作更符合用户的使用习惯，窗口切换关系也是活动相似度的重要度量。设切换关联度为 SR，$Activity_i$ 与 $Activity_j$ 的总频度分别为 $F1$ 和 $F2$，且：

$$Activity_i \xrightarrow{State(switch)} Activity_j=m \quad Activity_i \xrightarrow{State(switch)} Activity_i=n$$

则：$SR=\frac{\sqrt{m\times n}}{F1+F2}$

因此定义规则 3：

$$((Activity_i \xleftrightarrow{State(switch)} Activity_j)>Threshold) \Rightarrow RelationS(Activity_i, Activity_j)=SR$$

规则 3 表示：两个活动互相切换的次数超过某阈值则切换相关，相关度为 SR。

3.2.4 时间相关

相关活动有其时效性，若发生时间接近，则推断两个活动是相关的。较长的间隔看作活动断点。假设系统共有 N 断点，时间相关度为 TR，$Activity_i$ 与 $Activity_j$ 的总频度分别为 $T1$ 和 $T2$。若 $Together(k)=1(1<k<N)$，表示在第 k 时间段两个活动同时出现，$Together(k)=0$ 表示没有同时出现。且 $Activity_i(k)$ 与 $Activity_j(k)$ 分别代表在第 k

时间段 $Activity_i$ 与 $Activity_j$ 的频度。

则 TR：

$$TR=\sum_{k=1}^{N}\text{Together}(k)\times\frac{2\times\sqrt{\text{Activity}_i(k)\times\text{Activity}_j(k)}}{T1+T2}$$

因此定义规则 4：

$$(\text{IsRunTogether}(\text{Activity}_i,\text{Activity}_j)>\text{Threshod})\Rightarrow\text{RelationT}(\text{Activity}_i,\text{Activity}_j)=TR$$

规则 4 表示两个活动在同断点内出现次数超过某阈值，则时间相关，相关度为 TR。

3.2.5 活动相关性总公式

设活动相关值为 AS，综上给出 AS 表达式：

$$AS(\text{Activity}_i,\text{Activity}_j)=q\cdot\text{RelationV}(\text{Activity}_i,\text{Activity}_j)+w\cdot\text{RelationC}(\text{Activity}_i,\text{Activity}_j)+e\cdot\text{RelationS}(\text{Activity}_i,\text{Activity}_j)+r\cdot\text{RelationV}(\text{Activity}_i,\text{Activity}_j)$$

其中：q、w、e、r 表示各规则系数，系数根据经验调节。AS 大于阈值的保存在活动相关文档中。

4 数据关系的提取

4.1 关系提取

提取数据关系，首先要处理活动相关文档。其中数据文件窗口和网页窗口信息存在一定结构。依据结构，本文设计了基于规则的提取算法，将活动关系文档中可识别信息提取为数据关系(本地文件或网页)。下面以活动文档为例：

具有自适应邻域探测机制的简化 PSO 算法修改稿 20100711[兼容模式]-Microsoft Word

从数据库到数据空间，从服务于企业到服务于大众-Adobe Reader

以上软件信息常出现在“- ”后，例如“Microsoft Word”，系统依据软件信息生成文件类型。示例中文件类型为“.doc”和“.pdf”。系统依据文件类型作相应的处理，去除无关信息，生成完整文件名如下：

具有自适应邻域探测机制的简化 PSO 算法修改稿 20100711.doc

从数据库到数据空间，从服务于企业到服务于大众.pdf

提取是有损过程，有损原因如下：(1)活动相关文档所保存的活动关系对中，有一项以上为杂项或不明信息，提取算法无法识别。(2)若软件信息不常见，提取算法将忽略此关系对。

4.2 关系确认与更新

数据在计算机上有其生命周期。上述提取的数据关系需要确定，并删除无效关系。处理步骤如下：(1)系统维护本地文件列表，比对数据是否被删除。若不存在则删除。(2)将有效的数据关系保存为数据相关文档。

数据空间中数据关系不断变化，因此，下一次计算出的数据相关文档与旧文档合并，并依据新关系权重大、旧关系权重小的原则，对数据相关文档进行更新，突出用户数据关系的新变化。

5 实验与结论

实验 1 用户评判数据相关文档的准确率。实验 2 将数据关系发现子系统整合数据空间，邀请用户进行相关搜索，并依照关系的有用程度及相关搜索体验为子系统打分。

5.1 数据间关系评测

实验挑选了 5 位实验室研究人员，他们习惯于在电脑上完成日常工作。经过一段时间的收集、分析后，完成数据间关系的评测。表 1 为分析后各用户相关信息统计情况。

表 1 用户信息统计

用户	活动记录数	活动相关文档	数据相关文档	数据相关提取率/%
用户 1	2 324	2 221	1 403	63.17
用户 2	2 580	7 029	6 727	95.70
用户 3	882	641	263	41.02
用户 4	6 230	6 319	4 006	63.39
用户 5	4 068	4 608	3 071	66.64

用户 2 数据相关文档提取率较高的原因是其活动记录大多是网页浏览活动，减少了因本地文件删除等造成的数据关系流失。用户 3 活动相关文档数量较少，且包含大量即时通信窗口，提取率偏低。

评测显示用户对数据关系基本满意。用户 2 与用户 3 的准确率和召回率偏低的原因与其数据相关提取率有关，而且其活动中访问本地数据较少，影响了系统发现数据关系的能力。

5.2 数据关系发现与相关搜索评估

将数据关系发现子系统嵌入到课题组的初步数据空间模型中，利用已发现的数据关系进行用户体验评估。图 1 是数据空间系统的界面图。

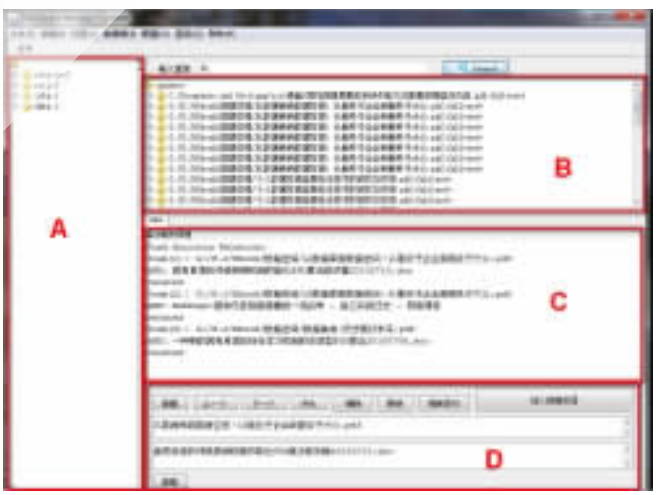

图 1 数据空间界面

其中，A 区为已导入数据列表，B 区显示已索引的搜索结果，C 区显示与 B 区结果相关的数据信息。当键入查询关键字后，B 区显示已索引数据信息。且传递已索引数据信息到后台系统准备 C 区与 D 区的数据相关信

息。当用户认为某项数据关系对自己有用时，可使用“加入数据空间”将关系和数据导入数据空间，数据空间系统将相应数据关系对进行集成与分析。

试验阶段，请用户在数据空间中进行多次搜索，每次搜索用户评判 C 区以及 D 区的活动数据关系的帮助。按照多次搜索的满意度 0-5 打分。如表 2 所示。

表 2 相关搜索评分

用户	相关搜索评分
用户 1	4.5
用户 2	3.5
用户 3	3.0
用户 4	4.0
用户 5	4.5

用户 1 由于本地数据多，抽取率适中，相关搜索时，返回较多的有用信息；用户 2 数据相关文档抽取率高原因是：其活动相关文档主要由网页浏览活动关系组成，在关系确认中损失极小，同时本地信息少，搜索时较少获得本地数据关系帮助；用户 3 其活动记录数较少，且较多即时聊天、设置等信息，提取了较有限的数据相关信息。由于关系过少，对用户相关搜索时的支持也偏少；用户 4 和用户 5 的数据相关提取率适中，且本地数据较多，因此可以提供较多的帮助供用户使用，取得了较好的效果。

参考文献：

[1] 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.

[2] 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.

[3] Nardi B. Context and consciousness: activity theory and human-computer interaction[M]. The MIT Press,1996.

[4] Jeffery S.Franklin M, Halevy A. Pay-as-you-go user feedback for dataspace systems[C].SIGMOD'08. Vancouver, BC, Canada:ACM,2008.

[5] Li Y. Meng X, Kou Y. An efficient method for constructing personal dataspace [C]. WISA 2009. Xuzhou, Jiangsu, China: IEEE,2009.

[6] Ning W. De X. Resource summary for pay-as-you-go dataspace systems[C]. ICSP 2008. Beijing, China: IEEE, 2008.

[7] 寇玉波，李玉坤，孟小峰，等.个人数据空间管理中的任务挖掘策略[J].计算机研究与发展，2009，46(2).

[8] 吴思颖，吴扬扬.基于中文 WordNet 的中英文词语相似度计算[J].郑州大学学报：理学版，2010，42(2)：66-69.

(收稿日期：2011-01-11)

作者简介：

崔晨，男，1986 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，主要研究方向：数据库及数据挖掘技术。