个人数据空间管理中的任务挖掘策略

寇玉波　李玉坤　孟小峰　张相於　赵婧
(中国人民大学信息学院，北京 100872)
(yubokou@gmail.com)

A Strategy for Task Mining in Personal Dataspace Management

Kou Yubo, Li Yukun, Meng Xiaofeng, Zhang Xiangyu, Zhao Jing
(Information School, Renmin University of China, Beijing 100872)

Abstract In personal dataspace management, users need to manage heterogeneous data such as emails, documents, pictures and so on. As user data rapidly increase in amount and variety, it has become a challenging problem to effectively manage these data and provide users with effective store and query service. Traditional data management tools such as file system and desktop search tools are not enough to satisfy users. Because personal dataspace is composed of three factors (data, user and service), while those old ones neglect the user factor. Thus they can only provide service based on paths or full-text index. Actually, there is close relationship between data and user. Personal data comes from user behavior. Meanwhile, user behavior is composed of many tasks. Mining tasks in user data could help establish semantic associations between data and provide users with task-based data management and query service. Therefore, we propose a method to mine tasks based on user behavior. We first analyze user behavior, and find time sequence relationships between personal data, then mine user tasks. The experiments show this method is effective.

Keywords task mining; user behavior

摘要　在个人数据空间管理过程中，用户需要处理大量异质数据如邮件、文档、图片等。随着用户数据在数量和种类上的增多，如何有效管理这些数据，为用户提供有效的存储及查询服务成为一个具有挑战性的问题。传统数据管理工具如文件系统、桌面搜索工具等并未给用户提供足够的管理能力。究其原因，个人数据空间是由数据、用户以及服务三要素组成。而传统数据管理工具却忽略了用户这一要素，因此仅能在存储路径或全文索引的基础上提供服务。实际上数据与用户之间具有密不可分的联系，个人数据空间中的数据正是来自于用户行为。而用户行为是由一个一个任务组成的。挖掘个人数据中的任务，可以建立起数据间由用户行为的语义关系，进而可以为用户提供任务视角的数据管理服务以及基于任务的查询服务。本工作正是基于这一思想，提出了基于用户行为挖掘用户任务的方法。本工作通过分析用户行为，发现个人数据由用户行为产生的时序关系，然后根据该时序关系生成用户的任务。本工作的实验证明该方法是有效的。

关键词　任务挖掘；用户行为

中图法分类号　TP393

计算机用户每天都需要处理大量异质数据如邮件、文档、图片等。如何有效管理这些数据，为用户提供有效的存储及查询服务成为一个具有挑战性的问题。然而，现阶段的数据管理工具对个人数据管理提供的支持比较有限。虽然桌面搜索工具为用户提供关键词搜索，操作系统的文件系统为用户提供路径查找等，然而这些功能并不能完全满足用户。例如，某用户因工作需要想找自己一年前写论文 a 时访问过的网页 b，然而由于论文 a 中并未有该网页的信息，且经过一年时间该用户已忘记网页 b 的网址。因此，及时且准确的找到网页 b 这种需求很难被现有的文件系统、桌面搜索工具或传统的数据库系统所满足。

如上所述，传统数据管理工具不能满足用户日

收稿日期：
基金项目：国家自然科学基金项目(课题号：60833005, 60573091), 国家 863 计划（课题号：2007AA01Z155，2009AA011904），教育部博士点基金项目（200800020002）
本文通讯作者：孟小峰（xfmeng@ruc.edu.cn）

益增长的需求。究其原因，是传统数据管理工具忽略了数据间的语义联系。在这种情况下，数据空间研究[1]日渐兴起。个人数据空间涉及用户、数据与服务三要素。个人数据空间考虑到用户这一要素并建立用户与数据之间的关联，才能克服传统工具所解决不了的问题。具体来讲，用户数据来自于用户一个个特定的任务，这里的任务可能是写论文、制作报告等。挖掘数据中隐藏的任务信息，可以将用户杂乱的数据以任务的形式组织起来，将个人数据空间管理中的用户要素与数据要素关联起来。这不仅仅是满足了用户查询某个文件的需求，更是可以为用户提供一种基于任务的新的数据管理方式。就上面的例子来说，用户的需求其实是查询与论文 a 有任务关系的网页 b。用户的这一需求无法被传统工具满足。若在个人数据空间中建立基于任务的管理，用户就可以通过查询论文 a 所在的任务，获得网页 b 的网址。

以此为出发点，本文提出在个人数据空间管理中一种基于用户行为的任务挖掘策略，旨在一方面提高用户管理数据能力，另一方面也为用户提供具有个人数据空间特色的数据管理服务。

本工作的贡献主要包括：给出了个人数据空间上任务的概念并进行了形式化定义；提供了建立文件间时序关联图的方法；提出了在时序关联图上进行任务挖掘的算法。

1. 相关工作

研究界目前尚未有在个人数据空间中基于用户行为挖掘任务的相关工作，然而很多研究小组已开始关注挖掘用户行为以建立用户数据之间的语义联系。

在语义桌面研究领域，有一些利用用户行为建立数据关联的工作[3，4]。Paul-Alexandru Chirita 等在[3]中提出一种方法通过用户行为日志，以用户对数据项操作的时序关系建立起数据项之间的关联。在[4]中，他们利用已建立的语义关联为用户的搜索结果提供更好的排序。

在邮件管理领域，[5，6]讨论了如何有效的管理电子邮件中的任务。需要注意的是这里的任务不同于本研究提出的任务。[5，6]中的任务指的是邮件承载的用户会议通知等，而本研究提出的任务是指一定数量文件的集合。

在人机交互领域，[7，8，9]致力于挖掘用户的任务，不过这里的“任务”是窗口的集合。他们的工作有利于提供更加友好的用户界面，与本工作不同之处在于他们的工作是希望利用窗口的切换来建立两个窗口之间的关联，而本文旨在研究自动挖掘个人任务及其关联文件的方法。

综上，本研究所定义的任务以及提出任务概念的背景都与上述工作不同。本工作旨在通过定义任务使用户更加有效的管理个人数据空间。据本研究所知，目前尚无工作关注个人数据空间中任务挖掘问题。

2. 问题定义

个人数据空间管理需要考虑到用户与数据这两个要素。个人数据空间中的数据来自于用户的行为。而任务就是对用户行为的描述方式。挖掘任务正是为了建立起了用户与数据，数据与数据之间的关联，以提高个人数据空间的数据管理能力。因此，本文研究了个人数据空间中的任务挖掘问题。为解决该问题，本文首先定义了个人数据空间中数据项、操作等基本概念，并在这些概念的基础上提出了任务挖掘的算法。相关概念如下：

定义 1. 数据项（Item）：用户数据中可以操作的最小单位。例如：一张图片、一个 word 文档或一个 pdf 文档等。

定义 2. 操作（Operation）：用户对单一数据项的一次操作行为。操作是用户行为的基本单位。一个操作具有如下结构：

Operation={Username, VistTime, Subject, Url, Type}, 其中：

- Username：用户名；
- VisitTime：用户访问该数据项的时间；
- Subject：用户访问的数据项名；
- Url：该数据项的路径；
- Type：该用户操作的类型，本文中操作类型有两种：“Access”和“Modify”。前者表示用户只是访问、参考了该文件，后者表示用户修改产生了该文件。

例如用户的一个操作{user1, 15:28:42, Readme.txt, C:\AccessMonitor, Access}表示用户 user1 在 15：28：42 时访问了 C：\AccessMonitor 下的 Readme.Txt 文件。在定义了数据项与操作的基础上，本研究给出了任务的定义。

定义 3. 任务（Task）：任务是用户通过参考一系列数据项，如网页、论文、图片等，最后生成或修改自己的数据项的过程。从数据管理的角度，将任务定为一系列数据项的集合。根据在任务中的功能不同，这些数据项可以分为两类：表示任务目标的核心数据项；为完成任务目标参考访问的数据项。一个任务可以形式化地定义为：

Task={Core; Reference},其中：

- Core：任务的核心数据项，定义为Core={Item1}。Core是特定任务中用户建立或修改产生的新数据项。核心数据项是任务中的“重要”文件，其重要性体现在该数据项代表了用户的任务目标。
- Reference：任务的参考数据项集用户在完成特定任务过程中参考过的数据项的集合。Reference={Item1，Item2，……}。

例如：将用户准备某个报告看作任务 T，他可能参考了数篇论文 A.pdf、B.pdf 等，使用了一些图片 1.jpg、2.bmp 等，最后生成了一个幻灯片文件 presentation.ppt。则该任务可以记为：

Task T;

T.Core={presentation.ppt};

T.Reference={A.pdf, B.pdf, 1.jpg, 2.bmp}。

为挖掘用户任务，本研究记录了用户操作的记录（见图 4）。在用户操作记录上观察发现：

观察 1：同一任务涉及的数据项在记录中经常相邻，即用户经常首先访问某用户中一数据项后马上访问该任务的另一数据项，这便产生了数据项间操作上的时序关联。

观察 2：在同一任务中关系越紧密的两个数据项，用户在这两个数据项上连续操作越多。

基于上述观察，本研究发现了用户操作记录中数据项间的时序关联这一现象，并提出时序关联图这一概念来建立数据项间的关系。

定义 4. 时序关联图（Time Sequence Graph）：本工作根据用户操作，建立了数据项间的时序关联图。建图规则是：1、对应每个单独的数据项，生成图上的一个点；2、根据两个数据项之间是否存在时序关联，建立两数据项对应图上的边。如果数据项 A 数据项 B 后被用户访问，则本研究认为数据项 A 与数据项 B 之间存在时序关联。示例见图 1。

time	item
9:10	A
9:20	B
9:25	A
9:26	C
9:27	D
9:28	A
9:34	B
9:50	E
9:55	F
10:00	G

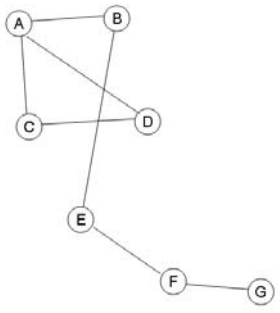

图 1 时序关联图的生成

根据本文的定义可知，时序关联图是一个连通图。基于上述定义，本文中个人数据空间管理中的任务挖掘问题是：给定基于用户行为生成的时序关联图，如何有效的挖掘该用户的任务。

根据时序关联图的特性，本研究提出了稠密块的概念，并进而提出了基于稠密块的任务挖掘算法。

定义 5. 稠密块（Dense Block）：一种特殊的块：在连通图G上，用H表示稠密块，则稠密块H不仅符合块的定义，而且满足条件：不存在边e_1，e_2，使得H-e_1-e_2由两个连通分支H_1和H_2组成，且H_1和H_2的节点数都大于 1。

例如，在图 2 中，G_1是一个稠密块，G_2是一个普通的块。对稠密块G_1来说，不存在两条边使得当这两条边被割掉时，G_1被分为两个度数均大于 1 的连通分支。在G_1中，割掉边e_1，e_2形成的两连通分支中有一个度数为 1。G_2是一个普通的块，它满足块的定义：连通且不存在割点。但是当e_3、e_4被去掉，该块可以被分成度数分别为 2 和 3 的两个连通分支，因此不是稠密块。

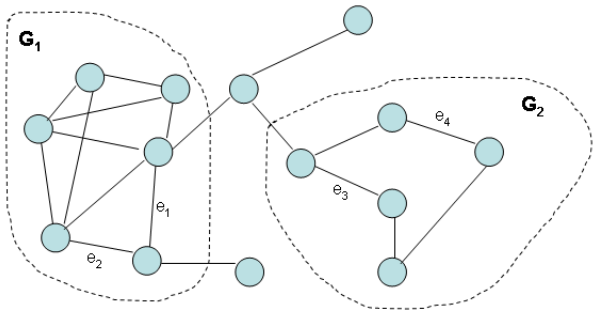

图 2 稠密块示意图

3. 任务挖掘算法

基于上一部分定义的相关概念，本研究提出了基于用户行为的任务挖掘算法。下面本文将介绍任务挖掘中两个关键的算法：时序关联图的预处理算法和基于稠密块的任务挖掘算法。

3.1 时序关联图的预处理算法

在生成的时序关联图上，观察到下列事实：

图中存在大量特殊结构，例如图 3 中数据项 A、B、C 形成的结构。在这种结构中，每个节点依次相邻，且度数均为 2。而这些结构所对应的数据项关联紧密，一并指向一个用户任务。不可能出现数据项 A、B 属于一个任务，而数据项 C 属于另一个任务的情况。为处理该结构以进行更精确的任务挖掘，本研究定义了这种结构为时序链结构，并在图上对该结构进行了处理。

定义 6.时序链结构(Time Sequence Chain)：在时序关联图G上，若存在点v_1,v_2,…，v_n($n\geq4$)使得v_i

（i=2,…,n-1 仅与v_{i-1}、v_{i+1}相邻，且v_1、v_n度数不等于 2，则v_2,…,v_{n-1}构成时序链结构。

在时序关联图的预处理阶段，本研究合并时序链结构为一个点。在图 3 中，本研究的预处理算法将 A、B、C 合并生成了 F。

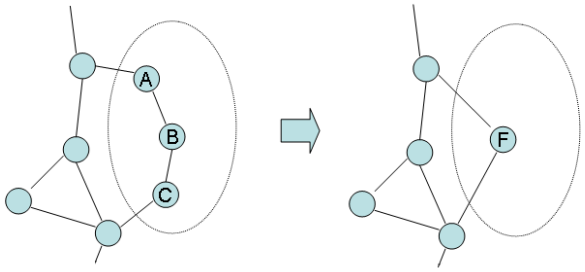

图 3 时序关联图的预处理

预处理过程的算法如下：

算法 1：图的预处理算法

名称：GraphPreprocess

输入：时序关联图 G(V, E).

输出：经过处理的时序关联图 G'(V' , E').

(1) for each node v in V{
(2) if v is not accessed Then
(3) create a new node t in V'
(4) If degree(v) = 2 Then
(5) set t as v's time sequence chain
(6) else
(7) set t as v }
(8) Build E' according to E
(9) return V'

在算法 1 中，步骤（1）到（7）完成了时序关联图上时序链结构的合并，步骤（8）用于建立新的时序关联图上的边，具体做法是对 G 中任意两点，如果它们之间存在边，则在新图中添加两点对应的新点间的边。

3.2 基于稠密块的任务挖掘算法

本文在经过预处理的时序关联图上进行任务挖掘。

本研究首先确定用户的核心数据项列表 Core List。如定义 3 的描述，本研究将数据项中被修改过的定为核心数据项 Core。其中，被修改过即意味着该文经历过操作类型为 Modify 的用户操作。

在获得了用户每个任务的核心文件 Core 后，接下来的目标就是寻找与每个任务核心数据项 Core 相关的参考数据项集 Reference 以生成任务。

本工作从每个核心数据项，寻找该核心数据项在图上的稠密块。本研究的方法是首先确定一个初始的点集，然后使用广度优先算法对与该点集 V 中与各点相连却不在 S 中的点进行判定，并将满足条件的点添加到点集 S 中，最后的点集 V 收敛到核心数据项所在的稠密块的点集上。本研究将每个核心数据项所在任务定为该核心数据项所在的稠密块包含的数据项。具体算法见算法 2。其中步骤（1）至（7）用于为每一个核心文件寻找参考过的文件集合；函数 FindDB（v）用于查找特定点所在的稠密块。

算法 2：挖掘核心数据项所在的稠密块。

名称：DenseBlock 算法

输入：经过处理的时序关联图 Ga(Va,Ea), 用户的核心数据项列表 Core List

输出：用户的所有任务的列表 Task List。

(1) create an empty Task List TL
(2) for each Core c in Core List{
(3) let v denote c in Va
(4) create a new task T
(5) T.Core=v
(6) T.Reference=FindDB(v)
(7) add T to TL}
(8) return TL

函数 FindDB(v){

(1) create an empty node set S
(2) add v to S;
(3) for each two node v_1, v_2 adjacent to v{
(4) if v_1,v_2 is adjacent to v
(5) add v_1,v2 to v}
(6) add node that is adjacent to at least 2 nodes in S till false
(7) return S-v;

4. 实验及评估

实验的软件环境是 Windows SP2 + JDK1.6 + Eclipse3.2，硬件环境是 P IV 3.2GHz，512MB SDRAM 内存。

本研究通过监视 Windows 环境的用户最近访问文件夹来获得用户连续的操作记录。本研究实验所用的数据集包括 5 个用户 3 个月的 11530 条数据，如图 4 所示。

表 1 展示的是每个用户操作过的数据项的数目以及核心数据项的数目。本工作所要提取的任务的准确答案来自于五个用户对各自的数据集进行分析标注的每个用户任务及其相关的数据项。

由于现阶段研究界尚未有在时序关联图上挖掘任务的相关工作，本工作选取了图上的两个初步方法：传统的 K-means 聚类算法（Cluster）以及取半径的算法（Radius）与本研究在本文第三部分中提

Username	VisitTime	Subject	Url	Type
User1	15:28:42	Readme.txt	C:\AccessMonitor	Access
user1	15:28:45	FileAccessLog.txt	C:\AccessMonitor	Access
user1	19:25:52	Blur.java	E:\Documents\Docume	Access
user1	19:41:39	Sams - Teach Yourself B	E:\Documents\Books\	Access
user1	19:52:34	CreditCardInterface.jav	E:\Documents\Books\	Access
user1	19:52:39	PassengerBean.java	E:\Documents\Books\	Access
user1	19:52:58	airplane.jpg	E:\Documents\Books\	Access
user1	10:40:13	zkmb.doc	E:\Documents\My Wor	Access
user1	10:40:15	NDBC2008 Draft.doc	E:\Documents\My Wor	Access
user1	10:46:35	index.pl	E:\Eclipse-Projects	Access
user1	14:22:14	%BF%CE%CC%C3%CC%D6%C2%D	E:	Access
user1	14:22:24	%BF%CE%CC%C3%CC%D6%C2%D	E:	Access
user1	14:23:13	NDBC2008 Draft.doc	E:\Documents\My Wor	Access
user1	14:23:15	NDBC2008 Draft.doc	E:\Documents\My Wor	Access
user1	14:24:18	NDBC2008 Draft.doc	E:\Documents\My Wor	Modify
user1	14:25:31	NDBC2008 Draft.doc	E:\Documents\My Wor	Modify
user1	15:57:44	lgpl.txt	D:\jnotify-lib-0.91	Access
user1	16:36:01	3.txt	E:\Eclipse-Projects	Access

图 4 用户操作记录

users	item number	core number
user1	813	86
user2	210	25
user3	374	44
user4	552	124
user5	1096	5

表 1 用户数据项信息

出的 DenseBlock 算法作比较，下面简单介绍一下这两种方法及参数设置：

K-means 算法：本文选取核心数据项数量作为 K 值。本研究将文件间的相似度定为在用户操作记录中两数据项出现时序关联的次数。

Radius 算法：以核心文件为起点，使用深度优先算法查找核心数据项附近深度为 3 的所有文件。需要说明的是，这里本研究选择半径为 3 是因为经过比较本研究发现半径为 3 时 Radius 算法的效果最好。

本工作的实验过程是通过任务挖掘获得用户每个任务的数据项列表，然后将每种方法获得的任务中的数据项与标准答案中的数据项进行比较。

图 5 展示的是三种算法在每个用户上的平均召回率（Recall），图 6 展示的是三种算法在每个用户上的平均准确率（Precision）。其中横坐标均表示参与实验的五个用户，分别为 user1、user2、user3、user4 和 user5。

由图 5、图 6 可明显看出，三种算法在不同用户数据上具有不同的表现。其中，DenseBlock 算法在大部分用户数据上准确率比其它两种算法高：在用户 user1、user2、user3 的数据上，DenseBlock 算法以损失少量召回率为代价，大幅提高了该算法的准确率。另外，实验效果与用户自身行为的特点有较大关系。例如：从图 6 可以看出，在用户 user4 的数据上 Radius 算法与 Cluster 算法的准确率高于 DenseBlock 算法，分析用户 user4 的操作记录发现用户 user4 在被记录的时期内频繁交互访问自己多

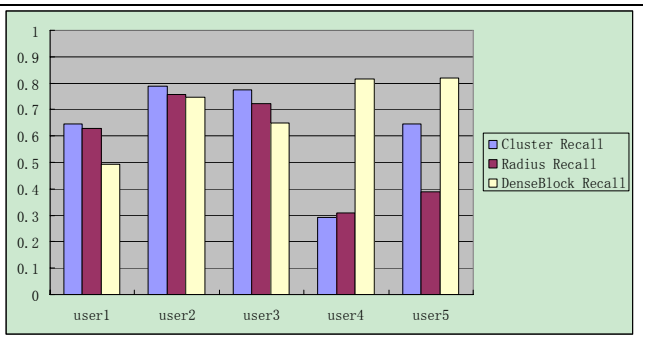

图 5 召回率(Recall)

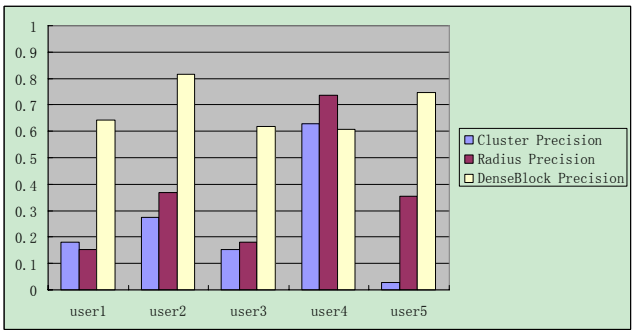

图 6 准确率(Precision)

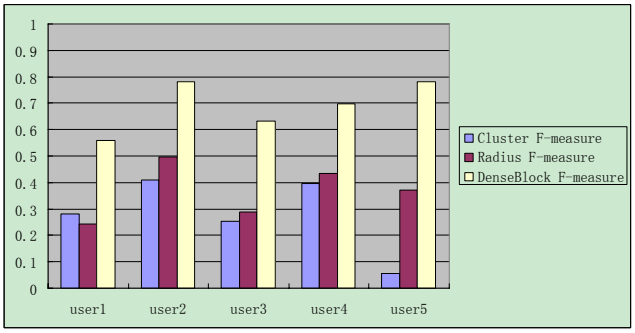

图 7 F-度量

个数据项，同一任务下数据项紧密相连，因此前两种方法的召回率较高。但同时这两种方法损失了较多的准确率。用户 user5 核心数据项数量较少，导致 Cluster 算法的准确率很低。而用户 user5 经常在各种数据项之间进行切换，因此导致各文件间距离较短，因此 Raidus 算法也效果较差。

同时，本文也使用了 F-度量来比较三种算法的挖掘效果，其度量公式如下：

F=2*(precision*recall)/(precision+recall)。

其中 precision 代表准确率，recall 代表召回率。具体结果见图 7。

F-measure 的比较表明本研究提出的基于稠密块的算法相对传统 Cluster 算法以及 Radius 算法更加有效。

5. 总结

本工作提出了在个人数据空间管理中的一种基

于用户行为进行任务挖掘的策略。本研究首先给出了个人数据空间上任务的概念并进行了形式化定义；其次提供了建立文件间时序关联图的方法；最后提出了在时序关联图上进行任务挖掘的算法。通过实验发现本研究提出的基于稠密块的算法是有效的，且具有较高的准确率和 F-measure。

作为个人数据空间中任务管理的初步工作，本工作进行了对任务的挖掘。在将来的工作中，本研究会在挖掘过程中考虑加入其他规则以提高挖掘效果，例如考虑同一路径下文件关联性等。另外，也会在本工作的基础上研究个人数据空间中基于任务的数据管理，为用户提供基于任务的存储及查询服务。

参 考 文 献

[1] Michael Franklin, Alon Halevy, David Maier: From databases to dataspaces: a new abstraction for information management [C]//*Proc of SIGMOD*'2005. *ACM SIGMOD Record* Vol.34, Issue.4, December 2005: 27−33

[2] Michael Bernstein, Max Van Kleek, David Karger, Schraefel: Information scraps: How and why information eludes our personal information management tools [J]. *ACM Transactions on Information Systems* Vol.26, Issue.4, September 2008: 1-46

[3] Paul-Alexandru Chirita, Wolfgang Nejdl: Analyzing User Behavior to Rank Desktop Items [C]//*Proc of SPIRE*'2006 11-13 Oct. 2006

[4] Paul-Alexandru Chirita, Stefania Costache, Wolfgang Nejdl, Raluca Paiu: Beagle++: Semantically Enhanced Searching and Ranking on the Desktop [C]//*Proc of ESWC'*2006, 11-14 Jun. 2006

[5] Victoria Bellotti, Nicolas Ducheneaut, Mark Howard, Ian Smith: Taking email to task: the design and evaluation of a task management centered email tool [C]//*Proc of the SIGCHI'2003*, 5-10 Apr. 2003

[6] Steve Whittaker: Supporting Collaborative Task Management in Email [J]. *Human-Computer Interaction* Vol.20, Issue 1, June 2005: 49 - 88

[7] Liam Bannon , Allen Cypher , Steven Greenspan , Melissa L. Monty: Evaluation and analysis of users' activity organization [C]//*Proc of the SIGCHI'1983*, USA, 12-15 Dec. 1983

[8] Nuria Oliver , Greg Smith , Chintan Thakkar , Arun C. Surendran: SWISH: semantic analysis of window titles and switching history. [C]//*Proc of IUI*'2006, 29 Jan.-01 Feb. 2006

[9] Jakob Bardram, Jonathan Bunde-Pedersen, Mads Soegaard: Support for activity-based computing in a personal computing operating system [C]//*Proc of the SIGCHI'2006*, 22-27 Apr. 2006

[10] Stefan Decker, Martin Frank: The Social Semantic Desktop [R] Galway: DERI Galway, 2004

[11] Richard O. Duda, Peter E. Hart, David G. Stork: Pattern classification [M] 2nd ed. New York: Wiley-IEEE Press, 2001

寇玉波，男，1985 年生，硕士研究生，研究方向：数据空间，个人数据管理。

李玉坤，男，1969 年生，博士研究生，研究方向：数据空间，个人信息管理。

孟小峰，男，1964 年生，教授，研究领域：Web 数据集成，XML 数据库，移动数据管理。

张相於，男，1986 年生，硕士研究生，研究方向：数据空间，个人数据管理。

赵婧，女，1985 年生，硕士研究生，研究方向：个人数据管理，Web 数据集成。