

PIM: 一个新的研究焦点

李玉坤 (Web组)

摘要: 科学技术的发展使个人信息量成倍增长,并成为影响个人生活秩序和生活质量的重要因素,于是产生了一个新的研究领域:PIM,即个人信息管理(Personal Information Management),本文对PIM历史发展、基本概念进行了系统介绍,分析了PIM的研究内容和关键的技术问题,特别对于个人信息的获取、存储、输出技术以及目前的研究情况进行了分析,对比分析了目前国内外的研究情况,对未来PIM研究进行了展望。

1. 引言

科学技术的发展为我们提供了巨大的信息量,报刊、手机、电视、电脑、互联网、人与人的交流,甚至个人的思考,都使我们时时地接触信息,每个人都处在信息的包围之中,信息在人们生活中的作用越来越重要,同时也使个人信息处理面临越来越多的问题。相信很多人都经历过这样的场景:

偶然的机会遇到自己非常感兴趣,或对于自己非常有价值的信息,如在网上偶然发现了一篇对自己非常有用的文章;在旅途中发现了一个市场需求信息;与人交流的时候,突然有了一个非常好的想法;等等,但是无法用有效的手段及时将其记忆下来;

自己积累的信息,往往随机记录在纸上,或保存在自己的计算机内,因为版本问题、存放位置问题、信息组织方法问题,往往不能随时随地访问自己要用到的信息,如某张照片,某个电话号码等;

记忆中确实将某张照片或某个电话号码保存到了自己的计算机或其他存储介质中,但是需要用到这个信息的时候,始终查询不到,或者查询成本非常高;

由于硬件损坏或丢失,如手机、笔记本电脑、硬盘等,造成自己重要信息的不可恢复性的丢失,给工作、生活带来严重影响;

忘记重要的日程安排,带来难以挽回的损失。

邮件系统受到容量限制、系统性能等各种问题的困扰,信息交流出现障碍,有时查询一封邮件往往成为很困难的事。

人们类似的经历还非常多,可以说,在这样一个充满信息的世界中,人们生活状态的好坏、工作效率的高低很大程度上依赖于信息处理的效率和及时性。特别是计算机技术、网络技术、web技术等的发展,为每个人提供了一个巨大的、共享的Web信息空间。使信息管理问题更加突出。据“网器”公司监测统计,2006年10月网站数量突破1亿,其中4700万或4800万家网站更是频繁更新网上信息^[1]。Web网页每周增长率是8%,每年新内容的增长率是50%^[2]。

除Web信息外,数据流、传感器、数字影像、数字电器、移动通信等技术的发展和应用,使我们每天所面临的信息更加丰富多样。如何将遇到的信息及时分析、保存;如何在需要的时候快速找到所需要的信息;如何在自己忘记的时候及时得到提醒;如何在信息管理中保护自己的隐私等等,这些问题变得越来越重要,处理的好坏直接影响到我们的生活质量和工作效率。如何解决这些问题,就引发产生了一个新的研究分支:个人信息管理(PIM)。

2. PIM 基本知识介绍

目前大家认同的最早提出PIM这一思想的是Vannevar Bush,他在1945年发表的文章“As

we may think”^[3]中第一次提出了个人信息管理-Memex的概念，他这样描述：Memex是一种能够记录所有书籍、唱片、交流信息的设备，它能够快速、自动、灵活的帮助人们找到所需要的信息。Bush只是为我们描述了一种远景，随着信息科学技术的发展，人们试图从不同视角对PIM给出一个准确的定义，在2005举办的第一届PIM Workshop的报告中，对这一概念进行了总结和阐述^[4]：

PIM是我们日常对于信息的处理、分类、访问---Lansdale (1988)。

- 为个人创建的供其在一个工作环境中使用的系统，其中包含人们获取信息的规则与方法；对信息进行组织与存储的机制，以及维持系统运行的一些规则与过程，以及对信息进行访问、处理、产生输出的方法机制。 ---Barreau (1995)
- 存储信息以使能够在以后被访问。 ---Boardman (2004)

由以上定义可以看出，PIM的定义与信息技术的发展有密切关系，Lansdale只是对PIM给出了一个宏观的描述；Barreau指出PIM中应包含获取信息的规则、方法，以及存储信息的策略、机制；到2004年，Web技术的成熟和存储技术的发展，使海量信息数据的存储成为可能，Boardman认为PIM的核心是数据的存储(store)和再访问(finding/refinding)。这些关于PIM的描述，成为进一步研究、定义PIM的基础。

2.1 PIM 基本概念

表面看来，对PIM定义是非常简单的，因为我们每天都接触它、使用它，其实PIM是很难定义的，以至于一直是一个挑战性的问题。首先，对PIM研究领域的界定比较困难，必须合理界定PIM与其他研究领域的关系。其次，从字面来看，PIM是一个包含主体、信息、工具的人机交互系统，涉及的概念很多。在PIM 2005 workshop上，与会专家对PIM的概念进行了讨论，对一些概念进行了阐述：

个人信息(Personal Information)：在PIM的研究中，我们聚焦于研究信息世界的一个信息子集，其中每个信息元素对于主体都有一定的影响能力。这样，就把PIM中的信息和我们平常的信息区分开来，这就是信息的相关性。或者称为信息的有用性，即PIM所研究的信息对于主体是有用的，这种有用性可能是现实的，也可以是潜在的。例如，我们到某地旅游，选择旅馆，关于旅馆的信息很多，如位置、价格、经理、员工数目、营业状况等，如果对主体做出选择产生影响的因素只有位置和价格，那么在PIM中关于旅馆的信息可以只包含旅馆的位置、价格。

因为主题需求是动态变化的，因此PIM的信息集合也是变化的，但具有相对稳定性。在PIM研究中，个人信息(PI)包括以下三层涵义：

- (1) 个人保存并为自己所用的信息。
- (2) 和一个人有关但被其它实体控制的的信息，例如，医疗保险机构掌握着我们的健康信息。
- (3) 一个人经历过的但不为自己所控制的信息，例如我们访问过的网页。

信息项(Information Item)：信息项是与主体相关的信息集合的一个单元，也可以叫做信息包。在传统的以纸为介质的PIM中，一篇文章，一封信都可以看作信息项；现在的信息中包含大量的数字信息，一个信息项可以是一封电子邮件，保存下来的一个网页、一个文件等。每个信息项有一个信息框(information form)，信息框与具体的应用和工具有关，这些应用和工具用来命名、移动、修改、复制、组织信息项，也可以为信息项赋予一些属性。

个人信息空间(PSI)：个人信息空间(Personal Space of Information)是指其所能够控制，或名义上能够控制的所有数据项的全体组成的集合（并不是指物理上对数据专属，例如邮件系统），一个PSI往往包括一个人的书籍、Paper文档，Email地址信息，Email文档、或其它存储在不同计算机上的与主体有关的文件，也包括网页链接。关于PSI的几点说明：

(1) 一个PSI是个人信息项组成的集合。PSI的大小是动态变化的。

(2) PSI包含的是主体记忆过的个人信息项。PSI不能包括我们访问过的，但是没有记录的信息（如尚在缓存中的网页）。

(3) 每个主体只有一个PSI。

(4) PSI是可供我们通过多种方法利用的潜在的数据源。对PSI中信息的有效重用，可以提高我们的工作效率。同时PSI的动态变动也引来了信任、安全问题。

个人信息管理 (PIM)：鉴于PIM的最终目的就是通过对数据的存储，以达到信息的有效重用，从这种数据存储的角度，PIM本质上是一系列操作行为的集合，其行为目的是建立、使用、保持信息及需求之间的映射。对PSI中有关的行为按照input-storage-output分类，可以归为三种：输入、存储和输出。在此基础上提出了一个如图1所示的PIM概念框架^[4]。

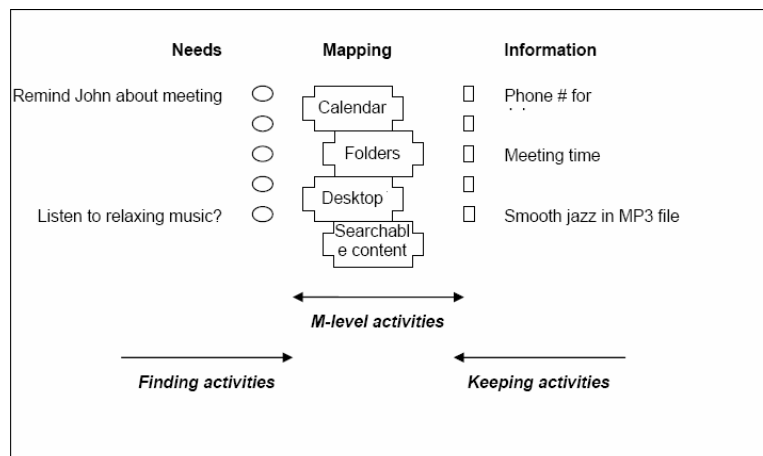


图1 PIM概念框架

由图1所示的PIM概念框架可以看出，PIM涉及的行为可以分为三类：

- **信息保持行为：**影响到PSI中数据输入的一系列行为。具体来说，是指完成从信息到需求所进行的行为，例如，当我们遇到信息的时候，如某人电话号码、会议时间等，我们要将这些信息保存下来以备将来之用，这类行为包括，信息的分析、分类、记忆、增强、记录等。

- **信息发现行为：**影响到PIM中信息输出的一系列行为。具体来说，是指完成从需求到信息所进行的行为。例如，当我们需要用到某项信息（如某人电话、一封邮件、一张图片等）的时候，将个人需求提交，并从PSI中得到该信息。这类行为包括查询语言、人机界面、搜索技术、信息分析、自动提醒等，需要区分的是：这里所说的信息发现，和通常的信息搜索不同，这里指的是从个人信息空间中发现自己曾经记忆过的信息，而不是在公共数据空间中搜索某项信息。

- **“M-level activities”：**影响PSI中数据映射的一系列行为。要高效地完成上面的两种映射，需要解决数据的存储、索引、安全性、一致性等一系列问题，这类行为就主要针对解决这些问题，其中核心的问题是个人数据空间的管理。

根据PIM的定义，作为一个研究分支，PIM聚焦于研究个人信息管理中的一系列行为，以提高各种行为的执行效率，最终提高个人信息管理的水平。

2.2 PIM 成为新的研究焦点

科学技术的发展使个人信息管理问题变得更突出，同时也为应对这一问题创造了条件，人们逐步关注到这一问题并开始进行这方面的研究，使PIM成为新的研究热点。

(1) PIM Workshop 的举办

PIM 研究引起了广泛的关注，其标志是 2005 年第一届 PIM 2005 Workshop 的举办，这是第一次专门针对 PIM 研究的专题研讨会，参加会议的有数据库领域的研究人员，也有微软、IBM 等公司的专家，从 PIM 基本概念、研究内容、研究目标等方面进行了讨论，取得了很大成果，我个人认为，最重要的是提出了很多重要的研究课题。

在 2006 年 PIM Workshop 上，收到论文 32 篇，其中有些针对 2005 Workshop 上提出的一些问题进行了深入的研究，有的论文针对个人信息管理中具体的问题进行了研究，这些论文有以下特点：

涉及面广。涉及到 PIM 研究的众多领域，包括基本理论、信息保存、信息分类、信息存储、隐私保护、邮件系统、行为分析、信息提醒等众多研究课题。即包括对基本概念、基本理论知识的研究，也包括针对特定应用需求的研究。

总体来说，处于起步阶段，这些论文很多针对 PIM 的某个具体课题，提出了自己的观点，并初步进行了论证，但对于具体的算法、数据的模型、系统的框架还没有作深入的研究和量化的实验分析，而且对于有些基本概念问题，也没有达成一致的看法。这些都为研究者提供了很好的机遇。

(2) 国际会议中开始出现有关 PIM 的论文

近两年关于 PIM 课题的研究论文也开始出现，如 2006 VLDB 关于个人数据空间研究的会议论文^[7]。在 CHI 2006 发表的会议论文中，有一些是关于 PIM 人机接口设计方面的问题的研究。

(3) PIM 为我们提出了许多新的具有创新意义的研究课题

跨学科研究成为 PIM 研究新的特点，PIM 与数据库技术、人机接口技术 (CHI/HCI)、认知科学、人工智能的结合，为我们提出了许多跨学科的研究课题，如 PIM 中的主体行为分析、个人信息挖掘、个人数据空间管理、特殊环境下的人机接口设计、数字记忆与信息自动提醒等。

同时，PIM 研究也面临重大挑战，一方面是由于信息的多样性，信息类型多样，位置各异，成为一个个“信息孤岛”；此外，主体在 PIM 中的关键作用使得信息的获取、组织、输出都充满个性化，PIM 的评测也受主体因素影响而变得复杂；但是，正是这些挑战的存在，给我们提供了进行创新性研究的课题和动力。

3. PIM 研究技术分析

由上节的 PIM 概念模型可以看出，对 PIM 的研究将重点放在对三类行为的研究方面：信息保持、信息存储、信息重用，每个环节都涉及很多技术问题，也有许多研究成果，本节从这三个方面对于 PIM 研究内容、关键技术问题做一分析。

3.1 PIM 中的信息保持技术

PIM 中的信息保持是指将所需要的信息项从公共信息空间复制到个人信息空间的一系列行为，以使主体能够重复访问、使用该信息项；同时也包括将无用的信息从个人信息空间中清除的行为，如图 2 所示。在 Barreau's definition 的框架模型中，将这一阶段定义为信息保持 (keeping)，这是为了区分传统的信息输入的概念，信息保持不仅仅包括主体通过手工输入的个人信。也包括其他偶然遇到并复制到个人信息空间中的信息，如浏览到的某个网页、收到的某个邮件、临时记录的某个电话号码等。

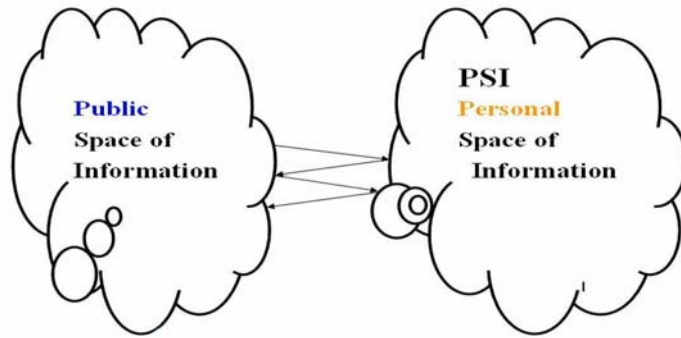


图2 PIM中的信息保持

由于个人信息形式的多样性，以及缺乏相应的工具，使得个人信息保持（输入）异常重要，也成一个技术难点。原因主要有：

(1) 信息的隐蔽性，很多数据以隐式数据的形式存在于公共数据空间中，在公共数据空间遇到、发现这些数据本身就是一个很难的过程；

(2) 当我们积极（寻找）或消极（偶然）遇到某些信息的时候，由于缺乏相应的技术和工具，也无法有效的将信息保持在我们的个人信息空间中；

(3) 信息的异构性，数据存在于众多的“信息孤岛”当中，如何将这些异构数据转化为PSI中的数据元素，也是一个困难的过程；

(4) 主体本身的因素。个人信息管理不仅包括数据等客体因素，主体本身也在这一系统中起重要作用，由于主体个性差异巨大，也为信息保持带来了很大难度，例如不同个体获取信息的能力是不同的；遇到信息后对其价值的判断也是一个难点问题。

信息保持主要通过以下几种行为实现：

(1) 手工输入方式

这种方式是指用户将自己所需要用到的信息项手工输入到个人信息空间中，如写邮件、输入自己的电话号码本、输入自己的个性化信息、写文章等，元数据输入需要采用这种方式。

(2) 信息集成

信息集成是指将现有的与个体有关的信息项输入到自己的信息空间中。信息集成的基础是信息感知（遭遇），只有遇到我们需要的信息后才能进行集成。信息感知又分为主动和被动两种：

被动信息感知是指我们无意偶遇到与我们有关的信息；

主动信息感知是指我们有意识的借助一些外部工具去寻找我们所需要的信息；主动信息感知往往要借助外部工具，如Web搜索工具，数据分析工具等等。从研究内容区分，感知到信息是PIM信息管理的起点。

与信息保持相关的技术问题：

信息记忆技术：由于信息遭遇的随机性，可能在任何时间、任何地点遭遇到有价值的信息，如何在这种情况下将信息快速的记忆下来是一个难点问题。

信息分类技术：遇到信息以后，保持信息的目的是为了以后信息的再利用，因此，“如何对信息进行分类、组织、存储，以保证用户需要该信息时能够快速找到”也是一个问题。

异构数据的处理：由于信息类型多样，需要研究不同的处理技术。

信息可用性的判定：遇到信息以后，要评估其价值以确定是否进行保存。由于主体个性差异巨大，如何判定信息的价值也是一个难点问题。

信息保持的人机接口技术：信息保持需要高效完成，主体的多样性也决定了接口设计的

难度。例如，针对不同的用户人群研究其适用的信息保持界面；针对用户所处的不同的物理位置环境研究其适用的用户界面；或者针对局部的具体PIM应用研究其界面。高效的工具和人机界面才能使PIM高效的为个人信息管理服务。

个人信息挖掘技术。信息挖掘也不是一个新名词，但是，不同领域的数据有其自身特点，对知识信息的需求也不同，PIM也是如此，PIM数据挖掘是指根据PSI中现有个人信息，挖掘尚未被主体认识到的信息的过程，而不是指从公共数据空间挖掘数据信息。首先个人数据信息有自身特点，包括数据分布、数据量大小等方面；另外，受主体差异的影响，个性需求差异也很大，例如，不同职业、不同行业的个体关注的信息是不同的，因此个人信息挖掘技术也是一个重要的研究方向。一方面是针对有共性需求的数据挖掘技术研究；一方面是针对个性需求的数据挖掘技术研究。

3.2 PIM中的数据存储服务

PIM中的数据存储服务研究，实际上就是个人数据空间管理系统（PDSMS）的研究。目前，因特网的发展，使人们对于数据资源的存储、访问等出现了新的特点，从而传统的数据库技术不能完全满足新的数据管理的需要，google、百度等网站对数据处理方式的改变也说明了这一特点，于是人们提出了一个新的概念：数据空间（DataSpace）。由于个人数据信息的特点，Dataspaces将是存储个人信息的理想方式，在PIM中称为PDS(Personal Data Space)。目前在这一研究领域。Jens-Peter Dittrich Marcos, Antonio Vaz Salles等人作了大量工作，他们在04-05年共发表了2篇关于个人数据空间管理的论文^{[5]-[6]}，在2006 VLDB论文中，系统阐述了个人数据空间管理的概念，提出了一个新的数据模型：*iMEMEX*，提出了数据源视图的概念，并基于此实现了一个PDS原型。他们的工作为个人数据空间管理的研究建立了一个框架模型，但是需要研究的问题还很多。

首先是对PSI中数据存储策略的研究，面对如此巨大的个人数据信息，在一台机器上进行数据的存储是不可能的，同时从安全性、可访问性等方面进行考虑，也不是很好的。因此，对于Dataspaces存储策略的研究成为一个基础性的问题。

因为Web数据变动的特点，有的数据会随时消失，这样就要求我们对数据的安全性策略进行研究，对数据安全性进行评估，以确定数据的处理策略。

数据模型的研究也非常重要，传统的数据库管理系统大都是基于关系模型的，在PDS中，传统的关系模型是否适用，随着PIM的使用，PSI中的数据量会越来越大，这样就会造成用户访问数据代价增高，因此，如何提高用户访问效率就会成为一个大问题。和传统数据库不同，在数据空间中，影响访问效率的关键因素可能不再是磁盘的I/O，那么这种情况下应改采用什么样的索引策略，在Dataspaces中索引的涵义是什么。这些都是需要研究的问题，也是下一代数据库系统所必须解决的问题。

目前国外在这方面的研究还不多，进行这些理论问题的研究，对于下一代的数据库研究意义重大。

3.3 PIM中的数据输出技术

PIM的数据输出技术是指从客户提出需求到获取信息结果的技术，即信息查询技术。即研究客户如何快速、高效的从PDS中获取自己想要的信息。这与传统DBMS中的查询类似。因此也涉及到查询语言、查询优化等一系列技术，又由于查询语言、查询优化都依赖于数据存储模型，因此不同的数据模型也会影响到数据输出技术的研究。

PIM中的数据输出，不仅包括传统的基于用户查询的方式，也包括自动提醒，类似于传

统数据库的触发器,提醒(reminding)技术也是重要研究内容之一,由于PIM系统的特殊性,提醒机制也要考虑到主体的各种情况,包括所处的环境,提醒的方式等,这与传统的触发器概念又有质的不同。信息提醒的研究是数据输出中的重要课题,也是难点问题,数据提醒的前提是数据的分析,因此人工智能(AI)技术和推出信息(Push)技术的应用会成为数据提醒技术的研究重点。

人机界面设计也是这一部分的重要内容,由于PIM的研究目的是使人们能够高效快捷的享受信息带来的巨大便利。客户个体的差别,所处环境的差别,要求系统能够将信息以用户最方便的形式展示出来。

4. PIM 研究现状

PIM研究是跨学科的研究,它涉及信息搜索、人机接口、认知科学、数据库技术、人工智能等众多研究领域。目前PIM的研究还处于起始阶段,国外对PIM的一些基本问题进行了研究,取得了一些成果。在PIM Workshop 2006的32篇会议论文中,其中多是阐明的关于PIM的一些观点,有些论文着眼于PIM中的某个具体问题,如个人邮件信息的处理^[7]、移动环境下的PIM管理^[8]、个性化的信息查询^[9]。也有一些文章涉及到PIM的一些理论问题^{[10][11]},如数据模型, PIM研究的一些前提等,这些文章提出了一些非常有价值的观点和概念。这些工作表明, PIM的研究已经起步,并将引起研究者的广泛关注。

与国外相比,目前国内专门针对PIM、PSI的研究还比较少,软件学报的两篇综述文章“数据库技术发展趋势”^[12]和“个性化服务技术综述”^[13],对于数据集成技术、用户界面技术、个性化服务技术进行了分析总结,此外,也有一些针对信息搜索、基于用户行为进行信息分析的研究成果,但是,总的来说,对一些基础性的、核心性的技术研究不多,如数据空间中的数据存储技术、数据模型、索引技术、优化技术等,相对国外的研究还有差距。由于对PIM的研究会涉及新一代数据库技术方面的一些问题,因此,应当引起国内该领域研究者的关注,不断跟踪新技术动向,争取在PIM研究方面取得一些高水平的成果。

5. PIM 研究展望

PIM 为我们提供了新的机遇和挑战。目前来看,近期 PIM 研究将主要围绕以下几个方面:

(1) 数据空间技术的研究, PIM 中将侧重于研究个人数据空间技术,具体包括:数据模型、数据存储、数据独立性、索引技术、查询优化等。

(2) 数据的保持技术。

(3) 数据的发现/再发现技术。数据的再发现是 PIM 研究的目的,特别是信息提醒技术的研究,需要应用人工智能的相关技术成果。

(4) 对于 PIM 技术、工具的评价方法学、框架、基准的研究。因为 PIM 实际上是一个包括主体、客体的系统,对其评价是非常复杂的,但这又是 PIM 研究的基础工作,因此会有很多研究围绕这方面展开。

(5) PIM人机接口技术的研究,

综上所述, PIM 研究具有重要的理论意义和现实应用意义。PIM 将研究 Web 环境中的个人数据管理问题,这些问题是下一代数据库技术必须解决的问题,同时为新的数据库技术提供了应用环境。PIM 是面向应用的,在研究过程中会不断开发出面向不同用户、不同领域的 PIM 工具软件,从而提高人们的信息管理水平,使个人从信息的枷锁中解放出来,产生巨大的社会效益。

参考文献

- [1] http://news.xinhuanet.com/world/2006-11/04/content_5290138.htm
- [2] Alexandros Ntoulas, Junghoo Cho, Christopher Olston: What's new on the web?: the evolution of the web from a search engine perspective. WWW 2004: 1-12
- [3] Vannevar Bush: As we may think. The Atlantic Monthly , July 1945
- [4] William Jones , Harry Bruce . Report on the NSF PIM Workshop, January 27-29, 2005, Seattle
A Report1 on the NSF-Sponsored Workshop on Personal Information Management, Seattle, WA, 2005
- [5] Dittrich, Jens, M. Salles, S. Karaksashian. *iMeMex: A Platform for Personal Dataspace Management.*, A SIGIR 2006 PIM Workshop Position paper.
- [6] VLDB 2006 regular paper. *iDM: A Unified and Versatile Data Model for Personal Dataspace Management*, JensPeter Dittrich, Marcos Antonio Vaz Salles
- [7] Yu, Xiaoyan, Mohammad Alkandari, Pengbo Liu, & Manuel A. Perez-Quinones, *Visualizing a Personal Social Network of Email Archives for Re-Finding*. A SIGIR 2006 PIM Workshop Position paper.
- [8] Singh, Gurminder, *PIM for Mobility*. A SIGIR 2006 PIM Workshop Position paper.
- [9] Cutrell, Edward, Susan Dumais, & Raman Sarin, *New directions in personal search UI.* , Personal Information Management.A SIGIR 2006 PIM Workshop Position paper.
- [10] Kirsh, David, *Personal information objects & Burden of multiple personal spaces*. A SIGIR 2006 PIM Workshop Position paper.
- [11]Spurgin, Kristina, *A Sense-Making Approach to Personal Information anagement*. A SIGIR 2006 PIM Workshop Position paper.
- [12] 孟小峰, 周龙骧, 王珊. 数据库技术发展趋势, 2004, Vol. 15, No. 12
- [13] 曾春, 邢春晓, 周立柱. 个性化服务技术综述, 2004, Vol. 13, No. 10