

Web 数据管理未来发展趋势的探讨

刘伟、李玉坤

1、引言

Web 作为 21 世纪最具生命力的信息媒体，每天都在以惊人的速度蓬勃发展，新的技术与应用层出不穷。它的迅猛发展使其成为一个巨大的网络平台，突破了国家和民族之间的界限，把世界各个角落联系在一起，渗透到了人们的日常生活、工作、学习以及娱乐等各个方面。根据美国市场调研机构 comScore Networks 提供的报告显示，到 2007 年 1 月，世界上的网民已接近 10 亿人，我国目前的网民数量位居第二，超过了 1 亿。但从这一方面的数字就可以看出 Web 所带来的巨大影响。

Web 犹如一个巨大庞杂的数据源，几乎包含了现实世界中各个领域的信息，逐渐成为了人们获取有用信息的最重要的途径。由于 Web 中蕴含的数据存在着高度异质、规模巨大以及动态变化等特点，使得人们从这个巨大的数据源中快速准确地获取自己需要的信息变得愈加困难，不幸的是，在企业中取得巨大成功的传统数据库系统却对复杂的 Web 数据难以管理。因此，对 Web 数据进行有效的管理成为了研究界和企业界人士一直关注的热点领域。比如 Web 数据的抽取、Web 模型与查询语言以及 Deep Web 数据集成等方面都已经提出了许多的研究工作和原型。然而，Web 数据管理中存在的问题远不止于此，而且随着其发展，还会有更多新的问题出现。

本文从研究者的角度，对 Web 数据管理未来的发展趋势进行深入浅出的探讨。需要指出的是，本文的重点不在于如何解决某一个具体问题，而是就未来 Web 数据管理这个研究方向提出一些抛砖引玉的观点。这些观点主要包括两个方面，一个是从数据管理角度，尝试提出一种新的数据管理方式，另一个方面是对一些新兴起的热点问题和技术进行探讨。

2、数据空间：一种新的 Web 数据管理方式

Web 日益成为一个巨大的信息源，无论企业还是个人，每天都从 Web 获取大量有价值的信息。这新信息来源于不同的数据源，如邮件、Deep web、网页，并且形式多样，有图片、word 文档、email 等，如何将这些数据高效的集成起来，使企业或个人能够便捷地共享这些数据，成为一个重要的问题。于是人们试图针对不同的应用领域，建立集成系统。但是从长远来看，集成系统由于有着建立代价大、数据支持有限、演化性能差等缺点，将不能适应日益增长的数据需要，所以不可能成为最终的解决方案。这样的现状催生了数据空间的提出，基于数据空间技术的数据集成为新的趋势。

2.1 数据空间基本特性

数据空间是对新的数据特点和数据管理技术的抽象与概括，其本质就是解决数据集成问题。数据空间定义为一个实体所拥有的所有数据的集合。数据空间与实体一一对应，数据具有时空特性，其空间特性表现在数据可以来自多个分布的数据源；时间特性表现在数据空间的不断演化。数据空间的主要特征包括[1]：

(1) 数据多样性

数据多样性包括数据格式的多样性和数据内容的多样性。一个数据空间中可能包含关系表、文本、电子邮件、图像、音频、视频等形式各异的数据；在一个数据空间中可能存在多份不同格式但是反映同样信息的数据，比如一份关系表和一份 Excel 表格可能表示的是同一份数据，也可能存在描述同样内容但是版本不同的数据。

(2) 先有数据、后有格式

这是数据空间和传统数据管理系统最大的不同。传统数据管理系统对数据的格式都是严格要求的，是一种“先有格式，再有数据”的数据管理方式。数据空间则不同，它对数据格式没有要求，数据能否保存到数

据空间的标准只有一个，那就是数据的内容必须是属于这个空间的。数据并不是一进入数据空间就被集成到某种模式，而是针对用户对数据操作的需求逐步进行数据模式的生成，也就是说，数据模式是在数据的基础上，根据用户需求总结出来的。

(3) Pay-as-you-go

理论上，数据空间应当包括与对应实体相关的所有数据，但这往往是不可能的，也是不必要的。因为数据本身就是不断变化的，新的数据源、数据项不断出现。系统无法将所有与实体相关的数据包括进来。而且用户的数据操作需求也是一步步产生的，因此数据空间的建立、完善、模式的生成也是一个逐步的过程。因此，相对于集成系统来说，这种数据管理方式成本比较低。

(4) 数据源不确定性

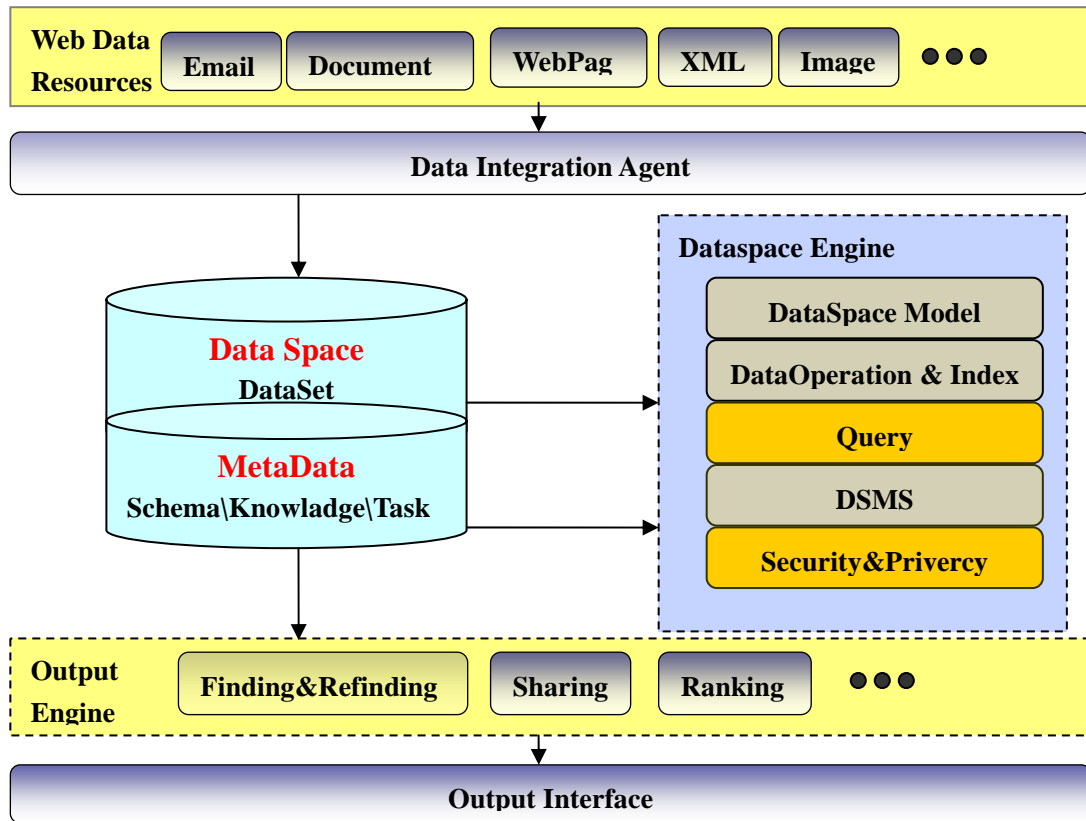
数据空间要管理的数据来源于多个数据源。数据源的多样性又引出了数据源不确定性这个特点。数据源的不确定性包括三个方面：数据源的未知性、数据源位置的不确定性、数据内容的不确定性。由于数据源分布在多台计算机上。这些数据源的物理位置和逻辑位置往往是不确定的。而且这些数据源中的数据结构和信息内容也往往是不确定的。这种数据源的不确定性也给数据空间带来了新的挑战。

(5) 持续演化性

持续演化指的是数据空间系统会随着时间以及应用的变化而不断自我进化，进化的准则是满足用户不断增加的应用需求。如上所述，Pay-as-you-go 的特性降低了构建数据空间的代价，但由于缺乏模式对数据关系的刻画，因此只能提供低水平的服务。所以数据空间需要演化，通过演化发现数据关系，生成数据模式。不断满足用户新的数据操作需求。

2.2 基于数据空间技术的 Web 数据管理

传统的数据管理技术不能适应 Web 数据集成的要求。传统 RDBMS 往往是建立在严格的数据模式之下的。而我们所要集成的 web 数据却是不断变化的。很难为这些数据建立统一不变的数据模式。因此在数据空间概念下进行数据集成将成为一种新的方向。图 1 为基于 Dataspace 的一种 web 数据集成框架。



在该数据集成框架中，主要包括三部分：

- (1) Web数据集成引擎 (Web Data Integration Agent)
- (2) 数据空间引擎 (Data Space Agent)
- (3) 数据输出接口 (Output Interface)

数据集成引擎负责对 web 数据源进行数据抽取、实体识别、数据转化等工作，每个 Agent 对应不同的数据源，例如针对 Email 数据集成的 Agent，针对 Web page 的 Agent，等等。这些异构的数据通过数据集成引擎，转化为数据空间所要求的统一的数据表示形式并保存到数据空间中。

数据空间引擎负责数据的存储、索引、数据访问、查询机制、数据的安全等。目前对于数据空间技术的研究引起广泛关注。[2][3]研究了数据空间的逻辑模型，提出了一种对于异构数据源的统一表示方法 iDM；[4]研究了数据空间的索引技术；这些工作对数据空间管理技术进行了研究，其宗旨是体现数据空间的设计思想，如持续演化、pay-as-you-go, from-data-to-schema 等。目前仍然有很多问题没有很好解决。例如数据空间存储模型，高效数据查询算法、数据安全隐私、数据演化模型与实现等等。

数据输出接口。这一部分研究如何对数据空间中的数据进行快速高效的访问。例如数据共享技术、高效的数据查询算法、界面设计等等。

2.3 相关工作和未来的研究问题

数据空间是一个新的研究领域。因此基于数据空间的数据集成也面临许多问题。

(1) **数据集成中的数据不确定性问题。** Web 数据本身就具有不确定性，如何判断并量化 Web 数据的不确定性，以及如何在数据空间中表示这种不确定性，如何建立基于不确定数据的数据操作等都是需要研究的问题。

(2) **数据空间模型、实现及数据操作。** 尽管对于数据空间进行了一些研究，但是还没有能够实用的系统出现，而且这些模型的数据操作效率如何，也没有定量的分析。因此在数据空间研究方面，还有很多问题需要解决。

3、未来 Web 数据管理中一些研究问题

Web 可以看作是一个巨大的、时刻向前演化的数据源。与一般的企业或研究机构的数据源不同，它具有社会性，是对现实世界的反映，覆盖了各个领域，因此 Web 数据管理也会面对各种各样的挑战。虽然目前 Web 中占主导地位的仍然是非结构化的文本信息，但不可否认的是，对 Web 数据管理的研究正逐渐转向结构化的文本信息以及各种非文本的信息（图片、视频等）。在这节我们对未来 Web 数据管理中正在出现以及可能会出现的一些问题和技术尝试着进行探讨和前瞻，希望能够起到抛砖引玉的作用。

3.1 Web 数据源的选择问题

随着 Web 的飞速发展，越来越多的 Web 数据源可以为我们提供服务，使得我们更多的选择来得到想要的信息。根据[5]在 2004 年进行的采样估算，目前 Web 中可访问数据源的数量已经超过了 45 万个。然而任何事物总是具有其两面性，Web 是如此巨大，一般而言提供某一特定服务的 Web 数据源也会以成千上万计，如果逐个去访问，不但用户难以忍受，对网络资源和 Web 数据源也是巨大的消耗。因此我们需要从中确定能够最适合的 Web 数据源，而不是逐个去访问。

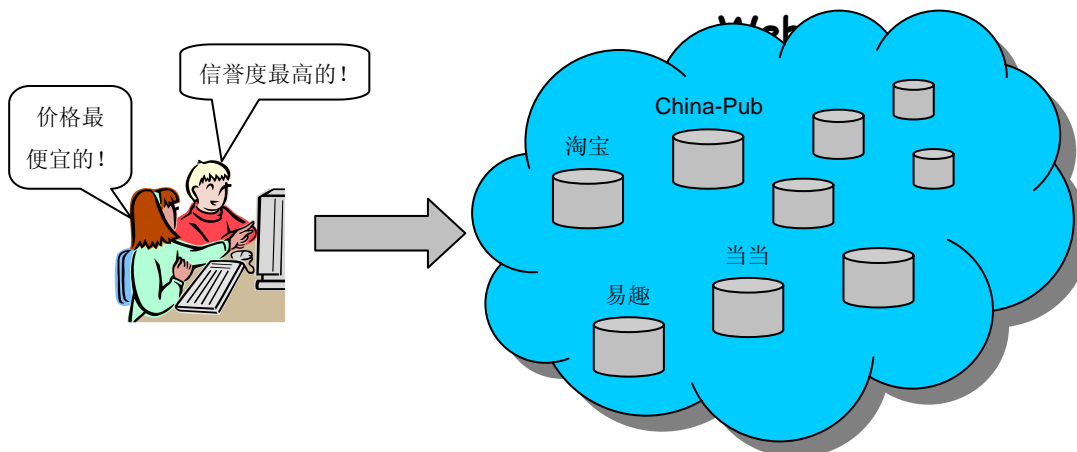


图 2 Web 数据源选择示例

举一个大家最熟悉的场景就是网上购物，有过网上购物体验的人们都可以感受到，现在几乎现实中所有的商品都可以通过购物网站购买，每个购物网站都可以看作一个 Web 数据源。但大家不得不承认的一个事实是，现在的购物网站是如此得多，难以计数。就象图 2 中的示例那样，如何知道哪些购物网站出售自己要购买的商品？又如何从这些购物网站中选择比如最便宜的或是距离最近的？如果花几个小时甚至是一天时间把每个购物网站都逐个查看一遍，显然是一件难以让人接受的事情。

由此可以看出，如何为用户一个特定的查询选择合适的 Web 数据源将成为 Web 数据集成系统中一个重要的问题。为了解决这个问题，需要在于是否能够对 Web 数据源有一个预先的了解，即获取 Web 数据源的有效特征。这里的特征是指 Web 数据源的大小、主题以及更新频率等方面。Web 数据源的大小是指其拥有的数据量，一般我们可以用记录的数量来表示。直观上，一个 Web 数据源越大，表明其包含的信息越丰富，也越有可能满足用户的查询，相应的也会造成查询代价高，返回记录数量过多等缺点。而一些较小的 Web 数据源包含的信息常常更加专业，适合专业人士进行查询。因此，Web 数据源的大小是 Web 数据源选择的一个重要依据。Web 数据源的主题是指其包含信息的内容属于现实世界中哪一方面的，比如经济、体育、政治等等。用户的一个特定查询，必然属于某一主题，只有向属于这一主题的 Web 数据源查询才有意义。需要指出的是，现实世界的主题划分不是水平的，而是层次结构的。比如计算机可以分为硬件和软件，而软件又包括数据库、编程语言等。因此对一个 Web 数据源的主题判断的越准确，就会对 Web 数据源选择的越准确。Web 数据源的更新频率是指其内容变化的频度，这是由于 Web 数据源自身具有动态变化的特点造成的。不同主题下的 Web 数据源其更新频率是不同的。比如经济主题下 Web 数据源的更新频率要远远大于生物和宗教主题下的 Web 数据源。对于更新频率快的 Web 数据源，对于用户的一个特定查询，不同时间的查询结果往往具有很大差异。我们这里只是简要介绍了最主要的三个特征，其实还有许多特征都有助于 Web 数据源的选择，就不再一一介绍。总之，获得 Web 数据源的特征越多，对 Web 数据源的选择就会越准确。

由于 Web 数据源一般是自治的，因此向其拥有者获取这些特征并非总是可行的，一种最有效解决方案就是对 Web 数据源的采样，即通过获取一些样本来得到该 Web 数据源的特征。然而这个方案的挑战之处在于：Web 数据源所提供的查询接口是查询受限的，无法用传统的采样方式获取样本记录。如何通过受限的查询接口获取随机的样本是今后重要的研究问题，这个问题已经开始受到研究者的关注。

3.2 Web 数据集成中查询结果的排序与展现

大量 Web 数据源使得用户有了更多的选择，也又有了更大的可能性来获得想要的信息。但是，随之而来的问题就是：对于用户的一个查询，把各个 Web 数据源的返回结果汇总在一起时通常会有大量的记录，而实际上用户真正需要的记录很可能只是少数几个，因此用户经常需要大海捞针似的在大量返回结果中寻找想要的那几个记录（图 3）。据调查发现，用户一般是不会把全部的返回结果浏览一遍的。

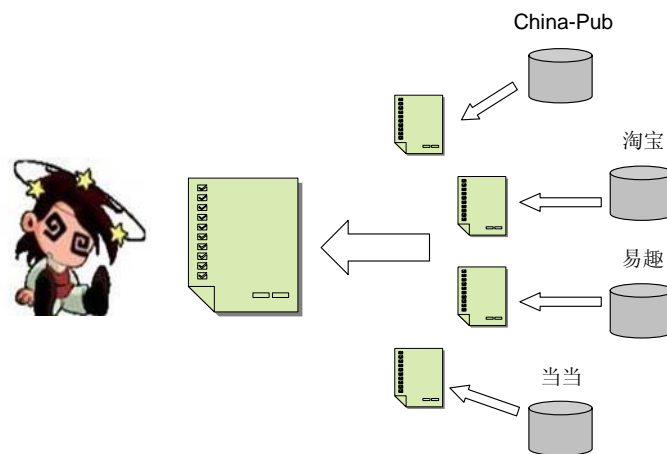


图 3 真正需要的记录被淹没在大量结果中

因此，Web 数据管理者们亟待需要解决的问题就是：如何寻求有效的方式来帮助用户在大量的查询结果中快速地找到真正需要的记录？针对这个问题，可以通过两种方式来解决：一是尽量将用户想要的记录排在前面的位置；二是采用新的展现方式来代替传统的罗列形式，使得用户可以快速地定位到想要的记录上。下面我们分别就这两种方式进行简要介绍。

如何对大量查询结果的排序在搜索引擎领域中已经得到了较好的解决，但我们却不能将它们照搬到结构

化的记录排序中来，这是因为结构化的记录是由一组属性值组成，不同属性上的值对记录的代表性并不相同，因此不能简单看作关键词的集合。这就需要提出一种新的排序方法，利用记录中结构化的信息以及来达到满意的排序效果。由于用户需要在 Web 数据源的查询接口上填写结构化的查询才能得到查询结果，我们可以通过用户所提交的查询来推测其感兴趣的记录。换句话说，对于满足两个不同用户查询的同一个查询结果集，会因为这两个查询的不同而产生不同的排序。举个例子，如果一个购买图书的用户在查询中只填写了书名和价格信息，说明他关注的是这两个属性，因此我们需要在这两个属性上对查询结果排序，而。当然，这只是一个非常直接的想法，要达到用户满意的排序效果，我们还必须在理论上进行深入的研究。

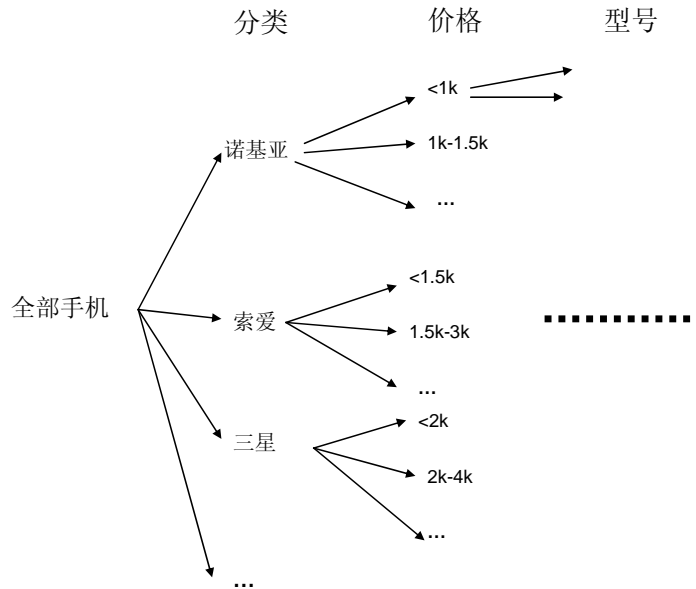


图 4 层次结构的查询结果展现方式

对查询结果的顺序罗列是最传统的浏览方式，这对于大量的查询结果有着明显的弊端。我们可以借鉴数据库索引的思想，不妨把查询结果集构建成一棵树状的层次结构。这样用户只需要从根节点开始，以导航的方式，只需少量的点击就可以找到想要的记录。这虽然是一个不错的方式，但是根据当前的查询结果如何构建这样的层次结构将是一个十分具有挑战性的问题，这其中需要考虑到属性的展现顺序以及每个属性的值域如何划分等因素。

3.3 公共 Web 数据中的个人隐私问题

随着 Web 发展，特别是 Web 2.0 的提出并迅速繁荣，其不再仅仅是人们被动获取信息的来源，它也成了普通网民们以个体为单位在 Web 中交流和发布信息的平台。博客是 Web 2.0 应用的典型代表，每个人，无论响当当的公众人物，还是默默无闻的草根，都可以在 Web 中拥有一块属于自己的空间，记录自己的日常生活，对热点人物或事件发表自己的观点和感想。还有 BBS 论坛，把兴趣相同的网民吸引到了一起，在某个兴趣上进行自由的交流。另外像 QQ、MSN 等聊天工具，也极大地推动个人数据在 Web 中的迅速增长。

在人们欣喜地享受这些好处的同时，却还不知道自己的隐私正在不知不觉间在 Web 中暴露出来。有人曾经做过这样真实的搜索尝试：1、从论坛上一篇帖子中获得这个用户公开的 QQ 号；2、在搜索引擎中搜索这个 QQ 号，在某个搜索结果中发现了此人的真实姓名、工作单位、联系电话以及 Email 等；3、在搜索引擎中搜索这个人的 Email，在搜索结果中发现此人常用的注册名（为了记忆的方便，似乎大家都习惯在注册时总是使用一个名称）；4、在搜索引擎中搜索这个人的常用注册名，发现了此人某个健康论坛发了一篇关于肝炎的帖子，其内容表明他患有乙型肝炎³。通过这个实例可以看出，把这些散碎的看似无关紧要的内容整合在一起就把个人的隐私信息暴露了出来。大家有兴趣的话可以仿照上面的例子，试一试自己的隐私是不是已经暴露了，比如兴趣爱好或者健康状况等。

³虽然这是一个真实的例子，但出于对个人隐私的保护，这里并没有透露他的姓名或 Email 等信息。

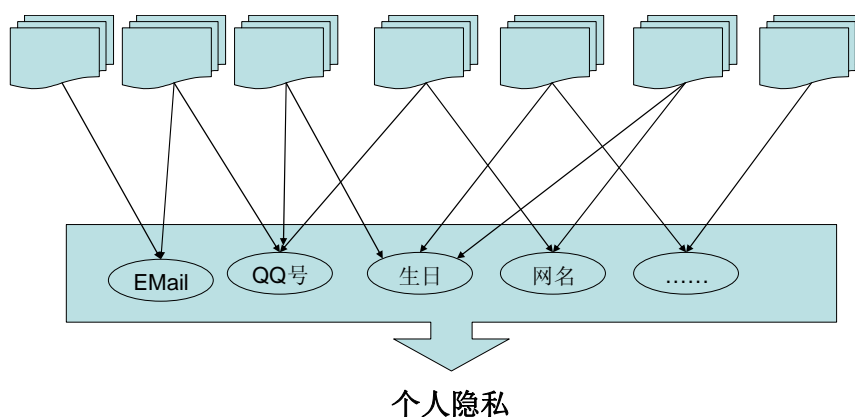


图 5 个人隐私的泄漏

这种在 Web 上通过搜索引擎搜 Email、QQ 号或者常用注册名的简单方式（图 5）就可以轻易地得到一些个人的隐私信息，让人有一种在 Web 中无处藏身的感觉。就连中情局首席女发言人迪伊克也不得不承认：“掩护身份是一个很复杂的事务，在互联网时代则变得更为复杂，以前适用的一些方法现在已不适用了。”有一点需要指出的是，我们这里所说的对隐私的获取途径是指通过利用正常的手段（比如搜索引擎）获取 Web 中公开的信息，而不是类似黑客采用不正当的手段获取非公开的信息。

由此可见，对 Web 数据中个人隐私的保护将是 Web 数据管理领域一个重要的研究课题。为了保护 Web 中人们的隐私，需要从隐私的发现、隐私的保护以及安全性评估三个方面来考虑。

隐私的发现是指如何利用正常合法的手段从 Web 公开的信息中获得个人的隐私，就象上面的例子那样。我们从隐私的发现角度来研究是为了总结或归纳出各种隐私发现的手段和可能性。首先要指出的是，隐私发现研究的最终目的并不是利用这些手段获得别人的隐私达到某种不良的目的，恰恰相反，而是通过分析这些手段来有针对性的防止个人隐私的泄露。上面举的例子就是一种大家容易想到的隐私的发现手段。可以大胆猜测，必然存在一些更加复杂的手段可以得到更多的个人信息，比如加入一些推测的方法。举个简单的例子：如果某个大学的 BBS 上的一个用户在自己帖子上透露自己的生肖是属鼠的，那么我们可以推测一他的年龄极有可能是 24 岁。

隐私的保护与隐私的发现相对，是指如何避免利用正常合法的手段从 Web 公开的信息中获得个人的隐私。隐私的保护可以从网站和搜索引擎三个方面来保证。

- 网站

网站的建设者有义务保护自己注册用户的个人隐私，因为这样对双方来讲都是受益的。问题的关键主要包括两点：一是如何是别哪些信息属于用户的隐私或者对用户的隐私具有潜在的威胁，比如 Email；二是如何保护这些信息，使得无法通过搜索引擎获得，比如将 Email 图片化。

- 搜索引擎

搜索引擎可以为人们获取 Web 中的信息带来极大的方便，但不正当的使用就会造成像隐私泄露这样的反面作用。在防止隐私泄露这方面，搜索引擎起着最关键的作用，因为就如图 4 所示，如果搜索引擎不允许像 Email、QQ 号这类敏感信息的搜索的话，那么就相当于切断了各个信息片段之间的联系，也就达到了隐私保护的日的。

遗憾的是，人们对此还没有引起足够的重视，无论是网站的建设者还是搜索引擎。这也正需要 Web 数据管理的研究者们提出有效的途径防止个人隐私在 Web 中的泄漏。

3.4 Mashups: 一种轻量级的 Web 应用构建方法

一种新型的基于 Web 的数据集成应用程序 Mashups 正在 Web 上悄悄出现并迅速繁荣起来。Mashups 是将多个不同的支持 Web API 的应用进行堆叠而形成的新型 web 服务。它利用其它 Web 服务来创建全新的 Web 应用，将来自不止一个数据源的内容进行组合，创造出更加增值的服务。Mashup 所能利用的外部数据源格式多种多样，表现出惊人的兼容性，它涵盖 public APIs, XML/RSS/Atom feeds, web services, HTML 等。

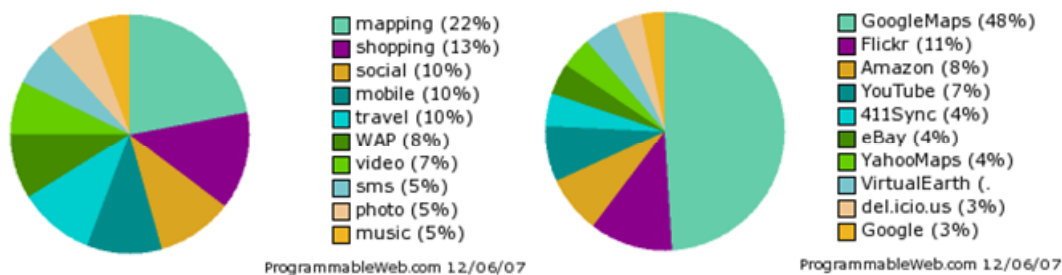


图 6 Programmableweb网站Mashups（左）和Web API（右）的相关统计信息

如果说博客是Web 2.0信息共享的代表，那么Mashups则是Web 2.0服务共享的代表。人们可以利用公开的Web API轻松地搭建自己的Mashups，颠覆了传统的完全从底层做起的开发方式。从某种角度讲，类似于使用程序员使用已有的类库编写新的程序。Programmableweb是目前最大的Mashups和Web API发布网站，它已经拥有了超过2500个Mashups和500个Web API，并且增长的速度也称加快的趋势，图6指出了Mashups的主题分布和Web API的来源分布。

直观看来，Mashups带给我们的似乎更多的是技术实现上的问题，但我们要从中发现背后隐藏的研究问题。虽然目前绝大多数的Mashups是由两个Web API搭建起来，但随着应用需求的不断提高，必然会有更加复杂的Mashups出现，由多个进行若干步搭建而成，类似一个树的层次结构：根节点是Mashups应用，叶节点是Web API。因此，如果要搭建一个应用确定的Mashups，理论上我们可能会面对下面几个问题。

Web API的选择：即如何从众多的Web API中选择出一组最合适的进行搭建。Web API选择不合适，会导致Mashups构建复杂甚至失败。另外，有时有多个Web API提供相同或类似的服务，比如地图API，这就需要根据实际应用场景、网络传输延迟等多方面因素来选择。

Mashups的维护：即当某个Web API出现意外事件的处理。由于Web API并不在本地，需要在Web中远程访问，因此非常容易出现网络中断或Web API不可访问等意外情况。这就需要采取措施进行应对，比如Web API的替换、根据现有可访问的Web API生成临时Mashups等。

4、结论

本文从探讨的角度，提出Web数据管理领域未来的研究方向，并非成熟的观点，其目的是希望能够起到抛砖引玉的作用。事实上，Web是新事物最有可能产生的平台，是新思想、新应用催生的载体。也许“在二十一世纪，唯一不变的是变化”这就话最能够在Web上得到体现。作为Web数据管理的研究者们必须要善于捕捉这种变化，走在变化的前头，成为Web发展的引导者。最后，我们引用大家耳熟能详的一句话作为结尾：“没有什么是不可能的！”

参考文献：

- [1] 数据空间：一种新的数据管理技术。孟小峰，李玉坤，张相於。计算机通讯，2007.8
- [2] L. Blunski, J.-P. Dittrich, O. R. Girard, S. K. Karakashian and M. A. V. Salles. A Dataspace Odyssey: The iMeMex Personal Dataspace Management System. In *CIDR*, 2007.
- [3] J. P. Dittrich and M.A.V. Salles. iDM: A Unified and Versatile Data Model for Personal Dataspace Management. In *VLDB*, 2006.
- [4] X. Dong and A. Halevy.. Indexing Dataspace. In *SIGMOD 2007*.
- [5] K. C.-C. Chang, B. He, C. Li, M. Patel, Z. Zhang: Structured Databases on the Web: Observations and Implications. *SIGMOD Record* 33(3): 61-70 (2004)