Pay-as-you-go information integration in personal and social dataspaces

Author(s):
Salles, Marcos Antonio Vaz

Publication Date:
2008

Permanent Link:

https://doi.org/10.3929/ethz-a-005716839 →

DISS. ETH NO. 18079

Pay-as-you-go Information Integration in Personal and Social Dataspaces

A dissertation submitted to
ETH ZURICH

for the degree of
Doctor of Sciences

presented by

MARCOS ANTONIO VAZ SALLES

Master in Informatics, Catholic University of Rio de Janeiro (PUC-Rio)
born 25th of April, 1977
citizen of Brazil

accepted on the recommendation of

Prof. Donald Kossmann, examiner
Prof. Jens Dittrich, co-examiner
Prof. Thomas Gross, co-examiner
Prof. David Maier, co-examiner

2008

Abstract

A personal and social dataspace is the set of all information pertaining to a given user. It includes a heterogeneous mix of files, folders, email, contacts, music, calendar items, images, among others, distributed among a set of data sources such as filesystems, email servers, network shares, databases, and web servers. In addition, it includes all connections this user has to other users in a number of online services such as social networking web sites. In spite of a personal and social dataspace being richly heterogeneous and distributed, only very limited tools are available to aid users manage their information and have a unified view over it. At one extreme, search engines allow users to pose simple keyword and path searches over all of their data sources. These systems, however, return only ranked lists of best-effort results and provide limited or no means for users to increase the quality of query results returned by the system over time. At the other extreme, systems built on top of classic database technology, such as traditional information-integration systems, provide precise query semantics for queries over a set of data sources. Although the quality of query results returned by these systems is high, they are typically restricted to a subset of the personal information of a user, given the need to specify complex schema mappings to integrate the data. As a consequence, these systems have limited coverage and provide equally limited support for non-expert users to refine their view of their personal information over time.

This thesis investigates a new breed of information-integration architecture that stands in-between the two extremes of search engines and traditional information-integration systems. We term this new type of system a *Personal Dataspace Management System (PDSMS)*. Like a search engine, when a PDSMS is bootstrapped, it provides a simple search service over all of the user's dataspace. In contrast to search engines, however, the PDSMS represents data not at the coarse-grained level of files (or text documents), but rather using a fine-grained graph-based data model. In addition, a PDSMS provides means for a user to increase the level of integration of her dataspace gradually, in a pay-as-you-go fashion. That is done by enabling users to provide simple integration "hints" that allow the PDSMS to improve the quality of query results. In contrast to traditional information-integration systems, however, at no point does the PDSMS require users to specify a global mediated schema for their information.

We make four main contributions to the design of PDSMSs. First, we propose the `iMeMex` Data Model (iDM), a simple, yet powerful, graph-based data model able to represent the heterogeneous data mix found in a personal and social dataspace. Our data model enables query capabilities on top of the user's dataspace not commonly found in state-of-the-art tools.

Second, we introduce iTrails, a technique for pay-as-you-go information integration in dataspaces. This technique enables users to increase the level of integration of their personal dataspaces by providing integration "hints", called trails, to the system. Trails specify relationships between arbitrary subsets of items in a user dataspace.

Third, we propose *association trails*, a technique to declaratively model fine-grained relationships among individual instances in a dataspace. With association trails, instances in a dataspace are connected in a graph intensionally. This graph may be used to explore the dataspace and find related instances according to different contexts (e.g. time, similar content, or same metadata).

Fourth, we integrate all of the previous contributions into the architecture of `iMeMex`, the first PDSMS implementation. The `iMeMex` PDSMS follows a layered and extensible architecture. The various layers of `iMeMex` provide increasingly higher levels of abstraction over a personal and social dataspace, bringing physical and logical data independence to this heterogeneous environment.

Zusammenfassung

Ein persönlicher und sozialer Datenraum umfasst die Menge aller Informationen, die einem bestimmten Benutzer zugeordnet sind. Hierzu gehören eine heterogene Mischung aus Dateien, Ordnern, E-Mails, Kontakten, Musik, Kalender-Einträgen, Bildern und anderem. Diese sind verteilt auf verschiedene Datenquellen wie Dateisysteme, E-Mail-Server, Netzwerk-Ablagen, Datenbanken und Webserver. Darüber hinaus umfasst der Datenraum alle Verbindungen, welche dieser Benutzer zu anderen Benutzern in Online-Portalen für Soziale Netzwerke hat. Obwohl ein persönlicher und sozialer Datenraum also sehr heterogen und auf verschiedene Datenquellen verteilt ist, gibt es doch nur eine sehr geringe Anzahl von Werkzeugen, welche einem Benutzer helfen, seine Informationen zu verwalten und eine einheitliche Sicht darüber zu erhalten. Auf der einen Seite ermöglichen Suchmaschinen einfache Schlüsselwortabfragen beziehungsweise Pfadabfragen zu machen. Diese Systeme liefern jedoch nur eine geordnete Liste von Ergebnissen und bieten dem Benutzer keine oder nur wenig Möglichkeiten, die Qualität der Suchergebnisse nach und nach zu verbessern. Auf der anderen Seite ist es möglich, mit Systemen, die auf klassischer Datenbanktechnologie aufbauen, wie zum Beispiel traditionellen Information-Integrationssystemen, eine präzise Abfrage-Semantik über verschiedenen Datenquellen zu ermöglichen. Die Qualität der Suchergebnisse dieser Systeme ist zwar hoch, aber üblicherweise ist die Suche auf eine Teilmenge aller Informationen beschränkt, da für jede Datenquelle komplexe Schema-Abbildungsregeln spezifiziert werden müssen. Als Folge davon bieten diese Systeme nur eine begrenzte Abdeckung der Daten und bieten auch nur eingeschränkte Unterstützung für Nicht-Experten, ihre Sicht über ihre persönlichen Informationen im Laufe der Zeit zu verfeinern.

Diese Arbeit geht einen neuen Weg. Wir untersuchen eine neue Art von Information-Integrationsarchitektur, welche zwischen den beiden Extremen von Suchmaschinen und traditionellen Information-Integrationssystemen steht. Wir nennen diese neue Art von System *Persönliches Datenraum-Management-System (PDSMS)*. Vergleichbar mit einer Suchmaschine bietet ein PDSMS anfänglich einen einfachen Suchdienst über alle Informationen eines Benutzers an. Im Gegensatz zu Suchmaschinen repräsentiert ein PDSMS die Daten jedoch nicht auf einer grobkörnigen Ebene wie Dateien (oder Textdokumente), sondern mit Hilfe eines feinkörnigen, grafischen Datenmodells. Zusätzlich bietet ein PDSMS die Möglichkeit, das Niveau der Datenintegration kontinuierlich zu erhöhen. Dies wird erreicht, indem ein Benutzer dem System einfache Integrations-“Tipps” gibt, welche dem System helfen, die Qualität der Suchergebnisse zu verbessern. Ungleich einem traditionellen Information-Integrationssystem, verlangt ein PDSMS kein kompletes Regelset, welches alle Datenquellen miteinander verbindet.

In dieser Arbeit machen wir vier wichtige Beiträge zur Gestaltung von PDSMS. Erstens präsentieren wir das iMeMex Daten-Modell (iDM), ein einfaches aber trotzdem mächtiges, grafisches Datenmodell, welches den heterogenen Datenmix in einem persönlichen und sozialen Datenraum darstellen kann. Unser Datenmodell ermöglicht Abfragen über den Datenraum eines Benutzers, welche in anderen modernen Werzeugen nicht möglich sind.

Zweitens stellen wir iTrails, eine Technik für Pay-as-you-go-Informationsintegration in Datenräumen vor. Diese Technik ermöglicht Benutzern, den Grad der Integration von Daten Schritt für Schritt durch Integrations-"Tipps" — so genannten Trails — zu erhöhen. Trails geben Hinweise auf Beziehungen zwischen beliebigen Teilmengen von Elementen im Datenraum eines Benutzers an.

Drittens schlagen wir *Assoziationstrails* vor, eine Technik, um deklarativ feinkörnige Beziehungen zwischen einzelnen Elementen in einem Datenraum zu modellieren. Mit Assoziationstrails sind Instanzen in einem Datenraum über die Auswertung eines Prädikates verbunden. Hierdurch entsteht ein assoziativer Overlay-Graph. Dieser Graph wird ausgenutzt, um einen Datenraum nach ähnlichen Kontexten zu durchforschen, zum Beispiel Zeit, ähnlichem Inhalt, gleichen Metadaten oder ähnlichen Interessen von Benutzern.

Viertens verbinden wir alle oben genannten Beiträge zu einem konkreten System: iMeMex. Dieses System ist eine der ersten Implementierungen eines PDSMS. Zudem ist iMeMex erweiterbar und besteht aus verschiedenen Systemebenen. Jede weitere Ebene in iMeMex führt zu höherer Abstraktion in einem persönlichen und sozialen Datenraum und ermöglicht damit physische und logische Daten-Unabhängigkeit, sowie Pay-as-you-go-Informationsintegration.