Provenance Meets Adaptive Hypermedia

Evgeny Knutov
Department of Computer Science
Eindhoven University of Technology
P.O. Box 513, 5600 MB
Eindhoven, the Netherlands
e.knutov@tue.nl

Paul De Bra
Department of Computer Science
Eindhoven University of Technology
P.O. Box 513, 5600 MB
Eindhoven, the Netherlands
debra@win.tue.nl

Mykola Pechenizkiy
Department of Computer Science
Eindhoven University of Technology
P.O. Box 513, 5600 MB
Eindhoven, the Netherlands
m.pechenizkiy@tue.nl

ABSTRACT

In this paper we consider provenance modelling in Adaptive Hypermedia Systems (AHS). We revisit adaptation and data provenance questions and bring up new and complementary aspects of adaptation and provenance, showing similar and supplementing characteristics. We also scrutinize the provenance importance and issues in Adaptive Hypermedia (AH). The aim of this paper is to extend the conventional AH classification questions with the notion of data lineage which essentially plays an important role in adaptation.

Categories and Subject Descriptors

H.5.4 [**Hypertext/Hypermedia**]: Architectures, Theory

General Terms

Design, Theory

Keywords

adaptive hypermedia, adaptation questions, provenance, W7 provenance model

1. INTRODUCTION

A hypermedia application offers its users navigational freedom within a large hyperspace. AH offers personalized content, presentation, and navigation support. Most AHS do so by building a User Model (UM) and using that to guide an Adaptation Engine (AE). The subject of UM scrutability has been studied extensively [4] because users want to be able to review (scrutinize) what the system knows about them. Adaptation scrutability still remains largely uncharted territory: most systems are not set up to explain to users why the content, presentation and navigation are adapted the way they are. We take a new approach towards offering scrutability by studying the parallels with the area of data provenance. In general provenance data aims at providing users with an explanation of the origin, history and evolution of data and processes.

In this paper we re-examine the adaptation questions stated in the very beginning of the AH era [1] in the context of data provenance. In fact we match the major questions of adaptation (*What, To What, Why, Where, When, How*) with a question-driven provenance model (section 2). Our major goal is to show how complementary to each other adaptation and provenance are (in terms of question-based classification) (see section 3). We also present a number of demonstrative examples of data lineage, harvesting and interpretation importance in AH (section 4). And finally we investigate what the problems of designing provenance support in AHS are (section 5).

Provenance.

Provenance is information about the origin, ownership, source, history, lineage and/or derivation of an information object or data. Provenance is important as it is vital for providing the detailed explanation of user action, system usage and data origin and inference, ensuring analysis of dependencies in the system and repeatability of user actions.

Provenance Modelling: There are several provenance modelling approaches:

- **Data-centric**: refers to meta-data models such as Dublin-core, Premis, OAIS, etc., where a metadata schema stores the provenance data; this was for instance presented in [7].
- **Process-centric**: refers to the description of the process with particular change steps through which this metadata is obtained. It collects not only the data about a particular step, but about the application processes as well [3].
- **Pipeline-centric**: as investigated in [2] a new pipeline-centric approach to provenance data was introduced for the class of workflow-based applications, which helps to determine the provenance of the application output based on the provenance graph of the application.

We believe that process and pipeline type of obtaining provenance data [2] can be complementary with the adaptation process shown in Figure 1. These two have very much in common with the process of collecting data that could be used for the provenance analysis. By *Adaptation Process* we mean the interaction in AHS which starts with the goal statement, exploits features of the user and domain mod-

HT'10, June 13–16, 2010, Toronto, Ontario, Canada..

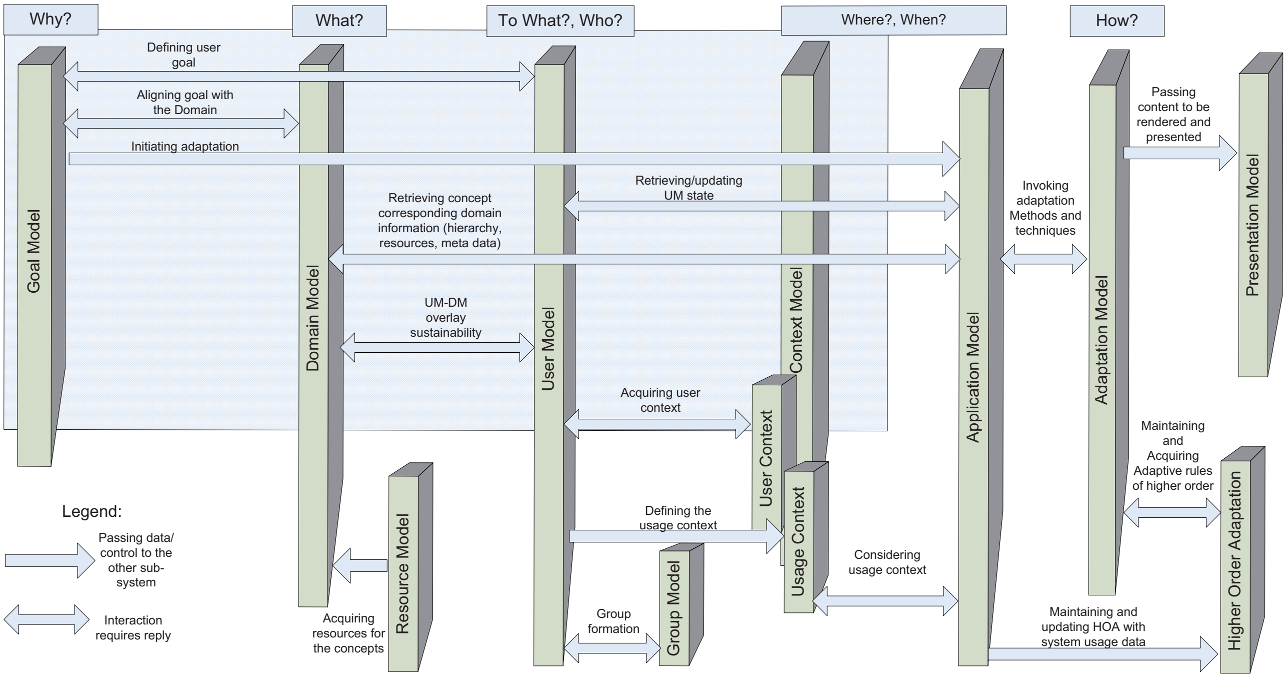
Why?
What?
To What?, Who?
Where?, When?
How?
Defining user goal
Aligning goal with the Domain
Initiating adaptation
Retrieving/updating UM state
Retrieving concept corresponding domain information (hierarchy, resources, meta data)
Invoking adaptation Methods and techniques
Passing content to be rendered and presented
UM-DM overlay sustainability
Acquiring user context
Maintaining and Acquiring Adaptive rules of higher order
Defining the usage context
Considering usage context
Group formation
Maintaining and updating HOA with system usage data
Acquiring resources for the concepts
Goal Model
Domain Model
User Model
Context Model
User Context
Usage Context
Group Model
Resource Model
Application Model
Adaptation Model
Presentation Model
Higher Order Adaptation
Legend:
Passing data/ control to the other sub-system
Interaction requires reply

Figure 1: Conceptual Adaptation Process Sequence

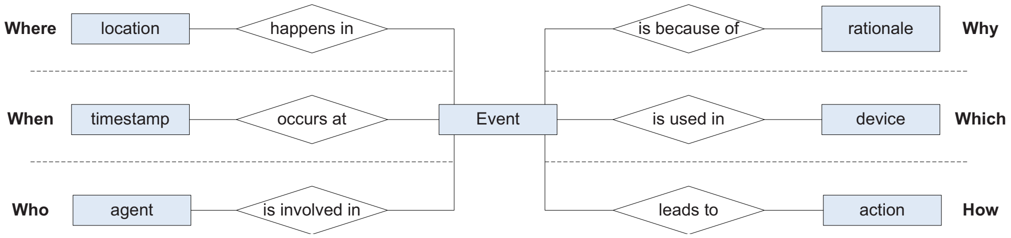

Figure 2: W7 Provenance Model [9]

els in different contexts and adapts various aspects of the system to the user. This sequence of user-adaptation actions could be aligned according to the classification of AH methods and techniques as well as Provenance questions and results in the adaptation sequence, coupling the 'layers' of AHS (Figure 1).

Possible alternative and less generic classification of provenance models includes but is not limited to:

- In the *Data Specific Provenance Model*, each data type has its own provenance model, carrying forward information covering the complete path of the data. This approach benefits in a way that all provenance metadata comes with each result, but lacks models interoperability.
- The *Generic Complete Provenance Model* retains provenance information in the form of prior data sets and transformations. This approach has the advantage of being very complete, however it requires storage of intermediate results and hardly be visualised and analysed in a simple way.
- In the *Hierarchical Provenance Model* only the provenance information covering the previous transformation is retained, but at the same time all the provenance data can be recursively returned. This type of retaining hierarchical data may correspond to a hierarchical structure of the Domain Model (DM). Thus it is complementary in such a way that it keeps track of the user following the conceptual structure in depth gaining more detailed knowledge on a particular subject represented with this hierarchy.

We leave out the question of a particular provenance modelling approach, but rather consider the generic model of provenance information. We will use "W7" which is a provenance model representing diverse information about the data produced in the system [9].

2. ADAPTATION AND PROVENANCE QUESTIONS

Adaptation Questions.

There is a number of adaptation questions that have to be answered in order to build an AH application [5]. Moreover they define the adaptation process aligning the structure of system sub-components. These questions also denote the adaptation process flow.

The core of adaptation is defined by stating and answering six major questions:

- What can we adapt? (*"What?"*)
- What can we adapt to? (*"To What?"*)
- Why do we need adaptation? (*"Why?"*)
- Where can we apply adaptation? (*"Where?"*)
- When can we apply adaptation? (*"When?"*)
- How do we adapt? (*"How?"*)

We want not just to revisit these questions, but to address the issue of aligning them (and corresponding answers) in a common, modular structure of a generic AHS architecture.

By answering the major adaptation questions we elaborate the adaptation process description outlined in Figure 1 and consider it in the context of the provenance model. This process is usually initiated by the user stating the adaptation goal and thus answering the "Why adaptation is needed?" question. Then in the process we consider the "What?" and "To What?" questions, which emphasize the Domain Model (DM) and the User Model (UM) description. "When?" and "Where?" in this process go further providing context and application area definitions. Lastly, the, "How?" question describing methods and techniques on conceptual and implementation level and finally all together resulting in an AHS description.

Provenance Questions.

Hereafter we will consider one of the most extensive definition of the provenance models and investigate the way it can be complementary to AHS, possibly after some extensions. According to [9] ***Provenance*** is defined as a *n-tuple P = (What, When, Where, How, Who, Which, Why, Occurs_at, Happens_in, Leads_to, Brings_about, Is_used_in, Is_because_of)*, which represents the "W7" model [9], where:

- *"What"* denotes the sequence of events that affect the data object;
- *"When"*, the set of event times;
- *"Where"*, the set of all locations;
- *"How"*, the set of all actions leading up to the events;
- *"Who"*, the set of all agents involved in the events;
- *"Which"*, the set of all devices;
- *"Why"*, the set of reasons for the events.

According to the model an action is taken by *agents* using *devices* for *reasons*, which is reflected by the various relationships existing between "what" and the elements "who", "which" and "why". The conceptual schema of the W7 provenance model is presented in Figure 2. In other words these questions may also describe such information as: event decision (what), duration (when), activity (how), method (which), person (who), arguments and justification (why).

This scheme shows a schematic representation of the W7 model. In Table 1 we will consider Provenance questions side by side with Adaptation questions and aim at aligning them hence extending AHS with the notion of provenance.

3. ALIGNING QUESTIONS

Considering the question-centric, extensive definition of the W7 Provenance Model [9] and the AH methods and techniques classification questions [5] we combine and align the questions and corresponding answers. Such an alignment will be able to provide complementary features description. Here we investigate commonalities and similarities in the semantics of the answers and meanings of these questions, emphasising the idea that provenance information can be very useful in AHS and at the same time provenance information can help to reason in AH, for example providing more explanations to the end user or making the system more trustworthy. In Table 1 we map questions and look for common understanding in-between Provenance and AH.

"What?" — answer to this question on the one hand describes the way domain information is represented in the system (hierarchy of concepts, ontology, etc.) and on the other hand shows what events in the system these data objects can affect.

"Who?" ("To Whom?"), "Which?" — answers to these questions give us an idea of the UM environment: *Which?* defines the device capabilities and in general *Who?* represents the user profile. They also describe the set of devices and agents involved in the process from the provenance point of view and can be used to select the target group of users, representing the high-level user division and defining the group adaptation parameters.

"To What?" — answer narrows down the user profile to a particular set of attributes involved in the adaptation process (accessed and updated by the system to retrieve or refresh the current state of the user knowledge, interest, competence, etc.). These are usually domain dependent attributes. There is no actual match on the provenance question here, however the history of UM attributes' access and updates directly refers to storing and harvesting provenance data from user logs.

"Why?" — answer determines the set (one-at-a-time or a sequence) of goals of adaptation and describes the set of reasons for initiating the concerned adaptation process. Thus, these two indicate the premises of the adaptation process in general, provide arguments and describe the way adaptation is initiated.

"When?" and ***"Where?"*** — answers are registered as a part of the provenance model events. The AHS keeps track of these changes and interprets this data to be used as the input for the reasoning component, which should take into account this time and place contextual information.

"How?" — answer provenance data records event-action sequences, describing mostly the syntax of these changes, on the other hand AHS describes the semantics (understanding of these cause-event relationships), contributing to the picture of the reasoning model. As a whole it describes AE functionality of the system.

Examples of Provenance Importance in AH

We have mentioned some of the motivating examples in the introduction, but we would like to extend this list and provide more insight on the importance of AH provenance. We consider the following examples to be significant:

- *Adaptation*: provenance data can be directly used in the adaptation process. Being interpreted by the AE to determine the result of the next adaptation step, it may extend the capabilities of the adaptive reasoning from a conventional pre-authored type to become more context and provenance/lineage dependent, taking into account not only UM values and updating them, but analysing the origin of these UM updates and thus adjusting them accordingly (e.g. AE may

Table 1: Aligning Adaptation and Provenance questions

Questions	AHS	Provenance Model	Comments
What?	Domain Model	denotes the sequence of events that affect the data object	answers describe the sequence of events when the user gets access to the domain information and acquires domain knowledge
Who? (To Whom?) Which?	describes the user profile selection (or/and device usage) (e.g. can be used to select a group or target users)	the set of all agents and/or devices involved in the process	
To What?	UM attributes (selecting particular attributes that are accessed and updated within the concerned adaptation process)	no actual representation in terms of provenance question, however historical information on accessing and updating UM represents provenance information	
Why?	stating the adaptation goal(s) (might be a domain concept, representing either a new goal to follow or a sequence of concepts)	the set of reasons for triggering a particular event (evidence of what has happened)	reasons and goals are complementary, indicating the premises of the adaptation process
When? Where?	Application Model (which serves as the core of the system: coupling other layers and dispatching information in AHS) and Context information keeps track and interprets the context information	the set of event times and locations	contextual information in general
How?	describing AH methods and techniques on a conceptual and implementation level (Adaptive Engine (AE) functionality); explains the sequence of event-actions; ***describes the semantics*** of cause-effect relations	the set of all actions leading up to the events (keeping track of the events, and corresponding action in the system); ***describes the syntax*** of events and actions recorded	in pair provenance and AH describe the syntax and semantics of AE functionality (record events and actions and show cause-effect relationship)

interpret the knowledge source properties and assign different scores to the user depending on the source trustworthiness).

- *Explanation and Analysis*: explanations not only include information about system usage (e.g. the prerequisite knowledge level is reached and the user can access new information) but also where this information comes from, how it was derived, etc. (e.g. the user is provided with an additional explanation about the origin of their 'knowledge' or 'interest' which may come from an update event issued by the system interpreted in a way the user can understand). This could be useful both in providing additional explanation and recommendations to the user and in the analysis of the system behaviour by the domain expert.
- *Usage Patterns Analysis*: could be helpful to discover and analyze certain abnormal user behaviour patterns (in combination with information retrieval methods, e.g. provide the dependency of unusual data and the source of it using the provenance information). Essentially provenance data can be mined to discover these patterns and used to analyze the origin of such a behaviour.
- *Information reliability*: provenance regarding how the adapted information was delivered to a user helps to ensure that it can be trusted so that the user understands the way he received the information and is explained why he gets this and where it comes from.
- *Information currency (prevalence/efficacy)*: capturing provenance such as when the update to the User Model is done could be used to avoid being misled (e.g. by outdated information).
- *Semantics of provenance*: provenance data provide a semantic extension to the system, expanding the description of the data to what is answered by the provenance questions. For instance the data providing the information of the data origin will extend the semantics of each particular delivered information unit. Thus provenance information "naturally" extends the semantics of the existing data model [8].
- *Adaptation Process*: considering process and pipeline centric types of provenance information we anticipate mapping the notion of provenance on the adaptation process shown in Figure 1. Each step of the adaptation sequence here represents a single data transformation from the data lineage point of view, answering aforementioned questions of Provenance and AH models.

4. PROVENANCE ISSUES AND PROSPECTIVE SOLUTIONS

In this section we would also like to indicate some of the issues that one may face while investigating provenance of data in AHS. These are only some of the most common problems:

Harvesting of provenance data. Since AHS usually has a rather complex structure the problem of data harvesting arises. That's why we proposed a layered structure (Figure 1), strictly distinguishing AH data (essentially adaptation questions) and process, using major AH classifications. This will provide transparency of the system functionality, reduce (by associating type of provenance data with the layer) places of provenance data origin and help in analysing the data.

Understanding the semantics of provenance. Understanding provenance enables users to share, learn the meaning and take advantage of data, facilitating collaboration and learning, reducing the number of deadlocks and providing mechanisms to conceptual modelling [9]. We have partially covered this issue in this paper trying to understand and match the semantics of provenance and AH data (in terms of questions).

Diversity in data types and many places of data origin. As mentioned above, considering a generic layered structure of AHS we try to put things in order and clearly distinguish between the major questions first of Adaptation and then Provenance, thus reducing and foremost matching overall places of origin diversity. Anticipating the layered structure of AHS we foresee the idea of clear separation in such a way that already mentioned harvesting of the provenance data becomes transparent and clear which simplifies data analysis and reduces the number of ambiguous places of data origin.

Moreover as investigated in [6] some of these provenance problems related to data dynamics, diversity and overload and in particular issues of storing, retrieving (harvesting) and analysing (interpreting) data in AH can be facilitated by using versioning techniques. There are many more problems to be considered, such as reusability and alignment of provenance data, different reasons for needing provenance, its representation and propagation. Finally implementing and using provenance could become an issue. Without fulfilling most of these requirements it may cause misuse and misinterpretation of the AH related data (e.g. misaligning user and a domain knowledge) and lead to unpredictable adaptation results (e.g. termination and confluence problems).

5. CONCLUSIONS AND FUTURE WORK

In this paper we revisited provenance modelling approaches and put them in the context of AHS which resulted in a complementary features (questions) description in Table 1. Moreover we investigated the questions of provenance importance and issues in AH field proposing prospective solutions.

The generalization table presented in the paper and the fact that premises for the 'event' in W7 system and the result of adaptation process in AHS show that they have much in common and to some extent complement each other, especially in the field of Adaptive Systems (e.g. describing syntax and semantics of different questions or describing the user profile, selecting the target group to determining the agents and devices involved in the adaptation process). The significance of the provenance data is getting more important. It can be used to provide the richer user experience, implement more sophisticated adaptation and recommendation techniques and at the same time to analyze and explain the system functionality (so as to facilitate scrutability of the adaptation). As a result we anticipate that provenance will be considered as an essential part of AHS (and in fact it is already partially there), and that an AHS itself will provide provenance information.

In our future work we will study the aforementioned examples of provenance importance in depth, analysing the impact on the adaptation process and comparing results with the conventional AHS to evaluate the utility of AH provenance.

6. ACKNOWLEDGEMENTS

This work has been supported by the **NWO GAF: Generic Adaptation Framework** Project.

7. REFERENCES

[1] P. Brusilovsky. Methods and Techniques of Adaptive Hypermedia. *User Modeling and User Adapted Interaction*, 6(1-2):87–129, 1996.

[2] P. Groth, E. Deelman, G. Juve, G. Mehta, and B. Berriman. Pipeline-centric Provenance Model. In *Proc. of WORKS '09: the 4th Workshop on Workflows in Support of Large-Scale Science*, pages 1–8, New York, NY, USA, 2009. ACM.

[3] F. James and B. Rajendra. Earth system science workbench: A data management infrastructure for earth science products. In *Proc. of SSDBM'01: the 13th Int. Conference on Scientific and Statistical Database Management*, pages 180–189, Washington, DC, USA, 2001. IEEE CS.

[4] J. Kay. Scrutable Adaptation: Because We Can and Must. In *Proc. of AH 2006: 4th Int. Conference on Adaptive Hypermedia and Adaptive Web-Based Systems*, pages 11–19, Berlin-Heidelberg, 2006. Springer.

[5] E. Knutov, P. De Bra, and M. Pechenizkiy. AH 12 years later: a comprehensive survey of adaptive hypermedia methods and techniques. *New Rev. Hypermedia Multimedia*, 15(1):5–38, 2009.

[6] E. Knutov, P. De Bra, and M. Pechenizkiy. Versioning in Adaptive Hypermedia. In *Proc. of DAH'09: Int. Workshop Dynamic and adaptive hypertext: generic frameworks, approaches and techniques*, pages 61–71, Aachen, 2009. CEUR-WS.org.

[7] C. Pancerella and et al. Metadata in the collaboratory for multi-scale chemical science. In *Proc. of DCMI'03: the 2003 Int. Conference on Dublin Core and Metadata Applications*, pages 1–9. Dublin Core Metadata Initiative, 2003.

[8] S. Ram. Data Provenance. Project accomplishments, The University of Arizona, Advanced Database Research Group, The University of Arizona, Mar 2006.

[9] S. Ram and J. Liu. Understanding the Semantics of Data Provenance to Support Active Conceptual Modeling. In *Proc. of ACM-L: 1st Int. Workshop on Active Conceptual Modeling of Learning*, pages 17–29, Berlin-Heidelberg, 2006. Springer.