User constructed data integration via mixed-initiative design

Anthony Tomasic, John Zimmerman, Ian Hargraves, Roderick McMullen

Carnegie Mellon University
{tomasic,johnz}@cs.cmu.edu

Abstract

Administrators frequently perform data integration "by hand" on the desktop as part of the execution of administrative tasks. This position paper discusses the application of mixed-initiative design to this problem. This design style leverages the interaction between a user and an intelligent assistant, minimizing the effort required to execute a task.

Introduction

As computational systems increase in capabilities and reasoning power, they become more difficult to effectively use. This difficulty has resulted in a fundamental design tension between direct manipulation interfaces and agent-based interfaces. In direct manipulation interfaces, the human is in control and directs the system in detail. The system passively accepts interaction. In agent-based systems, the agent is in control and the human passively accepts interaction. Each design approach has advantages and disadvantages. In an effort to provide the advantages of both design approaches, mixed initiative interfaces focus on a cooperative interaction between human and machine. When performing a task, control passes between human and machine in an effort to efficiently complete the task. We propose studying mixed-initiative design in the context of data integration.

Broadly stated, data integration is the creation of a consistent representation of a collection of heterogeneous data sources. With respect to databases, data integration has most extensively been studied in the context of querying multiple data sources simultaneously (federated database integration) and in the context of the integration of data sources by the construction of mappings between them (extract-transform-load database integration). Current data integration solutions typically require an entire system lifecycle to produce a consistent representation of a new set of sources. Thus, organizations are continually paying the cost of integration of data sources. We propose that users construct "desktop" data integration via a mixed-initiative interface that communicates with an intelligent assistant. Given input from the interface, the intelligent assistant learns new data integration scripts.

Contextual Inquiry

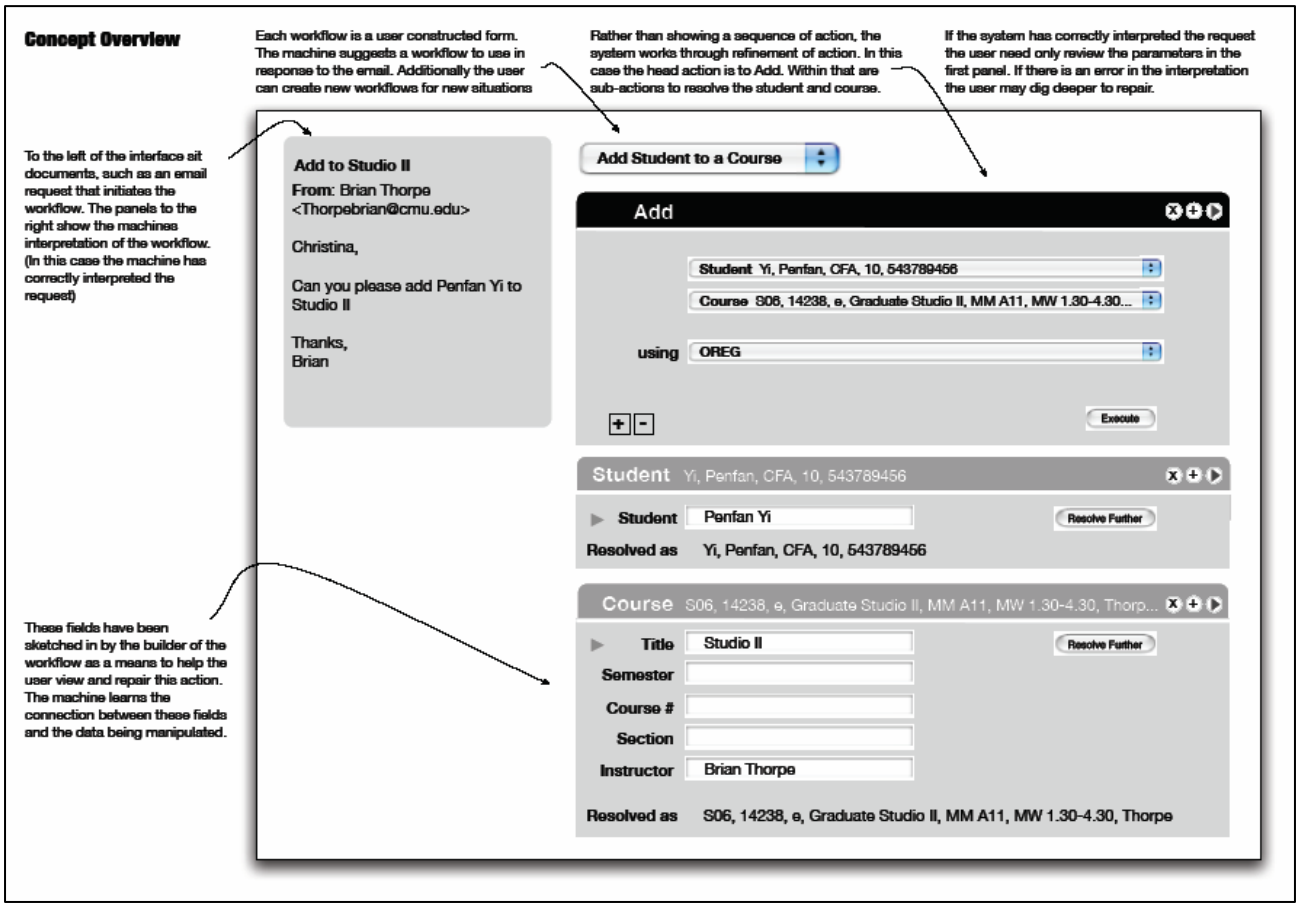
A contextual inquiry of administrators at Carnegie Mellon University (Hargraves 2006) reveals that many administrators construct ad-hoc, dynamic, data integration procedures to accomplish routine tasks.

One example procedure recorded in the contextual inquiry describes an administrator adding a student to a course. The procedure consists of ten steps as follows:

1. Receive email request "Could you please add Penfan Yi to Studio II?" from an instructor to add a student to a class roster. [E,T]
2. Determine that "Studio II" for this instructor is course "14-238" and section "e" by checking a handwritten list of course sections and instructors. [R]
3. Login into class administration system.[T]
4. Navigate on the administration system to the waitlist for the course section. [N]
5. If student is waitlisted, select student record from waitlist and go to step 7. [L,R,C]
6. Else, navigate, lookup, and resolve ambiguities in system to select the student in the directory of all students. [N,L,R]
7. Navigate to the student's record in the student management part of system. [N]
8. Add the course to the student, and specify that violated prerequisites should be ignored, if any. [T]
9. Email student from within the system. [T]
10. Reply to instructor e-mail via the email system. [N,T]

The operations performed by administrators were grouped into classes and sorted in order of frequency. The two most frequent classes were those that sent information to the administrator and those where the administrator communicated information to another user or system. This reflects the fact that operations are not simply transformations of data, they are initiated in the context of social connections and obligations and their conduct is of concern to parties beyond the immediate user. This class is indicated by the "T" (transaction) designation at the end of a step.

The next most frequent class identifies the tailoring of a particular workflow to the circumstances in hand. This class consists of extraction ("E"), lookup ("L"), and resolve ("R"). Extraction corresponds to classical information extraction. Lookup corresponds to a keyword based database query. The "resolve" operation refers to reference resolution. For example, in step 2 above, the administrator uniquely identifies a course and section. Note that all three operations translate unstructured information



into a structured form. Finally, navigation ("N") and conditions ("C") complete the set of operations performed. Once in a structured form, the intelligent assistant can execute the workflow. Note that these operators are precisely those that are concerned with shaping data integration to the circumstances of its use. To assist the user in constructing workflows from these operators, we have designed a mixed-initiative interface, illustrated above.

Conclusion

In this position paper we outlined the result of a contextual inquiry that documents data integration tasks performed manually on the desktop. We proposed to support data integration through mixed-initiative design for intelligent assistants. To better understand the intelligent assistant aspect of our vision, we have constructed two prototypes. The Virtual Information Officer (Tomasic, Simmons, Zimmerman, 2006) (Zimmerman, Tomasic, Simmons, 2007) analyzes e-mail messages and presents pre-filled forms that assist the user in processing the message. Workflow by Example (Tomasic, McGuire, Myers, 2006) constructs workflow scripts by observing demonstrations of workflows generated by the user. Future work will focus on constructing the intelligent assistant user constructed data integration tasks.

Acknowledgements

This material is based upon work supported by the Defense Advanced Research Projects Agency (DARPA), through the Department of the Interior, NBC, Acquisition Services Division, under Contract No. NBCHD030010, Delivery Order No. D0300100003.

References

Ian Hargraves, 2006. *Report on Design Investigations for User Constructed Workflows*, Unpublished Manuscript.

Anthony Tomasic, R. Martin McGuire, Brad Myers, 2006. *Workflow By Example: Automating Database Interactions via Induction*, Carnegie Mellon University Technical Report Nr. CMU-ISRI-06-103.

Anthony Tomasic, Isaac Simmons, John Zimmerman, 2006. *Processing Information Intent via Weak Labeling*, Conference on Integrated Knowledge Management (CIKM), Washington, D.C.

John Zimmerman, Anthony Tomasic, Isaac Simmons, Ian Hargraves, Ken Mohnkern, Jason Cornwell, Robert McGuire, 2007. *VIO: a mixed-initiative approach to learning and automating procedural update tasks*, CHI 2007, San Jose, CA.