Quality-driven Information Integration Evolution

Ph.D. in Computer Science Course - Series XXVIII

Riccardo Porrini

**Supervisor:** Dott. Palmonari
**Tutor:** Prof. Messina
Web Information Integration

Heterogeneous Information Sources

DataSpace Management System

Source Schema / Instances

Mappings

Target Schema / Instances

Application Front-end

Refinement

Domain Experts

Querying

Browsing

End-Users

Riccardo Porrini | Quality-driven Information Integration Evolution
Web Dynamicity

Schema and Instances Changing Over Time

Riccardo Porrini | Quality-driven Information Integration Evolution
Web Dynamicity

End-Users’ Information Needs Changing Over Time

Interest over time

tablet
Search term

netbook
Search term

Example in the eCommerce Domain

Taxonomy Integration and Evolution

Source Categories
- Netbooks
- Apple Ipad

Target Category
- Netbooks
- Netbooks

End-Users dynamic behavior
- Ipad: A netbook without keyboard
- Ipad: A tablet
- Netbooks? Come on, we are in 2013, get a tablet!

Application Front-end

Riccardo Porrini | Quality-driven Information Integration Evolution
Quality-driven Information Integration Evolution

DataSpace Management System
- Source Schema / Instances
- Target Schema / Instances
- Mappings

Evolution

Quality Assessment
- Conformance to Domain Experts Knowledge
- Fitness to End-users' Information Needs

Application Front-end

Heterogeneous Information Sources
- RDF
- XML
- CSV

Refinement
- Domain Experts

Feedback
- Querying
- Browsing
- End-Users
Related Works

- Explicit Mappings \([Iv09, Ca09]\)
- Implicit Interrelated Components
- Components in isolation (Mappings, Schema, Instances)
- Feedback Target
- Evolution:
  - Improvement
  - Query Logs [Ma12]
  - Query Answers Click-through [Ma07, Je08]
  - Query Answers [Ya13, He12]
  - Mappings [Iv09, Ca09]

Feedback Source: Explicit, Implicit, Behavior

- My Work

References:
- [Iv09] Ives et al. Interactive data integration through smart copy & paste. In CIDR, 2009
- [Ca09] Cafarella et al. Data integration for the relational web. In PVLDB, 2009
Goals

Hypothesis

Quality defined in terms of **fitness** to end-users information needs and **conformance** to domain experts knowledge can be used to evolve integration over time.
Goals

Hypothesis

Quality defined in terms of **fitness** to end-users information needs and **conformance** to domain experts knowledge can be used to evolve integration over time

Goal 1

Model the **quality** of a DataSpace Management System considering **end-user behavior**

Riccardo Porrini | Quality-driven Information Integration Evolution
Goals

Hypothesis
Quality defined in terms of **fitness** to end-users information needs and **conformance** to domain experts knowledge can be used to evolve integration over time.

Goal 1
Model the **quality** of a DataSpace Management System considering **end-user behavior**

Goal 2
Model mappings evolution considering **continuous quality evaluation**
Case Studies from Price Comparison Engines

Case Study ① Adaptive Mappings and Facets Management
Case Study ② Adaptive Mappings and Taxonomy Refinement

DataSpace Management System

Source Taxonomies / Instances

Target Taxonomy / Instances

Evolution

Mappings

Refinement

Domain Experts

Quality Assessment

End-Users

8M product offers

2000 sources

500 target categories

10M users/month

Riccardo Porrini | Quality-driven Information Integration Evolution
Case Studies from Price Comparison Engines

Case Study 1: Adaptive Mappings and Facets Management
Case Study 2: Adaptive Mappings and Taxonomy Refinement

Riccardo Porrini | Quality-driven Information Integration Evolution
Preliminary Work

Lightweight matching algorithm for instance mapping tasks [1]

[1] Porrini et al. COMMA: A Result-Oriented Composite Autocompletion Method for E-marketplaces. In Web Intelligence, 2012 (extended work to be published on Web Intelligence and Agent Systems Journal)
Ongoing Work

Case Study 1 Adaptive Mappings and Facets Management

Domain Experts Defined mappings

Inferred Mappings

Inferred Facets

Riccardo Porrini | Quality-driven Information Integration Evolution
Ongoing Work

Case Study ① Adaptive Mappings and Facets Management

Domain Experts Defined mappings

Deployed on TrovaPrezzi

Inferred Mappings

Inferred Facets

Wines

0.50 cl 0.75 cl

Cabernet

Grape

Cabernet

Bordaux

Fabiano

Intenso

Chateau Greysac 2009

Red Wines

Riccardo Porrini | Quality-driven Information Integration Evolution
Ongoing Work

Case Study ① Adaptive Mappings and Facets Management

Domain Experts
Defined mappings

Inferred
Mappings

Inferred Facets

Currently collecting end-users usage statistics

Riccardo Porrini | Quality-driven Information Integration Evolution
Next Steps

Case Study ① Adaptive Mappings and Facets Management

- Qualitative analysis of usage statistics
- Include usage statistics analysis into extraction and evolution phase
Next Steps

**Case Study 2** Adaptive Mappings and Taxonomy Refinement

- Model mappings and taxonomy management operations
- Include temporal information on mapping definition
- Model user behavior
- Link mappings and taxonomy management model to user behavior model
Past Activities

Publications

Attended Courses and Schools
- Advanced Analytics and Behavior Informatics - DISCo
- Foundations of Data Exchange and Integration - Politecnico di Milano
- Third ESWC Summer School - Kalamaky - Crete (GR)

Attended Workshops and Seminars
- WOA 2012: 13th National Workshop “Dagli Oggetti agli Agenti” - DISCo
- Big Data e la forza degli eventi - DISCo
- Semantic Constraints for Data Quality Assessment and Cleaning - DISCo
- Progettare e Fare Open Data. Metodologia e tools sviluppati in Evodevo a partire dallesperienza Open Data INPS - DISCo
- Phase Transitions in Social and Economic Systems - DISCo
- Optimum Hyperpaths in Directed Hypergraphs - DISCo

Teaching
- Lecturer (1 lecture) for the “*Artificial Intelligence*” course (MSc, 2nd year) - DISCo
- Tutor for the “*Distributed Systems*” course (BSc, 2nd year) - DISCo
- Co-Advisor of a student BSc thesis
Questions?