Quality-driven Evolution in Information Integration Systems

Ph.D. in Computer Science Course - XXVIII Series

Riccardo Porrini

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

Dataspace Management System

local classifications and instances

taxonomic classification

mappings

faceted classification

global classification and instances

refinement

domain experts

end-users

application front-end

used to effectively access integrated information
Challenges

Changes in the world

End-users

Local classifications and instances

Taxonomy

Facets

Application front-end

Changes in information needs

Changes in data

New products
New events
New trends
...

Propagate much more quicker than changes in data

Must be consistent with respect to both data and user information needs to be effective
Goal: Quality-driven Information Integration Evolution

Dataspace Management System

- local classifications and instances
- global classification and instances
- taxonomy
- mappings
- facets
- application front-end
- refinement
- domain experts
- querying
- browsing
- feedback
- end-users

Quality Assessment
Coverage
Fitness to end-users information needs
Adaptation in Data Integration

Related Work

- Based on data analysis
  - Mappings: [Iv09, Ca09]
  - Taxonomies: [Ne12, Pa12, ...]
  - Text corpora: [Me13, Da08, ...]
  - Ontology structure: [St04]
  - Ontology change logs: [Ja11]

- Based on information needs analysis
  - Query logs: [Bl06]
  - Expert query samples: [Ya13, He12]
  - Resource usage: [Lu09]
  - Query logs: [Ma12]
  - Query click-through: [Ma07, Je08]

My Work

Challenge

How techniques based on data analysis can benefit from information needs analysis ones and vice versa?
Evaluation using Real World Case Studies

Dataspace Management System

- local classifications and instances
- global classification and instances
- taxonomy
- facets
- mappings
- refinement

Quality Assessment
- Coverage
- Fitness to end-users information needs

Feedback

- querying
- browsing
- end-users
- domain experts

- 8M product offers
- 3000 sources
- 500 global categories
- 2000 global facets
- 10M users/month

- traffic data
- domain experts
- in-vivo experiments

Riccardo Porrini | Quality-driven Evolution in Information Integration Systems | 7
Preliminary Work

Usage of global classifications to fulfill end-users’ explorative information needs

Autocompletion algorithm that leverages both taxonomic and faceted classifications to suggest Dataspace instances [WI2012, WIAS2014]

Dataspace Management System

<table>
<thead>
<tr>
<th>global classification and instances</th>
</tr>
</thead>
<tbody>
<tr>
<td>taxonomy</td>
</tr>
<tr>
<td>facets</td>
</tr>
</tbody>
</table>

start typing a full text search query over integrated information

instances

end-users


Facet Extraction

Adaptation of the global faceted classification based on data analysis

Extraction of facets from heterogeneous source taxonomies [CAiSE2014]

Facet Extraction

Contributions

▶ Integration of heterogeneous taxonomies at Web scale
▶ **Taxonomy Layer Distance**: a new distance metric between heterogeneous taxonomic concepts
▶ Evaluation using real world data and comparison with existing metrics
▶ Currently deployed and used by TrovaPrezzi domain experts
Ongoing and Future Work

Dataspace Management System

Local classifications and instances

Global classification and instances

Taxonomy

Mappings

Facets

Reconcile extracted facets to shared vocabularies (e.g., DBPedia ontology)

(@Advis Lab - UIC - Chicago)

Adaptation from data analysis

Reconcile extracted facets to shared vocabularies (e.g., DBPedia ontology)

(@Advis Lab - UIC - Chicago)

Adaptation from data analysis

Study of the correlation between experimental economic indicators and user search behavior mined by search logs

Correlating dynamic information needs with "external" factors

Quality Assessment

Coverage

Fitness to end-users information needs

Application front-end

Querying

Browsing

Domain experts

End-users

Riccardo Porrini | Quality-driven Evolution in Information Integration Systems | 10
Ongoing and Future Work

Dataspace Management System
local classifications and instances
data space for end-users
natural language processing
application front-end
global classification and instances
mappings
taxonomy
facets

Quality Assessment
Coverage
Fitness to end-users information needs

1. Reconcile extracted facets to shared vocabularies (e.g., DBPedia ontology) (@Advis Lab - UIC - Chicago)
   Adaptation from data analysis

2. Suggest effective classification changes by mining search logs
   Adaptation from information needs analysis

Riccardo Porrini | Quality-driven Evolution in Information Integration Systems | 10
Ongoing and Future Work

1. Reconcile extracted facets to shared vocabularies (e.g., DBPedia ontology) (@Advis Lab - UIC - Chicago)
   Adaptation from data analysis

2. Suggest effective classification changes by mining search logs
   Adaptation from information needs analysis

3. (side activity)
   Study of the correlation between experimental economic indicators and user search behavior mined by search logs
   Correlating dynamic information needs with “external” factors
Scientific Activities

Publications


Conference Paper Reviews (as sub-reviewer)

▶ AAAI-15: 23th AAAI Conference on Artificial Intelligence
▶ EKAW 2014: 19th International Conference on Knowledge Engineering and Knowledge Management
▶ ISWC 2014: 14th International Semantic Web Conference
▶ WI 2014: 2014 IEEE/WIC/ACM International Conference on Web Intelligence
▶ AAAI-14: 22th AAAI Conference on Artificial Intelligence
▶ ESWC 2014: 11th Extended Semantic Web Conference
▶ CAiSE 2014: 26th International Conference on Advanced Information Systems Engineering
▶ ODBASE 2013: 12th International Conference on Ontologies, DataBases, and Applications of Semantics
▶ WI 2013: 2013 IEEE/WIC/ACM International Conference on Web Intelligence
Educational Activities

Courses and Schools

▶ The Impact of Logic: from Proof Systems to Databases - Politecnico di Milano (evaluation pending)
▶ Cluster Analysis - DISCo
▶ Third ESWC Summer School - Kalamaky - Crete (GR)
▶ Foundations of Data Exchange and Integration - Politecnico di Milano
▶ Advanced Analytics and Behavior Informatics - DISCo

Seminars

▶ From Sentiment Analysis to Continuous Learning - DISCo
▶ Introduzione alla Logica nella Rappresentazione della Conoscenza - DISCo
▶ La Misurazione Della Felicità al Tempo dei Big Data - DISCo
▶ Optimum Hyperpaths in Directed Hypergraphs - DISCo
▶ Phase Transitions in Social and Economic Systems - DISCo
▶ Progettare e Fare Open Data. Metodologia e tools sviluppati in Evodevo a partire dall’esperienza Open Data INPS - DISCo
▶ Semantic Constraints for Data Quality Assessment and Cleaning - DISCo
▶ Big Data e la forza degli eventi - DISCo
▶ WOA 2012: 13th National Workshop “Dagli Oggetti agli Agenti” - DISCo

Teaching

▶ Tutor for Data and Web Semantics course (MSc and PhD, Fall 2014) - UIC
▶ Co-Advisor of two student BSc thesis
▶ Lecturer (2 lectures) for the “Artificial Intelligence” course (MSc A.A. 2012/2013 and 2013/2014) - DISCo
▶ Tutor for Distributed Systems course (BSc A.A. 2012/2013) - DISCo
Thank You

Questions?