Facet Extraction, Annotation and Alignment in Dataspaces

Riccardo Porrini

DISCo, University of Milano-Bicocca
7Pixel s.r.l.
riccardo.porrini@disco.unimib.it
About Me

- PhD Student @ UniMiB (Milan) from January 2013
  supervisor: Matteo Palmonari

- currently visiting ADVIS Lab @ UIC
  under the supervision of Prof. Isabel Cruz

- 5 years as back-end software developer for 7Pixel
  eCommerce and price comparison domain
  mainly facing Data Integration issues

research interests (broadly speaking):

- information systems
- data integration
- semantic web
- linked data
- ...follow the rest of the presentation :)

Riccardo Porrini | Facet Extraction, Annotation and Alignment in Dataspaces | 3
Dataspace Management System

local classifications and instances

global classification and instances

taxonomy-based classification

mappings

facet-based classification

refinement

domain experts

querying

end users

enable effective access to integrated information
**Taxonomy-based Classification**

**taxonomy**: categories organized through a hierarchical structure *(informal)*

---

Coarse-grained classification helps end users to recall the “class” of instances.

- **Country of Origin**
  - USA (320)
  - France (91)
  - Italy (40)
  - Spain (18)
  - Australia (17)
  - Bulgaria (10)
  - Chile (8)
  + See more

- **Vintage**
  - No Vintage (96)
  - 2013 (72)
  - 2012 (102)
  - 2011 (84)
  - 2010 (50)
  + See more

---

Renwood Winter Reds Port, Syrah, Primitivo Mixed Pack, 3 x 750 mL

**$63.91** $79.89

Eligible for 1¢ Standard Shipping See Details

Show only Renwood items
Facet-based Classification

**facet**: a clearly defined, mutually exclusive, and collectively exhaustive aspect, property, or characteristic of a class or specific subject [Taylor 2004]

Multiple Classifications in Action: Product Autocomplete

* translated from Italian

mobile phones
- features: android, samsung
  - samsung galaxy ace s5830 white
  - samsung galaxy 5 i5500 black
  - samsung galaxy 3 i5800
  - samsung galaxy s i9000 8gb ceramic white
  - samsung galaxy s i9000 8gb metallic black
  - samsung galaxy 5 i5500 white

tablet
- features: android, samsung
  - samsung galaxy tab gt-p1000 16gb white

* translated from Italian

Multiple Classifications in Action: Product Autocomplete

**mobile phones**
- features: android, samsung
  - samsung galaxy ace s5830 white
  - samsung galaxy 5 i5500 black
  - samsung galaxy 3 i5800
  - samsung galaxy s i9000 8gb ceramic white
  - samsung galaxy s i9000 8gb metallic black
  - samsung galaxy 5 i5500 white

**tablet**
- features: android, samsung
  - samsung galaxy tab gt-p1000 16gb white

* translated from Italian

Multiple Classifications in Action: Product Autocomplete

* translated from Italian

**mobile phones**
- features: android, samsung
  - samsung galaxy ace s5830 white
  - samsung galaxy s5 i5500 black
  - samsung galaxy 3 i5800
  - samsung galaxy s i9000 8gb ceramic white
  - samsung galaxy s i9000 8gb metallic black
  - samsung galaxy 5 i5500 white

**tablet**
- features: android, samsung
  - samsung galaxy tab gt-p1000 16gb white

* support for explorative keyword queries

**facets** and **categories** are considered in result-driven completion of the query

- Popularity
- Syntactic Relevance
- Semantic Relevance
- Facet Salience

* translated from Italian

Mappings

source taxonomy

Wines
Lombardy
Cabernet

global taxonomy

Beverages and Foods
Wines
Root
Mappings

usually, source taxonomies provide much more **granular classification**
Faceted Classifications Maintenance

- Faceted classification bootstrapping
  - time and effort consuming
  - requires detailed knowledge about dataspace instances
Faceted Classifications Maintenance

faceted classification bootstrapping
  ▶ time and effort consuming
  ▶ requires detailed knowledge about dataspace instances

however
  ▶ source taxonomies usually provide much more granular classification
  ▶ this granular information is lost when mapping specific source categories to generic global ones
faceted classification bootstrapping

- time and effort consuming
- requires detailed knowledge about dataspace instances

however

- source taxonomies usually provide much more granular classification
- this granular information is lost when mapping specific source categories to generic global ones

how about extracting facets from those lost fine-grained source taxonomies?
Facet Extraction Problem

given a global category $g$, a set of mappings $M$ from source categories $s_1, \ldots, s_n$ to $g$, extract a set $\mathcal{F}_g$ of facets $F_g$
Problem Statement

Facet Extraction Problem

given a global category $g$, a set of mappings $M$ from source categories $s_1, \ldots, s_n$ to $g$,
extract a set $F_g$ of facets $F_g$

Wines
Problem Statement

Facet Extraction Problem
given a global category $g$, a set of mappings $M$ from source categories $s_1, \ldots, s_n$ to $g$, extract a set $\mathcal{F}_g$ of facets $F_g$

Wines

<table>
<thead>
<tr>
<th>Winery Country of Origin</th>
<th>Wine Alcohol By Volume</th>
<th>Grape Variety</th>
<th>Wine Bottle Volume</th>
</tr>
</thead>
<tbody>
<tr>
<td>USA</td>
<td>Under 10%</td>
<td>Blend - White</td>
<td>375 mL</td>
</tr>
<tr>
<td>China</td>
<td>10% to 12%</td>
<td>Blend - Other</td>
<td>500 mL</td>
</tr>
<tr>
<td>Australia</td>
<td>12% to 14%</td>
<td>Fruit</td>
<td>750 mL</td>
</tr>
<tr>
<td>Italy</td>
<td>14% &amp; Up</td>
<td>Muscadine</td>
<td></td>
</tr>
<tr>
<td>Specialty Wine Type</td>
<td></td>
<td>Cabernet Sauvignon</td>
<td></td>
</tr>
<tr>
<td>Sustainable</td>
<td></td>
<td>Pinot Noir</td>
<td></td>
</tr>
<tr>
<td>Small Lot</td>
<td></td>
<td>Chardonnay</td>
<td></td>
</tr>
<tr>
<td>Kosher</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Gluten-Free</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Wine Vintage</th>
<th>Grape Variety</th>
<th>Wine Bottle Volume</th>
</tr>
</thead>
<tbody>
<tr>
<td>2011</td>
<td>Blend - White</td>
<td>375 mL</td>
</tr>
<tr>
<td>2010</td>
<td>Blend - Other</td>
<td>500 mL</td>
</tr>
<tr>
<td>2009</td>
<td>Fruit</td>
<td>750 mL</td>
</tr>
<tr>
<td>2008</td>
<td>Muscadine</td>
<td></td>
</tr>
<tr>
<td>2007</td>
<td>Cabernet Sauvignon</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Pinot Noir</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Chardonnay</td>
<td></td>
</tr>
</tbody>
</table>
Challenges

Source taxonomies are:

▶ many
e.g., 3900 within the TrovaPrezzi Italian price comparison engine
Challenges

Source taxonomies are:

- **many**
  - e.g., 3900 within the TrovaPrezzi Italian price comparison engine

- **noisy**
  - type > white > by vine > chardonnay > producer > firriato
Challenges

**Source taxonomies are:**

- **many**
  e.g., 3900 within the TrovaPrezzi Italian price comparison engine

- **noisy**
  type > white > by vine > chardonnay > producer > firriato

- **heterogeneous**
  type > white > by vine > chardonnay > producer > firriato
  wines > white wines > greco di tufo
Challenges

**Source taxonomies are:**

- **many**
  - e.g., *3900* within the TrovaPrezzi Italian price comparison engine

- **noisy**
  - type > white > by vine > chardonnay > producer > firriato

- **heterogeneous**
  - type > white > by vine > chardonnay > producer > firriato
  - wines > white wines > greco di tufo

- **ambiguous** different meaning in different domains
  - *red* is a wine type for wines
  - and a color for shirts
Facet extraction

[Suggested Facets]

1. vino bianco
2. vino rosato
3. vino rosso
4. bianco
5. rosato
6. rosso
7. abruzzo
8. argentina
9. basilicata
10. calabria
11. campania
12. emilia romagna
13. francia
14. friuli
15. friuli venezia giulia
16. lazio
17. liguria
18. lombardia

Porrini et al. 2014

Facet extraction

Facet extraction

[Suggested Facets]

- vino bianco
- vino rosato
- vino rosso
- bianco
- rosato
- rosso
- abruzzo
- argentina
- basilicata
- calabria
- campania
- emilia romagna
- francia
- friuli
- friuli venezia giulia
- lazio
- liguria
- lombardia


Riccardo Porrini | Facet Extraction, Annotation and Alignment in Dataspaces | 12
**Taxonomy Layer Distance**

**principle**

the more two values refer to mutually exclusive categories, the more they should be grouped together into the same facet

Source Taxonomy A

- **Wines**
  - Lombardy
  - Tuscany
  - Sicily
  - Cabernet
  - Chianti
  - Cantina Firriato
  - Cantina Almeria
  - Cabernet
  - Nero d’Avola

Source Taxonomy B

- **Wines**
  - White Wines
    - Vermentino
    - Verdicchio
  - Red Wines
    - Cabernet
    - Chianti
  - Lombardy
  - Sicily

\[
\text{TLD}(\text{cabernet}, \text{chianti}) = 1 - \frac{|\text{L}_{\text{cabernet}} \cap \text{L}_{\text{chianti}}|}{|\text{L}_{\text{cabernet}} \cup \text{L}_{\text{chianti}}|} = 1 - \frac{2}{3} = \frac{1}{3}
\]
Taxonomy Layer Distance

\[
TLD(s_1, s_2) = 1 - \frac{|L_{s_1} \cap L_{s_2}|}{|L_{s_1} \cup L_{s_2}|}
\]

Jaccard Distance between the two sets of taxonomy layers where two categories \( s_1 \) and \( s_2 \) occur
Taxonomy Layer Distance

$$\text{TLD}(s_1, s_2) = 1 - \frac{|L_{s_1} \cap L_{s_2}|}{|L_{s_1} \cup L_{s_2}|}$$

Jaccard Distance between the two sets of taxonomy layers where two categories $s_1$ and $s_2$ occur

cabernet
Taxonomy Layer Distance

\[ TLD(s_1, s_2) = 1 - \frac{|L_{s_1} \cap L_{s_2}|}{|L_{s_1} \cup L_{s_2}|} \]

Jaccard Distance between the two sets of taxonomy layers where two categories \( s_1 \) and \( s_2 \) occur

Source Taxonomy A

- Wines
  - Lombardy
  - Tuscany
  - Sicily
  - Cabernet
  - Chianti
  - Cantina Firriato
  - Cantina Almeria
  - Cabernet
  - Nero d'Avola

Source Taxonomy B

- Wines
  - Red Wines
  - White Wines
  - Vermentino
  - Verdicchio
  - Cabernet
  - Chianti
  - Lombardy
  - Sicily

\textbf{cabernet} \hspace{1cm} \textbf{chianti}
Taxonomy Layer Distance

$TLD(s_1, s_2) = 1 - \frac{|L_{s_1} \cap L_{s_2}|}{|L_{s_1} \cup L_{s_2}|}$

Jaccard Distance between the two sets of taxonomy layers where two categories $s_1$ and $s_2$ occur

cabernet  chianti

$TLD(\text{cabernet}, \text{chianti}) = 1 - \frac{|L_{\text{cabernet}} \cap L_{\text{chianti}}|}{|L_{\text{cabernet}} \cup L_{\text{chianti}}|} = 1 - \frac{2}{3} = \frac{1}{3}$
### Example of Extracted Facets

<table>
<thead>
<tr>
<th>Facet</th>
<th>LC $F^g_1$</th>
<th>WP $F^g_1$</th>
<th>TLD $F^g$</th>
<th>Gold Standard $F^g$</th>
</tr>
</thead>
<tbody>
<tr>
<td>LC</td>
<td>${\text{Wine, Red Wine, White Wine, \ldots, Piedmont, Lombardy, \ldots, Sicily, Donnafugata, Cusumano, \ldots, Alessandro di Camporeale, \ldots, France}}$ (98)</td>
<td>${\text{Wine, Red Wine, White Wine, \ldots, Piedmont, Lombardy, \ldots, Sicily, Donnafugata, Cusumano, \ldots, France}}$ (100)</td>
<td>${\text{Piedmont, Tuscany, Sicily, \ldots, France}}$ (14)</td>
<td>${\text{Piedmont, Lombardy, \ldots, Sicily}}$ (21)</td>
</tr>
<tr>
<td>WP</td>
<td>${\text{Wine, Red Wine, White Wine, \ldots, Piedmont, Lombardy, \ldots, Sicily, Donnafugata, Cusumano, \ldots, France}}$ (100)</td>
<td>${\text{Wine, Red Wine, White Wine, \ldots, Piedmont, Lombardy, \ldots, Sicily, Donnafugata, Cusumano, \ldots, France}}$ (100)</td>
<td>${\text{Red, White, Rosé}}$ (3)</td>
<td>${\text{Red Wine, White Wine, Rosé Wine}}$ (14)</td>
</tr>
<tr>
<td>TLD</td>
<td>${\text{Piedmont, Tuscany, Sicily, \ldots, France}}$ (14)</td>
<td>${\text{Piedmont, Tuscany, Sicily, \ldots, France}}$ (14)</td>
<td>${\text{Red Wine, White Wine, Rosé Wine}}$ (3)</td>
<td>${\text{Red Wine, White Wine, Rosé Wine}}$ (14)</td>
</tr>
<tr>
<td></td>
<td>${\text{Red, White, Rosé}}$ (3)</td>
<td>${\text{Red, White, Rosé}}$ (3)</td>
<td>${\text{Red Wine, White Wine, Rosé Wine}}$ (3)</td>
<td>${\text{Red Wine, White Wine, Rosé Wine}}$ (14)</td>
</tr>
<tr>
<td></td>
<td>${\text{Moscato, Chardonnay, \ldots, Merlot}}$ (13)</td>
<td>${\text{Moscato, Chardonnay, \ldots, Merlot}}$ (13)</td>
<td>${\text{Tuscany Wine, Sicily Wine}}$ (2)</td>
<td>${\text{Tuscany Wine, Sicily Wine}}$ (2)</td>
</tr>
<tr>
<td></td>
<td>${\text{Donnafugata, Cusumano, \ldots, Principi di Butera}}$ (27)</td>
<td>${\text{Donnafugata, Cusumano, \ldots, Principi di Butera}}$ (27)</td>
<td>${\text{Donnafugata, Cusumano, \ldots, Principi di Butera}}$ (27)</td>
<td>${\text{Donnafugata, Cusumano, \ldots, Principi di Butera}}$ (27)</td>
</tr>
</tbody>
</table>

- **LC**: Leacock and Chodorow similarity [Leacock and Chodorow 1998]
- **WP**: Wu and Palmer similarity [Wu and Palmer 1994]
Are we really done?

Suggested Facets

- vino bianco
- vino rosato
- vino rosso

- bianco
- rosato
- rosso

- abruzzo
- argentina
- basilicata
- calabria
- campania
- emilia romagna
- francia
- friuli
- friuli venezia giulia
Are we really done?

so far

- only facet values were extracted
Are we really done?

so far

▶ only facet values were extracted
▶ what about facet roles?

HammerSky 2010 Red Handed

Red

2010

U.S.A.

2009 Tignanello, Tuscany 750 mL

Red

2009

Italy

2012 Paulinshof Urstueck 750 mL

White

2012

Germany

2012 Sobremesa Vineyards VRM

White

2012

Argentina

Riccardo Porrini | Facet Extraction, Annotation and Alignment in Dataspaces | 15
Are we really done?

so far

- only facet values were extracted
- what about facet roles?

recurrent problem in different domains

- only partially tackled by previous work on facet extraction [Dou et al. 2011, Kong and Allan 2013] . . .
Are we really done?

so far

- only facet values were extracted
- what about facet roles?

recurrent problem in different domains

- only partially tackled by previous work on facet extraction [Dou et al. 2011, Kong and Allan 2013] . . .
- interpretation of results from clustering algorithms [Carmel et al. 2009] . . .
Are we really done?

so far
- only facet values were extracted
- what about facet roles?

recurrent problem in different domains
- only partially tackled by previous work on facet extraction [Dou et al. 2011, Kong and Allan 2013] . . .
- interpretation of results from clustering algorithms [Carmel et al. 2009] . . .

<table>
<thead>
<tr>
<th>HammerSky 2010 Red Handed</th>
<th>Red</th>
<th>2010</th>
<th>U.S.A.</th>
</tr>
</thead>
<tbody>
<tr>
<td>2009 Tignanello, Tuscany 750 mL</td>
<td>Red</td>
<td>2009</td>
<td>Italy</td>
</tr>
<tr>
<td>2012 Paulinshof Urstueck 750 mL</td>
<td>White</td>
<td>2012</td>
<td>Germany</td>
</tr>
<tr>
<td>2012 Sobremesa Vineyards VRM</td>
<td>White</td>
<td>2012</td>
<td>Argentina</td>
</tr>
</tbody>
</table>

Riccardo Porrini | Facet Extraction, Annotation and Alignment in Dataspaces | 15
ongoing joint work with Prof. Isabel Cruz, Advis Lab - UIC

**Goal**

link a **facet** $F$ to a suitable representation of the role that **facet values** play in the characterization of instances from a specific **facet domain** $D$
ongoing joint work with Prof. Isabel Cruz, Advis Lab - UIC

**goal**

link a **facet** $F$ to a suitable representation of the role that **facet values** play in the characterization of instances from a specific **facet domain** $D$

again, **ambiguity** is challenging:

2011
2012
2013
...

Riccardo Porrini | Facet Extraction, Annotation and Alignment in Dataspaces | 16
Facet Role Linking

ongoing joint work with Prof. Isabel Cruz, Advis Lab - UIC

goal

link a facet $F$ to a suitable representation of the role that facet values play in the characterization of instances from a specific facet domain $D$

again, ambiguity is challenging:

vintage

2011
2012
2013
...

wines
Facet Role Linking

ongoing joint work with Prof. Isabel Cruz, Advis Lab - UIC

**goal**

link a facet $F$ to a suitable representation of the role that facet values play in the characterization of instances from a specific facet domain $D$

again, **ambiguity** is challenging:

<table>
<thead>
<tr>
<th>release year</th>
<th>music albums</th>
</tr>
</thead>
<tbody>
<tr>
<td>2011</td>
<td>2012</td>
</tr>
<tr>
<td></td>
<td>2013</td>
</tr>
<tr>
<td></td>
<td>...</td>
</tr>
</tbody>
</table>
Facet Role Linking

intuitions adapted from Web table annotation approaches

- properties from existing ontologies provide the semantics that we are looking for
- a knowledge base can provide evidence of certain properties holding between entities from domains similar to the facet domain and facet values
bonus points

- machine readable semantics
- facets published on the Web can be annotated with it (e.g., RDFa)
given a facet $F$ with $f_1, \ldots, f_n$ facet values and a facet domain $D$
Approach Overview

basic idea

given a facet $F$ with $f_1, \ldots, f_n$ facet values and a facet domain $D$
select a set $P$ of properties from triples $\langle s, p, o \rangle$ such that
- the facet domain $D$ matches the type(s) of $s$
- one or more facet values $f_1, \ldots, f_n \in F$ match $o$ (entity or literal)
Approach Overview

basic idea

given a facet \( F \) with \( f_1, \ldots, f_n \) facet values and a facet domain \( D \)

select a set \( P \) of properties from triples \( \langle s, p, o \rangle \) such that

- the facet domain \( D \) matches the type(s) of \( s \)
- one or more facet values \( f_1, \ldots, f_n \in F \) match \( o \) (entity or literal)

then, rank \( p \in P \) according to several criteria

- facet values coverage
- weighted frequency w.r.t. \( D \)
- specificity w.r.t. \( D \)
Future Work and Interesting Directions

facet annotation

▶ what if a facet role is already provided (in natural language)?

syntactic-based property alignment techniques [Cheatham and Hitzler 2014]

annotation of facets already published on the Web
Future Work and Interesting Directions

facet annotation

▶ what if a facet role is already provided (in natural language)?
  syntactic-based property alignment techniques [Cheatham and Hitzler 2014]
  annotation of facets already published on the Web

extension to Web table annotation

▶ apply the facet annotation technique to Web Table columns
  in literature, little study of relation annotation compared to type/entity
Future Work and Interesting Directions

extension to property alignment in LOD

- incorporate insights from facet annotation into property alignment techniques for LOD, by considering their usage
Future Work and Interesting Directions

extension to property alignment in LOD

▶ incorporate insights from facet annotation into property alignment techniques for LOD, by considering their usage

classification evolution

▶ study how to evolve facets (and also their values) over time
  all mobile phones have a digital camera: that facet is not important anymore
  a new operating system for mobile phones is released
Thank You

Questions?

riccardo.porrini@disco.unimib.it
http://rporrini.info
References


