KDI : **Knowledge and Data Integration**

Evaluation

iTelos Inception & Informal Modeling Phase

Fausto Giunchiglia

Contents

Contents

Evaluation purpose on Inception phase

In **schema level**, we have a set of Competency queries (CQs) and several collected ontologies.

We aim to measure:

- If the collect ontologies cover CQs, using metric **coverage.**
- If the collect ontologies bring additional information to CQs, using metric **Extensiveness.**

Examples: CQ vs Ont

Given a set of Competency Query (*CQ*) , the coverage (*Cov*) of the aligned ontology (*Ont*) is:

Etype Coverage

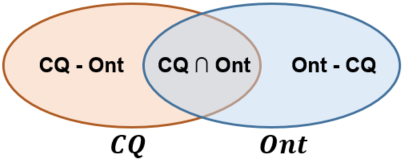
$$Cov(CQ_c) = \frac{|\, CQ_c \cap Ont_c \,|}{|\, CQ_c \,|}$$

$Cov = 1$ *Full coverage*
$0.6 < Cov < 0.8$ *Ideal*
$Cov = 0$ *No coverage*

Property Coverage

$$Cov(CQ_p) = \frac{|\, CQ_p \cap Ont_p \,|}{|\, CQ_p \,|}$$

$Cov = 1$ *Full coverage*
$0.6 < Cov < 0.8$ *Ideal*
$Cov = 0$ *No coverage*



Examples: CQ vs Ont

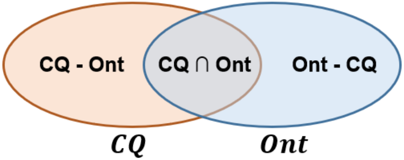
Given a set of Competency Query (*CQ*) , the Extensiveness (*Ext*) of the aligned ontology (*Ont*) is:

Etype Extensiveness $$Ext(CQ_c) = \frac{|\ Ont_c - CQ_c\ |}{|\ CQ_c \cup Ont_c\ |}$$

$$Ext = 1 \quad \textit{Full Extensiveness}$$
$$Ext \simeq 0.5 \quad \textit{Ideal}$$
$$Ext = 0 \quad \textit{No Extensiveness}$$

Property Extensiveness $$Ext(CQ_p) = \frac{|\ Ont_p - CQ_p\ |}{|\ CQ_p \cup Ont_p\ |}$$

$$Ext = 1 \quad \textit{Full Extensiveness}$$
$$Ext \simeq 0.5 \quad \textit{Ideal}$$
$$Ext = 0 \quad \textit{No Extensiveness}$$



Examples: CQs

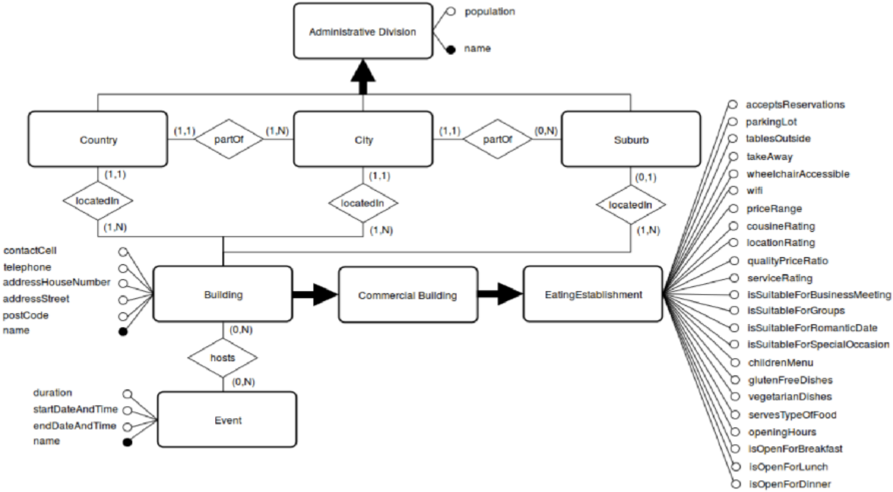
- List all the **places** where one can eat which have all rating higher than 4 stars.
- List all the **eating establishments** which are suitable for groups and also provide children menu.
- Give me **facilities** with vegetarians menu and price range medium-low.
- Give me the contacts of all the **eating establishments** that accept reservations and have a parking lot.
- Where to eat **pizza** for lunch?

Classes in CQs: C_c={place, eatingEstablishment, facilities, pizza}
(Num class = 4)

Properties in CQs: C_p={reservation, lunch, rating, suitableForGroup, childrenMenu, Vegetarians, contact, address, parkingLot }
(Num property = 8)

*Notice, words in purple means that class/property is aligned.

Examples: Reference Ontologies



Examples: Reference Ontologies

Classes/Etypes in ontology:

C_c={AdministrativeDivision, Country, City, Suburb, Building, CommercialBuilding, place, eatingEstablishment}
(Num class = 8)

Properties in ontology:

C_p = {rating, suitableForGroup, childrenMenu, Vegetarians, contact, address, parkingLot,. . . .. OpenForlunch}
(Num property = 38)

*Notice, words in purple means that class/property is aligned.

Examples: CQ vs Ont

Given the example *CQ* , and the example reference ontology *Ont*, we have:

Etype Coverage $$Cov(CQ_c) = \frac{|\ CQ_c \cap Ont_c\ |\ (2)}{|\ CQ_c\ |\ (4)} = 0.5$$

Etype Extensiveness $$Ext(CQ_c) = \frac{|\ Ont_c - CQ_c\ |\ (6)}{|\ CQ_c \cup\ Ont_c\ |\ (12)} = 0.6$$

Property Coverage $$Cov(CQ_p) = \frac{|\ CQ_p \cap Ont_p\ |\ (6)}{|\ CQ_p\ |\ (8)} = 0.75$$

Property Extensiveness $$Ext(CQ_p) = \frac{|\ Ont_p - CQ_p\ |\ (32)}{|\ CQ_p \cup Ont_p\ |\ (40)} = 0.8$$

Notice that:

- Intersection information (coverage) should more likely belongs to common or core category.
- Additional information (extensiveness) should be core and contextual information.

Evaluation purpose on Inception phase

In **data level**, we have a set of CQs and several collected datasets/schema.

We aim to measure:

- If the collect datasets cover CQs, using metric **coverage.**
- If the collect datasets are much different from CQs, using metric **Sparsity**.

Examples: CQ vs Dataset

Given a set of Competency Query (*CQ*) , the coverage (*Cov*) of the aligned dataset (*D*) is:

Etype Coverage

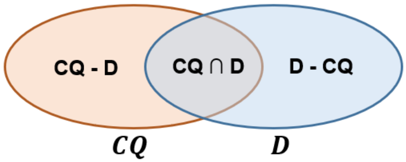
$$Cov(CQ_c) = \frac{|\, CQ_c \cap D_c \,|}{|\, CQ_c \,|}$$

$Cov = 1$ *Full coverage*
$0.6 < Cov < 0.8$ *Ideal*
$Cov = 0$ *No coverage*

Property Coverage

$$Cov(CQ_p) = \frac{|\, CQ_p \cap D_p \,|}{|\, CQ_p \,|}$$

$Cov = 1$ *Full coverage*
$0.6 < Cov < 0.8$ *Ideal*
$Cov = 0$ *No coverage*



Examples: CQ vs Dataset

Given a set of Competency Query (*CQ*) , the sparsity (*Spr*) of the aligned dataset (*D*) is:

Etype sparsity

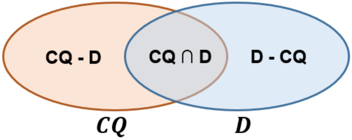
$$Spr(CQ_c) = \frac{\mid CQ_c - D_c \mid + \mid D_c - CQ_c \mid}{\mid CQ_c \cup D_c \mid} = 1 - \frac{\mid CQ_c \cap D_c \mid}{\mid CQ_c \cup D_c \mid}$$

$Spr = 1$ *Full Sparsity*
$Spr \simeq 0.5$ *Ideal*
$Spr = 0$ *No Sparsity*

Property sparsity

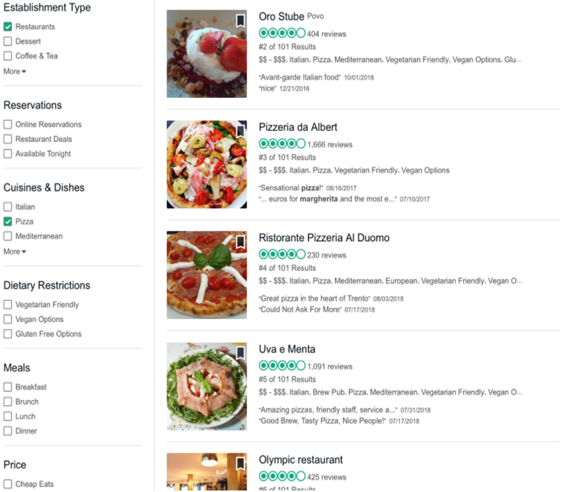
$$Spr(CQ_p) = \frac{\mid CQ_p - D_p \mid + \mid D_p - CQ_p \mid}{\mid CQ_p \cup D_p \mid} = 1 - \frac{\mid CQ_p \cap D_p \mid}{\mid CQ_p \cup D_p \mid}$$

$Spr = 1$ *Full Sparsity*
$Spr \simeq 0.5$ *Ideal*
$Spr = 0$ *No Sparsity*



Examples Dataset

Classes in dataset: (Num = 11, 2 of them aligned with CQ.)
Properties in dataset: (Num = 40, 6 of them aligned with CQ.)



Examples: CQ vs Dataset

Given the example *CQ*, and the example collected Dataset *D*, we have:

Etype Coverage $Cov(CQ_c) = \frac{|CQ_c \cap D_c|\,(2)}{|CQ_c|\,(4)} = 0.5$

Etype Sparsity $Spr(CQ_c) = 1 - \frac{|CQ_c \cap D_c|\,(2)}{|CQ_c \cup D_c|\,(13)} = 0.84$

Property Coverage $Cov(CQ_p) = \frac{|CQ_p \cap D_p|\,(6)}{|CQ_p|\,(8)} = 0.75$

Property Sparsity $Spr\left(CQ_p\right) = 1 - \frac{|CQ_p \cap D_p|\,(6)}{|CQ_p \cup D_p|\,(42)} = 0.86$

Notice that:

- Intersection information (coverage) should more likely belongs to common or core category.
- Different information (sparsity) should be core or contextual information.

Contents

Evaluation purpose on Informal modelling phase

In **schema level**, we have the proposed informal ER model and a set of CQs. We aim to measure:

- If the proposed informal ER model cover CQs, using metric **coverage.**
- If the proposed informal ER model properly extend CQs, using metric **extensiveness.**

Examples: ER model vs CQs

Given a set of Competency Query (*CQ*) , the coverage (*Cov*) of the ER model (*ER*) is:

Etype Coverage

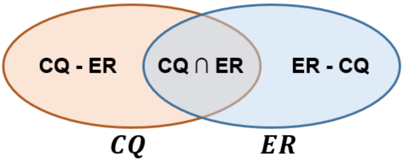
$$Cov(CQ_c) = \frac{| CQ_c \cap ER_c |}{| CQ_c |}$$

$$\begin{array}{rl} Cov = 1 & \textit{Full coverage} \\ 0.6 < Cov < 0.8 & \textit{Ideal} \\ Cov = 0 & \textit{No coverage} \end{array}$$

Property Coverage

$$Cov(CQ_p) = \frac{| CQ_p \cap ER_p |}{| CQ_p |}$$

$$\begin{array}{rl} Cov = 1 & \textit{Full coverage} \\ 0.6 < Cov < 0.8 & \textit{Ideal} \\ Cov = 0 & \textit{No coverage} \end{array}$$



Examples: ER model vs CQs

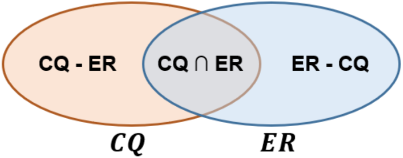
Given a set of Competency Query (*CQ*) , the Extensiveness (*Ext*) of the ER model (*ER*) is:

Etype Extensiveness $Ext(CQ_c) = \frac{| ER_c - CQ_c |}{| CQ_c \cup ER_c |}$

$Ext = 1$ *Full Extensiveness*
$Ext \simeq 0.5$ *Ideal*
$Ext = 0$ *No Extensiveness*

Property Extensiveness $Ext(CQ_p) = \frac{| ER_p - CQ_p |}{| CQ_p \cup ER_p |}$

$Ext = 1$ *Full Extensiveness*
$Ext \simeq 0.5$ *Ideal*
$Ext = 0$ *No Extensiveness*



Examples: ER model

Classes/Etypes in ontology:

C_c={AdministrativeDivision, City, Suburb, Building,, EatingEstablishment, . . . , Event}
(Num class = 18, 3 of them aligned with CQ.)

Properties in ontology:

C_p = {rating, Vegetarians, contact, reservation, parkingLot,. . . .. OpenForlunch}
(Num property = 36, 5 of them aligned with CQ.)

Examples: CQs

- List all the **places** where one can eat which have all rating higher than 4 stars.
- List all the **eating establishments** which are suitable for groups and also provide children menu.
- Give me **facilities** with vegetarians menu and price range medium-low.
- Give me the contacts of all the **eating establishments** that accept reservations and have a parking lot.
- Where to eat **pizza** for lunch?

Classes in CQs: C_c={place, eatingEstablishment, facilities, pizza}
(Num class = 4)

Properties in CQs: C_p={rating, suitableForGroup, childrenMenu, Vegetarians, contact, reservation, parkingLot, lunch}
(Num property = 8)

Examples: ER model vs CQs

Given the example CQ, and the example ER model ER, we have:

Etype Coverage $Cov(CQ_c) = \frac{| CQ_c \cap ER_c | (3)}{| CQ_c | (4)} = 0.75$

Etype Extensiveness $Ext(CQ_c) = \frac{| ER_c - CQ_c | (15)}{| CQ_c \cup ER_c | (19)} = 0.79$

Property Coverage $Cov(CQ_p) = \frac{| CQ_p \cap ER_p | (5)}{| CQ_p | (8)} = 0.625$

Property Extensiveness $Ext(CQ_p) = \frac{| ER_p - CQ_p | (31)}{| CQ_p \cup ER_p | (39)} = 0.79$

Notice that:

- Intersection information (coverage) should more likely belongs to common or core category.
- Additional information (extensiveness) should be core and contextual information. We should find the balance on extensiveness, since too much hard to maintain, too less not properly extend

Evaluation purpose on Informal modelling phase

In **data level**, we have the proposed informal ER model and several collected datasets.

We aim to measure:

- If the informal ER model align with collect datasets, using metric **coverage**.
- If the informal ER model is much different from collect datasets, using metric **Sparsity**.

Examples: ER model vs Dataset

Given the dataset (*D*), the coverage (*Cov*) of the ER model (*ER*) is:

Etype Coverage

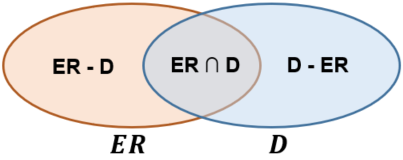
$$Cov(D_c) = \frac{|\, ER_c \cap D_c \,|}{|\, D_c \,|}$$

$$\begin{array}{rl} Cov = 1 & \textit{Full coverage} \\ 0.6 < Cov < 0.8 & \textit{Ideal} \\ Cov = 0 & \textit{No coverage} \end{array}$$

Property Coverage

$$Cov(D_p) = \frac{|\, ER_p \cap D_p \,|}{|\, D_p \,|}$$

$$\begin{array}{rl} Cov = 1 & \textit{Full coverage} \\ 0.6 < Cov < 0.8 & \textit{Ideal} \\ Cov = 0 & \textit{No coverage} \end{array}$$



Examples: ER model vs Dataset

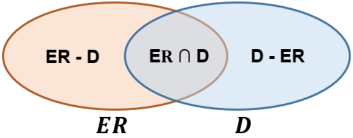
Given the dataset (*D*), the sparsity (*Spr*) of the ER model (*ER*) is:

Etype Sparsity $Spr(D_c) = \frac{|ER_c - D_c| + |D_c - ER_c|}{|ER_c \cup D_c|} = 1 - \frac{|ER_c \cap D_c|}{|ER_c \cup D_c|}$

$Spr = 1$ *Full Sparsity*
$Spr \simeq 0.5$ *Ideal*
$Spr = 0$ *No Sparsity*

Property Sparsity $Spr(D_p) = \frac{|ER_p - D_p| + |D_p - ER_p|}{|ER_p \cup D_p|} = 1 - \frac{|ER_p \cap D_p|}{|ER_p \cup D_p|}$

$Spr = 1$ *Full Sparsity*
$Spr \simeq 0.5$ *Ideal*
$Spr = 0$ *No Sparsity*



Examples: ER model

Classes/Etypes in ontology:

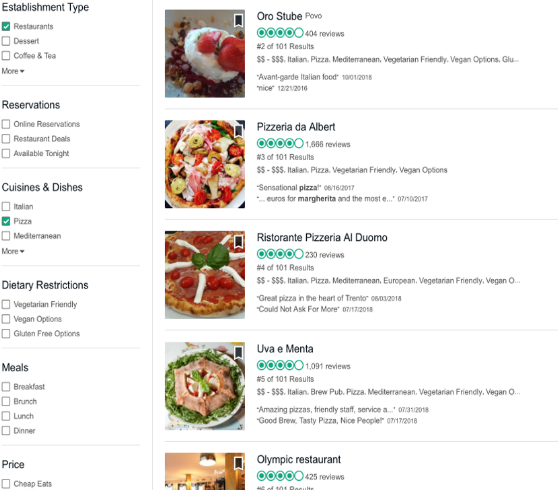
C_c={AdministrativeDivision, City, Suburb, Building,, EatingEstablishment, . . . , Event}
(Num class = 18, 8 of them aligned with Dataset.)

Properties in ontology:

C_p = {rating, Vegetarians, contact, reservation, parkingLot,. . . .. OpenForlunch}
(Num property = 36, 16 of them aligned with Dataset.)

Examples Dataset

Classes in dataset: (Num = 11, 8 of them aligned with ER model.)
Properties in dataset: (Num = 40, 16 of them aligned with ER model.)



Examples: ER model vs Dataset

Given the dataset (*D*), and the example ER model (*ER*), we have:

Etype Coverage $Cov(D_c) = \frac{|ER_c \cap D_c|\ (8)}{|D_c|\ (11)} = 0.73$ **Etype Sparsity** $Spr(D_c) = 1 - \frac{|ER_c \cap D_c|\ (8)}{|ER_c \cup D_c|\ (19)} = 0.58$

Property Coverage $Cov(D_p) = \frac{|ER_p \cap D_p|\ (16)}{|D_p|\ (40)} = 0.4$ **Property Sparsity** $Spr(D_p) = 1 - \frac{|ER_p \cap D_p|\ (16)}{|ER_p \cup D_p|\ (60)} = 0.73$

Notice that:

- Intersection information (coverage) should more likely belongs to common or core category.
- Different information (sparsity) should be core or contextual information. The sparsity should also keep a balance because if ETG model and dataset are very different, they will be hard to align.

KDI Knowledge and Data Integration

Fausto Giunchiglia

Evaluation

iTelos Inception & Informal Modeling Phase