

# Identifying Basic Level Entities in a Data Graph

Expanding user's knowledge while exploring data graphs

By

**Marwan Al-Tawil**

**Supervisors**

Vania Dimitrova

Brandon Bennett



# Outline of the talk



UNIVERSITY OF LEEDS

- **Introduction**
- **Problem and Focus**
  - What makes some entities more important to provide better paths?
  - How to develop automatic ways to identify knowledge anchors?
- **Identifying Basic Level Entities**
- **Experimental Study**
- **Experimental Results**
- **Conclusion and Future work**

# Introduction



UNIVERSITY OF LEEDS

## We usually explore data to learn new things

Often, learning involves exploring large amount of data and deciding what to explore next

The Web is our first choice [Marchionini 2006]

The Web is becoming powerful with linked data graphs [Berners-Lee 2009]



## Two developments in exploration over data graphs motivates this research



Entities and paths are different

Cognitive structures  
may not match the  
semantic structures



Users have different  
cognitive structures

# Problem & Focus



UNIVERSITY OF LEEDS

**Aiding users' exploration over data graphs to expand knowledge**



- ➔ **Serendipitous learning occurs while moving between individual entities. However, not all paths are beneficial for knowledge expansion.**
- ➔ **Some entities are important for knowledge expansion (knowledge anchors).**
- ➔ **Identifying these entities is important to nudge the user through better paths.**

# Problem & Focus

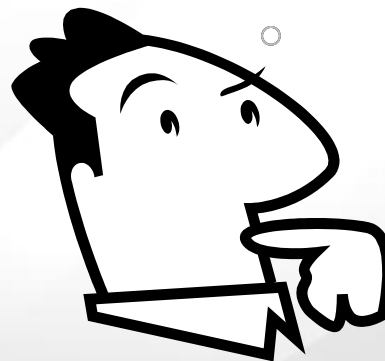


UNIVERSITY OF LEEDS

Two important questions we need to answer:

What makes some entities more important to provide better paths?

How to develop automatic ways to identify knowledge anchors?





# Problem & Focus

## What makes some entities more important to provide better paths?


### Exploratory study

Dataset : *MusicPinta*

- 876 Instruments
- 71k performances (albums, tracks)
- 188k artists
- 2.4M entities and 38M triples.

Description	Features	Relevant Information	Reviews	Link History
Features are items extracted from the data sources and immediately related to the search term.				
<b>Harp is:</b>				
<a href="#">Instrument</a>				
<b>Harp belongs to:</b>				
<a href="#">Celtic musical instruments</a> <a href="#">Composite chordophones</a> <a href="#">Harps</a> <a href="#">Instrument</a> <a href="#">Irish musical instruments</a> <a href="#">Lyre</a> <a href="#">National symbols of Ireland</a> <a href="#">Plucked string instruments</a> <a href="#">String instruments</a>				

Description	Features	Relevant Information	Reviews	Link History
Relevant information provides links to other items of interest in the data sources.				
<b>The following are Harp:</b>				
No items found.				
<b>The following share features with Harp:</b>				
<a href="#">Concert harp</a> <a href="#">Electric harp</a> <a href="#">Folk harp</a> <a href="#">German harp</a> <a href="#">Harpsichord</a> <a href="#">Irish harp</a> <a href="#">Kora</a> <a href="#">Psaltery</a> <a href="#">Wire-strung harp</a>				

Home	Semantic Search	Contribute	Help	
<a href="#">Home</a> > <a href="#">Semantic Search</a> > <a href="#">Harp</a>				
<h2>Harp</h2>				
Description	Features	Relevant Information	Reviews	Link History
Description is extracted from Wikipedia when available.				
 <p>The harp is a multi-stringed instrument which has the plane of its strings positioned perpendicularly to the soundboard. Organologically, it is in the general category of chordophones (stringed instruments) and has its own sub category (the harps). All harps have a neck, resonator and strings. Some, known as frame harps, also have a pillar; those without the pillar are referred to as open harps.</p>				

## What makes some entities more important to provide better paths?

### Exploratory study

- Three strategies were examined (*Density, Familiarity, Unfamiliarity*). Figure
- Observations:
  - **Central** entities with many subclasses are good potential anchors
  - **Recognition** is a key enabler for knowledge expansion
  - Encourage connections to discover **new** entities linked to **recognized** ones

### Subsumption for meaningful learning

- **Familiar** and basic entities are used as anchors to subsume **new** knowledge into users' cognitive structure.
- **Basic** entities are deliberately introduced prior bringing **new** knowledge.



# Problem & Focus



UNIVERSITY OF LEEDS

Two important questions we need to answer

What makes some entities more important to provide better paths?

How to develop automatic ways to identify knowledge anchors?







# Problem & Focus

## How to develop automatic ways to identify knowledge anchors?

- We utilize the Cognitive Science notion of basic level object introduced by Rosch.
- Category objects that are commonly used in our daily life (e.g. Chair and Dog).
- Category objects that carry the most information, possess the highest category cue validity, and are, thus, the most differentiated from one another.

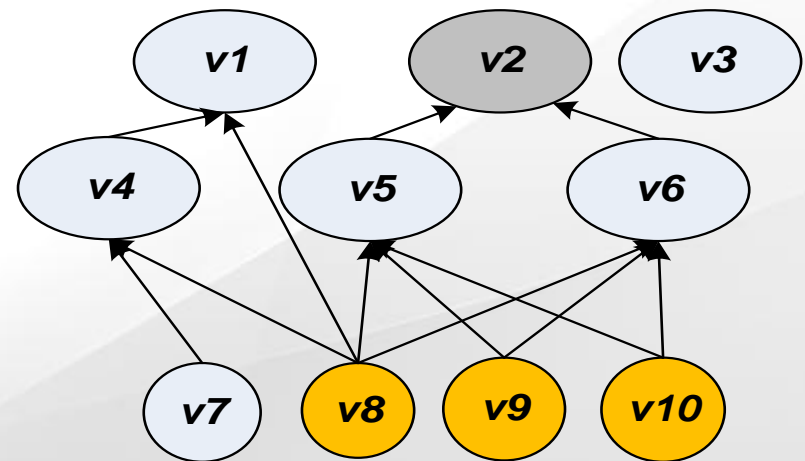
## We adopt two approaches

### Distinctiveness

Identifies most differentiated entities whose cues linked to its members, and not linked to other entities

### Homogeneity

Identifies entities whose members have high similarity values.  
(i.e. complementary with the distinctiveness)





# Identifying Knowledge Anchors

## Preliminaries

A **Data Graph** is a labeled directed graph  $DG = \langle V, E, P \rangle$

$V = \{v_1, v_2, \dots, v_n\}$  is a finite set of vertices

$E = \{e_1, e_2, \dots, e_m\}$  is a finite set of edges

$P = \{p_1, p_2, \dots, p_k\}$  is a set of triples  $\langle v_s, e_i, v_o \rangle$

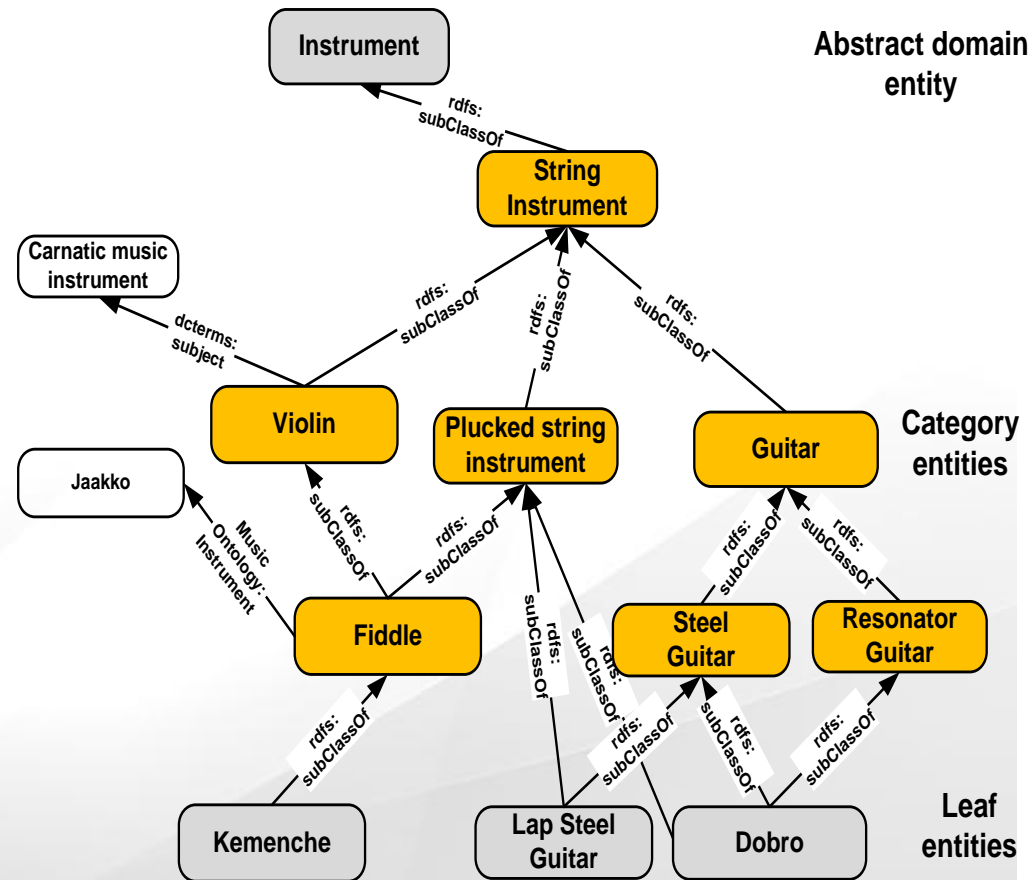
**Entities set  $V$  in the graph can be divided into:**

Category entities and leaf entities

**Relationship types;**

- **Hierarchical:** denote category membership between the *Subject* and *Object* entities (e.g. *rdfs:subClassOf*; *dcterms:subject*)
- **Domain-specific:** other than hierarchical relationships (e.g. *performances*)

**Candidate entities are category entities**





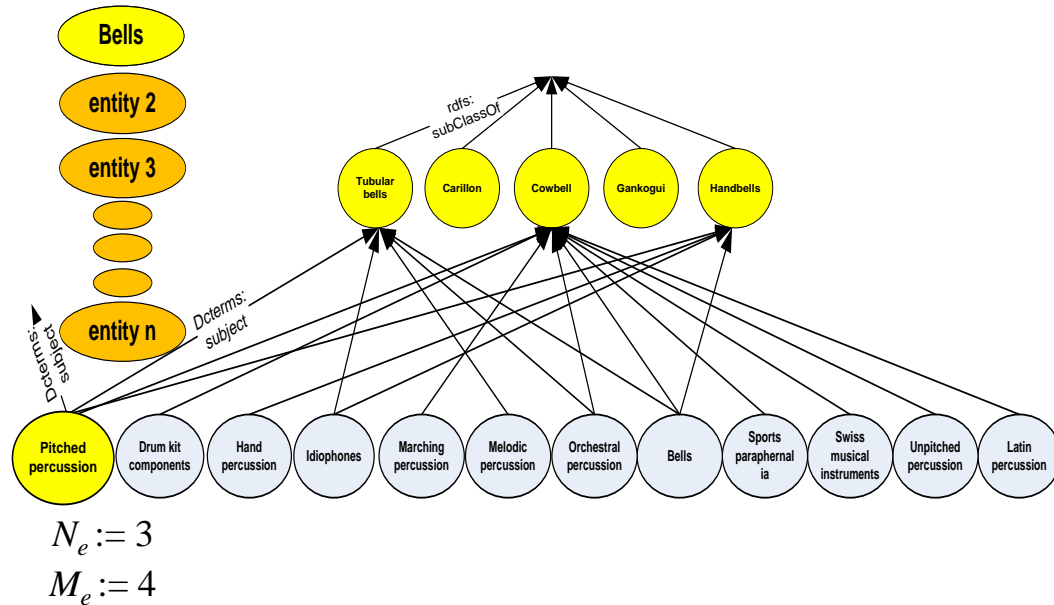
# Identifying Knowledge Anchors

## Algorithm I: Distinctiveness Metrics – adopted from formal concept analysis

Input:  $DG = \langle V, E, P \rangle, e \in E$  e.g.  $e = dterms:subject$

1. for all  $v \in \{C\}$  do
2.      $V' :=$  the set of all  $v' : v' \subseteq v$
3.     for all  $v'_e : \exists \langle v'_e, e, v' \rangle$  do
4.          $N_e :=$  set of all  $\langle v'_e, e, v' \rangle : v' \in V'$
5.          $M_e :=$  set of all  $\langle v'_e, e, v_a \rangle : v_a \in V$
6.          $AV_{v'_e} := |N_e| / |M_e|$
7.          $CAC_{v'_e} := (|N_e| / |M_e|) \cdot (|N_e| / |V'|)$
8.          $CU_{v'_e} := (|N_e| / |V'|)^2 - (|M_e| / |V|)^2$
9.          $AV_v := AV_v + AV_{v'_e}$
10.          $CAC_v := CAC_v + CAC_{v'_e}$
11.          $CU_v := CU_v + CU_{v'_e}$
12.     end for
13. end for

Output:  $AV_v, CAC_v, CU_v$  for all  $v \in \{T \cup C\}$





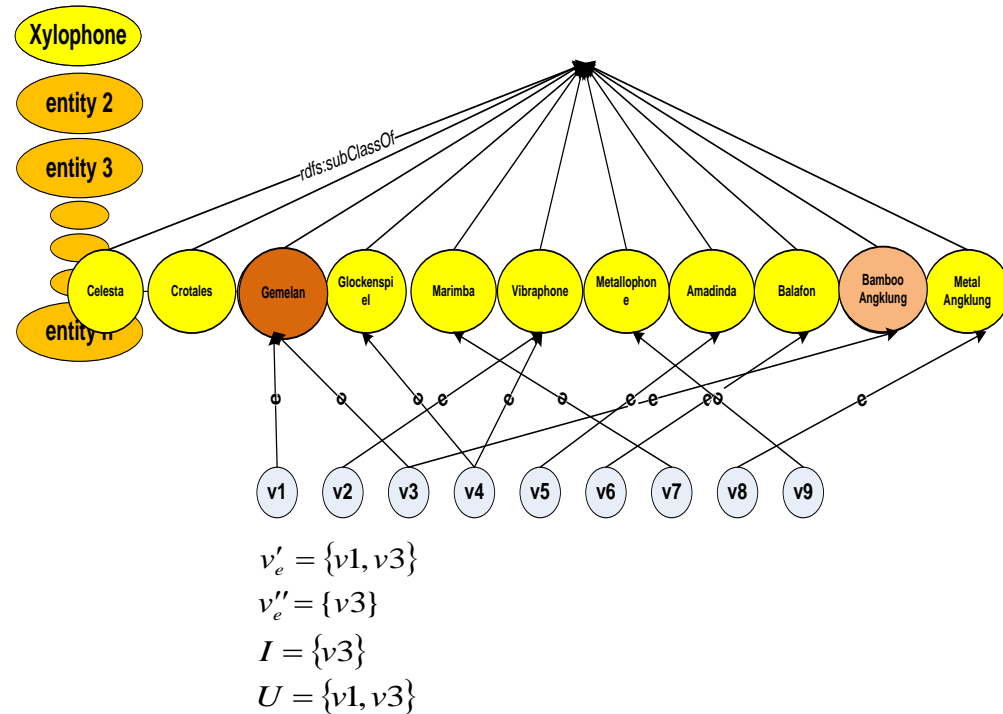
# Identifying Knowledge Anchors

## Algorithm II: Homogeneity Metrics – use set based similarity metrics

**Input:**  $DG = \langle V, E, P \rangle, e \in E$

1. **for all**  $v \in \{T \cup C\}$  **do**
2.      $V' :=$  the set of all  $v' : v' \subseteq v$
3.     **for all**  $(v', v'') : v' \in V' \wedge v'' \in V'$  **do**
4.          $V'_e := \{v'_e : \exists \langle v'_e, e, v' \rangle\}$
5.          $V''_e := \{v''_e : \exists \langle v''_e, e, v'' \rangle\}$
6.          $I := V'_e \cap V''_e$
7.          $U := V'_e \cup V''_e$
8.          $CN_{v',v''} := |I|$
9.          $Jac_{v',v''} := |I| / |U|$
10.          $Cos_{v',v''} := |I| / (\sqrt{|V'_e|} \cdot \sqrt{|V''_e|})$
11.          $CN_v = CN_v + CN_{v',v''}$
12.          $Jac_v = Jac_v + Jac_{v',v''}$
13.          $Cos_v = Cos_v + Cos_{v',v''}$
14.     **end for**
15.      $CN_v = CN_v / (|V'| \cdot (|V'| - 1) / 2)$
16.      $Jac_v = Jac_v / (|V'| \cdot (|V'| - 1) / 2)$
17.      $Cos_v = Cos_v / (|V'| \cdot (|V'| - 1) / 2)$
18. **end for**

**Output:**  $CN_v, Jac_v, Cos_v$  for all  $v \in \{T \cup C\}$



# Problem & Focus



UNIVERSITY OF LEEDS

## Two important questions we need to answer

What makes some entities more important to provide better paths?

How to develop automatic ways to identify knowledge anchors?



**How to evaluate?**

## Experimental Setup

- We follow earlier Cognitive Science studies (free naming tasks – 10s).
- Images of all taxonomical entities linked via *rdfs:subClassOf* were presented
  - Eight surveys presented 256 leaf entities.
  - Two surveys presented 108 top/category entities.

## Benchmarking Sets Identified

### Accuracy and frequency were used

- **Set 1:** accurate naming of a category entity(parent) when leaf entity is seen
  - Violotta → Violin.
- **Set 2:** accurate naming a category entity with its exact name, children or parent.
  - Violin → Violin.
  - Fiddle → Violin.
  - Plucked String Inst → String Inst

### Strong Anchors (Set1 $\cap$ Set2)

{Accordion, Bell, Bouzouki, Clarinet, Drum, Flute, Guitar, Harmonica, Harp, Saxophone, String instrument, Trumpet, Violin, Xylophone}

### Weak Anchors (Set1 $\cup$ Set2)

{Accordion, Banjo, Bell, Bouzouki, Cello, Clarinet, Drum, Electric piano, Flute, Gong, Guitar, Harmonica, Harp, Lute, Lyre, Organ, Recorder, Saxophone, String Instrument, Trombone, Trumpet, Tuba, Violin, Xylophone}

## Quantitative Analysis

- For each metric, we aggregate (union) entities for hierarchical relationships.
- Identify a cut-off threshold point (normalize values and take 60<sup>th</sup> percentile).
- Precision values were poor (*0.17-0.29 for StrongAnchors & 0.27-0.37 for WeakAnchors*)
- Inspecting the *False Positive* entities, we noticed two reasons for the poor precision.
  - Picking entities with a low number of subclasses [ $SN_v = 1 - (1 / |\{v' : v' \subseteq v\}|)$ ]
  - Returning FP entities which had long label names [use weighted median of labels]
- Baseline:
  - 0.25 using Strog Anchors.
  - 0.41 using WeakAnchors
- Precision results were improved noticeably (*lowest value 0.41 to highest value 0.62*).

# Experimental Results



UNIVERSITY OF LEEDS

## Hybridization

- **Analyze algorithms performance at different taxonomical levels.**
  - Eight taxonomical levels
  - First and last levels are excluded (**Figure**)
- **Two heuristics:**
  - **Use *hierarchical Jaccard* for most specific categories in the graph**  
(FP entities were rich in domain-specific relationships)
  - **Take the majority voting for all other taxonomical levels.**  
(Most of the entities at the middle and top taxonomical level will be well represented in the graph hierarchy and may include domain-specific relationships)

<b>Benchmarking sets</b>	<b>Precision</b>	<b>Recall</b>
<i>StrongAnchors</i>	0.48	0.79
<i>WeakAnchors</i>	0.65	0.63



## Qualitative Analysis of Hybridization – examine FP and FN entities

### Observations

- **Missing basic level entities due to unpopulated areas in the data graph**
  - None of the metrics picked FN entities (Cello, Banjo) that belonged to the bottom quartile of the taxonomy (small number of subclasses and limited performances).

**We argue that these entities would take the user to 'dead-ends' with unpopulated areas which may be confusing for navigation**



- **Selecting entities that are superordinate of basic level entities**
  - The FP included entities, such as which are well presented in the graph hierarchy (e.g. Reeds has 36 subclasses linked to 60 DBpedia categories).
  - Also, their members participate in many domain-specific relationships (e.g. Reeds members are linked to 606 performances).

**We argue that such entities could be seen as 'good picks' because they can provide navigation bridges to reach basic level entities**



# Contribution



UNIVERSITY OF LEEDS

- **We uniquely provide formal description of metrics and corresponding algorithms for identifying knowledge anchors in a data graph.**
- **Implementation of the algorithms to identify basic level entities in a data graph in the music domain.**
- **The performance of the algorithms is examined using a benchmarking set with basic level entities identified by humans.**
- **This work has been accepted in HT2016 conference.**

# Conclusion and Future work



UNIVERSITY OF LEEDS

## Conclusion

- We develop algorithms for identifying knowledge anchors in a data graph.
- Knowledge anchors can provide bridges to subsume new knowledge.
- Our approach can be validated with Crowdsourcing in other domains.
- Our approach can be useful for the cold start problem.
- Our approach can be also applied to ontology summarization for capturing a lay person's view of the domain.

## Future work

- Utilize the algorithms to generate navigation paths by following Ausubel's subsumption strategies for meaningful learning.
- Apply the algorithms in another domain.



UNIVERSITY OF LEEDS

**Thank you, Questions ?**

**Marwan Al-Tawil**

***Contact: [scmata@leeds.ac.uk](mailto:scmata@leeds.ac.uk)***