Using Basic Level Concepts in a Linked Data Graph to Detect User's Domain Familiarity

Marwan Al-Tawil, Vania Dimitrova, Dhavalkumar Thakker
School of Computing, University of Leeds, United Kingdom.

Abstract. We investigate how to provide personalized nudges to aid a user's exploration of linked data in a way leading to expanding her domain knowledge. This requires a model of the user's familiarity with domain concepts. The paper examines an approach to detect user domain familiarity by exploiting anchoring concepts which provide a backbone for probing interactions over the linked data graph. Basic level concepts studied in Cognitive Science are adopted. A user study examines how such concepts can be utilized to deal with the cold start user modelling problem, which informs a probing algorithm.

Keywords: Linked data, knowledge utility, user modeling, basic level concepts.

1 Introduction

The recent growth of the Web of Linked Data[1](LD), which provides access to big data graphs representing domain entities and their relationships, has opened a new avenue of research on developing computational models to facilitate data exploration by layman users [9]. This has brought together research from Semantic Web and HCI to shape novel tools for interactive exploration of semantic data[2]. One of the key challenges is ensuring that the interaction with linked data *brings benefits for the users*. Hence, personalization and adaptation can play a crucial role. Research in personalized exploration of linked data is still in an embryonic stage. Current work includes improving search efficiency by considering user interests [4, 7] or diversifying the user exploration paths with recommendations based on the browsing history [8].

Our research brings a new dimension of personalization and adaptation to enhance the benefits of linked data exploration, namely *knowledge utility*. We investigate how to aid a user's exploration of linked data in a way leading to expanding her domain knowledge. This can have a broad implication for facilitating sense making while exploring linked data. Learning is an inevitable part of exploratory search, as users are discovering new connections and associations. Our earlier research has shown that although linked data exploration can promote domain knowledge expansion ('serendipitous learning' effect), *not all paths can be beneficial*. We derived empirically strategies to nudge the user to beneficial paths [10]. The user familiarity with the entities in the linked data graph (LDG) was identified as a crucial input for the nudging

[1] http://linkeddata.org/

[2]See the series of IESD workshops, e.g. IESD2014 held @ ISWC: https://iesd14.wordpress.com/

strategies we aim to develop – profitable exploration sequences include a start (anchor) in a familiar entity followed by bringing a new (unexpected/interesting) entity.

Identifying the user familiarity with the domain entities (domain concepts or instances) from the LDG is not a trivial task because LDGs usually include thousands of entities at different levels of abstraction. This brings forth the well-known *cold start problem* of user modelling, which is aggravated by the sheer number of LDG entities. One way to address cold start is via a probing dialogue. While LDG can provide a knowledge pool to implement a probing dialogue for user modelling (c.f. [6]), it is not clear what domain entities to select from the vast amount of possibilities for probing. Consequently, the interactions can be too long and may refer to entities that do not bring high value for modelling a user's domain familiarity.

This paper examines an approach to detect user domain familiarity by using anchoring concepts in the LDG around which a probing dialogue can be developed. We adopt the Cognitive Science notion of *basic level concepts* (BLCs) – domain concepts that are highly informative and can be easily retrieved from memory. An example of a *basic level concept* in the Music domain is Guitar [3]; it has Musical Instrument as a *superordinate* concept (more abstract) and Classical Guitar as a *subordinate* concept (more specific). BLCs are likely to provide knowledge bridges to learn new concepts in big information spaces and to serve as indicators for user modelling. Cognitive science research has shown that the use of BLCs may indicate domain familiarity, e.g. experts tend to recognise subordinate concepts [5].

To get insights into how BLCs can be utilized to identify user's domain familiarity, we conduct a user study that adopts earlier Cognitive Science methods which derive BLCs in a specific domain [3, 5] to identify the BLCs in a LDG. Based on the study, we derive heuristics how BLCs can be related to user domain familiarity. We then suggest a user modelling probing algorithm that utilizes BLCs.

2 Identifying Basic Level Concepts in a Linked Data Graph

We conducted a user study to examine how BLCs in a LDG *can be utilized to model a user's domain familiarity.*

2.1. Study Design

Dataset. We have used a dataset from the music domain which underpins a linked data browser (MusicPinta) developed by us in an earlier research [2]. The MusicPinta LDG is fairly large and diverse, yet of manageable size for experimentation. It contains 2.4M entities and 38M triple statements, and includes facts about 876 musical instruments from various categories, including many country-specific instruments. Musical instruments, which have been used by Cognitive Science studies in BLC, provide a suitable domain for cognitive activities linked with BLCs [11].

Participants. The study involved 40 participants recruited on a voluntary basis, varied in Gender (28 male and 12 female), cultural background (1 Belgian, 10 British, 5 Bulgarian, 1 French, 1 German, 5 Greek, 1 Indian, 2 Italian, 6 Jordanian, 1 Libyan , 2 Malaysian, 1 Nigerian, 1 Polish, and 3 Saudi Arabian), and age (18 – 55, mean = 25).

Method. We follow the experimental set up in earlier Cognitive Science studies which derived BLCs using free-naming tasks in a specific domains [3, 5], including the Music domain. Participants were asked to freely name objects that were shown in image stimuli, under limited response time (10s). 364 taxonomical musical instruments were extracted from the MusicPinta dataset by running SPARQL queries from the MusicPinta SPARQL endpoint to get all musical instrument concepts linked via the rdfs:subClass relationship. The musical instrument concepts were classified either into *leaf* (*l*) instruments (total=265) or *category* (*c*) instruments (total= 108). Leaf instruments are found at the bottom of a hierarchy and do not have children, whereas category instruments have at least one child. For each leaf instrument *l*, a representative image (stimuli) was collected from the Musical Instrument Museums Online (MIMO)[3] and Wikipedia[4]. For a category *c*, all leafs from that category were shown as a group. Following the Cognitive Science studies, additional objects, outside the domain, were included to minimize response bias - 64 *additional* images were randomly chosen from the most occurring concepts in artificial and natural categories from the Battig and Montague category norms [1], including vehicles, clothing, furniture, tools, fruits, vegetables, animals and birds.

Ten online surveys were run adopting two strategies: (i) *Strategy 1 – leaf instruments*: eight surveys presented the leaf instruments – each survey presented 32 *leaves* and 8 *additional* images. (ii) *Strategy 2 – category instruments*: two surveys presentedthe category instruments- each survey showed 54 *categories* and 14 *additional* images. The image allocation in surveys was random. Every survey had 4 participants; each participant conducted one survey following an online link, including:

- *Pre-task questionnaire*-collecting information about user profile (e.g. age-group, nationality, and gender).
- *Free-naming task*- Each image was shown for 10 seconds on the participant's screen and he/she was asked to type the name of the given object(s) in the image as quickly as possible. Figures 1-4 show example instrument images and participant's answers from the study. For this task, we recorded accuracy (i.e. the participant answered correctly) and frequency (i.e. how many times a particular instrument name was mentioned correctly) of their accurate answers.
- *Post-task questionnaire*- collected information about the participant's familiarity level for the six top level musical instrument categories (String Instruments, Wind Instruments, Percussion Instruments, Electronic Instruments, and Other Instruments). Participants were asked to rate their knowledge in these categories on a scale of 1 to 7 (1=No Knowledge and 7=Expert).

2.2. Basic Level Concepts Identified

To extract BLCs from the MusicPinta dataset we considered accuracy and frequency of the participants' answers [5], grouping the answers into:

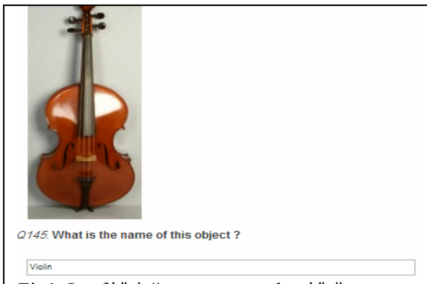
- ***Group1**: Naming a leaf instrument with its category instead of its own name*. In this group, we calculated the frequency of exact matches between the partici-

[3] http://www.mimo-international.com/MIMO/

[4] Wikipedia images were used only in the cases when a MIMO image did not exist.

pants' answers and the category of instruments seen. For example, as shown in Fig.1, a participant has named the leaf instrument Violotta with its parent category Violin. We counted how many times Violin was named when its leaves were seen.

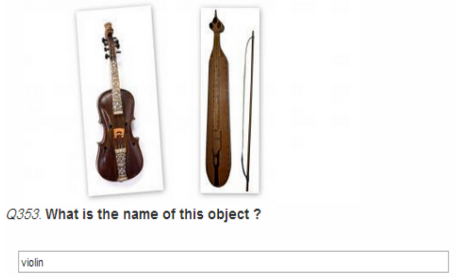
- ***Group 2****:Exact naming of categories.* In this group, we considered the cases when participants were able to exactly name the category of the instrument they saw, e.g.Fig. 3 shows a response where the category Violins was seen and named.
- ***Group 3****: Naming a category level instrument with its parent or children instrument name.* This is illustrated in Fig. 2 and Fig. 4 - the participant saw a category level instrument (Fiddle and Plucked String Instruments*)*and named its parent (Violin and *String Instruments*, respectively).



Q145. What is the name of this object ?

Violin

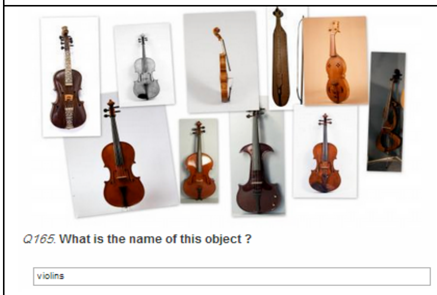
Fig1. Leaf Violotta seen, named as Violin.



Q353. What is the name of this object ?

violin

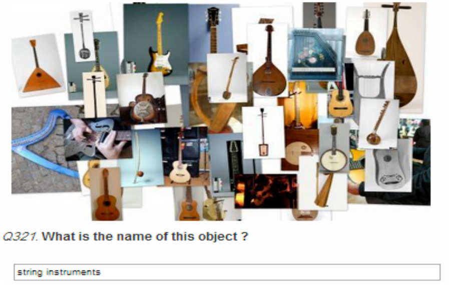
Fig2. Category Fiddle seen, named as Violin.



Q165. What is the name of this object ?

violins

Fig3. Category Violins seen, named as Violins.



Q321. What is the name of this object ?

string instruments

Fig4. Category Plucked String Instruments seen, named as String Instruments.

In each of the groups, LDG entities with frequency above 2 (i.e. they were named by at least two users) were included. The entities identified in Group 1 are derived from Strategy 1, while Group 2 and Group 3 give complementary output for the LDG entities derived by Strategy 2 (i.e. when the participants saw categories of instruments). Hence, the union of Group 2 and Group 3 gives the likely BLCs identified with Strategy 2, which is then intersected with the output from Strategy 1 to obtain the final BLC list. This included: Accordion, Bells, Bouzouki, Clarinet, Drums, Flute, Guitars, Harmonica, Harp, Saxophone, String instruments, Trumpet, Violins and Xylophone.

3 Using Basic Level Concepts for User Modeling

We compared the user survey answers and user's familiarity for the six top level musical instrument categories. The findings were used to derive probing heuristics.

When the user is able to name an instance (leaf instrument as seen in Strategy 1) instead of its corresponding BLC, she has high familiarity in the corresponding top *level category.* There were 27 cases (out of 41) where the participants could name leaf instruments rather than using their BLCs (as the majority of users did). For example, a participant named the leaf *Electric cello* instead of naming its BLC *Violin.* In 67% of these cases users had high familiarity with the top level instrument category.

When the user successfully names children of a basic level concept from images of the corresponding categories (as seen in Strategy 2), she has high familiarity in the corresponding top level category. There were 34 cases where the participants named children that belonged to the basic level. For example, one participant named the child *Cello* instead of naming it with its BLC *Violin*. In 62% of these cases participants had high familiarity with the top level instrument category.

When the user cannot name a basic level concept from the corresponding BLC images (as seen in Strategy 2), she has low familiarity in the corresponding top level category. There were 11 cases (out of 64) where participants were shown a BLC and were unable to name it. In these cases, the participants had indicated low or no knowledge with the top level instrument category.

*When the user can name a basic level concept from the corresponding images for the BLC category (as seen in Strategy 2), she is likely to have high familiarity with the corresponding top level category.*There were 43 (out of 64) cases where participants were shown a BLC and named it correctly. In half (58%) of these cases the participants had high familiarity with the top level instrument category of the BLC.

Based on the above heuristics, we propose a probing algorithm based on BLC.

Input

Domain - a linked data graph $G=(V,E)$ where $V=\{v_1,v_2,\dots,v_n\}$

Set of Images $I=\{i_1,i_2,\dots,i_n\}$ and a function $image: V \rightarrow I$ assigning an image *i* to each vertex *v*

Set of Basic level concepts: $B=\{b_1,b_2,\dots,b_k\}$

User diagnosis is a mapping that overlays G's vertices with a familiarity level - none, low, medium, high. $familiarity: V \rightarrow W$ where $V=\{v_1,v_2,\dots,v_n\}$ and $W=\{none,low,medium,high\}$

For every vertex from G, we define the following functions which are implemented with simple inferences using hierarchical relationships:

- P(v) - returns all parent concepts for v
- C(v) - returns all children concepts (including the leafs) for v
- L(v) - returns the leaves (instances) for v
- T(v) - top level categories for v

Output

User model $U=V \times W$ where $d: V \rightarrow W$

Processing

```
//initialization
for all $v \in V$ do $d(v) = none$

for all $b \in B$ do //BLC naming
  show $image(b)$ and ask to name it
  if user_answer≠$b$ do //cannot name BLC
    $familiarity(b) = none$
    for all $t \in T(b)$ $familiarity(t) = low$
  else do //names BLC
    $familiarity(b) = medium$
    for all $t \in T(v)$ $familiarity(t) = medium$
    for all $c \in C(b)$ do //check subordinate
      show $image(c)$ and ask to name it
      if user_answer== $c$ do
        $familiarity(c) = high$
        $familiarity(b) = high$
      end if
    end for
    if $familiarity(b) == high$ do//check leaves
      for all $l \in L(b)$ do
        show $image(l)$ and ask to name it
        if user_answer== $l$ do
          $familiarity(l) = high$
          for all $t \in T(l)$ $familiarity(t) = high$
        end if
      end for
    end if
  end if
end for
```

4 Current State and Future Work

In this work, we examine the advantage of using basic level concepts to detect user familiarity in a linked data graph. The user study identifies the BLCs in a Music domain in a free-naming task and illustrates how these concepts can be utilized to detect the user familiarity with a subset of entities from the LDG. Obviously, these findings can only be applied if it is possible to automatically detect BLCs from a LDG. Following the Cognitive Science definition of BLCs–domain concepts that carry the most information, possess the highest category cue validity, and are, thus, the most differentiated from one another are highly informative [3]-we have implemented eight algorithms for extracting BLCs from the LDG. The algorithms search for basic categories at the most inclusive level at which attributes are common to most categories' members and basic categories which are most differentiated from other categories (categories with highest cue validity, i.e. their members have attributes common to the category and not belonging to other categories). We have implemented appropriate SPARQL queries over the MusicPinta dataset adopting several semantic relationships and similarity measures. The set of BLCs identified in the study is used as a ‘ground truth’ to benchmark the algorithms. Current results show that the best performing algorithms achieve precision of 0.48, which is promising but insufficiently high.

Our immediate future work is to tune the BLC algorithms and explore various fusion methods to improve the precision results. We will then be able to implement the probing algorithm and utilise it in developing the nudging strategies derived in [10].

References

1. Van Overschelde, J. P., Rawson, K. A., &Dunlosky, J. Category norms: An updated and expanded version of the Battig and Montague (1969) norms. Journal of Memory and Language, 2004, 50, 289-335.
2. Thakker, D., Dimitrova, V., Lau, L., Yang-Turner, F. &Despotakis, D. Assisting User Browsing over Linked Data: Requirements Elicitation with a User Study. In proceedings of ICWE 2013, pp. 376-383.
3. Rosch, E., Mervis, C. B., Gray, W. D., Johnson, D. M., &Boyes-Braem, P. Basic objects in natural categories. Cognitive Psychology, 1976, 8, 382-439.
4. Sah, M. & Wade, V. Personalized Concept-based Search and Exploration on the Web of Data using Results Categorization. In ESWC 2013.
5. Tanaka, J., & Taylor, M. Object Categories and Expertise: Is the Basic Level in the Eye of the Beholder? Cognitive Psychology, 1991, 23(3). 457-482.
6. DhavalThakker, Lydia Lau, Ronald Denaux, VaniaDimitrova, Paul Brna, Christina M. Steiner:Using DBpedia as a Knowledge Source for Culture-Related User Modelling Questionnaires. UMAP 2014.
7. Rossel,O. Implemention of a “search and browse” scenario for theLinkedData. In Intelligent Exploration of Linked Data (IESD), 2014.
8. Vocht1, et, al. A Visual Exploration Workflow as Enablerfor the Exploitation of Linked Open Data. In Intelligent Exploration of Linked Data (IESD), 2014.
9. MC Schraefel, What does it look like, really? Imagining how citizens might effectively, usefully and easily find, explore, query and re-present open/linked data. In ISWC 2010.
10. Al-Tawil, M., Thakker, D. and Dimitrova, V. Nudging to Expand User’s Domain Knowledge while Exploring Linked Data. In Intelligent Exploration of Linked Data (IESD), 2014, @ ISWC2014.
11. Palmer, F., Jones, K., Hennessy, L., Unze, G., & Pick, A. D. How is a trumpet known? the “basic object level” concept and the perception of musical instruments. American Journal of Psychology, 102, 1989.