CLOCQ: A Toolkit for Fast and Easy Access to Knowledge Bases
Philipp Christoph, Rishiraj Saha Roy, and Gerhard Weikum
Max Planck Institute for Informatics, Germany

Public API and source code at clocq.mpi-inf.mpg.de

KNOWLEDGE BASES STORE VAST AMOUNTS OF FACTUAL KNOWLEDGE

- Curated knowledge bases (KBs) store factual knowledge in structured way and have many use-cases for search, entity linking, etc.
- Qualifiers express n-ary relationships in Wikidata; similar concepts used in other KBs such as DBpedia or YAGO
- Real-world KBs have billions of facts, with millions of entities and thousands of predicates, consuming multiple terabytes of disk space
- KBs used in question answering (QA) systems, to answer factoid questions like Who wrote Harry Potter? or Who scored an own goal in the 2018 final?

LIMITATIONS OF EXISTING TRIPLE-CENTRIC KB INTERFACES

- Existing KB interfaces allow general-purpose access via queries (e.g., SPARQL)
- Access requires detailed knowledge and understanding of KB schema
- Interfaces not designed for accessing n-ary facts
- Treat KB as pure set of triples and integrate qualifiers via reification
- Leads to expensive querying and post-hoc processing

Graph-based definition of KB distance (with triple-centric view)

CKB distance (Mario Mandžukić, 2018 FIFA World Cup Final) = 4 (follow 4 edges)
CKB distance (Mario Mandžukić, Luzhniki Stadium) = 6 (follow 6 edges)

Fact-based definition of KB distance (Proposed)

CKB distance (Mario Mandžukić, 2018 FIFA World Cup Final) = 4 (follow 4 edges)
CKB distance (Mario Mandžukić, Luzhniki Stadium) = 2 (fact-query)

Legend

Entity node
Predicate node

CLOCQ APPROACH

- Take fact-centric view of KBs (vs. triple-centric)
- Establish intuitive definitions for vaguely defined concepts, such as
  - KB graph, KB neighborhood, KB distance, shortest path between KB items
- Implement fact-centric KB index that enables more efficient implementation of core KB functionalities utilized in many IR and NLP systems
- Provide public API to conveniently access Wikidata at clocq.mpi-inf.mpg.de

CLOCQ FUNCTIONALITY

- Direct lookups
  - Label, aliases, description, types, or most frequent type of KB item
- More complex functionalities
  - k-hop neighborhood of KB item
  - Frequency of KB item
  - Connectivity / shortest path between two KB items
  - Entity linking (text)
  - Relation linking (text)

CLOCQ improves runtime over traditional triple-centric KB interfaces

RUNTIME EXPERIMENTS

Baselines

- HDT [1]: Efficient triple lookups using bitmap encodings
- QueryService [2]: Publicly available SPARQL query interface for Wikidata

Large-scale runtime analysis for key KB functionalities and randomly chosen KB items.

Table: CLOCQ vs. HDT and QueryService

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Neighborhood</td>
<td>20.0 s</td>
<td>&quot;n/a&quot;</td>
<td>1.07 x 10^-4 s</td>
</tr>
<tr>
<td>Neighborhood (n)</td>
<td>2.990 x 10^-4</td>
<td>&quot;n/a&quot;</td>
<td>15.6 s</td>
</tr>
<tr>
<td>Neighborhood (n)</td>
<td>15.2 x 10^-6</td>
<td>&quot;n/a&quot;</td>
<td>3.56 x 10^-4 s</td>
</tr>
<tr>
<td>Frequency (n)</td>
<td>2.05 x 10^-6</td>
<td>&quot;n/a&quot;</td>
<td>5.34 x 10^-4 s</td>
</tr>
<tr>
<td>Frequency (n)</td>
<td>5.20 x 10^-5</td>
<td>&quot;n/a&quot;</td>
<td>3.99 x 10^-4 s</td>
</tr>
<tr>
<td>Frequency (n)</td>
<td>5.20 x 10^-5</td>
<td>&quot;n/a&quot;</td>
<td>5.44 x 10^-4 s</td>
</tr>
<tr>
<td>Connectivity (n)</td>
<td>6.13 x 10^-5</td>
<td>&quot;n/a&quot;</td>
<td>5.37 x 10^-4 s</td>
</tr>
<tr>
<td>Connectivity (n)</td>
<td>60.3 x 10^-5</td>
<td>&quot;n/a&quot;</td>
<td>5.21 x 10^-4 s</td>
</tr>
<tr>
<td>Connectivity (n)</td>
<td>60.3 x 10^-5</td>
<td>&quot;n/a&quot;</td>
<td>5.10 x 10^-4 s</td>
</tr>
<tr>
<td>Shortest path (n)</td>
<td>118 x 10^-5</td>
<td>&quot;n/a&quot;</td>
<td>8.42 x 10^-4 s</td>
</tr>
<tr>
<td>Shortest path (n)</td>
<td>120 x 10^-5</td>
<td>&quot;n/a&quot;</td>
<td>8.99 x 10^-4 s</td>
</tr>
<tr>
<td>Shortest path (n)</td>
<td>5,260 x 10^-5</td>
<td>&quot;n/a&quot;</td>
<td>0.178 s</td>
</tr>
</tbody>
</table>

[1] Based on HDT representation for publication and exchange (HDT) Ferdinand et al., Journal of Web Semantics 2015