Knowledge Bases

Lidia Pivovarova

Based on:

Building, Maintaining, and Using Knowledge Bases: A Report from the Trenches by Deshpande et. al  SIGMOD’13
Introduction

- Knowledge base – machine-readable way to store human knowledge.
- Usually consists of concepts, instances and relations

![Diagram](image)

Figure 1: A tiny example of a KB
Examples
Applications

- search engines such as Google and Bing use global KBs to understand and answer user queries
- So do ecommerce Web sites, such as amazon.com and walmart.com, using product Kbs.
- iPhone voice assistant Siri uses KBs to parse and answer user queries
- echonest.com builds a large KB about music, then uses it to power a range of applications, such as recommendation, playlisting, fingerprinting, and audio analysis
- using KBs to find domain experts in biomedicine, to analyze social media, to search the Deep Web, and to mine social data...
This paper

- Describe an end-to-end process on building, maintaining and using KBs in industry
  - “how do we maintain a KB over time?”,
  - “how do we handle human feedback?”,
  - “how are schema and data matching done and used?”
  - “the KB will not be perfectly accurate, what kinds of application is it good for?”,
  - “how big of a team do we need to build such a KB, and what the team should do?”.

- The team:
  - Kosmix startup, later Walmart-Labs
  - working on product search, customer targeting, social mining, and social commerce
Preliminaries

- a set of concepts $C_1, \ldots, C_n$,
- a set of instances $I_i$ for each concept $C_i$,
- a set of relationships $R_1, \ldots, R_m$ among the concepts
- \textit{is-a} – special relation, that imposes a taxonomy

Figure 1: A tiny example of a KB

- Domain-Specific KBs vs. Global Kbs
- Ontology-like KBs vs. Source-Specific KBs
The main processes

● BUILDING THE KNOWLEDGE BASE
  – Constructing the Taxonomy Tree from Wikipedia
  – Constructing the DAG on top of Taxonomy
  – Extracting Relationships from Wikipedia
  – Adding Metadata
  – Adding Other Data Sources

● MAINTAINING THE KNOWLEDGE BASE
  – Updating the Knowledge Base
  – Curating the Knowledge Base

● USING THE KNOWLEDGE BASE
The main processes

- **BUILDING THE KNOWLEDGE BASE**
  - Constructing the Taxonomy Tree from Wikipedia
  - Constructing the DAG on top of Taxonomy
  - Extracting Relationships from Wikipedia
  - Adding Metadata
  - Adding Other Data Sources

- **MAINTAINING THE KNOWLEDGE BASE**
  - Updating the Knowledge Base
  - Curating the Knowledge Base

- **USING THE KNOWLEDGE BASE**
Constructing the Taxonomy Tree from Wikipedia

Figure 2: Two main kinds of Wikipedia pages - article page (left) and category page (right)

- Ideally: categories are concepts, articles are instances
- In reality: cycles, too general categories
Too general categories

Figure 4: Constructing the top levels of our taxonomy and the list of verticals
Cycles

- **Solution:**
  - First build a graph
  - Then use a pruning:
  - Edmonds’ algorithm, Tarjan implementation
  - Finds optimal branching using edge weights
  - **Weights:**
    - artcat, wsubcat, warticle
    - co-occurrence count
    - name similarity
    - manually assigned weights
The main processes

- BUILDING THE KNOWLEDGE BASE
  - Constructing the Taxonomy Tree from Wikipedia
  - Constructing the DAG on top of Taxonomy
  - Extracting Relationships from Wikipedia
  - Adding Metadata
  - Adding Other Data Sources

- MAINTAINING THE KNOWLEDGE BASE
  - Updating the Knowledge Base
  - Curating the Knowledge Base

- USING THE KNOWLEDGE BASE
Constructing the DAG on top of Taxonomy

- Ronald Reagan – *U.S. President*
  - *American actor*

- Go back to Wikipedia graph and preserve as many relations as possible without having cycles
The main processes

- **BUILDING THE KNOWLEDGE BASE**
  - Constructing the Taxonomy Tree from Wikipedia
  - Constructing the DAG on top of Taxonomy
  - *Extracting Relationships from Wikipedia*
  - Adding Metadata
  - Adding Other Data Sources

- **MAINTAINING THE KNOWLEDGE BASE**
  - Updating the Knowledge Base
  - Curating the Knowledge Base

- **USING THE KNOWLEDGE BASE**
Figure 5: Extraction of relationships from a Wikipedia page
The main processes

- **BUILDING THE KNOWLEDGE BASE**
  - Constructing the Taxonomy Tree from Wikipedia
  - Constructing the DAG on top of Taxonomy
  - Extracting Relationships from Wikipedia
  - Adding Metadata
  - Adding Other Data Sources

- **MAINTAINING THE KNOWLEDGE BASE**
  - Updating the Knowledge Base
  - Curating the Knowledge Base

- **USING THE KNOWLEDGE BASE**
Add Metadata

- Adding Synonyms
  - Redirect pages: e.g. Sokrat → Socrates

- Adding Homonyms
  - Disambiguation text: e.g. Socrates the philosopher, Socrates a Brazilian football player, Socrates a play, Socrates a movie...

- Adding Metadata per Node
  - Web urls, Twitter Ids,
  - Co-occurring concepts and instances,
  - Wikipedia page traffic
  - Frequency of concept mentions in Wiki and social
  - Web-signature, social signature
The main processes

• BUILDING THE KNOWLEDGE BASE
  – Constructing the Taxonomy Tree from Wikipedia
  – Constructing the DAG on top of Taxonomy
  – Extracting Relationships from Wikipedia
  – Adding Metadata
    – Adding Other Data Sources

• MAINTAINING THE KNOWLEDGE BASE
  – Updating the Knowledge Base
  – Curating the Knowledge Base

• USING THE KNOWLEDGE BASE
Adding other Data Sources

**Main principles:**
- Handle as many simple cases as possible
- In difficult cases alert human expert
- Remember and re-use all human actions

<table>
<thead>
<tr>
<th>Name</th>
<th>Domain</th>
<th>No. of instances</th>
</tr>
</thead>
<tbody>
<tr>
<td>Chrome</td>
<td>Automobile</td>
<td>100K</td>
</tr>
<tr>
<td>Adam</td>
<td>Health</td>
<td>100K</td>
</tr>
<tr>
<td>Music-Brainz</td>
<td>Music</td>
<td>17M</td>
</tr>
<tr>
<td>City DB</td>
<td>Cities</td>
<td>500K</td>
</tr>
<tr>
<td>Yahoo! Stocks</td>
<td>Stocks and companies</td>
<td>50K</td>
</tr>
<tr>
<td>Yahoo! Travel</td>
<td>Travel destinations</td>
<td>50K</td>
</tr>
</tbody>
</table>

- Extract a taxonomy from a new source
- Merge taxonomies using concordance ("car" = "auto", "movie" = film")
- Merge taxonomies
- Extract instances and attributes
- Try to merge as many instances as possible automatically
- Alert experts in other cases
The main processes

• BUILDING THE KNOWLEDGE BASE
  – Constructing the Taxonomy Tree from Wikipedia
  – Constructing the DAG on top of Taxonomy
  – Extracting Relationships from Wikipedia
  – Adding Metadata
  – Adding Other Data Sources

• MAINTAINING THE KNOWLEDGE BASE
  – Updating the Knowledge Base
  – Curating the Knowledge Base

• USING THE KNOWLEDGE BASE
Updating the Knowledge Base

- Incremental update may cause difficulties in handling inconsistencies
- Thus, the whole KB is rebuilt from scratch
  - a single machine with 256G RAM, 0.8GHz processor, and 32 processors, takes roughly 12.5 hours to complete the construction pipeline
- To preserve manual changes
  - All human curations are saved in a form of commands in a special language that can be rerun after update
The main processes

• BUILDING THE KNOWLEDGE BASE
  – Constructing the Taxonomy Tree from Wikipedia
  – Constructing the DAG on top of Taxonomy
  – Extracting Relationships from Wikipedia
  – Adding Metadata
  – Adding Other Data Sources

• MAINTAINING THE KNOWLEDGE BASE
  – Updating the Knowledge Base
  – Curating the Knowledge Base

• USING THE KNOWLEDGE BASE
Curating the Knowledge Base

- Evaluating the quality
  - random sample of paths (from root to leave)
  - nodes with more than 200 children

- Curating by writing commands
  - Adding/deleting nodes and edges
  - Changing edge weights
  - Changing the assignment of an instance-of or an is-a relationship
  - Recommending an ancestor to a node
  - Assigning preference to a subtree in the graph
The team

• 25-30 developers.
• a core team of 4 persons was in charge of the KB
• A data analyst performed quality evaluation and curated the KB
• A developer wrote code, developed new features, added new signals on the edges, and so on.
• A system person worked 50% of the time on crawling the data sources, and maintaining the in-house Wikipedia mirror and the Web corpus.
• An UI specialist worked 50% of the time on the look and feel of the various tools.
• A team lead designed, supervised, and coordinated the work.
Applications

- Understanding User Queries
- In-context Advertising
- Social Mining
- Event Monitoring in Social Media
- Product search
- Social gifting
Thanks for your attention!