GRAPH-BASED KNOWLEDGE MODELS

JELENA JOVANOVIĆ
Email: jeljov@gmail.com
Web: http://jelenajovanovic.net
OVERVIEW

• Graphs and semantic networks for knowledge representation
• Data and knowledge graphs in the business domain
• Open data and knowledge graphs
• Gigantic Global Graph
  • Vision of the Web as a gigantic global (data and knowledge) graph
  • Creation of gigantic knowledge bases through automated data collection from the Web
Graphs and Semantic Networks for Knowledge Representation
AN EXAMPLE OF A GRAPH OF ENTITIES AND THEIR MUTUAL RELATIONS

The example illustrates a tiny segment of the Freebase knowledge base

Image source: http://inspirehep.net/record/1286695/plots
AN EXAMPLE OF A GRAPH OF COMMONSENSE KNOWLEDGE

The example illustrates a small segment of the ConceptNet knowledge base.

Image source: http://www.opasquet.fr/omcsnet/
AN EXAMPLE OF A GRAPH WITH 2 KINDS OF KNOWLEDGE:

1) META-KNOWLEDGE (CLASSES/CONCEPTS)
2) DOMAIN SPECIFIC KNOWLEDGE (DOMAIN ENTITIES)

DATA AND KNOWLEDGE GRAPHS IN THE BUSINESS WORLD
“…Google users will able to browse through the company’s ‘knowledge graph,’ or its ever-expanding database of information about ‘entities’ – people, places and things – the ‘attributes’ of those entities and how different entities are connected to one another.”

What Google’s Search Changes Might Mean for You
Wall Street Journal, March 14, 2012
“...Every piece of information that we crawl, index, or search is analyzed in the context of Knowledge Graph.”

“...Almost all the structured data from all of our products like Maps and Finance and Movies and Music are all in the Knowledge Graph, so we can reasonably say that everything we know about is in this canonical form.”
Google’s Knowledge Graph

“[Google Now] works by using machine learning algorithms to determine what you’re doing, then matches this understanding with information stored in what the company calls the Google Knowledge Graph—a database of semantic data describing more than 1 billion people, places, and things. ‘To be able assist you,’ says Aparna Chennapragada, who oversees Google Now, ‘we have to understand the world.’”

Startup Unleashes Its Clone of Google’s 'Knowledge Graph'
Wired, April 6, 2014
Facebook’s Entity Graph

“Facebook is building a rich stock of knowledge that could make its software smarter and boost the usefulness of its search engine…

…Entities such as colleges and employers are learned from data typed in profile pages; businesses, movies, fictional characters, and other concepts are learned from fan pages created by Facebook users. … analyzing many employment histories on the site allows Facebook’s search engine to know that a search for “software engineers” should also return people who say they are “coders.”
Microsoft's Bing seeks enlightenment with Satori
CNET News, July 30, 2013

"At the core of Microsoft's work to create a state-of-the-art Bing digital assistant is Satori, a knowledge repository of more than a billion objects digested in the past 3.5 years...

...Satori catalogs entities and the associated data and relationships among them...

...Satori is a self-learning system that is running every day and learning more, adding 28,000 DVDs of content every day...

...Bing search and Windows already are using Satori's knowledge repository..."
Bing’s Knowledge and Action Graph

“Bing has over a billion entities (people, places, and things) and the number is growing every day. For those entities, we have over 21 billion associated facts, 18 billion links to key actions and over 5 billion relationships between entities.

Millions of Bing users around the globe use this rich information every day, in bing.com, Cortana, Xbox, Office and more”

“knowledge and action graph will be available to developers via a new API”

Bing announces availability of the knowledge and action graph API
Bing Blogs, 20 August 2015
“Spark [a semantic search assistance tool] takes a large **entity graph** as input, … consisting of the most important entities, their most important related entities, and their respective types. This entity graph is drawn from a larger Yahoo! Knowledge Graph, a unified **knowledge base** that provides **key information** about all the **entities** we care about, and **how they relate** to each other.”
“[Economic graph is] a digital mapping of the global economy, comprised of a profile for every professional, company, job opportunity, the skills required to obtain those opportunities, every higher education organization, and all the professionally relevant knowledge associated with each of these entities”

“With these elements in place, we can connect talent with opportunity at massive scale”

Announcing The LinkedIn Economic Graph Challenge
Oct 14, 2014
Open data and knowledge graphs
Machine interpretable version of Wikipedia

Data collected from Wikipedia are:

- **Structured**: represented in the form `{subject-predicate-object}`
  triplets suitable for further processing

- **Semantically annotated**: semantics of each triplet element is explicitly defined => it can be directly interpreted by a computer program
Data and knowledge representation in DBpedia

Image source: http://www.accessola2.com/olita/insideolita/wordpress/?p=60281
http://en.wikipedia.org/wiki/San_Francisco

<table>
<thead>
<tr>
<th>Country</th>
<th>State</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>United States</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Founded</th>
<th>Incorporated</th>
<th>Founded by</th>
</tr>
</thead>
<tbody>
<tr>
<td>June 29, 1776</td>
<td>April 15, 1850[9]</td>
<td>José Joaquin Moraga Francisco Palóu St. Francis of Assisi</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Government</th>
</tr>
</thead>
<tbody>
<tr>
<td>Type</td>
</tr>
<tr>
<td>Body</td>
</tr>
<tr>
<td>Mayor of San Francisco</td>
</tr>
<tr>
<td>Board of Supervisors</td>
</tr>
<tr>
<td>Supervisors</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Supervisors</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tom Ammiano (D)</td>
</tr>
<tr>
<td>Phil Ting (D)</td>
</tr>
<tr>
<td>Mark Leno (D)</td>
</tr>
<tr>
<td>Nancy Pelosi (D)</td>
</tr>
<tr>
<td>Barbara Lee (D)</td>
</tr>
<tr>
<td>Jackie Speier (D)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Area</th>
</tr>
</thead>
<tbody>
<tr>
<td>231.89 sq mi (600.6 km²)</td>
</tr>
<tr>
<td>46.87 sq mi (121.4 km²)</td>
</tr>
</tbody>
</table>

```xml
<http://dbpedia.org/resource/San_Francisco>
  db:country dbpedia:United_States ;
  ...
  db:foundingDate "1776-6-29"^^xsd:date ;
  dbpprop:namedFor
    dbpedia:Francis_of_Assisi ;
  db:governmentType
    dbpedia:Mayor–council_government ;
  ...
```
"Wikipedia’s data is buried in 30 million Wikipedia articles in 287 languages from which extraction is inherently very difficult."

“Population numbers for Rome, for example, can be found in English and Italian articles about Rome but also in the English article “Cities in Italy.” The numbers are all different.”

Main objectives of the WikiData project:

- Turn Wikipedia data into a machine interpretable format, suitable for direct processing
- Sustain the accuracy and ‘freshness’ of the data

“Wikidata is a project of the Wikimedia Foundation: a free, collaborative, multilingual, secondary database, collecting structured data to provide support for Wikipedia, Wikimedia Commons, the other Wikimedia projects, and well beyond that.”

“A secondary database: Wikidata can record not just statements, but also their sources, thus reflecting the diversity of knowledge available and supporting the notion of verifiability.”

“Collecting structured data: [to] allow easy reuse of that data by Wikimedia projects and third parties, and enable computers to easily process and “understand” it.”

Wikidata:Introduction
August 2015
Data and knowledge representation in WikiData knowledge base
Data and knowledge representation in WikiData knowledge base

"Linked Data - San Francisco" by Jeblad - https://commons.wikimedia.org/wiki/File:Linked_Data_-_San_Francisco.svg
GIGANTIC GLOBAL GRAPH
GIGANTIC GLOBAL GRAPH (1):
VISION OF THE WEB AS A GIGANTIC GLOBAL REPOSITORY (GRAPH) OF DATA AND KNOWLEDGE
BY SIR TIM BERNERS-LEE
Phase 1: *International Information Infrastructure (III)*
- network/graph of computers known as *Internet* or *Net*
- "*It isn't the cables, it is the computers which are interesting*"

Phase 2: *World Wide Web (WWW)*
- network/graph of documents known as *Web*
- "*It isn't the computers, but the documents which are interesting*"

Phase 3: *Gigantic Global Graph (GGG)*
- network/graph of entities (resources) and data that describe the entities
- "*It's not the documents, it is the things they are about which are important*"

“…when I book a flight it is the flight that interests me. Not the flight page on the travel site, or the flight page on the airline site, but the URI (issued by the airlines) of the flight itself. … And whichever device I use … it will access a situation-appropriate view of an integration of everything I know about that flight from different sources. The task of booking and taking the flight will … be primary things in my awareness, the websites involved will be secondary things, and the network and the devices tertiary.”
WWW (= Web of documents)

GGG (= Web of data)

Alice

is a friend of

BOB

is interested in

The Mona Lisa

is about

La Joconde à Washington

was created by

Leonardo Da Vinci

is born on

14 July 1990

Person

WWW (Web of documents)
# Web of Docs vs. Web of Data

<table>
<thead>
<tr>
<th></th>
<th>Web of Documents</th>
<th>Web of Data</th>
</tr>
</thead>
<tbody>
<tr>
<td>Designed for</td>
<td>Human consumption</td>
<td>Humans served by computer programs</td>
</tr>
<tr>
<td>Primary objects</td>
<td>Documents (including multimedia)</td>
<td>Resources* (and descriptions of resources)</td>
</tr>
<tr>
<td>Links between</td>
<td>Documents</td>
<td>Resources</td>
</tr>
<tr>
<td>Degree of structure</td>
<td>Fairly low</td>
<td>High</td>
</tr>
<tr>
<td>Semantics of content and links</td>
<td>Implicit</td>
<td>Explicit</td>
</tr>
<tr>
<td>Analogy</td>
<td>A global file system</td>
<td>A global database</td>
</tr>
</tbody>
</table>

* Resource is anything that can be uniquely identified (has its URI); e.g., a resource is Belgrade, and its description is [DBpedia entry for Belgrade](https://dbpedia.org/blank).
WEB OF (OPEN) DATA

Source: http://lod-cloud.net/

Animation illustrating the evolution of LOD-a: http://goo.gl/49p9Eh
Gigantic Global Graph (2)

Gigantic graph-based knowledge bases that

• contain structured data extracted from Web pages
• continuously grow and evolve so that their content properly reflects the data and knowledge of the Web

Features:

• based on automated learning systems
• combine different Machine Learning methods to assure continuous improvement of the data/knowledge extraction process
• subject of extensive research (both in industry and academia) aimed at the improvement of the knowledge extraction process
READ THE WEB

Research project at the Carnegie Mellon University

- [http://rtw.ml.cmu.edu/rtw/](http://rtw.ml.cmu.edu/rtw/)

Objectives:

- Develop a never-ending machine learning system for extracting structured information from unstructured Web pages
- The development of the world largest structured KB that
  - reflects the factual content of the Web,
  - continually grows in terms of both predicates and instances,
  - could be useful to many AI efforts
NEVER ENDING LANGUAGE LEARNER (NELL)

NELL is an implementation of the Read the Web approach. It performs 2 tasks each day, 7 days per week:

- **Reading task**: extract new instances of categories and relations from texts on the Web, and thus extend the KB.
- **Learning task**: learn to ‘read’ better each day, as evidenced by the ability to extract more information more accurately.
  - the learning components continuously retrain themselves using the growing KB as a set of training examples.
One can follow NELL while it ‘reads’, and help it *learn* to ‘read’ better.

**Recently–Learned Facts**

<table>
<thead>
<tr>
<th>instance</th>
<th>iteration</th>
<th>date learned</th>
<th>confidence</th>
</tr>
</thead>
<tbody>
<tr>
<td><em>estonians</em> is an <em>ethnic group</em></td>
<td>956</td>
<td>23–oct–2015</td>
<td>97.9</td>
</tr>
<tr>
<td><em>jane_ krakowski</em> is an <em>architect</em></td>
<td>955</td>
<td>20–oct–2015</td>
<td>100.0</td>
</tr>
<tr>
<td><em>cups spoons</em> is an <em>item found in the kitchen</em></td>
<td>956</td>
<td>23–oct–2015</td>
<td>99.6</td>
</tr>
<tr>
<td><em>voight park</em> is a <em>zoo</em></td>
<td>959</td>
<td>07–nov–2015</td>
<td>90.5</td>
</tr>
<tr>
<td><em>michael mayer</em> is <em>american</em></td>
<td>955</td>
<td>20–oct–2015</td>
<td>100.0</td>
</tr>
<tr>
<td><em>bob</em> is a U.S. politician who <em>holds the office of president</em></td>
<td>959</td>
<td>07–nov–2015</td>
<td>99.2</td>
</tr>
<tr>
<td><em>karl rove</em> works for <em>fox</em></td>
<td>956</td>
<td>23–oct–2015</td>
<td>93.8</td>
</tr>
<tr>
<td><em>david toseland</em> died in <em>the country england</em></td>
<td>960</td>
<td>23–nov–2015</td>
<td>100.0</td>
</tr>
<tr>
<td><em>cavaliers</em> is a sports team that <em>won</em> the <em>nba finals</em></td>
<td>955</td>
<td>20–oct–2015</td>
<td>98.4</td>
</tr>
<tr>
<td><em>logan was born in chicago south</em></td>
<td>960</td>
<td>23–nov–2015</td>
<td>99.6</td>
</tr>
</tbody>
</table>

Source: web site of the Read the Web project: [http://rtw.ml.cmu.edu/rtw/](http://rtw.ml.cmu.edu/rtw/)
Google’s Knowledge Vault (KV)

Envisioned as a probabilistic knowledge base which would contain all the factual knowledge of the Web, and would grow and evolve as the Web grows and evolves.

Knowledge representation:
- relies on \{subject-predicate-object\} triplets (like DBpedia),
- each triplet has its confidence score, which represents the estimated probability of the triplet’s validity / accuracy

The stored knowledge includes:
- facts automatically extracted from Web pages (uncertain, unverified knowledge)
- knowledge gathered from existing knowledge bases (verified, validated knowledge)
**GOOGLE’S KNOWLEDGE VAULT (KV)**

Comparison of KV with other state-of-the-art knowledge bases

<table>
<thead>
<tr>
<th>Name</th>
<th># Entity types</th>
<th># Entity instances</th>
<th># Relation types</th>
<th># Confident facts (relation instances)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Knowledge Vault (KV)</td>
<td>1100</td>
<td>45M</td>
<td>4469</td>
<td>271M</td>
</tr>
<tr>
<td>DeepDive [32]</td>
<td>4</td>
<td>2.7M</td>
<td>34</td>
<td>7M*</td>
</tr>
<tr>
<td>NELL [8]</td>
<td>271</td>
<td>5.19M</td>
<td>306</td>
<td>0.435M*</td>
</tr>
<tr>
<td>PROSPERA [30]</td>
<td>11</td>
<td>N/A</td>
<td>14</td>
<td>0.1M</td>
</tr>
<tr>
<td>YAGO2 [19]</td>
<td>350,000</td>
<td>9.8M</td>
<td>100</td>
<td>4M*</td>
</tr>
<tr>
<td>Freebase [4]</td>
<td>1,500</td>
<td>40M</td>
<td>35,000</td>
<td>637M*</td>
</tr>
<tr>
<td>Knowledge Graph (KG)</td>
<td>1,500</td>
<td>570M</td>
<td>35,000</td>
<td>18,000M*</td>
</tr>
</tbody>
</table>

Table 1: Comparison of knowledge bases. KV, DeepDive, NELL, and PROSPERA rely solely on extraction, Freebase and KG rely on human curation and structured sources, and YAGO2 uses both strategies. Confident facts means with a probability of being true at or above 0.9.

NELL (Never Ending Language Learner) is a software system that has been developed in the previously introduced Read the Web project

DiffBot’s Global Index

“...in recent months Diffbot has been analyzing websites to build its index at a rate of up to 15 million pages a day.

Its Global Index now contains more than 600 million objects (this can be anything from a celebrity to an Ikea chair model) and 19 billion facts.

‘Our approach is fairly radical in that there’s no human behind the curtain’

Diffbot ... [is] enhancing other search engines including Microsoft’s Bing and DuckDuckGo, and powering apps for companies such as Cisco and AOL”

Diffbot Challenges Google Supremacy With Rival Knowledge Graph
Xconomy, June 4, 2015
RECOMMENDED VIDEOS, ARTICLES

- [article] Diffbot Bests Google's Knowledge Graph To Feed The Need For Structured Data (link)

- [video] Mike Tung, DiffBot CEO, Turning the Web into a Structured Database (link)

- [video] Knowledge Vault: A Web-Scale Approach to Probabilistic Knowledge Fusion (link)
  - this lecture explains the overall idea and concept of the KV, the principles its knowledge collection process is based upon, and the like

- [video] Tom Mitchell, Never-Ending Learning to Read the Web (link)

- [video] From Structured Data to Knowledge Graph, Google I/O 2013 (link)