Deep Text Analytics
2019 NKOS Workshop

How to extract hidden information and ‘aboutness’ from text using SKOS, ontologies, corpus analysis and linked data

Andreas Koller
Co-founder and CIO
Semantic Web Company
Introduction

Andreas Koller

active at

Developer & CIO of

W3C Software Ltd

Managing director of

London

2004

located

serves customers

>200

headquartered

Vienna

founded

version

7.0

developer & vendor of

based on

used for

Text Mining

Ontologies

standard for

Knowledge Graphs

part of

Semantics Web

standard for

Ontologies

part of

Taxonomies

used for

integrates with

PoolParty

manages

Search engine

CMS/DMS/DAM/.. version

Gartner

PKMWorld

Graph database
At a glance

PoolParty Semantic Suite

- PoolParty is the most complete and most secure semantic middleware on the global market.

- PoolParty integrates with leading graph database technologies, NoSQL stores, and search engines.

- PoolParty was launched in 2009.

Semantic Web Company

- SWC is developer, licensor and vendor of PoolParty Semantic Suite.

- Established in 2004, SWC has pioneered the areas of Enterprise Semantics & Semantic AI.

- SWC is dedicated to use W3C standards like SKOS, OWL, SPARQL, etc.
Why Knowledge Graphs?

WHAT’S THE PROBLEM?

Many Challenges in (Enterprise) Information Management & AI
“Search” is still about documents only

…and in enterprises it’s painful

wind farms OR wind parks OR wind power plants OR wind power station
Missing Context

Market Selector | Jaguar | View the site in your preferred

https://www.jaguar.com

Discover the different language sites we have to make browsing our vehicle range's easier. We have over 100 different language options available. Learn more

Luxury saloons, sports cars & performance SUV | Jaguar Cars

https://www.jaguar.com/index.html

The official website of Jaguar. Discover the Art of Performance with our range of vehicles from the XF, XJ, XE, F-Type and, our luxury SUV, the F-PACE.

Jaguar - Wikipedia


The jaguar (Panthera onca), is a wild cat species and the only extant member of the genus Panthera native to the Americas. The jaguar's present range ...

Etymology · Evolution · Taxonomy · Characteristics · Distribution and habitat

Images of jaguar

bing.com/images

See more images of jaguar

Jaguar Model Prices, Photos, News, Reviews and Videos

www.autoblog.com › Automobiles

Research Jaguar car and SUV pricing and get news, reviews, specs, photos, videos and more - everything for Jaguar owners, buyers and enthusiasts.
Context-free data models and User-agnostic

Do machines understand user intent? Do they have enough context?

1. Intent recognition
2. Entity linking
3. Background knowledge

Which Sport-utility vehicle from France provides enough space for my family with 3 kids?

<table>
<thead>
<tr>
<th>Car</th>
<th>Loadspace</th>
<th>Max no. of seats</th>
</tr>
</thead>
<tbody>
<tr>
<td>KIA Sorento</td>
<td>550 litres</td>
<td>7</td>
</tr>
<tr>
<td>Peugeot 5008</td>
<td>823 litres</td>
<td>7</td>
</tr>
<tr>
<td>BMW X3</td>
<td>550 litres</td>
<td>5</td>
</tr>
</tbody>
</table>
Perth is one of the most isolated major cities in the world, with a population of 2,022,044 living in Greater Perth.

Australia is a member of the OECD, United Nations, G20, ANZUS, and the World Trade Organisation.
How does it work?

KNOWLEDGE GRAPHS

Bring your metadata to the next level!
Google Knowledge Graph

Jaguar Cars
Luxury vehicles company

Jaguar is the luxury vehicle brand of Jaguar Land Rover, a British multinational car manufacturer with its headquarters in Whitley, Coventry, England. Wikipedia

Headquarters: Coventry, United Kingdom
Founded: September 4, 1922, Blackpool, United Kingdom
CEO: Ralf Speth (Feb 18, 2010–)
Tagline: "The Art of Performance"

Parent organizations: Tata Motors, British Leyland, British Motor Holdings, Jaguar Land Rover Holdings Limited
Founders: William Lyons, William Walmsley

Profiles

Instagram  LinkedIn  YouTube  Facebook  Twitter

People also search for

Tata Motors  Aston Martin  Mercedes-Benz  Bentley Motors Limited  Land Rover
Example Knowledge Graph

- URIs
- Concepts
- Labels
- Relations

Demo
‘Things’ but no strings:
Using a ‘Semantic Knowledge Graph’

Retina

The retina is the light-sensitive layer of tissue at the back of the eyeball. Images that come through the eye’s lens are focused on the retina. The retina then converts these images to electric signals and sends them along the optic nerve to the brain.

The retina usually looks red or orange because there are many blood vessels right behind it. An ophthalmoscope allows a health care provider to see through your pupil and lens to the retina. Sometimes photos or special scans of the retina can show things that the provider cannot see just by looking at the retina through the ophthalmoscope. If other eye problems block the provider’s view of the retina, ultrasound can be used.

Anyone who experiences these vision problems should get a retinal examination:

- Changes in sharpness of vision
- Loss of color perception
- Flashes of light or floaters
- Distorted vision (straight lines look wavy)

Watch this video about: Retina

Funduscope

Ophthalmoscope

Diagnostic Equipment
Men’s SuperBoots are warm, waterproof slip-on Snow boots. They are perfect for quick trips out to grab the paper or all-day escapades in the snow. They’re built with all-over insulation and rich, durable, waterproof leather, with a seam-sealed construction for bonus protection from the wet and rugged traction to keep you stable on wet, slippery surfaces.
Traditional vs. Graph-based Metadata Management

Traditional approach vs. Graph-based approach

Show me all documents about *Carnivores*
Traditional vs. Graph-based Metadata Management

Traditional approach

Graph-based approach

Show me all documents about Carnivores

Jaguar
Elephant
Wolf
Rabbit

Carnivore

Jaguar
Elephant
Wolf
Rabbit
Traditional vs. Graph-based Metadata Management

Traditional approach vs. Graph-based approach

Show me all documents about **Carnivores**

- **Jaguar**, **Cat**, **Carnivore**
- **Elephant**
- **Wolf**, **Carnivore**
- **Rabbit**

Show me all documents about **Cats**

- **Jaguar**, **Elephant**, **Wolf**, **Rabbit**

© Semantic Web Company 2019
Traditional vs. Graph-based Metadata Management

Traditional approach

Show me all documents about *Felidae*

Show me all documents about *Carnivores*

Graph-based approach

Jaguar
Cat
Felidae
Carnivore

Elephant
Carnivore

Wolf
Carnivore

Rabbit

Cat = Felidae

Jaguar

Elephant

Wolf

Rabbit

Traditional vs. Graph-based Metadata Management
Traditional vs. Graph-based Metadata Management

Traditional approach

Graph-based approach

Show me all documents about *Felidae*

Show me all documents about *Carnivores*

Keystone Species?

Jaguar
Cat
Felidae
Carnivore
Keystone Species

Elephant
Keystone Species

Wolf
Carnivore

Rabbit

Carnivore

Cat = Felidae

Keystone Species

Jaguar

Elephant

Wolf

Rabbit
Traditional vs. Graph-based Metadata Management

Traditional approach

Metadata per document

1. No or little network effects
2. No reuse of metadata
3. Metadata resides in silos
4. Data quality hard to measure
5. Not machine-readable

Graph-based approach

Knowledge about metadata

1. Explicit knowledge models
2. Reusable and measurable
3. Metadata is machine-processable
4. Standards-based metadata
5. Linkable metadata opens silos
Knowledge Graphs as a result of human-machine cooperation

Manually created parts of graph

Supervised learning

Automatically created parts of graph (corpus analysis, RDF transformation, machine learning, ....)
How does it work?

CORPUS ANALYSIS

Semantic Information Retrieval
Bionics
How do we learn from a lot of text?

- **Taxonomical model → is-a abstractions**
  - Switched on devices are dangerous devices.
  - The stove is on. The stove is hot!

- **Ontological model → reasoning**
  - Switched on devices are dangerous, only if the operating temperature is above 100 degrees and the automatic shutdown mechanism is broken.

- **Bionics**
  - The stove is on. The stove is hot!
  - Bla bla bla bla.

- **Statistical model/cooccurences → is related**
  - Bla stove bla bla.
  - Bla bla bla hot.
  - The stove is on. The stove is hot!
# Extraction of the non-obvious

<table>
<thead>
<tr>
<th>Mechanism</th>
<th>How does it work?</th>
</tr>
</thead>
<tbody>
<tr>
<td>Entity extraction based on concepts instead of simple term-based extraction</td>
<td>Make use of synonyms in a taxonomy; disambiguation when necessary</td>
</tr>
<tr>
<td>Extraction of super classes and broader concepts</td>
<td>Make use of hierarchical structures in a knowledge graph</td>
</tr>
<tr>
<td>Extraction of related terms and concepts</td>
<td>Co-occurrence models and word embeddings</td>
</tr>
<tr>
<td>Extraction of ‘Shadow Concepts’</td>
<td>Combining co-occurrence models and knowledge graphs</td>
</tr>
<tr>
<td>Semantic Classifier</td>
<td>Enrichment of training documents with metadata from a knowledge graph</td>
</tr>
<tr>
<td>Deep Text Analytics</td>
<td>Derivation of new metadata and classifications based on mechanisms as described above combined by rules</td>
</tr>
</tbody>
</table>
Entirety extraction based on concepts instead of simple term-based extraction


Austria’s capital, lies in the country’s east on the Danube River. Its artistic and intellectual legacy was shaped by residents including Mozart, Beethoven and Sigmund Freud. The city is also known for its Imperial palaces, including Schönbrunn, the Habsburgs’ summer residence. In the MuseumsQuartier district, historic and contemporary buildings display works by Egon Schiele, Gustav Klimt and other artists.
Extraction of super classes and broader concepts
Make use of hierarchical structures in a knowledge graph

BMW has designed a car that is going to drive Jaguar X1 out of the market

Word Sense Induction
From Word to Sense Embeddings

Unsupervised sense representation (Word sense induction)

Knowledge-based sense representation

The main problem: Meaning conflation deficiency (Word sense ambiguity)
Extraction of related terms and concepts

- Document Corpus
  - Websites
  - PDF, Word, …
  - Abstracts from DBpedia
  - RSS Feeds

- Relevant terms and phrases
- Relevancy of terms
- co-occurrence between terms and terms
Extraction of ‘Shadow Concepts’

Use co-occurrences between concepts and terms to extract ‘shadow concepts’

Example:

This site is a 15th-century Inca site located 2,430 metres above sea level. It is located in Cusco, Peru. It is situated on a mountain ridge above the Sacred Valley through which the Urubamba River flows. Most archaeologists believe that it was built as an estate for the Inca emperor Pachacuti. Often mistakenly referred to as the "Lost City of the Incas", it is the most familiar icon of Inca civilization. The Incas built the estate around 1450, but abandoned it a century later at the time of the Spanish Conquest.

In addition to explicitly used concepts and terms, Machu Picchu is extracted from the article as a Shadow Concept. As a prerequisite, one has to provide and analyze a representative text corpus first.
Semantic Classifier

ML based on semantically enriched training data

**PoolParty Semantic Classifier** combines machine learning algorithms (SVM, Deep Learning, Naive Bayes, etc.) with Semantic Knowledge Graphs.
Bain Capital is a venture capital company based in Boston, MA. Since inception it has invested in hundreds of companies. In 2018, Bain had $75b AUM.

Give me all paragraphs in documents about “US based Private Equity firms with AUM higher than $20B”
Benefit from Knowledge Graphs and Corpus Analysis

USE CASES

See how it can be used in your environment!
Some domains use text that doesn’t always call a spade a spade. With ‘shadow concept extraction’ those ‘masked’ concepts still can be surfaced.

Since these technologies would have become conventional technologies that are made into products and introduced into market at the time of their introduction, it would be difficult to differentiate them as innovative environmental and energy technologies from other global warming prevention technologies that have already been put to practical use in the industrial, commercial, residential, and energy conversion sectors.

- The Innovative Global Warming Prevention Technology Working Group under the Research and Development Subcommittee - Council assessed that innovative global warming prevention technologies would bring about a reduction effect of 7.49 million t-CO₂ case of average emissions factor for all power sources of carbon dioxide in 2010. In view of the difficulty in putting innovative carbon dioxide sequestration technology into practical use by 2010, the Working Group reassigned it as an issue of global warming prevention technology to be tackled by 2030.

The Central Environment Council, however, has not had the opportunity to examine the contents of these technologies in detail. (Promotion of climate change prevention activities by every social actor)

- The Programme encourages every social actor to take actions to prevent global warming. The actions include measures undertaken by the public sector.

Since these technologies would have become conventional technologies that are made into products and introduced into market at the time of their introduction, it would be difficult to differentiate them as innovative environmental and energy technologies from other global warming prevention technologies that have already been put to practical use in the industrial, commercial, residential, and energy conversion sectors.

- The Innovative Global Warming Prevention Technology Working Group under the Research and Development Subcommittee - Council assessed that innovative global warming prevention technologies would bring about a reduction effect of 7.49 million t-CO₂ case of average emissions factor for all power sources of carbon dioxide in 2010. In view of the difficulty in putting innovative carbon dioxide sequestration technology into practical use by 2010, the Working Group reassigned it as an issue of global warming prevention technology to be tackled by 2030.

The Central Environment Council, however, has not had the opportunity to examine the contents of these technologies in detail. (Promotion of climate change prevention activities by every social actor)

- The Programme encourages every social actor to take actions to prevent global warming. The actions include measures undertaken by the public sector.
Mini Countryman

And it’s probably more of a crossover than ever, with the design to match. Being a Mini, the Countryman is clearly meant to be the driver’s car among small crossovers. The suspension is sophisticated, and there are lots of chassis options (a stiffer sports setup, variable damping, the electronically controlled ALL4 all-wheel-drive).

But it’s also the crossover for people who’ve bags of cash to blow on personalisation and luxury.

There’s been a lot of effort on ramping up the cabin quality, but the outgoing Countryman was a sad let-down in that department.

On the outside, plastic wheel-arch extensions, as well as roof bars and sill protectors all add to the Mini with angular rather than oval headlamps on in the lower face.

There are eight versions at launch, and they’re Cooper or Cooper S, each fuelled by petrol or diesel, each of them auto, too, if you count that as a separate choice. The Cooper petrol is a three-cylinder, the rest’s fours.

You get extra kit as standard versus the old car, and park sensors. Upgrades include a bigger touch-screen nav with high-definition traffic, various posher seats, a HUD, and driver aids. Oh and a cushion thingy that folds out from the boot so you can sit on the rear bumper without getting your clothes mucky.

In June 2017 a Cooper E will launch, which has the Cooper three-cylinder petrol driving the front wheels, and an electric motor for the rears, with a capacity to do a claimed 25 miles of gentle all-electric running. So it has the performance of a Cooper S ALL4 with the tax-busting advantages of a plug-in hybrid. And you wouldn’t use any fuel if you commuted a short distance.

The platform is BMW’s contemporary transverse-engined hardware, in the bigger of its two sizes. That means it shares a lot with the BMW X1. The 4WD system is more sophisticated than the previous Countryman’s. The proportion of drive to the rear is computed by a controller that takes into account parameters including grip, steering angle and throttle position, as well as whether you’ve got the sports mode and sports traction systems selected.
Use Case Recommender System

Connecting

- content to content
- people to content
- people to people

I am interested in ...

Languedoc-Roussillon

Languedoc is a significant producer of wine, and a major contributor to the surplus known as the "wine lake". Today it produces more than a third of the grapes in France, and is a focus for outside investors.

The region contains the historic cities of Carcassonne, Toulouse, Montpellier, countless Roman monuments, medieval abbeys, Romanesque churches, and old castles.

Occitanie, vineyard, sea, church, village

© Semantic Web Company 2019
Example EIP on Water

At the center of the marketplace is the matchmaking function.

https://www.eip-water.eu/
Semantic analysis and pre-selection of relevant clauses in contracts
Connect

Andreas Koller
CIO & Managing Partner, Semantic Web Company

- andreas.koller@semantic-web.com
- https://www.linkedin.com/in/ankoller/
- https://twitter.com/semwebcompany
Q & A