SocialLink

Linking DBpedia Entities to Corresponding Twitter Accounts

Yaroslav Nechaev (speaker), Francesco Corcoglioniti, Claudio Giuliano
Linking Knowledge Bases to Social Media Profiles

- The goal is to bridge Linked Open Data cloud and social media

**dbr:SpaceX**

**Properties:**
- foaf:name: SpaceX
- dbo:locationCity: Hawthorne, CA, USA
- rdfs:label: SpaceX
- foaf:homepage: http://www.spacex.com
- dbo:foundedBy: Elon Musk
- dbp:industry: Aerospace Engineering
- dbp:type: Private company
Linking Knowledge Bases to Social Media Profiles

• Trying to get the best of both worlds:
  • Knowledge Bases provide high-quality, structured, easily accessible knowledge
    • Cover wide range of entity types
    • Can contain obsolete knowledge
  • Social media is a vibrant source of up-to-date knowledge with an unparalleled coverage
    • More than 2B users in Facebook alone
    • Mainly covers living people, organisations and brands
    • Mostly unstructured
Use Cases

User profiling, entity linking, event detection, community detection and labelling

Enrichment, validation, referencing (providing sources to claims in KB)
SocialLink

• Aligns people and organisations from 120 DBpedia language chapters to Twitter

• Considers 2.5 million entities

• Proposes candidates for 906k entities and aligns 271k entities

• The code and the resource are publicly accessible and updated
  • http://w3id.org/sociallink/
  • https://github.com/Remper/sociallink

• Released in RDF, JSON and CSV, accessible through SPARQL endpoint
SocialLink Pipeline

**Data Acquisition**
- **KB:** DBpedia
  - KB merging
  - KB Entity index

- **INDEX:** Stream API
  - Indexing
  - Index User index

**Candidate Acquisition**
- Target entity:
  - dbr:Barack_Obama
- Search query:
  - barack obama
  - барак обама

**Candidate Selection**
- List of candidates:
  - @BarackObama
  - @ObamaNews
  - @ThePresObama
  - @BarackObamaRF

- Neural Network
  - Training
  - Inference
  - Scored Candidates:
    - @BarackObama: 0.80
    - @ObamaNews: 0.59
    - @ThePresObama: 0.54
    - @BarackObamaRF: 0.19
SocialLink Pipeline

- RDFpro\textsuperscript{[1]} tool is used to download 120 DBpedia chapters and merge them along \texttt{owl:sameAs} links

- Resulting Entity Index contains 1.4B triples

- Considers living people (2M) and currently existing organisations (550k)

- 58,745 gold standard alignments were extracted from \texttt{foaf:isPrimaryTopicOf} and \texttt{wikidata:P2002} properties

SocialLink Pipeline

Data Acquisition

- 3TB of Twitter data is processed and indexed using Apache Flink and stored in Postgres
- Produced User Index contains 450GB worth of indexed data
- Covers more than 240M Twitter users

SocialLink Pipeline

- Produces a short list of candidate profiles for every entity in the Entity Index
  - 240M → ~300
- Name-based queries to User Index
  - Similar to our previous Twitter-based approach\[2\]
- Multiple queries for better recall

SocialLink Pipeline

Candidate Selection

Candidates
@BarackObama
@ObamaNews
@ThePresObama
@BarackObamaRF

Entity index

Training
Inference

Neural Network

Scored Candidates
@BarackObama 0.80
@ObamaNews 0.59
@ThePresObama 0.54
@BarackObamaRF 0.19

SocialLink Pipeline

Candidate Selection

Entity index

Training

Inference

Neural Network

Candidates
@BarackObama
@ObamaNews
@ThePresObama
@BarackObamaRF

Scored Candidates
@BarackObama 0.80
@ObamaNews 0.59
@ThePresObama 0.54
@BarackObamaRF 0.19

SocialLink Pipeline

Candidate Selection

Candidates
@BarackObama
@ObamaNews
@ThePresObama
@BarackObamaRF

Entity index

Training

Inference

Neural Network

Scored Candidates
@BarackObama 0.80
@ObamaNews 0.59
@ThePresObama 0.54
@BarackObamaRF 0.19

[dbr:Barack_Obama](#)

Textual
Popularity
Entity

match / not match classifier

0.80 confidence score

SocialLink Pipeline

Candidate Selection

- Candidates are ranked based on the confidence score
- Alignment can be chosen aiming at the desired precision/recall balance
- Default thresholds exhibit 90% precision / 41% recall evaluated on the gold standard dataset

SocialLink Pipeline

Data Acquisition
- KB: DBpedia
  - Entity index
- Stream API
  - User index
  - Search query: barack obama, барак обама

Candidate Acquisition
- Target entity: dbr:Barack_Obama
- List of candidates:
  - @BarackObama
  - @ObamaNews
  - @ThePresObama
  - @BarackObamaRF

Candidate Selection
- Neural Network
  - Inference
  - Scored Candidates:
    - @BarackObama: 0.80
    - @ObamaNews: 0.59
    - @ThePresObama: 0.54
    - @BarackObamaRF: 0.19

SocialLink website
- RDF, CSV, JSON
- SPARQL endpoint

GitHub
RDF Representation

SELECT ?account {?e owl:sameAs <E>; foaf:onlineAccount ?account}
SocialLink Applications

- User profiling
- Entity linking
- Knowledge base enrichment
SocialLink for User Profiling

• Simplifies profiling pipelines.
  • Inference of user interests on Re-coding Black Mirror workshop\(^3\)
  • Usage of linking was explored in other papers\(^4\)[5]\(^5\)

• Enables an automatic creation of gold standard datasets for the large spectrum of attributes (instead of manual annotation)

• Paper in the works

\(^3\) Nechaev, Y., Corcoglioniti, F., Giuliano, C.: Concealing Interests of Passive Users in Social Media. In: Re-coding Black Mirror workshop @ ISWC2017
SocialLink for Entity Linking

• Enables named entity linking to social media profiles

• Implemented as part of the Social Media Toolkit[6]

• Directly disambiguates profile mentions in tweets against DBpedia


This week, @esa & @NASA’s SOHO satellite will see comet 96P/Machholz swing by the Sun for the 5th time! ☀☄

SocialLink for Knowledge Base Enrichment

• Direct import of up-to-date data to the knowledge base

• Validation and referencing of the existing data

• Proposed to Wikimedia community as a project – soweego[9]

Grants: Project/Hjfocs/soweego

Project idea [edit]

soweego (solid catalogs and weeekee go together) is a fully automatic robot that links existing Wikidata items about people to a set of reliable external catalogs.

Why: the problem [edit]

Data quality in a broad knowledge base (K3) like Wikidata is a vital aspect to secure confidence on its content, thus encouraging the development of effective ways to curate it. Wikidata is not meant to tell us the absolute truth: instead, we can see it as a container for different points of view (read claims) about every little fact of our world. These claims should always be verified against at least one reference to a trusted external source. However, this still does not seem to be the case. Despite specific efforts such as WikidataRef[1] and Streff[2] as well as more extensive community endeavors like Wikidata[3] the lack of references is a critical issue that remains open. Recently, the problem was further highlighted by the Wikidata product manager in her presentation at WikiCite 2017,[4] where the following aspects about the KB statements clearly emerged:

- roughly half of them is totally unreferenced;
- less than a quarter of them has references to non-wiki sources;
- most reference values are internal links to other Wikidata items.

How: the solution [edit]

Alignment of Wikidata to structured databases can represent a complementary alternative to reference mining from unstructured data.[5][6] Think of an identifier as a reference to a Wikidata item: we can give force to the trustworthiness of that item if we manage to match it with the corresponding entry of an authoritative
Future Work

• Extension to Facebook and Instagram
  • Consistency checking across social media to improve results overall
• Candidate selection algorithm improvements
  • Better feature space using user and entity embeddings
  • Learning to rank instead of classification
• Enable alignment to multiple accounts
Conclusions

• Introduced a resource for aligning DBpedia entities to Twitter profiles
• Applications for social media analysis and knowledge base enrichment
• Will be supported with new releases including:
  • Recent social media data
  • Pipeline and algorithm improvements
Thanks!

http://w3id.org/sociallink/

https://github.com/Remper/sociallink

Suggestions and/or pull requests are always welcome!