RdfLiveNews
—
Real-time RDF Extraction from Unstructured Data Streams

Daniel Gerber, Sebastian Hellmann, Lorenz Bühmann, Tommaso Soru, Ricardo Usbeck and Axel-Cyrille Ngonga Ngomo

University of Leipzig

ngonga@informatik.uni-leipzig.de

October 24th, 2013
60% of LD is static
Most data is encyclopedic
- LinkedTCGA ($7.6 \times 10^9$ T)
- LinkedGeoData ($10^9$ T)
- ...
Lack of actuality
Live tools gather data from semi-structured data
Motivation

1. 60% of LD is static
2. Most data is encyclopedic
   - LinkedTCGA \((7.6 \times 10^9 \ \text{T})\)
   - LinkedGeoData \((10^9 \ \text{T})\)
   - ...
3. Lack of actuality
4. Live tools gather data from semi-structured data

Goal

Provide actual data in RDF that allow answering questions such as Give me all news of the last week from the New York Times pertaining to the director of Nokia.
Requirements

- Timeliness needs to run 24/7
- Has to handle large amounts of data
- Needs to handle unstructured data
- Extraction has to be precise
- Open information extraction
Assumptions

1. Unstructured data source $S_i$ emits continuous data streams $D_i$ (e.g., RSS feeds)
2. Each data stream $D_i$ consists of atomic elements $d^i_j$
Assumptions
1. Unstructured data source $S_i$ emits continuous data streams $D_i$ (e.g., RSS feeds)
2. Each data stream $D_i$ consists of atomic elements $d^i_j$

Approach
1. Gather all elements of $D_i$ emitted between $t$ and $t + d$ in $D_i^{[t,t+d]}$
2. Apply sentence splitting and use sentences as atoms for processing
3. Perform string-similarity-based deduplication
Search & Filtering

- Apply Part Of Speech Tagging or Named Entity Recognition (Loc, Per, Org, Misc)

POS Tagging

... Smith/NNP ,/, is/VBZ the/DT manager/NN of/IN ABC/NNP.
Search & Filtering

- Apply Part Of Speech Tagging or Named Entity Recognition (Loc, Per, Org, Misc)

**POS Tagging**

... Smith/NNP ,/, is/VBZ the/DT manager/NN of/IN ABC/NNP.

**Example: Pattern**

\[ p = ([, is the manager of], (John Smith, ABC)) \]

- Patterns must
  - Contain a verb or noun, at least one non stop-word & shorter than 50 characters
  - Sort by number of elements in support set and take top n% 

- Where does “John” come from?
Pattern Refinement

- Refine patterns by disambiguating support set to
  - Complete entity names ("John Smith")
  - Find suitable $rdfs$-range & $rdfs$-domain

- Score each candidate URI based on 4 measures (local/global context, apriori & string similarity)

- Use normalized and weighted features to find best URI or create new URI

- Use $rdf$-types of URIs in support set to find $rdfs$-domain/range:
  MostGeneric vs. MostSpecific

Example: Refined Pattern

$p_r = ([, is the manager of], (John Smith <dbr:John_Smith>, ABC <dbr:ABC>), dbo:Person, dbo:Company)$
Idea

- Merge patterns to groups of patterns that express a certain relation
- Group of patterns stand for a predicate
## Idea

- Merge patterns to groups of patterns that express a certain relation
- Group of patterns stand for a predicate

- Baseline (string similarity) often fails, e.g., “attorney” vs. “lawyer”
Clustering & Labeling

Idea
- Merge patterns to groups of patterns that express a certain relation
- Group of patterns stand for a predicate

Baseline (string similarity) often fails, e.g., “attorney” vs. “lawyer”

Approach
1. Preprocess patterns by using lemmatization and stop word removal
2. Use linear combination of Wordnet and string similarity to generate similarity graph
3. Cluster using BorderFlow with hardening
4. Majority vote to determine cluster labels

Example: Clustering
\[ \text{sim}(\text{[\text{is(be) the manager of}]}, \text{[\text{was(be) the executive director of}]}) = 0.8 \]
RDF Generation

- Facts and properties are mapped to RDF
- Metadata: extraction/publication date and provenance links to text
- Mint URIs to provide linked data
- Link predicates to DBpedia via LIMES
Example

### cluster information

```
rlno:directorOf a owl:ObjectProperty ; skos:prefLabel "director of"
  , skos:altLabel " , director of " ; owl:equivalentProperty dbp:director .
```

### extracted facts:

```
rlnr:Rolf_Heuer a dbo:Person ; rdfs:label "Rolf Heuer"@en ; rlno:
  directorOf dbpedia:CERN .
dbpedia:CERN a owl:Thing ; rdfs:label "CERN"@en .
```

### provenance tracking with NIF:

```
<char=0,10> itsrdf:taClassRef dbo:Person ; itsrdf:taIdentRef rlnr:
  Rolf_Heuer .
<char=14,18> itsrdf:taIdentRef dbpedia:CERN .
<char=11,24> nif:anchorOf ", director of"^^xsd:string ; itsrdf:
  taPropRef rlno:directorOf .
```

### detailed NIF output with context, indices and anchorOf

```
<char=0,> a nif:String , nif:Context , nif:RFC5147String ; nif:
  isString "Rolf Heuer , director of CERN , said the newly
discovered particle is a boson …. — an extremely fine
distinction." ;
nif:sourceUrl <http://www.necn.com/07/04/12/Scientists—discover—new
```

Daniel Gerber, Sebastian Hellmann, Lorenz B

RdfLiveNews

October 24th, 2013 13 / 21
Experiments

1. How good is the RDFLiveNews URI disambiguation?
2. Can we find clusters of similar patterns?
3. Which quality does the generated RDF have?
4. Is near real-time RDF generation possible?
Experimental Setup

- Crawled 1457 RSS feeds for 76 hours
- 38 time slices of 2 hours and 11.7M sentences = 100%
- Average number of sentences/article is about 26.5
- Average number of articles/timeslice is about 3445
- Two subsets: 10% and 1%
URI Disambiguation

- Generated gold standard from 1%, 473 entity pairs
- Hill climbing with random initialization
  - $P_{max} = 0.85$ / $R = 0.407$ / $F = 0.55$
  - $F_{max} = 0.665$ / $P = 0.67$ / $R = 0.66$
- AIDA achieves $F = 0.6$
Pattern Clustering

- Manually clustered patterns from gold standard
- Sensitivity ($Se$, similar to recall)
- Positive Predictive Value ($PPV$, similar to precision)
- Accuracy = $\sqrt{Se \times PPV}$
Pattern Clustering

- Manually clustered patterns from gold standard
- Sensitivity ($Se$, similar to recall)
- Positive Predictive Value ($PPV$, similar to precision)
- Accuracy = $\sqrt{Se \times PPV}$
- Wordnet-String-Similarity: $S = 0.712$, $PPV = 0.955$, $A = 0.825$
- String-Similarity: $A = 0.691$
- Wordnet-Similarity: $A = 0.789$
Random selection of 100 triples

5 different evaluation sets, each set may only contain properties of clusters with at least $i = 1 \ldots 5$ patterns

<table>
<thead>
<tr>
<th>$E_i$</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
</tr>
</thead>
<tbody>
<tr>
<td>$S_{Acc}$</td>
<td>0.81</td>
<td>0.88</td>
<td>0.86</td>
<td>0.857</td>
<td>0.804</td>
</tr>
<tr>
<td>$P_{Acc}$</td>
<td>0.86</td>
<td>0.89</td>
<td>0.90</td>
<td>0.935</td>
<td>1.00</td>
</tr>
<tr>
<td>$O_{Acc}$</td>
<td>0.93</td>
<td>0.91</td>
<td>0.90</td>
<td>0.948</td>
<td>0.941</td>
</tr>
<tr>
<td>$Total_{Acc}$</td>
<td>0.86</td>
<td>0.892</td>
<td>0.885</td>
<td>0.911</td>
<td>0.906</td>
</tr>
<tr>
<td>$</td>
<td>E_i</td>
<td>$</td>
<td>100</td>
<td>100</td>
<td>100</td>
</tr>
<tr>
<td>$</td>
<td>P</td>
<td>\in</td>
<td>E_i</td>
<td>$</td>
<td>28</td>
</tr>
</tbody>
</table>
Scalability

Runtimes for different components and corpora (1% left, 10% middle, 100% right) per iteration.

- Average time slice processing: 20min (2.2min for 10%, 30s for 1%)
- Deduplication and merging largest bottlenecks
Presented RdfLiveNews, a framework for extraction of RDF from unstructured data streams

- URI disambiguation of 85%, cluster patterns with accuracy of 82.5% and achieve a total accuracy of about 90%
- Can handle 2h timeslices (300,000 sentences) in 20min

Google said Motorola Mobility contributed revenue of US$ 1.25 billion for the second quarter.
Conclusion and Future Work

- Presented RdfLiveNews, a framework for extraction of RDF from unstructured data streams
- URI disambiguation of 85%, cluster patterns with accuracy of 82.5% and achieve a total accuracy of about 90%
- Can handle 2h timeslices (300,000 sentences) in 20min
- Requires extension to support datatype properties
- Need for reification
  - ..., Google said Motorola Mobility contributed revenue of US$ 1.25 billion for the second quarter.
- Integrate support for temporal logics
Thank you!
Questions?

Axel Ngonga
University of Leipzig
AKSW Research Group
Augustusplatz 10, Room P616
04109 Leipzig, Germany
ngonga@informatik.uni-leipzig.de