Data journalism and fact-checking: a content management perspective

Ioana Manolescu

Inria Saclay and LIX (CNRS and Ecole Polytechnique)

25/10/2017
Plan

1. Data journalism and fact-checking
2. ... are content management problems
3. Some works I have been involved to in this area (2013--)


Thanks to: S. Laurent, M. Vaudano, A. Senecat, M. Ferrer from Les Décodeurs
Data journalism

Investigative journalism based on complex and/or large data

http://abonnes.lemonde.fr/les-decodeurs/portfolio/2017/04/18/les-fractures-francaises-1-5-le-logement-les-raisons-de-la-crise_5112859_4355770.html
Data journalism

Panama Papers (International Consortium of Investigative Journalism, ICIJ)
Data journalism

Panama Papers (International Consortium of Investigative Journalism, ICIJ)
Fact-checking (since 1930 approx.)

Fact-checking: verification of facts mentioned in media content

- To protect media reputation and avoid legal action
- Verification supposes the existence of a reference dataset

The day I became a fact-checker at The New Yorker, I received one set of red pencils [...] for underlining passages on page proofs of articles that might contain checkable facts. [...] confirmed with the help of reference books from the magazine’s library”

Fact-checking (2012 – ongoing)

Not everyone agrees, however, that Democrats are not flip-flopping on the issue. Mark Krikorian, executive director of the Center for Immigration Studies, a think tank that champions lower immigration, said that because the public doesn’t know exact numbers, Mulvaney’s statement is a “tendentious” comparison. But, he said, to dismiss it simply on that basis would be “tendentious.”

“The fact is that, other than the ‘Mexico will pay for it’ stuff, Trump is simply cherry-picking information from the 2006 Secure Fence Act, and Schumer et al. who voted for it out of political calculus are wrong if they’re calling Mulvaney hypocrases for opposing the attempt to finally bring that law to fruition,” he told us via email.

At the surface level, it is true in a broad sense that Democrats including Schumer and Clinton have in the past supported border fencing. All three voted for the Secure Fence Act of 2006, and all three supported the 2013 Senate immigration overhaul that passed out of committee and which called for tougher border security including some additional fencing. But those measures are the same as what Trump is proposing is a stretch.
Fake news: what we're up against

In a former porcelaine factory town in Macedonia, young software developers get rich by publishing on Facebook stories such as:
• "Michelle was caught cheating with Eric Holder – Obama is FURIOUS!!"
• "Bill Clinton loses it in an interview and admits he's a murderer"

Apolotical; work mostly for US clients; claim $2500/month in revenue
A veteran organizes training school for publishing fake news for a cost
• Estimates 100 former pupils are very active on US social networks

http://money.cnn.com/interactive/media/the-macedonia-story/
Fake news: what we're up against

Russian propaganda, 2017:
• mixing true and false content
• building a reputation to better push fake content
• "Orwell was right: “We have always been at war with Eastasia” really does work, if there are enough people repeating it”.
• Ask me about the Elisa story!

Chinese propaganda, 2017
• On days when critical events occur in China, official and semiofficial gov't representative flood social networks with positive posts
• (children gaining prizes, production above expectations...)

http://jonathanstray.com/networked-propaganda-and-counter-propaganda

Defense Against the Dark Arts: Networked Propaganda and Counter-Propaganda

http://jonathanstray.com/networked-propaganda-and-counter-propaganda
Fake news: what we're up against

Chinese propaganda, 2017

"Posts spiked around political events (CCP Congress) and emergencies that the government would rather citizens not talk about, such as riots and a rail explosion. This “cheerleading” propaganda was... a precisely controlled strategy designed to drown out undesirable narratives."

"Fact checking is really after-the-fact-checking"
http://jonathanstray.com/networked-propaganda-and-counter-propaganda
Fact-checking is a content management problem

Claim to be checked (text or data)

Human actors (journalists, experts, crowd workers)

Verification tool (query, match, source search...)

Analysis result + proof « True / rather true / rather false / false »

See sources: http://dataref.com...
Fact-checking is a content management problem

- **Claim extraction**
  - Claim to be checked (text or data)
  - Human actors (journalists, experts, crowd workers)

- **Social network analysis**
  - Source selection
  - Reference information source 1
  - Reference information source 2
  - Reference information source n

- **Verification tool** (query, match, source search...)

- **Reconciliation, reputation**

- **Analysis result + proof**
  - « True / rather true / rather false / false »

Reference source construction, refinement, integration
Fact-checking is a content management problem

Claim extraction

Claim to be checked (text or data)

Human actors (journalists, experts, crowd workers)

Reconciliation, reputation

Social network analysis

Verification tool (query, match, source search...)

Source search / source selection

Reference information source 1

Reference information source 2

... Reference information source n

Reference source construction, refinement, integration

Analysis result + proof « True / rather true / rather false / false »

See sources: http://dataref.com... »
A point of caution: it’s not just checking

Most aspects of modern reality are complex

From a journalistic perspective, explaining may be as important and useful as checking.

Also: future is hard(er) to check.
Toward automated fact-checking

FactMinder demo [SIGMOD 2013]

Browser plug-in

Bringing up rich context for a Web page, from document and knowledge bases

Source search

Supported by template queries over XML documents and RDF graphs

« Second screen »
Content management tools for data journalism

tatooine demo [VLDB2016]

So many data sources, so little time!

data lake

Can we group tweets of politicians by political current and analyze their most frequent topics? »

Can we classify articles by their main concept and do a tag cloud? »

Can we industrialize answering such quests?
Content management for data journalism


So many data sources, so little time!

Data journalism; reference source enrichment

Dataviz modules

Tatooine query engine

Custom processing
Classification, ...

Data sources

JSON sources (SOLR, Postgres...)
Full-text search Indexing

Relational sources (RDBMS)
SQL Indexing

RDF sources (Jena, Corese...)
SPARQL Reasoning

Streams (Twitter, RSS)
Subscribe
Query
Modeling facts, statements and lies

Ongoing work with F. Goasdoué, PhD L. Duroyon

We need to represent:

1. Penelope worked for the National Assembly (NA) from 2002 to 2012
2. In 2002 the employee database showed that she was going to work there from 2002 to 2005
   On Jan 21, 2017, he corrects himself to state that Penelope worked for the NA from 2002 to 2008.
4. On Jan 22, 2017, Le Canard Enchaîné wrote that François had stated on Jan 21 that Penelope worked for the NA from 2002 to 2008.
5. Charles and Marie both worked for the NA at some point, but Charles' contract was after Marie's (they never overlapped)
Modeling facts, statements and lies

Ongoing work with F. Goasdoué, PhD L. Duroyon

Approach:
1. **Facts** = RDF
2. **Timed facts** = bitemporal RDF (transaction time, validity time) [Gutierrez, Hurtado 2005]
3. **Beliefs (viewpoints)**: A states that B states that… [Gatterbauer, Suciu 2009]
4. **Timed beliefs**: at time T1, A states that at time T2, B …

Defined:
- Saturation (set of all consequences of a set of statements)
- Query answering

TBD: query answering algorithm
# Improving access to reference data

Ongoing work with X. Tannier, PhD D. Cao

Linked Open Data extraction from INSEE statistic spreadsheets

Ongoing: fine-granular search in the RDF
To assist live fact-checking

## Table: Mother’s age at the time of the birth

<table>
<thead>
<tr>
<th>Region</th>
<th>Department</th>
<th>16-20</th>
<th>21-25</th>
<th>26-30</th>
<th>31-35</th>
<th>36-40</th>
<th>41-45</th>
<th>46-50</th>
</tr>
</thead>
<tbody>
<tr>
<td>Île-de-France</td>
<td>Essonne</td>
<td>215</td>
<td>1230</td>
<td>5643</td>
<td>4320</td>
<td>3120</td>
<td>1514</td>
<td>673</td>
</tr>
<tr>
<td></td>
<td>Val-de-Marne</td>
<td>175</td>
<td>987</td>
<td>4325</td>
<td>3156</td>
<td>2989</td>
<td>1740</td>
<td>566</td>
</tr>
<tr>
<td></td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>Rhône-Alpes</td>
<td>Ain</td>
<td>76</td>
<td>1103</td>
<td>3677</td>
<td>2897</td>
<td>1976</td>
<td>1464</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Ardèche</td>
<td>45</td>
<td>954</td>
<td>2865</td>
<td>2761</td>
<td>1752</td>
<td>1653</td>
<td>523</td>
</tr>
<tr>
<td></td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
</tbody>
</table>

The data reflects children born alive in 2015...
Digging for Interesting Aggregates in RDF graphs

[ISWC2017]

From one RDF graph...

To **quantitative insights**

- **Automatic** identification of candidate facts, dimensions, measure, aggregations
- Ranked by statistic **interestingness**

Number of conferences listed in DBLP every year from 1936 to 2006
Much more to do

Detecting fake news based on the content, the social context, the language
- "Michelle was caught cheating with Eric Holder – Obama is FURIOUS!!!"

Statistic / ML / IR / graph (social) analysis approaches

Publishing content with proofs or provenance
- Restricted to well-behaved content.. not those we worry about

Building fact bases: crowd-sourcing, knowledge base construction, extraction

Formalizing explanations and proofs; argumentation theory

Calling Bullshit in the Age of Big Data course at U. Washington:

https://www.youtube.com/playlist?list=PLPnZfKID1Sje5jWxt-4CSZD7bUI4gSPS
Merci / questions?