Fact-Checking the Web: Content Management Technologies for a Democratic Decision Process

Part 1

Ioana Manolescu, OAK
Inria Saclay and U. Paris Sud

Joint work with
François Goasdoué, Konstantinos Karanasos, Yannis Katsis, Julien Leblay and Stamatis Zampetakis

5/29/13
Plan

• OAK group
• Context: fact-checking the Web
  – The need, the variants, the projects
• Our proposal: FactMinder
  – Architecture (SIGMOD 2013 demo)
  – Data model and query language
  – Algorithms and open problems
• Conclusion
OAK group
Inria Saclay and Université Paris-Sud
OAK group
Inria Saclay and Université Paris-Sud
OAK team

• Database optimizations and architectures for complex large data

• Inria researcher: Ioana Manolescu

• U. Paris Sud professors
  – Nicole Bidoit (Database team leader), Dario Colazzo, François Goasdoué, Melanie Herschel, Bogdan Cautis

• Post-doc: Zoi Kaoudi

• Inria engineers
  – Andrés Aranda-Andújar
  – Tushar Ghosh

PhD students:
  – Jesús Camacho-Rodríguez
  – Asterios Katsifodimos
  – Julien Leblay
  – Alexandra Roatis
  – Aikaterina Tzompanaki
  – Stamatis Zampetakis
# Oak research at a glance

<table>
<thead>
<tr>
<th>Document (XML) data</th>
<th>Static analysis &amp; logical optim. (query, update, transform)</th>
<th>Storage optimization through views and indices</th>
<th>Parallel data management in the cloud</th>
</tr>
</thead>
<tbody>
<tr>
<td>Semantic (RDF)</td>
<td>Oak</td>
<td>Oak</td>
<td>Oak</td>
</tr>
<tr>
<td>Other complex data</td>
<td>Oak</td>
<td>Oak</td>
<td>Oak</td>
</tr>
</tbody>
</table>

5/29/13

Ioana Manolescu / WOD 2013
From democracy to fact-checking
What do we know?
Informed consent in an information society
What do we know?

Informed consent in an information society

Democracy

Voting
What do we know?
Informed consent in an information society

Democracy

Voting

Opinion

Sentiments/Values
What do we know?
Informed consent in an information society

Democracy → Voting

Opinion → Sentiments/Values

Facts/Information → Democracy
What do we know?
Informed consent in an information society
What do we know?
Informed consent in an information society

Democracy

Voting

Opinion

Sentiments/Values

Meaning/interpretation

Facts/Information

Data

5/29/13
Ioana Manolescu / WOD 2013
What do we know?
Informed consent in an information society

Democracy ➔ Voting

Opinion ➔ Sentiments/Values

Facts/Information ➔ Data

Meaning/interpretation

Studied by social scientists, pollsters, media consultants (aka spin doctors) etc.

Studied by social scientists, historians, writers etc.
What do we know?
Informed consent in an information society

Democracy → Voting

Opinion → Sentiments/Values

Facts/Information → Data

Meaning/interpretation

Studied by social scientists, pollsters, media consultants (aka spin doctors) etc.
What do we know?
Informed consent in an information society
What do we know?
Informed consent in an information society

• Democratic societies are based on citizen voting
• Vote = choice among alternatives
• Choice requires knowing at least what are the options
  • And the consequences of each option
  • « Ay, there’s the rub! » (Hamlet, « To be or not to be » monologue)
• What do we know, really?
What do we know?
Data is on the Web

Opinion

Facts/Information

Data

Social media

Web

5/29/13
Ioana Manolescu / WOD 2013
What do we know?
Data is on the Web

Opinion

Facts/Information

Data

Web
What do we know?
Data is on the Web

Opinion

Facts/Information

Ioana Manolescu / WOD 2013
What do we know?
Data is on the Web

Fact-checking

Opinion

Facts/Information

Data

Newspapers

TV

Web

Social media

Radio

5/29/13
Ioana Manolescu / WOD 2013
Fact-checking

Wikipedia 28/5/13:
“A fact checker is the person who checks factual assertions in non-fictional text, usually intended for publication in a periodical, to determine their veracity and correctness. The job requires general knowledge and the ability to conduct quick and accurate research.”

Since: 1930 approx.

Purpose: publish true information; for periodical reputation and to avoid legal threats...
Fact-checking before the Web

“The day I became a fact-checker at The New Yorker, I received one set of red pencils and one set of No. 2 pencils. [...] The red pencils were 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, including Merriam-Webster’s Geographical Dictionary, the New Grove Dictionary of Music and Musicians and Burke’s Peerage and Gentry.”

Fact-checking the Web: what’s new? (1)

• **New data** is on the Web
  – Media, organization and government Web sites, social networks, blogs etc.
  – Recall **data ≠ facts**

• **Reference data** is on the Web
  – General: Google, Wikipedia (despite caveats)
  – Domain-specific: specialized data sets

• **Fact-checking** can be made on the Web
  – What about automating it as much as possible?
Fact-checking the Web: what’s new? (2)

- New data is on the Web
- Reference data is on the Web
- Fact-checking can be made on the Web

→ Goes outside and beyond newsrooms
→ Anyone with a camera/keyboard and Internet
→ That makes a lot of us!
... and we have different interests!

• National politics: well covered by traditional media
  – But we may have interesting information 😊 and it is becoming ever easier to share it...
  – Hard to **check what politicians are saying** (not a cultural tradition in France, but growing)
  – The modern **world is complex and check results may not be black/white**

• Local administration (not important enough for big media)
  – Local distribution of money, favors
... and we have different interests!

• Health/environment hazards
  – GM food, epidemics, nuclear power, fracking, diesel cars, wood fires, ...

• Local decisions
  – Where to put a parking lot, whether a piece of land can be built upon etc.

• My favorite for a while:
  – « How much $ did Europe give to the car industry in 2008-2010 and what is the evolution of the car making jobs in Europe in 2008-2013 »?

• What is yours?
Some well-known fact-checking projects
Reduce the level of deception and confusion in US politics
FactCheck.org

- Rich activity during US 2012 elections campaign
  – "We're not going to let our campaign be dictated by fact-checkers. » 😊
    Neil Newhouse, pollster for Republican nominee Mitt Romney

- Activity continuing after the elections
Reviewing some key statistical measures of Barack Obama’s presidency so far, we find:

• The economy has added more jobs since Obama took office than it did in his predecessor’s entire eight years in office.
• [...] more people have been added to the food-stamp rolls under Obama than under any single previous president. [...]  
• Obama has ordered seven times more drone strikes than Bush in the covert conflicts in Pakistan and Yemen  
• Domestic oil production has soared; oil imports have dropped by one-third; [...]
PolitiFact

• Truth-o-meter
  @ PolitiFact.com

5500 facts checked as of 2012...
TruthGoggles interface on top of PolitiFact

Being a media consumer is tough.

The Internet is full of statements and opinions. We're bombarded with them every day.

How do you decide what's true?

What goes through your head when you see these types of statements? Do you trust some claims and focusing on others? How do you know what's been presented by the author. Your content tastes?

Often you'll resort to making a judgment call on whether or not statements are perfectly normal.

because of the new health care law, 17 million kids can no longer be denied for a pre-existing condition.

Are you sure this is accurate?

PolitiFact Says: Barely True
Cherry-picking highest estimate

A new film from President Barack Obama's re-election campaign makes the case that his health care law has expanded coverage and stopped insurance companies from denying benefits to people because of previous illnesses.

We're checking several claims from the film, including this one, that "17 million kids can no longer be denied for a pre-existing condition."

It's true that the law prohibits excluding children with pre-existing conditions from private health insurance, effective six months ...

More >>

Close
« Novation », préjudice moral, démissions : les imprécisions de François Bayrou

François Bayrou, président d'honneur du MoDem et ancien candidat à la présidentielle, était l'invité de "Tous politiques", sur France Inter, dont Le Monde est partenaire. Comme chaque semaine, nous avons passé au crible ses déclarations.

Les décodeurs, mode d'emploi

Ce blog du Monde.fr se propose de passer au crible les propos des hommes et femmes publiques pour y démêler le vrai du faux. Envoyez-moi vos interrogations sur les propos tenus dans les médias à l'adresse lesdecodeurs@gmail.com ou sur le compte Twitter du blog. Nous vérifierons ensemble la véracité des déclarations, à partir de sources fiables et transparentes et d'interlocuteurs de référence.

Le service Politique du Monde

A lire aussi

- Sur le mariage homosexuel, Roselyne Bachelot dit vrai
- Le dealer de cocaïne de Jean-Luc Delarue décrit les
May 10, 2013

Wayne LaPierre testifies at Senate hearing on gun violence

Thank you intrusion it's not a today on behalf of the more than four-point-five million moms and dads sons and daughters in every state across our nation who make up the national rifle association of america there are four-point-five million active members of the NRA and they're joined by tens of millions of supporters throughout the country it's on behalf of those millions of decent hard-working law-abiding citizens that I'm here today to voice their concerns. The title of today's hearing is what should america do about gun violence we believe the answer that question is to be honest about what works and honest about what doesn't work. Teaching safe and responsible gun ownership works and the
Thank you instrument it's not a today on behalf of the more than four-point-five million moms and dads sons and daughters in every state across our nation who make up the national rifle association of america there are four-point-five million active members of the NRA and they're joined by tens of millions of supporters throughout the country it's on behalf of those millions of decent hard-working law-abiding citizens that I'm here today to give voice to their concerns. The title of today's hearing is what should america do about gun violence we believe the answer that question is to be honest about what works and honest about what doesn't work. Teaching safe and responsible gun ownership works and the
Thank you instrument it's not a today on behalf of the more than four-point-five million moms and dads sons and daughters in every state across our nation who make up the national rifle association of america there are four-point-five million active members of the NRA and they're joined by tens of millions of supporters throughout the country, it's on behalf of those millions of decent hard-working that I'm here today to give voice to... today's hearing is what should americans do today's hearing is what should americans do about gun violence we believe the answer that question is to be honest about what works and honest about what doesn't work teaching safe and responsible gun ownership works and the...
Washington Post’s TruthTeller

The source is on FactCheck.org
Bobards en stock sur les plans de sauvetage européens

«Les Français vont devoir donner 2 à 3 milliards d'euros pour des banques à Chypre. D'un côté on supprime les infirmières, on surtaxe les PME en France [...] mais quand il s'agit de donner de l'argent de l'UE, c'est-à-dire les Français, à des banques pourries à Chypre, on le donne.»

Nicolas Dupont-Aignan, le 25 mars sur France Bleu

«C'est nous qui allons encore verser de l'argent puisque le MES s'est engagé à hauteur de dix milliards d'euros, dont deux milliards au titre de la France et des contribuables français.»

Florian Philippot, le 25 mars sur France Info

«Comment peuvent-on imposer aux Français, aux classes populaires et moyennes, aux petites entreprises, de nouvelles taxes, des taxes sur les sodas, une hausse de la CSG de 550 millions d'euros, et d'un autre côté alimenter de 15 milliards d'euros supplémentaires en Grèce l'incendie de la zone euro ? On a même pensé à une taxe sur les parcs d'attraction !»

Marine Le Pen, en septembre 2011, au moment du deuxième plan d'aide à la Grèce
Saving Cyprus: how much does it cost?

- The article goes on saying:
  - The money is not **given** but **lent**
  - The European Mechanism of Stability will lend $9bn
  - Out of which France contributes 20% or a bit less than $2bn
Saving Cyprus: how much does it cost?

• The article goes on saying:
  – The money is not **given** but **lent**
  – The European Mechanism of Stability will lend **$9bn**
  – Out of which France contributes 20% or a bit less than **$2bn**

• However, things are complicated because
  – the initial contribution of France to the EMS ($16 bn) **is counted** toward French public debt
  – the remainder contribution ($124) **is not**
Saving Cyprus: how much does it cost?

• The article goes on saying:
  – The money is not given but lent
  – The European Mechanism of Stability will lend $9bn
  – Out of which France contributes 20% or a bit less than $2bn

• However, things are complicated because
  – the initial contribution of France to the EMS ($16 bn) is counted toward French public debt
  – the remainder contribution ($124) is not

INTOX En période de rigueur budgétaire, les plans de sauvegarde successifs offrent depuis trois ans un boulevard aux détracteurs de l’euro et de l’Europe. Nicolas Dupont-Aignan et Florian Philippot ont ainsi, de concert, dénoncé récemment les milliards déversés à Chypre au moment où les Français se serrent la ceinture.

« Les Français vont devoir donner 2 à 3 milliards d’euros pour des banques à Chypre. D’un côté on supprime les infirmières, on surtaxe les PME en France […] mais quand il s’agit de donner de l’argent de l’UE, c’est-à-dire des Français, à des banques pourries à Chypre, on le donne », dénonçait Nicolas Dupont-Aignan. En version Florian Philippot, cela donne : « C’est nous qui allons encore verser de l’argent puisque le MES s’est engagé à hauteur de 10 milliards d’euros, dont 2 milliards au titre de la France et des contribuables français ». Un propos qui fait écho -notamment- à celui entendu quelque deux ans avant, dans la bouche de Marine Le Pen, au sujet cette fois du plan d’aide à la Grèce : « Comment peut-on imposer aux Français, aux classes populaires et moyennes, aux petites entreprises, de nouvelles taxes, des taxes sur les sodas, une hausse de la CSG de 500 millions d’euros, et d’un autre côté alimenter de 15 milliards d’euros supplémentaires en Grèce l’incendie de la zone euro ?

DESINTOX Quel impact ont donc les plans de sauvegarde successifs de l’euro? Et le dernier en date, en direction de Chypre, va-t-il contraindre les Français à allonger 2 milliards d’euros?
Saving Cyprus: how much does it cost?

• The article goes on saying:
  – The money is not given but lent
  – The European Mechanism of Stability will lend $9bn
  – Out of which France contributes 20% or a bit less than $2bn

• However, things are complicated because
  – the initial contribution of France to the EMS ($16 bn) is counted toward French public debt
  – the remainder contribution ($124) is not

The initial statements are mostly false
The complete picture is complicated
Saving Cyprus: how much does it cost?

- The article goes on saving:
  - The money is not given but lent
  - The European Mechanism of Stability will lend $9bn
  - Out of which France contributes 20% or a bit less

- However, things are complicated because:
  - the initial contribution of France to the EMS ($16 bn) is counted toward French public debt
  - the remainder contribution ($124) is not

The initial statements are mostly false

The complete picture is complicated

I would not have been able to give the answer
Saving Cyprus: how much does it cost?

• The article goes on saying:
  — The money is not given but lent
  — The European Mechanism of Stability will lend $9bn
    Out of which France contributes 20% or a bit less than $2bn

• However, things are complicated because:
  — the initial contribution to the EMS ($16 bn) is already counted toward French public debt
  — the remainder contribution ($124) is not

The initial statements are mostly false
The complete picture is complicated
I would not have been able to give the answer
Yet I vote!

Ioana Manolescu / WOD 2013
Saving Cyprus: how much does it cost?

• The article goes on saying:
  - The money is not given but lent
  - The European Mechanism of Stability will lend $9bn
  - Out of which France contributes 20% or a bit less than $2bn
• However, things are complicated because:
  - the initial contribution to the EMS ($16 bn) is counted toward French public debt
  - the remainder contribution ($124) is not

The initial statements are mostly false
The complete picture is complicated
I would not have been able to give the answer
Yet I vote!
Don’t even mention EU politics (commission, parliament...)
Old-style fact-checking: Radio Yerevan jokes

Famous in the Eastern block during communism

**Q1:** *Is it true* that in Moscow, Mercedes cars are being given to citizens?"

**A1:** "*Yes, but* it is not Moscow but Leningrad, not Mercedes but Ladas, and not given to but stolen from."

**Q2:** *Is it true* that in Romania...

**A2:** Yes it is.
Are children of immigrants doing badly in school?

- Guéant (then minister of Interior) says in 2011: « 2/3 of the schoolchildren lagging behind in French schools are children of immigrants »
- The only available numbers are from PISA (EU study)
- The number is 20%
- INSEE had to interfere
Are children of immigrants doing badly in school?

- Guéant (French minister of Interior) says in 2011: « 2/3 of the schoolchildren lagging behind in French schools are children of immigrants »
- The only available numbers are from PISA (EU study)
- The number is 20%
- INSEE had to interfere

Many useful statistic databases are gathered on public money and thus belong to the public and should be open

Reference data for fact-checking (and for political discourse in general)
Fact-checking the Web:

a Web data management perspective
Technologies involved in fact-checking

• Getting the data in digital form
  – It may involve OCR... Big problem noted in CACN « Computational Journalism » article (2010)
  – Transcription from audio and video is harder

Obama’s Speech on Race
Interactive video and transcript of Senator Barack Obama’s speech in Philadelphia on March 18.
Technologies involved in fact-checking

• Getting the data in digital form
  – It may involve OCR... Big problem noted in CACN « Computational Journalism » article (2010)
  – Transcription from audio and video is harder

• Analyzing and understanding the text (NLP)
  – Pretty good at that now: OpenCalais and more
  – Named entity recognition \(\rightarrow\) domain detection
  – Understanding phrases (works well for specialized text, hard in general for irony, style...)
Analyzing and understanding the text

• Extraction of tabular data
  – Research from Google and many others (Inria Saclay/Cachan, Télécom ParisTech, Inria Sophia/LIRMM…)
  – Machine learning (classifiers…) may help

• Extraction of classes (concepts) and instances (objects), e.g. YAGO @ MPI
  – When this is over we have: partially complete, potentially overlapping, potentially contradictory object descriptions (aka semistructured data)
Semistructured data management for fact-checking

• Data integration / entity resolution (Melanie’s tutorial) to consolidate reference and/or new data

• Storing, indexing, searching (with keywords), querying to exploit reference data
  – Querying particularly suited for structured facts
  – Queries generally unpleasant to write
  – NLP interfaces (e.g. SAP for BI), or expert users (DBAs)

• Scalable query processing (database research around the world; Zoi’s tutorial)
Other pertinent DM areas

• Stream and real-time data processing (and fact-checking! It’s just a join anyway 😊)
  – Imagine a Truth-o-meter displaying in real time on the TV screen, during political shows
  – Maybe then we would get adult conversations
Other pertinent DM areas

• Stream and real-time data processing (and fact-checking! It’s just a join anyway 😊)
  – Imagine a Truth-o-meter displaying in real time on the TV screen, during political shows
  – Maybe then we would get adult conversations

  – Such a good idea that Politifact is said to be working on it... it’s been a year though...
    http://slifty.com/2012/06/truth-goggles-study-results/
Other pertinent DM areas

• Data forensics: tracing the route of an information item / document/decision...
  – Many results in neighbor areas of workflow and provenance (not finished)
  – Some early research on forensics, DB for legal investigations

• Crowdsourcing:
  – Exploit the human expertise of many users
  – Build a community of contributors (mechanisms of trust, prestige, reward etc.)
  – Crowdsourcing with workers vs. by independent contributors
Toward generic, semi-automated fact-checking

We aim at usable tools based on open standards to gather and enrich a topic-oriented warehouse

1. **Open standards**: to facilitate interoperability and collaboration and prevent vendor lock-in
   – I don’t mean doing this on Facebook
   – We picked W3C standards

2. **Gather**: a whole research area in itself
   – Structured data, Web extraction... NLP, ontologies
   – We assume an NLP tool such as OpenCalais
Toward generic, semi-automated fact-checking

Simple, usable tools based on open standards to gather and enrich a topic-oriented warehouse

3. **Enrich**: annotate (manually or automatically), edit, connect (in a data / knowledge graph), produce derived data

4. **Topic-oriented**: probably most useful
   – Budget vs. immigration vs. air pollution

5. **Warehouse**: facts and evidence
   – Where did you find this
   – Who said that when
End of Part 1