The Human Factor in Big Data Analysis

Karl Aberer, EPFL
Distributed Information Systems Laboratory
lsir.epfl.ch
Big Data = Volume, Velocity, Variety, Veracity

Variety = Semantics = Meaning
Retrieval, data integration, information extraction, ...

Veracity = Pragmatics = Utility
Data quality, credibility, authority, trust, ...

Stating the obvious: every semantic and pragmatic information processing task related to human concerns requires human input

For example: Google is a huge relevance feedback engine
Big Data analysis today

- **Key innovation**: capacity to automatically process and analyse huge volumes of data
- **Key bottleneck**: human input to make the processing meaningful

Example: recent progress in machine translation and image recognition with deep learning
- Rely on huge corpuses with “ground truth”
Example: automatic generation of image captions

An extreme perspective

The end of theory: the data deluge makes the scientific method obsolete

No models!
No causality!
No understanding!
But often no ground truth available, in particular for applications with “not so big data” and involving expert knowledge

Three case studies

1. Web credibility
   *How human input enables machine learning*

2. Data integration
   *How humans and machines cooperate efficiently in a problem solving task*

3. Social Media Analytics
   *How humans interpret latent structures found by machine learning*
CASE STUDY 1: WEB CREDIBILITY
How to Evaluate Web Credibility?

When you browse a webpage, how do you know it’s content is valid and accurate?
Increasingly difficult to assess credibility of Web content

- Economic incentives to manipulate information
  - Marketing, fraud, political motives, etc.
- Enormous volume of web information

User: Believe or not?

Adversary: Put $$$ to make it look credible?
Which features indicate credibility?

- Numerous candidate features could indicate credibility?
- How to determine?
- Let experts annotate a collection of documents
**Credibility Features**

- Corpus of 1000 documents
- Evaluated by domain experts
- (prepared by MS Research)

**Statistical tests**

- Identification of features providing the signals on credibility

### Table: Credibility Features

<table>
<thead>
<tr>
<th>Topic</th>
<th>Query Terms</th>
<th>Expert URL Filters</th>
<th># of Users</th>
</tr>
</thead>
<tbody>
<tr>
<td>Health</td>
<td>Atkins diet effectiveness, P90x exercise program, H1N1 vaccine side effects, Alzheimer’s genes, Autism warning signs</td>
<td>ncbi.nlm.nih.gov/pubmed, pubmedcentral.nih.gov</td>
<td>254,175</td>
</tr>
<tr>
<td>Finance</td>
<td>Is it a good time to invest in gold? What mutual funds to invest in, Reduce personal debt, Mortgage refinancing, Is it a good time to invest?</td>
<td>bloomberg.com, edgar-online.com, hoovers.com, sec.gov</td>
<td>201,014</td>
</tr>
<tr>
<td>Politics</td>
<td>Iran election rigged, Cash for clunkers eligibility, Obama birthplace, Death Panels, Tea Party</td>
<td>foreignaffairs.com, theatlantic.com, foreignpolicy.com, harvard.edu, economist.com</td>
<td>86,135</td>
</tr>
<tr>
<td>Celebrity News</td>
<td>Lady Gaga, Adam Lambert, Nadya Suleman, Floyd Landis, Michael Jackson</td>
<td>ew.com, usmagazine.com, people.com</td>
<td>692,611</td>
</tr>
<tr>
<td>Environmental</td>
<td>Renewable energy, Green jobs, Climate change, Cap-and-trade, Organic farming</td>
<td>newsclimate.org, epa.gov, eff.org, nrdc.org, whitehouse.gov/administration/ceq</td>
<td>83,476</td>
</tr>
<tr>
<td>All Users</td>
<td>(total)</td>
<td></td>
<td>50,473,530</td>
</tr>
</tbody>
</table>

### Informativeness

- Google Search Ranking
  - Domain Type (.gov, .edu)
  - Use of Punctuation
  - Web Graph Structure
  - Browsing Patterns

### Readability

- Number of Bookmarks
- Ads Prominence
- Objectivity
- Text Complexity
- Webpage Topic
The content of a webpage as well as the social popularity offer signals for credibility.
A credible source?
Analysing the content

Treatments to Explore

Conventional medicine treats the symptoms of autism. Biomedical treatment addresses the root cause.

There is a wealth of biomedical therapies that treat the underlying issues of autism inside the body.

The following is a list of biomedical treatments to explore with a physician in order to help heal the body:

1. Follow the gluten-free, casein-free, soy-free diet and remove other food allergens.

The yeast-autism connection can be a result of candida (type of yeast) overgrowth in the system. This leads to many different behaviors such as, fogginess, sensory issues, negative behaviors.

"Gluten and Dairy seem to affect a lot of our children with autism and thus we see a
1. Follow the gluten-free, casein-free, soy-free diet and remove other food allergens.

The yeast-autism connection can be a result of candida (type of yeast) overgrowth in the system. This leads to many different behaviors such as, fogginess, sensory issues, negative behaviors.

"Gluten and Dairy seem to affect a lot of our children with autism and thus we see a lot of children respond terrifically when these are removed from the diet. The goal behind changing diets is to remove chemicals, toxins and potential neurotransmitters, which are liberated when food are broken down. These substances could be toxic for the brain and cause behavioral trouble in kids who are sensitive. Whether kids test as allergic or not, often they are causing a negative effect on the child and they must be removed. Each child has his or her own set of sensitivities that he or she can't deal with properly. When we change their diets, 80 percent of the kids with autism seem to respond." - Dr. Jerry Kartzinel, from "Healing and Preventing Autism" by Jenny McCarthy and Dr. Jerry Kartzinel.

- Effectiveness of the gluten-free, casein-free diet for children diagnosed with autism spectrum disorder: based on parental report.

- Nutrition Guide and how to implement the GF CF diet;

- Dr. Jerry's blog on why to implement the gluten free, casein free diet - Parts 1 & Part 2

More Resources:

- GFCFDiet.com
- The role of Clostridia and Autism
- The Yeast Problem and Bacteria By-products
- Improved Diet Helps Children with Autism
Transfer evaluation to semantically similar statements (claims)

Reconcile project: http://reconcile.pl/
The data fusion algorithm

Expectation-Maximization Algorithm

- Page credibility
- User credibility
- Statement credibility
- User reputation
- Features
- Rating
- Statement
- Web page
- Human evaluation is at the origin of every automated credibility evaluation task
- Same is true for any semantic or pragmatic task (e.g. translation, image labeling etc.)
- The Big Question: where is the ground truth?
- The answer: ask the crowd or experts

Supervised Learning
CASE STUDY 2: DATA INTEGRATION
• Integration of heterogeneous data sources
  – Every project on Big Data analysis first has to integrate data from different, heterogeneous data sources
  – One of the long-standing open problems in data management (both industry and research)

  ▪ How to find good “matches”?
  ▪ How to choose the “best matches”?

![Diagram with nodes labeled Name, Person_name, Company_name and edges labeled c1 and c2.]

Example: Schema Matching
Approaches for Identifying Correspondences

- **Manual matching**
  - still common practice today

- **Schema matching tools**
  - Based on structural and content features
    - names, domains, structure, values, ...
  - Establish correspondences and rank according to quality
    - Errors are frequent and unavoidable
    - Works well for small schemas
Wisdom of the Network

Data integration networks: different experts may contribute partial matches
Which one would you choose?

Instead of considering only one mapping, consider whole networks of mappings: **leverage knowledge from the network!**
By combining different matches in a network we can construct evidence for the correctness of those matches.

- For example, a matching contributing to a “bad cycle” less likely to be correct.

Idea: combine all this evidence and use probabilistic reasoning to select the most likely matchings.

variable $x$ to local factor $f$:

$$\mu_{x \rightarrow f}(x) = \prod_{h \in n(x) \setminus \{f\}} \mu_{h \rightarrow x}(x)$$

local factor $f$ to variable $x$:

$$\mu_{f \rightarrow x}(x) = \sum_{\sim \{x\}} \left( f(X) \prod_{y \in n(f) \setminus \{x\}} \mu_{y \rightarrow f}(y) \right)$$

matches

features
**Empowering the User**

- Probabilistic reasoning results in reasonable improvement of matching quality, but
  - a posteriori analysis can only identify potentially bad choices by experts, but not correct them
- Better approach
  - Let experts make better local decisions by providing them information on global consistency and asking targeted questions

Asking the right questions is important

Two possible solutions: {c1,c2,c3} and {c1,c4,c5}

- Ask c₁ first
  - the network is unchanged
  - no uncertainty reduction.
- Ask c₂ first
  - only 1 solution left
  - the network becomes certain.

Idea: optimize information gain with each question

![Network Diagram]
- Information gain ordering strategy achieves savings of up to **48% user effort** compared to random ordering.
- Outperforms the baseline with an average difference of 15% (precision) and 14% (recall).
Data Integration is a task that combines human and machine intelligence

The Big Question: How to minimize human effort and maximize information gain?

Active Learning
CASE STUDY 3: SOCIAL MEDIA ANALYSIS
Social Media (e.g. Twitter) contains many (hidden) signals on the public perception of issues of general interest—nutrition, health, politics, environment etc.

Goal: identify influencers, their communities, their topics of interest and their stance towards given issues

Methods

– **Semantic content analysis** to capture and classify relevant content
– **Social network analysis** to capture and analyze social influence
1. Describe the interest (keywords, users, time, geographic)
2. Select (or collect) the data
3. Extract the key Concepts, Entities and Categories
4. Identify Topics and Communities
5. Select relevant Issues, Influencers and Events
6. Produce insights (correlations)
Human-Machine Interaction

Initial question

Collect and select data

Find hidden structures

Identify interesting correlations

Identify interesting terms

Select and characterize structures and data of interest

Extend Terminology

New Terminology

Correlations

New questions

Terminology

Data

Structures
Creating Terminology

**Input:**
~50 Mio tweets

**Analysis method:**
Word embedding (word2vec)

<table>
<thead>
<tr>
<th>food</th>
<th>food ingredients</th>
</tr>
</thead>
<tbody>
<tr>
<td>Exploring the topic “food” the system suggest related terms. We find terms related to food ingredients.</td>
<td>We create a category “food ingredients”. The system proposes more related terms.</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>food ingredients</th>
<th>food ingredients</th>
</tr>
</thead>
<tbody>
<tr>
<td>We select all terms that are related. We may repeat this step.</td>
<td>Finally we have a clean list of all terms on food ingredients.</td>
</tr>
</tbody>
</table>

**Discovery of interesting dimensions**

<table>
<thead>
<tr>
<th>fat, food ingredients, 32,655</th>
<th>(low fat, low fat, 822)</th>
</tr>
</thead>
<tbody>
<tr>
<td>calories, food ingredients, 865</td>
<td>(added sugar, added sugar, 590)</td>
</tr>
<tr>
<td>sugar, (sugar), 34,228</td>
<td>(protein, protein), 8042</td>
</tr>
<tr>
<td>fructose, (fructose), 3915</td>
<td>processed foods, (processed foods), 1019</td>
</tr>
<tr>
<td>sugars, (sugars), 1484</td>
<td>trans fats, (trans fats), 680</td>
</tr>
<tr>
<td>sodium, (sodium), 2293</td>
<td>(hfcs, hfcs), 1,772</td>
</tr>
<tr>
<td>saturated fat, (saturated fat), 1146</td>
<td>calcium, (calcium), 2453</td>
</tr>
<tr>
<td>fiber, (fiber), 2222</td>
<td>trans fat, (trans fat), 732</td>
</tr>
<tr>
<td>fats, (fats), 3886</td>
<td>carb, (carb), 2232</td>
</tr>
<tr>
<td>(sats, fats), 437</td>
<td>calorie, (calorie), 4801</td>
</tr>
<tr>
<td>(salt, salt), 5769</td>
<td>(added sugars, added sugars), 233</td>
</tr>
<tr>
<td>grains, (grains), 1861</td>
<td>toxins, (toxins), 2033</td>
</tr>
<tr>
<td>(nutrients, nutrients), 2517</td>
<td>(sweeteners, sweeteners), 2030</td>
</tr>
</tbody>
</table>

**Instantiation of dimensions with terminology**

| (artificial sweeteners, (artificial sweeteners), 978) | (sat fat, (sat fat), 175) |
| (additives, food ingredients, 1750) | (antioxidants, antioxidants), 998 |
| (caffeine, (caffeine), 802) | (vitamins, (vitamins), 225) |
| (pesticides, (pesticides), 53) | (diet soda, (diet soda), 1) |
| (fruit juice, (fruit juice), 1) | (bacteria, (bacteria), 1) |
| (hormones, (hormones), 1282) | (additives, food ingredients, 1750) |
Using a semantically expanded terminology increases coverage significantly!
The system clusters the terms on food ingredients according to similarity.

The expert sees:
- A clear distinction between positive and negative terms.
- Distinction between natural and artificial ingredients.
- Clusters of related terms, e.g., vitamins, additives etc.

We may use this to create sub-categories of interest.
Analyzing social interactions we can identify clear communities

For each community we can identify
- Their influencers
- Their main concepts
- Potentially new interesting terminology
Findings: Company influencers
Findings: role of food ingredients in different countries
Machine learning applied to Big Data can reveal surprising hidden structures with valuable insights

Big questions:
- How to guide the machines to the right data and analysis
- How to make the resulting structures human-interpretable

Unsupervised Learning
Big Data has impressive potential to create insights and solve hard problems

Human intervention in the analysis processes is essential for obtaining meaningful results

Three main types of intervention
  - A priori: supervised learning
  - Interactive: active learning
  - A posteriori: unsupervised learning

No one size fits all: their specific implementation depends strongly on the use case