Big (User) (Health) Data Management

Sihem Amer-Yahia

DR CNRS @ LIG

Sihem.Amer-Yahia@imag.fr

Journées Big Data Mining & Visualization

Juin 2014
Data Management

• **What it is.** It is the development and execution of architectures, policies, practices and procedures that properly manage the full data lifecycle needs of an enterprise.” DAMA International}
Data Management

• **What it is.** It is the development and execution of architectures, policies, practices and procedures that properly **manage the full data lifecycle needs of an enterprise.**” DAMA International}

• **What it isn’t.** It is not about deciding which data to gather and how useful applications are (need for a domain expert)

• **What makes it a science.** It develops principled and reusable methods for gathering, organizing and exploiting data
  • **Gathering:** dumps, crawlers, Application Programming Interfaces (APIs)
  • **Organizing:** pre-processing and storing (because data is made persistent on disk) structured and semi-structured content
  • **Exploiting:** via exploration (search and querying languages and algorithms) and recommendation (functions and algorithms)
(Traditional) data management

Separation between physical and logical layers

- Application specification
- Access optimization
- Data storage and index creation

<table>
<thead>
<tr>
<th>Ward No.</th>
<th>Ward name</th>
<th>Type</th>
<th>No. of Beds</th>
</tr>
</thead>
<tbody>
<tr>
<td>3</td>
<td>Carey</td>
<td>Medical</td>
<td>8</td>
</tr>
<tr>
<td>6</td>
<td>Bracken</td>
<td>Medical</td>
<td>16</td>
</tr>
<tr>
<td>7</td>
<td>Brent</td>
<td>Surgical</td>
<td>12</td>
</tr>
<tr>
<td>8</td>
<td>Meavy</td>
<td>Surgical</td>
<td>10</td>
</tr>
</tbody>
</table>
User (health) data
User health data

- Electronic Health Records (EHRs)
- User-Generated Content (UGC)
EHRs

- Physiological monitoring
- Genomics
- Anatomical imaging
- Also include demographics, medical history, medication and allergies, immunization status, laboratory test results, personal statistics like weight, and billing information
  - Called Dossier Medical Personnel (DMP) in France (since August 2004) with a lot of controversy for adoption
  - 381,015 DMPs have been created so far according to www.dmp.gouv.fr
User data

- EHRs
- UGC
  - Mobile health apps
  - Online forums
mHealth apps categories

- General healthcare and fitness
  - Fitness & nutrition
  - Health tracking tools
  - Managing medical conditions
  - Medical compliance
  - Wellness (traditional and corporate)

- Medical information
  - Diagnostic Tools including predispositions
  - Continuing Medical Education (CME)
  - Alerts and Awareness

- Remote monitoring, collaboration and consultation

- Healthcare management
  - Logistical & payment support
  - Patient health records
Devices and tools

- Medical devices as wearables or as mHealth apps
  - thermometers, scales, blood pressure cuffs, electrocardiogram, sleep patterns
  - MyFitnessPal (> 40M users): tracks nutritional intake and monitors weight goals, connects to friends, API to connect to apps (e.g., Withings Scale and RunKeeper)

- In January 2012, the X Prize Foundation published a $10-million award for creating a medical tricorder capable of scanning, analyzing and recording a wide array of data, including physiological data
  - Walter De Brouwer's Scanadu SCOUT to read the brain
  - Eugene Chan's Universal Blood Sensor
The AliveCor handheld heart monitor and a sample ECG that can be emailed to doctors and patients

Mobisante's smartphone-based ultrasound imaging system (costs 1/10th of a regular ultrasound)

Quantified self with Withings
Online forums

• > 20% American adults posted on an online health-related forum

• **General:** WebMD.com (> 80M/month), health.nih.gov, healthfinder.gov, intelihealth.com, mayoclinic.org

• **Personalized:** HealthTap, “triage” system, where consumers ask doctors for the most effective way to get specific care

• **Collaborative:** YellowCard (public), PatientsLikeMe (private)
  • help patients with a chronic condition answer: “Given my status, what is the best outcome I can hope to achieve, and how do I get there?”
  • experience sharing via patient, reported outcomes, finding similar patients matched on demographic and clinical characteristics, and aggregated stats
  • CureTogether, Diabetic Connect
Categories

- ALS
- Are you sleeping?
- Chronic Fatigue Syndrome/ME
- Conditions
- Conferences/Events
- Drug Safety
- Epilepsy
- Fibromyalgia
- Genetics
- HIV/AIDS
- Idiopathic pulmonary fibrosis
- Media Coverage
- Mood Conditions
- Multiple Sclerosis
- One for All
- Openness
- Organ Transplants
- Parkinson's Disease
- Patient Choices
- Pulmonary fibrosis
- Rare Diseases

Who's like you?

Share your experience.
The more you share, the easier it will be to find patients like you. Start by adding a condition, symptom or treatment.
Bipolar Type II

**About this Condition:** Bipolar II disorder is a bipolar spectrum disorder characterized by at least one hypomanic episode and at least one major depressive episode; with this disorder, depressive episodes are more frequent and more intense than manic episodes. It is believed to be under-diagnosed because hypomanic behavior often presents as incredibly high-functioning behavior. Indeed, to a physician or psyc... Read more

**Synonyms:** Bipolar 2, Manic-Depressive Disorder

---

**Who has this Condition?**

- **11,879** patients have this condition
- **405** New patients this month
- For **6,789** this is their primary condition

---

**Age**

- <20: 30%
- 20-29: 10%
- 30-39: 15%
- 40-49: 15%
- 50-59: 15%
- 60-69: 10%
- 70+: 0%

**Gender**

- 80% Women
- 20% Men

---

**Top Treatments**

<table>
<thead>
<tr>
<th>Treatment</th>
<th>#Patients</th>
<th>#Evaluations</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ibuprofen</td>
<td>113</td>
<td>32</td>
</tr>
<tr>
<td>Acetaminophen</td>
<td>97</td>
<td>20</td>
</tr>
<tr>
<td>Topiramate</td>
<td>67</td>
<td>0</td>
</tr>
<tr>
<td>Excedrin Migraine</td>
<td>42</td>
<td>16</td>
</tr>
<tr>
<td>Butalbital-acetaminophen-caffeine</td>
<td>31</td>
<td>0</td>
</tr>
<tr>
<td>Chiropractic</td>
<td>24</td>
<td>0</td>
</tr>
</tbody>
</table>

**Top Symptoms**

<table>
<thead>
<tr>
<th>Symptom</th>
<th>#Patients</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mania</td>
<td>113</td>
</tr>
<tr>
<td>Depression</td>
<td>97</td>
</tr>
<tr>
<td>Headache</td>
<td>67</td>
</tr>
<tr>
<td>Migraine</td>
<td>42</td>
</tr>
<tr>
<td>Fatigue</td>
<td>31</td>
</tr>
<tr>
<td>Insomnia</td>
<td>24</td>
</tr>
</tbody>
</table>

---

**Find Patients Just Like You**

Join Now! (It's free!)

**Patient Spotlight**

**Rachel44**

Female, 57 years, Lancaster, PA

Hi, I am Rachel and I am married to my wonderful husband of 20 years, and we have 2 wonderful sons and a daughter in law. I was diagnosed back in '84, wh... See Profile

**Links**

- **PatientsLikeMe Research Updates**
  - Take a minute to read recently published reports from the PatientsLikeMe R&D team

- **What are others saying about PatientsLikeMe?**
  - Tune in to Twitter
  - Find us In the News
  - Read patient Testimonials

- **PatientsLikeMe Videos**
Managing and mining UGC is different from data management and data mining

- UGC is heterogeneous and sources diverse
  - does not all fit in a single store, does not fit the same model
- UGC contains loads of valuable information for business intelligence, public health, ...
- But is also noisy, unreliable, incomplete, uncertain and subjective!
  - how to extract valuable information from raw user data?
Let’s examine a canonical example
You Are What You Tweet
Analyzing Twitter for Public Health
Michael J. Paul, Mark Drezde (Johns Hopkins U.) ICWSM 2011

Analyze Twitter messages (1.5 million) to measure public health by discovering a dozen ailments (allergies, flu, insomnia, ...), tracking them over time and location and analyzing medication usage
Twitter and public health

- Web search: users express *need for information* flu medicine
- Social media: users express self information sick with the flu
Data preparation for tweets

• Step 1: find health-related tweets
  – 2 billion tweets from May 2009 to October 2010 (O’Connor et al. 2010)
  – supervised machine learning: SVM
  – Used 5128 labeled tweets to train a high precision SVM binary classifier (0.90) to identify health related messages, which produced a corpus of 1.63 million English tweets (removed punctuation, stop words, URLs and hashtags)

  

  I FEEL LIKE I'M GOING TO DIE OF BIEBER FEVER.
  NO JOKE.

  Web design class gives me a huge headache everytime.

• Step 2: group tweets by disease
  – unsupervised topic models: ATAM
  – 20 ailments from WebMD.com
  – paired each tweet with an ailment

• Step 3: validate with human evaluators
  – manual labeling of each output <ailment, symptom, treatment> by 2 evaluators
  – each evaluator labeled the 20 ailments with diseases they chose or as incoherent
Output to be validated by human evaluators

<table>
<thead>
<tr>
<th>Ailment</th>
<th>Allergies</th>
<th>Aches/Pains</th>
<th>Dental</th>
</tr>
</thead>
<tbody>
<tr>
<td>General Words</td>
<td>allergies</td>
<td>body</td>
<td>meds</td>
</tr>
<tr>
<td></td>
<td>stop</td>
<td>head</td>
<td>killers</td>
</tr>
<tr>
<td></td>
<td>eyes</td>
<td>need</td>
<td>dentist</td>
</tr>
<tr>
<td></td>
<td>allergic</td>
<td>hurts</td>
<td>teeth</td>
</tr>
<tr>
<td>Symptoms</td>
<td>sneezing</td>
<td>pain</td>
<td>pain</td>
</tr>
<tr>
<td></td>
<td>cold</td>
<td>aches</td>
<td>toothache</td>
</tr>
<tr>
<td></td>
<td>coughing</td>
<td>stomach</td>
<td>sore</td>
</tr>
<tr>
<td>Treatments</td>
<td>medicine</td>
<td>massage</td>
<td>braces</td>
</tr>
<tr>
<td></td>
<td>benadryl</td>
<td>“hot bath”</td>
<td>surgery</td>
</tr>
<tr>
<td></td>
<td>claritin</td>
<td>ibuprofen</td>
<td>antibiotics</td>
</tr>
</tbody>
</table>
Flu trends on Twitter
# Drug use analytics

<table>
<thead>
<tr>
<th>Medicine</th>
<th>Entropy</th>
<th>Most Common Ailments</th>
</tr>
</thead>
<tbody>
<tr>
<td>tylenol</td>
<td>1.57</td>
<td>Headache (39%), Insomnia (30%), Cold (9%)</td>
</tr>
<tr>
<td>ibuprofen</td>
<td>1.54</td>
<td>Headache (37%), Dental (21%), Aches (17%)</td>
</tr>
<tr>
<td>advil</td>
<td>1.08</td>
<td>Headache (61%), Cold (6%), Dental (5%)</td>
</tr>
<tr>
<td>aspirin</td>
<td>1.04</td>
<td>Headache (69%), Insomnia (10%), Aches (10%)</td>
</tr>
<tr>
<td>vicodin</td>
<td>1.33</td>
<td>Dental (61%), Injuries (11%), Headache (10%)</td>
</tr>
<tr>
<td>codeine</td>
<td>1.94</td>
<td>Cold (25%), Dental (19%), Headache (17%)</td>
</tr>
<tr>
<td>morphine</td>
<td>1.17</td>
<td>Dental (59%), Infection (22%), Aches (9%)</td>
</tr>
</tbody>
</table>
Geographic behavioral risk factors

- **Goal**: extract geographically linked health statistics from Twitter
- **Data collection**: phone interviews of over 350,000 US adults in 2009 with questions such as “During the past month, did you participate in any physical activities?” with the exercise ailment for each state
- **Approach**: 
  - measured the Pearson correlation coefficient between the ailments discovered in each US state with the state’s risk factor rate
  - assigned a US state to each tweet according to user location by extracting this information from the location field, part of a user’s profile
  - location information obtained for 12% of health related tweets (196,000 tweets)
  - computed the number of each ailment occurrences per capita: occurrences divided by the total number of tweets per state
- **Results**: correlations between each risk factor and the related ailments measured for each state that had at least 500 tweets (42 states in total.) The strongest correlation was for tobacco use (proportion of residents in each state who are smokers) with cancer
### Risk factors

<table>
<thead>
<tr>
<th>Risk Factor</th>
<th>Ailments</th>
<th>Correlation ATAM+</th>
<th>Correlation ATAM</th>
</tr>
</thead>
<tbody>
<tr>
<td>Asthma</td>
<td>Allergies</td>
<td>0.241</td>
<td>0.195</td>
</tr>
<tr>
<td>Diabetes</td>
<td>Obesity</td>
<td>0.073</td>
<td>0.203</td>
</tr>
<tr>
<td>Exercise</td>
<td>All ailments</td>
<td>-0.352</td>
<td>-0.394</td>
</tr>
<tr>
<td>Exercise</td>
<td>Exercise</td>
<td>0.140</td>
<td>-</td>
</tr>
<tr>
<td>Exercise</td>
<td>Obesity</td>
<td>-0.201</td>
<td>-0.248</td>
</tr>
<tr>
<td>Health Care Coverage</td>
<td>All ailments</td>
<td>-0.253</td>
<td>-0.319</td>
</tr>
<tr>
<td>Heart attack</td>
<td>Obesity</td>
<td>0.244</td>
<td>0.341</td>
</tr>
<tr>
<td>Heart attack</td>
<td>Cancer</td>
<td>0.613</td>
<td>0.291</td>
</tr>
<tr>
<td>Obesity</td>
<td>Obesity</td>
<td>0.280</td>
<td>0.203</td>
</tr>
<tr>
<td>Obesity</td>
<td>Exercise</td>
<td>-0.267</td>
<td>-</td>
</tr>
<tr>
<td>Tobacco use</td>
<td>Cancer</td>
<td>0.648</td>
<td>0.320</td>
</tr>
</tbody>
</table>
Spatio-temporal analytics for allergies
Geographic Syndromic Surveillance

Big Data Mining & Visualization 2014
Managing and mining UGC is different from data management and data mining

- **UGC is heterogeneous and sources diverse**
  - does not all fit in a single store, does not fit the same model

- **UGC contains loads of valuable information for business intelligence, public health, ...**

- **But is also noisy, unreliable, incomplete, uncertain and subjective!**
  - how to extract valuable information from raw user data?
  - how to help data scientists prepare raw user data?
User data management stack

Application logic

Search and Recommendation

Analytics

Data preparation

raw data
SOCLE: A framework for social data preparation

with N. Ibrahim, C. Kamdem-Kengne, F. Uliana, M.C. Rousset

Accepted at Ingénierie des Systèmes d'Information (ISI 2014)
Managing and mining user data is different from data management and data mining

- **User data is heterogeneous and sources diverse**
  - does not all fit in a single store, does not fit the same model

- **User data contains loads of valuable information for business intelligence, public health, …**
  - how to extract valuable information from raw user data?
  - how to help data scientists *prepare* raw user data?
  - how to help data scientists *label data and validate findings*?
Crowd4U today

>900 tasks/day (Dec., 2013)
CrowdHealth (CNRS MASTODONS)

• **Goal**
  – enable *health and nutrition-related hypothesis testing* in physical and virtual spaces
  – extract observations from Twitter and verify them on Nutrinet
  – a collaboration between 2 CNRS institutes (INS2I and SHS) and Paris Descartes and UREN (Univ. de Villetaneuse)

• **Partners**
  – researchers on *scalable algorithms for mining personal data and times series*, *nutritionists*, and researchers studying *individuals in geographical spaces*

• **Duration:** 7 months -> end of this year
• **Budget:** 30K
• **Looking for interns!**
Current tasks on Crowd4U

- Tweet annotation
- Restaurant reviews
- Translation tasks
- Others to come
Summary

• Managing UGC is a Big Data problem
• New opportunities for domain experts
  • Large-scale analytics on user-generated content
  • Can public health officials and researchers replace current more expensive and time consuming method?
• New opportunities for researchers in data management, data mining and related areas
  • Data preparation
  • Crowdsourcing for targeted human involvement
scalable models and algorithms for the discovery and exploitation of information and knowledge from data

• Data acquisition and enrichment
  – Big data preparation
  – Web data linkage
  – Crowd data sourcing

• Large-scale data mining
  – Advanced pattern mining
  – Distributed join algorithms
  – Social media and health analytics

• Information exploration
  – Ontology-based data access
  – Interactive pattern exploration