De la Business Intelligence aux Big Data

Marie-Aude AUFAURE
Head of the Business Intelligence team
Ecole Centrale Paris
Agenda

• Evolution of Business Intelligence

• Semantic Technologies for Big Data

• Social networks

• Big Data opportunities : use cases
## Evolution of Business Intelligence

<table>
<thead>
<tr>
<th>Classical Business Intelligence</th>
<th>Semantic Business Intelligence</th>
<th>Real-time Business Intelligence</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Output</strong></td>
<td><strong>Visual analytics</strong></td>
<td><strong>Real-time analytics</strong></td>
</tr>
<tr>
<td><strong>User Interaction</strong></td>
<td><strong>Flexible queries / SPARQL</strong></td>
<td><strong>Knowledge enrichment</strong></td>
</tr>
<tr>
<td><strong>Store</strong></td>
<td><strong>Triple Store</strong></td>
<td><strong>Continuous queries/Business rules</strong></td>
</tr>
<tr>
<td><strong>Gathering Information</strong></td>
<td><strong>Semantic ETL/Batch processing</strong></td>
<td><strong>Semantic ETL stream processing</strong></td>
</tr>
<tr>
<td><strong>Data sources</strong></td>
<td><strong>ETL/Batch processing</strong></td>
<td><strong>Load shedding</strong></td>
</tr>
</tbody>
</table>

- **Data sources**: databases
- **Gathering Information**: structured/unstructured data
- **Store**: Data Warehouse
- **User Interaction**: Ad-hoc queries Analytics
- **Output**: Static report
Evolution of Business Intelligence

Classical Business Intelligence
- Static report
- Ad-hoc queries
- ETL/Batch processing
- Databases

Semantic Business Intelligence
- Visual analytics
- Flexible queries / SPARQL
- Semantic ETL/Batch processing
- Structured/unstructured data

Real-time Business Intelligence
- Real-time analytics
- Knowledge enrichment
- Continuous queries / Business rules
- Semantic ETL / stream processing
- Load shedding

Output

User Interaction

Store

Gathering Information

Data sources

Data Warehouse

ETL/Batch processing

databases

Structured data

Real-time visual-analytics

Sensors

Static data
And now?

Big Data

A fundamental shift in the way companies conduct business and interact with customers

Open Data / Linked Data

Connected objects
Data sources

• Traditional data sources (transactions, behavior models, etc.) + new ones (social media, mobile information)
  – Extract content and consumer sentiment from social stream to get consumer insights, product feedback and market trends
  – Use GPS to deliver targeted real-time messaging based on consumer’s location

• Structured and unstructured data must be properly formatted, integrated and cleansed to extract actionable knowledge in real-time
  – Using real-time processing can help for personalizing a customer’s online website visit, enhancing her overall experience
  – Monitoring of transactions in real-time has benefits for security
BIG DATA: TECHNOLOGICAL CHALLENGES

• **Data infrastructure tools and platforms**: data centers, cloud infrastructures, noSQL databases, in-memory databases, Hadoop/Map Reduce Ecosphere

• **New generation of front-end tools for BI and analytic systems**: data visualization and visual analytics, self-service BI, Mobile BI

• **Data processing**: supercomputers, distributed or massively parallel-computing
Evolution of Business Intelligence

Classical Business Intelligence

- Static report
- Ad-hoc queries/Analytics
- Data sources: databases
- Gathering Information: Warehouse ETL Batch processing
- Store: Data

Semantic Business Intelligence

- Visual analytics
- Flexible queries/SPARQL
- Triple Store ETL Batch processing
- Structured/unstructured data
- User Interaction: Ad-hoc queries/Analytics

Real-time Business Intelligence

- Real-time analytics
- Continuous queries/Business rules
- Real-time visual-analytics
- Knowledge enrichment
- SemanticETL stream processing
- Real-time visual-analytics
- Load shedding
- Data sources: Real-time visual-analytics
- Data stream: Static data
- User Interaction: Ad-hoc queries/Analytics
- Store: Data
Semantic Technologies for Big Data
Linked Data / Web of Data

- **Linked Data** is a set of principles that allows publishing, querying and consumption of RDF data, distributed across different servers.
- Not necessarily free / open data.
- Exponential growth -> a Big Data approach: enriching Big Data with metadata & semantics, interlinking Big Data sets.
Semantic Data Aggregation and Linking for Big Data

- Transforming unstructured content into a structured format for later analysis is a major challenge.
- The value of data explodes when it can be linked with other data, thus data integration is a major creator of value.
- Data aggregation from various sources can establish the veracity.
- Semantic technologies are a way of addressing variety.
Semantic Data Aggregating and Linking for Big Data

**DATA LAYER**
- Database
- Sensor data

**KNOWLEDGE LAYER**
- Semantic aggregation
- Semantic Enrichment and disambiguation
- Linking data

Linked Open Data

**Structured**
- Documents
- Web pages

**Non-structured**
- Textual content

**Social Media**
- Foursquare
- LinkedIn
- Vimeo
- Facebook
- YouTube
- Twitter
- Google+
Value of Semantic Technologies

- Semantic Technologies provide opportunities for reducing the cost and complexity of data integration
- Common metadata layer
- Powerful solutions to find and explore information
- Semantic Technologies are a good fit for Big Data’s Variety
- Velocity and Volume: challenging issues for Semantic Technologies
- Linked Data will grow into Big Linked Data, but Big Data will also benefit from evolving into Linked Big Data
Social Networks
Benefits for enterprises

• A technology for internal communication, information sharing and collaboration
• A technology for information communication towards clients
  – Vote for the best product,
  – Understand the clients needs
• A technology for watching the gossip
  – E-reputation, opinion mining
• A technology for creating collective intelligence
  – Collaborative common knowledge
  – Wikis and blogs associated to social networks
Benefits for public administrations

• Public administrations need social networks:
  – As enterprises:
    • To analyze internal networks (projects, organization...)
    • To analyze external networks (suppliers, clients, partners...)
  – As an interface for citizens:
    • To be well-understood by citizens (who does what)
    • To understand citizens (who says what)

• Scenarios examples:
  – Need to look over the organizational structure (employees, departments, transversal projects) and identify costs
  – Need for citizens to understand the impact of public politics (offered services, available resources for each district of the city, which projects are the most relevant, citizens complains)
  – Opinion analysis from external social networks (Twitter for example)
Sentiment Analysis – Opinion mining

*Find out what other people think. Is it possible?*

**Recommender systems** *(avoid recommending items that received a lot of negative feedback).*

**Information Filtering**

**Business Intelligence** *(why aren’t consumers buying my laptop?)*

**Question answering** *(what did you want to say?)*

**Clarification of politicians positions!**

**eDemocracy…and so on**
Detecting feature sentiment in user-generated reviews

It is not possible to summarize everything with a unique vote polarity ⇒ detect local polarities expressed about the salient features of a considered domain.

Combine NLP+statistics:

1) We identify the most characterizing aspects of one domain (hotels, restaurant, products) by analyzing the domain corpus and extracting the most frequent terms (eventually structuring them as a vocabulary and/or ontology)

2) We formalize the content of each review as a dependency tree among its terms and retrieve (if they exist) the features discussed within it. Then, by using the tree, we aim at discovering all the other terms that vehiculate some polarity linguistically connected to them.
**English Corpus** $R_E$

**Raw text**

**Dep. Graph G**

**Feature Extractor**

**Feature Set** $F$

**Linguistic Parser**

**SensiWordNet**

**Synset Polarity computation**

**Sentiment Computation**

**Subset of features** $F_i$ in $G$

**Polarity for feature1**

**Synsets in G, carrying some sentiment, referred to a feature in $F_i$**

**English Corpus** $R_E$

**Raw text**

**Dep. Graph G**

**Feature Extractor**

**Feature Set** $F$

**Linguistic Parser**

**SensiWordNet**

**Synset Polarity computation**

**Sentiment Computation**

**Subset of features** $F_i$ in $G$

**Polarity for feature1**

**Synsets in G, carrying some sentiment, referred to a feature in $F_i$**
Availability of data on the Social Web allows us to ask new questions about social behavior

- How does information spread on networks?
- How far and how fast does information flow?
- What is the structure of the network?
- Who are the influential users and communities?
- How does network structure affect the flow of information?
- What is the collective behavior of users?
- How does individual behavior affect collective behavior?
Information flow

- Users are the generators of information.
- Wide range of domains (*daily chats*, *politics*, *sports*, *celebrity gossips*, etc).
- Influence depends on the domain?

A fundamental challenge is to trace the **spread** of the information (also termed as *contamination* in literature) in the network along with the cause of its spread.
The goal of our work was to provide an unsupervised technique for estimating influence of users of social network, in a particular domain (or set of domains).

**What does it mean influence?**

The influence is intended as the capacity of a user to make the posting activity of the others similar to his/her own one.

**Why domain-based influence?**

Because an author can be very authoritative in some domain and not in others *(Is Obama an influential user when he talks about sport?)*
Following this intuition, instead of only taking into account the structural connections among users (i.e. the following/follower relationship in Twitter), we aim at discovering the nature of their relationship (the domain on which the relationship is based) based on their exchange of information about a particular domain.

3-Steps algorithm:
- Domain Classification of Tweets
- Creation of a Domain-based information Exchange Graph
- Estimation of the influence of each for each considered domain

Is Twitter just a mirror of mass sentiment or is it also able to influence opinion?
Big Data opportunities

Use cases
Big Data opportunities

Opportunities: big data use cases

360° view of the customer

- Integration of data from social networks, CRM, transactional data, etc.
- Example: T-Mobile, telecom operator -> Reduction of the customer leave of 50% in a quarter

E-reputation

- Sentiment analysis, proactive monitoring of social networks
- Example: Nestlé, food group -> Gain of 4 places in the Reputation Institute’s Index due to an interaction 24/7

Optimisation

- Predictive analysis for anomalies detection, processes optimization using sensors and operational data
- Example: Union Pacific Railroad, reduce train derailments, increase train shipment, carbon emission reduction

Public security

- Monitoring social networks, integration of spatial data and sensors
- Example: Serious Request 2012 -> monitoring of crowd movements with Twitter and sensors, localization of public force, integration with GIS
Fraud detection

• With Big Data, additional criteria may be considered: e.g. how often a user typically accesses an account from a mobile device or PC, how quickly the user types in a username and password

• New data sources: mobile data, social data
  o credit card fraud detection: customer location can be inferred from social media data (and even from their contacts’ data e.g. coworkers)
  o insurance fraud detection: mine social networks for fraud suspicions, e.g. people who claim sick leave or disability benefits while engaging in incompatible leisure / sport / travel activities
Big Data for the oil & gas industry

• More upstream data (structured / unstructured / real-time)
• Traditional data sources (e.g. equipment sensor data, maintenance records) can now be analyzed on long ranges
• Insights can be gained from new data sources, e.g. geophysical data (seismic in 2D, 3D, & 4D, weather, soil, ocean currents, ice flows), image & video data, market data, social media data, but also production / transport costs...
• Big Data analytics is the next challenge facing the development of the digital oilfield:
  – seismic image processing (4-24 TB)
  – intelligent modeling and visualization of the Earth’s structure
  – faster exploration
  – well activity characterization
  – etc.
Transports (Union Pacific Road)

- Automatic rescheduling system relying on event streaming
  - Sensors, RFID, GPS, unplanned maintenance, accidents, weather forecast, etc.
  - Displays real-time information on:
    - Trains speeds and locations, fuel consumption, problems and repair status, etc.
- Trains derailments being reduced using predictive analysis from thermometers, acoustic and visual sensors
- Dangerous conditions known in advance (40 M noise patterns evaluated every day and alerts given for any anomaly)
Conclusion: challenges

• **Semantic Information aggregation**
  – Pattern extraction from streams and cross-analysis
  – Information extraction from Linked Open Data: concepts and relations linked to the streams patterns
  – Opinion aggregation from social media and web
  – Social aspects for collaboration
  – Information aggregation: “too much data to assimilate but not enough knowledge to act”

• **Distributed and real-time processing**
  – Design of real-time and distributed algorithms for stream processing and information aggregation
  – Storage and indexation of a knowledge base
  – Integration of business processes with aggregated information
  – Distribution and parallelization of data mining algorithms

• **Visual analytics and user modeling**
  – Dynamic user model
  – Novel visualizations for very large datasets
**Our Value proposition**

- Semantic aggregation from textual and non-textual streams
- Manage semantic heterogeneity, real-time and distributed processing
- Ensure data quality and veracity
- Visual analytics
QUESTIONS?