A Methodology for SBI

MATTEO FRANCIA, MATTEO GOLFARELLI, STEFANO RIZZI
UNIVERSITY OF BOLOGNA, ITALY

Introduction to Social BI
An architecture for SBI
Methodological Framework
  - Macro-Analysis
  - Ontology Design
  - Source Selection
  - Crawling Design
  - Semantic Enrichment Design
  - ETL & OLAP Design
  - Execution & Test
Case Studies
Conclusions
User Generated Contents

- **User-generated content (UGC)** refers to a variety of media content available in a range of modern communications technologies. UGC is often produced through open collaboration.

- UGC is raising an increasing interest from decision makers:
  - Give a **fresh** and **timely** perception of the **market mood**
  - Can be used to **deliver** important messages to potential **customers**
  - **Social events** are perceived by:
    - traditional information systems when they impact on the company processes (e.g. sales reduction).
    - SBI systems when they start happening

Social BI: a Definition

- **Social Business Intelligence (SBI)** is the discipline that applies DW and OLAP approaches to the analysis of user-generated content to let decision-makers improve their business based on the trends perceived from the environment.

- As in traditional BI the goal of SBI is to enable powerful and **flexible analysis** even for decision makers with limited technical skills.
Our architecture

- **Crawling**
- **Enrichment**
- **ETL & Data cleaning**
- **DM**
- **OLAP & Dashboards**
- **Simulation**
- **Data Mining**

### Tools and Technologies
- **Apache OPEN NLP** for German
- **Brandwatch + Gnip**
- **OSS**
- **Talend DI + PL/SQL**
- **Apache OPEN NLP**
- **Ad Hoc UI** + **SAIKU**
- **8 CPUs 64 GB RAM 1TB RAID 10 (10,000Rpm)**

### Semantic Analysis
- **SynTHEMA** for Italian and English
- A well-known commercial suite that enables a linguistic and semantic analysis of any piece of textual information based on its morphology, syntax, and semantics using logical-functional rules.

### Storage and Infrastructure
- **1TB RAID 10**
- **MongoDB**
- **Oracle 11g**

### Domain Ontology
- **Crawling Design**
- **Execution**
- **ETL & Data cleaning**
- **DM**
- **OLAP & Dashboards**
- **Simulation**
- **Data Mining**

### Execution Steps
- **Domain Ontology**
- **Crawling Design**
- **Execution**
- **ETL & Data cleaning**
- **DM**
- **OLAP & Dashboards**
- **Simulation**
- **Data Mining**

### Key Enriched Clips

### Performance
- **10,000 Rpm**
Macro Analysis

- **Goal:**
  - **Project Scope**
    - domain of interest for the users
  - **Inquiries**
    - captures an informative need of a user (What? How? Where?)
    - drive the definition of Themes and Topics
  - **Activities:**
    - Interview/non technical meeting with users

Ontology Design

- **Goal:**
  - Describes the project scope.
  - Key input for almost all process phases
- **Activities:**
  - Detecting domain-relevant
    - topics
    - alias
    - themes
  - and organizing them into a hierarchy
Source Selection

- Activities:
  - Identify as many relevant Web domains as possible to crawl.
  - Backlinks analysis
  - Primary sources
  - Searching the Web using keywords
  - Communication channels
  - Themes
  - Generalist sources (online versions of major publications)

- Finding set:
  - Reducing the set of clips by a trade-off
  - Optimizing the effort for analyzing the retrieved clips
  - Very focused on the project scope

Crawling Design

- Goal:
  - Retrieving in-topic clips by filtering off-topic clips out

- Activities:
  - Template Design (clipping)
  - Query Design
  - Content Relevance Analysis
  - Sometimes is useful to release some constraints
    - Filter clip at a later stage
Semantic Enrichment Design

- **Goal:**
  - Increase the accuracy of text analytics

- **Activities:**
  - Dictionary enrichment
  - entity
  - alias
  - entity/multi-word polarization
  - Inter-word relation definition

ETL & OLAP Design

- **Goal:**
  - Design and develop the analytics front-end and specific analysis metrics

- **Activities:**
  - ETL&OLAP design, depends on
    - semantic engine features
    - presence of specific data acquisition channels (CRM, enterprise db, etc.)
  - KPI design, depends on
    - Users informative needs
  - Both depends on metadata richness and availability
Execution & Testing

- Has a basic role in the methodology
- **Coverage Analysis**
  - Measure the **ontology maturity level**
  - percentage of clips that include at least one ontology topic
- **Correctness Analysis**
  - Measure **actual improvements** in the overall ability of the process in understanding a text
- **Crawling Coverage Analysis**
  - wrong query may lead to losing relevant clips
  - is a daily and critical task

Social BI Projects

- Social BI projects are characterized by:
  - Quickly changing requirements and environment
  - **Data sources** are not known a priori
    - Neither their structure
  - Project overall quality heavily depends on crawled content quality
  - **Keyword query** are in some situations rough tool
  - Cubes schema is **project independent**, mainly related to the **project domain**
In the table below activities executed in projects of higher levels are carried out in lower levels too.

<table>
<thead>
<tr>
<th>Project Type</th>
<th>Crawling</th>
<th>Semantic Enrichment</th>
<th>Storing &amp; Analysis</th>
</tr>
</thead>
<tbody>
<tr>
<td>Level 1: Best-of-Breed</td>
<td>template design</td>
<td>dictionary enrichment, inter-word relat., det.</td>
<td>ETL design and impl.</td>
</tr>
<tr>
<td>Level 2: end-to-end</td>
<td>source selection, query design, content rel.</td>
<td>polarization, correctness analysis, ontology coverage analysis</td>
<td>ontology design, KPI &amp; dashboard design</td>
</tr>
<tr>
<td>Level 3: Off-the-Shelf</td>
<td>macro-analysis</td>
<td>macro-analysis</td>
<td>macro-analysis</td>
</tr>
</tbody>
</table>

- **Case Studies**
  - **PR-CG:**
    - Level 2 (end-to-end) project
    - Domain: large consumer goods company
    - Team guided by previous experiences (not SBI)
  - **PR-Pol:**
    - Level 1 (Best of Breed) project
    - Domain: Italian politics
    - Methodology applied and enforced

- In both project the iterative approach were adopted.
### Case Studies

<table>
<thead>
<tr>
<th>Activity / Task</th>
<th>PR-CG</th>
<th>PR-POL</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1st iteration</td>
<td>Maint. iteration</td>
</tr>
<tr>
<td>Macro Analysis</td>
<td>10</td>
<td>9</td>
</tr>
<tr>
<td>Ontology Design</td>
<td>4</td>
<td>7</td>
</tr>
<tr>
<td>Topics Definition</td>
<td>2</td>
<td>5</td>
</tr>
<tr>
<td>Inter-Topic Relation Definition</td>
<td>2</td>
<td>5</td>
</tr>
<tr>
<td>Source Selection</td>
<td>3</td>
<td>5</td>
</tr>
<tr>
<td>Semantic Enrichment Design</td>
<td>7</td>
<td>5</td>
</tr>
<tr>
<td>Crawling Design</td>
<td>n.a.</td>
<td>n.a.</td>
</tr>
<tr>
<td>Template Design</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Query Design &amp; Content Relevance Analysis</td>
<td>10</td>
<td>1</td>
</tr>
<tr>
<td>ETL &amp; OLAP Design</td>
<td>15</td>
<td>24</td>
</tr>
<tr>
<td>ETL Design &amp; Implementation</td>
<td>5</td>
<td>10</td>
</tr>
<tr>
<td>KPI Design</td>
<td>5</td>
<td>7</td>
</tr>
<tr>
<td>Dashboard Design</td>
<td>5</td>
<td>7</td>
</tr>
<tr>
<td>Execution &amp; Test</td>
<td>3</td>
<td>5</td>
</tr>
<tr>
<td>Total</td>
<td>52</td>
<td>84</td>
</tr>
</tbody>
</table>

In charge of the customer: 15, 0.85, 84, 5

---

### Outcomes

- **Responsiveness** in an SBI project is not a choice but rather a necessity
  - the frequency of changes requires
  - Continuous and tight involvement of domain experts
  - Change in project managing
  - Huge effort to both end users and developers
- If a proper methodology is not adopted the main problems are:
  - a lack of synchronization between the activities, that reduced their effectiveness
  - an insufficient control on the effects of changes (side effects)
- With our methodology we tried to solve such problems through:
  - A clear organization of goals and tasks for each activity.
  - A protocol and a set of templates (not for brevity) to record and share information between activities to support collaboration
  - A set of tests to be applied
Outcomes

- Big Data raises many questions
- Storing
- OLAP with Big Data is far to be an explored topic
- Deep semantic analysis may largely increase the size of the data to be handled (70x)
- The polarization correctness has still a statistic value
  - is typically less than 70% when web/social sources are involved
  - May be higher than 90% on very specific sources, topics and vocabulary

Thank you for your attention!

matteo.francia3@unibo.it

More informations and demos on:
big.csr.unibo.it/SBI