Social Business Intelligence

5th European Business Intelligence Summer School
July 8, 2015

Matteo Golfarelli

University of Bologna - Italy

Summary

- Introduction to Social BI
- An architecture for SBI
- Data Modeling in SBI
  - MetaStar
- Our prototype
- A methodology for SBI Projects
- Conclusions
The Business Intelligence Group@UNIBO

- The Business Intelligence Group has been carrying out its research activity since 1997, mainly aiming at studying methodologies, techniques and technologies in the field of Data Analysis
  - Currently 5 researchers are involved
- More in details:
  - Business Intelligence
  - Data Warehouse
  - Simulations
  - Pervasive BI
  - Collaborative BI
- Our current research topics are related to:
  - Social BI
  - Big Data & NOSql DBMS
  - Semantic Data Warehousing
  - Data mining

Introduction to SBI
User Generated Contents

- **User-generated content (UGC)** refers to a variety of media content available in a range of modern communication technologies. UGC is often produced through open collaboration [wikipedia].
- Although, a UGC can be an audio or a video in the following we will refer to textual ones, that can be classified according to the media that generated it:
  - Internet forums
  - Blogs
  - Wikis
  - Social networks
  - Customer review sites
- UGCs do not carry only news, preferences, opinions, etc. but they are also rich of meta-data such as:
  - Geo-localization
  - Information about the authors
  - Information about the media

UGC Exploitation

- UGC is raising an increasing interest from decision makers because it can give them a fresh and timely perception of the market mood (inbound) and can be used to deliver important messages to potential customers (outbound)
  - Social events are perceived by traditional information systems when they impact on the company processes (e.g. sales reduction). Social events are perceived by SBI systems when they start happening, that can be several days/weeks/months before their effects impact the company information system
- Exploiting such opportunities requires the companies to adapt their business model to the new market that implies
  - new ways to communicate
  - new competitors
  - new consumers
- Such model is often called **Social Business Model (SBM)** since business processes are influenced by the internet user behaviors that can be captured and influenced through the analysis and production UGCs.
**UGC Exploitation**

- Many SBM related software are sailing the wave but the route is still very long

**UGC Exploitation: Digital Marketing**

- The division that is more affected by UGC is often the marketing one
  - Digital marketing divisions are growing their budgets and their relevance within the company strategy
- Marketing in the era of social network is based on **Word of Mouth** and is aimed at making the customer the main actor in communicating the value of a product or service

Source: (Malloy, 2012)
Digital Marketing: activities breakdown

COMMUNICATE
PERSONALIZE
GATHER DATA
PREDICT
ANALYZE

Source: (Malloy, 2012)

Digital Marketing: Activities Breakdown

COMMUNICATE
PERSONALIZE
GATHER DATA
PREDICT
ANALYZE

Social engagement
Search campaign management
Web site management
Email campaign management
(Other) campaign management
Ad targeting
Web site personalization
Visitor intent
Clickstream analysis
Web site analysis
Search optimization
(Other) campaign optimization
Social influence identification
Trend/anomaly detection
Segmentation
Community reaction

Source: (Malloy, 2012)
Digital Marketing: Technology Building Blocks

- Search campaign management
- Website management
- Email campaign management
- Ad targeting
- Social engagement
- Website personalization
- Ad targeting
- Consumer profile management
- Privacy management
- Text mining
- Visualization
- Search optimization
- Machine Learning
- Ontology building
- Ontology building
- Text mining
- Web site analysis
- (Other) campaign optimization
- PREDICT
- Analyze
- NLP
- Statistical Methods
- Social influence identification
- Social anomaly detection
- Source: (Malloy, 2012)

Social-Media Monitoring tools

- Many commercial tools and platforms are available for analyzing the UGC
  - Brandwatch
  - Tracx
  - Clarabridge
- They typically rely on a large but fix set of glitzy dashboards that analyze the data from set of points of view…
  - Topic usage
  - Topic correlation
  - Brand reputation
- … and using some ad-hoc KPIs
  - Topic counting (e.g. Top topic, Trending topic)
  - Sentiment and polarization
- Rely on a cloud architecture and are oriented to business users with limited capabilities in managing data
Social-Media Monitoring tools

- Social-Media Monitoring tools are often offered as-a-service
  - Are project-oriented (typically with a narrow time-horizon)
  - Lack in providing a sufficient verticalization/personalization of the system in term of dictionaries, rules, etc..
  - Provide limited capabilities for data cleaning and data enrichment
  - The historical depth of data is limited or expansive
  - Data reworking in presence of new requirements is unfeasible
  - Are perceived by companies as self-standing applications, so UGC-related analyses are run separately from those strictly related to business
  - Does not allow integration with corporate data (Grimes, 2014)
  - Lack in providing flexible and user-driven analysis (Grimes, 2014)

From Social-Media Monitoring to Social Business Intelligence

- Social-Media Monitoring process can be seen as a DW process
  - Extract semi-structured data from the web/data provider/CRM
  - Transform, enrich and clean data
  - Load data in a system oriented to data analysis

- The process is much more complex since
  - Crawling the web is not as easy as accessing the enterprise DBs
  - Data are semi-structured
  - Enrichment is based on text-mining and NLP techniques
  - Data volume could be huge
Social BI: a Definition

- **Social Business Intelligence** 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.
- In a SBI system
  - OLAP-like operators allow flexible, detailed and user-driven analysis
  - Social data becomes an asset of the company
    - Verticalization/Personalization improves analysis effectiveness
    - Data can be reworked in order to clean and enrich data as much as needed
    - Social data can be integrated with corporate data in order to better analyze the effect of social behaviors on the enterprise
Verticalization of an SBI system refers to the possibility of tuning the system based on the specific domain of listening the application is running on:

- Dictionary enrichment: words specific of the domain of listening are added to the system dictionary in order to make it able to recognize and analyze them.
- Polarization changes: refer to changing the general polarization of a word (positive or negative) in order to better capture its understanding in the specific domain of listening.
  - Fried is not negatively polarized but if we are running a project for a company working in the pre-cooked food market and we find "the fish smell of fried".
- Semantic Enrichment tuning: is aimed at improving the effectiveness of semantic enrichment phase helping the system to understand more text:
  - Syntactic relationships definition: typical of NLP approaches, new rules and principles that govern the sentence structure are added in order to allow the system to understand more clips.
    - The simplest example are multi-words such as `carta di credito`.
  - Textual patterns definition: typical of text mining approaches, improve capability of the engine by adding new text patterns.
Integration with Enterprise Data

There is no reason for considering social data different from owned one. Managers and data analysts must make a step forward and start working with unstructured, possibly imprecise data. This is just because such data are now available and have an indisputable information value.

- In a digital enterprise all the available information should freely flows within the information system.
- Directly handling the social data:
  - Adds an asset to the information system.
  - Enables a full exploitation of the social data since
    - Makes it possible to different users to analyze them according to their own goals. The value is in the data not in the reports!
    - Enable business processes to be triggered by events captured by social data.
- Directly handling the social data:
  - Requires new expertizes to be hired.
  - An enterprise-wide digital strategy to be defined.

Integration with Enterprise Data

Carrying out cross-analysis between enterprise and social data is fundamental to properly understand the impact of social events on the enterprise.

- Ex-post analysis:
  - Coupling the trend of a product sentiment with its sales.
  - Coupling customer complaints with the churn rate.
  - Coupling the level of appreciation of a marketing campaign with the sales increase.

- Ex-ante (real time):
  - Situational Awareness: understand which enterprise facts (contract, plant, delivery) are affected by an external event (strike, earthquake, accident).

Social and enterprise data integration requires:

- Use the same platform for BI and SBI analyses.
- Associate enterprise objects (i.e. products) to social ones.
Some Non-Technical Remarks

- The interest around Social-Media Monitoring and Social BI is high since companies cannot ignore UGC-derived information.

- Nowadays project costs are largely determined by the licenses of commercial tools:
  - Crawling engines for gathering data from the web
  - Semantic engines for data enrichment

- The future of Social BI will be determined by:
  - The spreading of open source software or with more affordable prices (e.g. Google Prediction API?)
  - A more clear understanding of the impact of the social environment on the performance of real enterprise
Social BI related research challenges

Crawling
- Open data
- Crawling tech.
- Linked data

Architecture
- Data Modeling
- Architecture & storage
- Ontology
- Big data
- Text-search engine

Entity recognition
- Document clustering
- Ontology
- Sentiment analysis
- Text mining

Querying expressiveness
- NLP
- Text-mining

Visual experience
- Methodology
- Business Model & value exploitation

Understanding
- Methodology
- Business Model & value exploitation

Software mash-up
- Testing
- Software mash-up
- Methodology

Managing
- Testing
- Software mash-up
- Methodology

Social BI related research challenges

Crawling
- Open data
- Crawling tech.
- Linked data

Architecture
- Data Modeling
- Architecture & storage
- Ontology
- Big data
- Text-search engine

Entity recognition
- Document clustering
- Ontology
- Sentiment analysis
- Text mining

Querying expressiveness
- NLP
- Text-mining

Visual experience
- Methodology
- Business Model & value exploitation

Understanding
- Methodology
- Business Model & value exploitation

Software mash-up
- Testing
- Software mash-up
- Methodology

Managing
- Testing
- Software mash-up
- Methodology
An Architecture for SBI

**Functional Modules**

- **ODS (Operational Data Store)** that stores all the relevant data about clips, their topics, their authors, and their source channels. For the topic ontology a triple store repository is adopted.
- **Document DB** that enables efficient free-text search.
- A **data mart** that stores clip and topic information in the form of a set of multidimensional cubes to be used for decision making.
- A **crawling component** that runs a set of keyword-based queries to retrieve the clips (and the related meta-data) that lie within the subject area.
Project types

- The components mentioned above are normally present, though with different levels of sophistication, in most current commercial solutions for SBI.

- **Level 1: Best-of-Breed** A best-of-breed policy is followed to acquire tools specialized in one of the steps necessary to transform raw clips in semantically-rich information.
  - Followed by those who run a medium to long-term project to get full control of the SBI process by finely tuning all its critical parameters.
  - Typically aimed at implementing ad-hoc reports and dashboards to enable sophisticated analyses of the UGC.

- **Level 2: End-to-End** A single software/service is acquired and tuned.
  - Customers only need to carry out a limited set of tuning activities that are typically related to the subject area, while a service provider or a system integrator ensures the effectiveness of the technical (and domain-independent) phases of the SBI process.

---

Project types

- **Level 3: Off-the-Shelf** Consists in adopting, typically in a as-a-service manner, an off-the-shelf solution supporting a set of reports and dashboards that can satisfy the most frequent user needs in the SBI area (e.g., average sentiment, top topics, trending topics, and their breakdown by source/author/sex).
  - With this approach the customer has a very limited view of the single activities that constitute the SBI process, so she has little or no chance of positively impacting on activities that are not directly related to the analysis of the final results.
The Crawling module

- This module is in charge of capturing UGCs through a set of keyword-based queries
- While this task is quite easy when the UGC sources are controlled by the enterprise (e.g. Enterprise CRM), it becomes very hard when the listening domain is the web
  - Parsing XML or JSON data
  - Split content from advertising
  - Discard duplicate contents/clips
  - Collect meta-information about sources, authors, etc.
- Approaches are mainly based on
  - Templates of the web-source pages
  - API provided by the web-source

The Semantic Enrichment module

- This module is in charge of extracting from the raw text as many information as possible
- A large amount of research has been carried out on this issue. The main tasks carried out are:
  - **Entity extraction**: locates and classifies elements in text into pre-defined categories (e.g. names of persons, organizations, locations)
  - **Relation extraction**: identifies relations between named-entities
  - **Sentiment analysis**: identifies the positive, negative or neutral polarization text. Depending on the adopted technique can be roughly computed at the clip level, or can be detailed for each words group. (Liu, 2012)
  - **Clip clustering**: identifies clusters of clips related to the same topics.
- Approaches range in
  - Statistical and Text mining
  - Machine-learning
  - Natural Language Processing
NLP – Natural Language processing

Try to achieve a complete understanding of the text

- **Morphological analysis**: analyzes morphems, the "minimal unit of meaning", that build up words.
  - **il termine “unhappyness” is made up by the prefix “un” (i.e. not), by the free morphem “happy” and by the suffix “ness” (i.e. being in a state or condition)**

- **Lexical Analysis**: associate to each word the
  - Matteo [proper noun, singular]
  - gives [to give, 3rd person singular, present]
  - a [art]
  - book [common noun, singular]
  - to [preposition]
  - Fabio [proper noun, singular]

- **Syntax Analysis**: determines the role of the terms in the sentence

- **Semantic Analysis**: determines the sentence meaning by choosing, for each word, the correct meaning exploiting syntactic relationships between terms and the sentence context

NLP – Natural Language processing

Ensures an high effectiveness when texts are properly written and do not contain error or use web dialects, etc.

- Blog
- On-line newspapers

Natural languages are ambiguous

"I'm glad I'm a man, and so is Lola!", Lola is a man or is she happy?

"John saw the man on the mountain with a telescope", Who has the telescope?

“Stolen painting found by tree”, Either a tree found a stolen painting, or a stolen painting was found sitting next to a tree.

- **Polysemy**

  John spilled coffee on the newspaper
  The physical newspaper
  The **newspaper** fired its editor
  The company that publishes the newspaper
  The newspaper has decided to change its format
  The newspaper as an edited work
The problem of storing textual documents in multidimensional form to enable OLAP analyses has been explored in the literature to some extent.

In (Lee, 2000) the authors propose a cube for analyzing term occurrences in documents belonging to a corpus.

- The categorization of terms is obtained from a thesaurus or from a concept hierarchy such as Wordnet.
- Term extraction has been carried out using a pre-processing phase.
- Document dimension includes structured bibliographic information obtained from the document XML.
- Terms categorization is very simple and nothing is said about performances.
Analysis of textual UGC

- In (Ravat, 2008) the authors propose *textual measures* as a solution to summarize textual information within a cube.

- Textual measures are classified as:
  - A raw textual measure is a measure whose content corresponds to the textual content of a document for a document fragment (e.g. the content of a scientific article in XML format stripped of all the XML tags that structure it).
  - An elaborated textual measure is a measure whose content is taken from a raw textual measure and has undergone a certain amount of processing. A textual measure such as a keyword measure is an elaborated textual measure. This kind of measure is obtained after applying processes on a raw textual measure such as withdrawing stop words and keeping the most significant ones regarding the document’s context.

The top-keywords are calculated associating a weight to the terms $t$ belonging to texts associated with a specific cell $c_{ij}$ of an aggregated cube (i.e. the $c_{ij}$ becomes the corpus)

- Pre-processing of texts eliminate stop-words
- A scoring function, such as tf-idf, is needed
  - $TF$ Term-frequency of a term $t$ within documents belonging to cell $i,j$
  - $IDF$ Inverse document frequency: representativeness of a term $t$ within documents belonging to cell $i,j$

\[
    w_{ij} = tf_{ij}(t) \times idf_{ij}(t)
\]

\[
    tf_{ij}(t) = \frac{n_{ij}(t)}{n_{ij}}
    \quad idf_{ij}(t) = \log \frac{d_{ij} + 1}{d_{ij}(t)}
\]

- $n_{ij}$ is the number of terms in $c_{ij}$ documents
- $n_{ij}(t)$ is the number of occurrence of $t$ in $c_{ij}$ documents
- $d_{ij}$ is the number of documents in $c_{ij}$
- $d_{ij}(t)$ is the number of documents including $t$ in $c_{ij}$

Top-keyword is an holistic operator, thus its computation requires the original data (i.e. materialized views cannot be used to optimize performances)
Analysis of textual UGC

- In (Rehman 2012) the authors propose a complete architecture for OLAP analysis of tweets
  - A multidimensional model is proposed
    - Only Twitter meta data are exploited and no-topic hierarchy is included
    - The cube is loaded with tweets related to a specific hashtag or topic. The selection of the relevant tweet is carried out at the ETL level.

- In (Garcia-Moya 2013) the authors propose a complete architecture for Social BI
  - Corporate and social cubes are integrated through conformed hierarchies
Analysis of textual UGC

- In (Garcia-Moya 2013) the authors propose a complete architecture for Social BI
  - Corporate and social cubes are integrated through conformed hierarchies
  - Automatic Semantic Annotation is carried out in order to semantically enrich both the corporate and social data

Corporate data and meta-data related to Product are annotated with wikipedia semantic information
Analysis of textual UGC

In (Garcia-Moya 2013) the authors propose a complete architecture for Social BI

- Corporate and social cubes are integrated through conformed hierarchies
- Automatic Semantic Annotation is carried out in order to semantically enrich both the corporate and social data

The lexicon related to semantically annotated concepts is stored \( L_{\text{Product}} \) together with those corporate terms that are not found on Wikipedia \( L_{\text{Miss}} \)
In (Garcia-Moya 2013) the authors propose a complete architecture for Social BI:

- Corporate and social cubes are integrated through conformed hierarchies
- Automatic Semantic Annotation is carried out in order to semantically enrich both the corporate and social data

Generated information are semantically annotated too.
In (Garcia-Moya 2013) the authors propose a complete architecture for Social BI

- Corporate and social cubes are integrated through conformed hierarchies
- Automatic Semantic Annotation is carried out in order to semantically enrich both the corporate and social data

![Diagram of Social BI architecture]

Exploiting all the information collected so-far semantic annotation are simplified and cleaned in order to store only the useful and complete ones.

In (Dayal 2012) the authors propose a complete architecture for Social BI

- A first advanced solution for modeling the topic hierarchy is proposed
- The authors stress the need to couple enterprise data with social one in order to achieve situational awareness for enterprise related events.
- Topic hierarchies are recognized to be different from traditional OLAP hierarchy from several point of views
- Topic hierarchies are modeled as parent-child tables
Analysis of textual UGC through relevant topics

A key role in the analysis is played by topics, meant as specific concepts of interest within the subject area.
Analysis of textual UGC through relevant topics

A key role in the analysis is played by topics, meant as specific concepts of interest within the subject area.

Topics are an obvious candidate dimension of the cubes for Social BI, but:
- Trending topics are heterogeneous and change quickly over time.
- A classical dimension table with a static hierarchy is not suitable.

Positive sentiment expressed

Topic hierarchy schema

Consider a mobile-oriented scenario.
Consider a mobile-oriented scenario

Most topics can be classified into levels, that correspond to aggregation levels in traditional hierarchies

Relationships between topics highlight roll-up relationships
Topic hierarchy schema

How is a topic hierarchy different from a traditional hierarchy?

1. **Dynamicity**: new topics, relationships and aggregation levels might be added at any time.

2. **Mixed granularity** (facts associated to non-leaf-topics) and **unbalanced hierarchies**.
How is a topic hierarchy different from a traditional hierarchy?

1. **Dynamicity**: new topics, relationships and aggregation levels might be added at any time
2. **Mixed granularity** (facts associated to non leaf-topics) and unbalanced hierarchies
3. **Integration**: some topics are also part of business hierarchies and require a direct connection with the enterprise cube
4. **Semantics**: roll-up relationships between topics can have different semantics
The Meta-Star approach

- Meta-Stars overcome these issues by using a combination of modeling strategies (Gallinucci, 2013; Gallinucci, 2015)

- Navigation tables
  - Support hierarchy instances with different lengths and non-leaf facts
  - Allow different roll-up semantics to be explicitly annotated

- Meta-modeling
  - Enable hierarchy heterogeneity and dynamicity to be accommodated

- Traditional dimension tables
  - Easy integration with standard business hierarchies

---

The Meta-Star approach

- Implementation of a Meta-Star requires two components:

  - A Topic Table
    - Stores all the topics of the hierarchy
    - Topic levels can be modeled in a static way (i.e., like in a classical dimension table)

  - A Rollup Table
    - Stores every relationship between two topics in the transitive closure
The Meta-Star approach

Implementation of a Meta-Star: the **topic table**

- One row for each topic

<table>
<thead>
<tr>
<th>ID</th>
<th>Topic</th>
<th>Level</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>8MP Camera</td>
<td>Component</td>
</tr>
<tr>
<td>2</td>
<td>Galaxy III</td>
<td>Product</td>
</tr>
<tr>
<td>3</td>
<td>Galaxy Tab</td>
<td>Product</td>
</tr>
<tr>
<td>4</td>
<td>Smartphone</td>
<td>Type</td>
</tr>
<tr>
<td>5</td>
<td>Tablet</td>
<td>Type</td>
</tr>
<tr>
<td>6</td>
<td>Mobile Tech</td>
<td>Category</td>
</tr>
<tr>
<td>7</td>
<td>Samsung</td>
<td>Brand</td>
</tr>
<tr>
<td>8</td>
<td>Finger Path.</td>
<td>-</td>
</tr>
<tr>
<td>9</td>
<td>Touchscreen</td>
<td>-</td>
</tr>
</tbody>
</table>
The Meta-Star approach

- Implementation of a Meta-Star: the topic table
  - One row for each topic
  - Columns for each static level, like in a classical dimension table

<table>
<thead>
<tr>
<th>id</th>
<th>Topic</th>
<th>Level</th>
<th>Product</th>
<th>Type</th>
<th>Category</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>8MP Camera</td>
<td>Component</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>2</td>
<td>Galaxy III</td>
<td>Product</td>
<td>Galaxy III</td>
<td>Smartphone</td>
<td>Mobile Tech</td>
</tr>
<tr>
<td>3</td>
<td>Galaxy Tab</td>
<td>Product</td>
<td>Galaxy Tab</td>
<td>Tablet</td>
<td>Mobile Tech</td>
</tr>
<tr>
<td>4</td>
<td>Smartphone</td>
<td>Type</td>
<td>-</td>
<td>Smartphone</td>
<td>Mobile Tech</td>
</tr>
<tr>
<td>5</td>
<td>Tablet</td>
<td>Type</td>
<td>-</td>
<td>Tablet</td>
<td>Mobile Tech</td>
</tr>
<tr>
<td>6</td>
<td>Mobile Tech</td>
<td>Category</td>
<td>-</td>
<td>-</td>
<td>Mobile Tech</td>
</tr>
<tr>
<td>7</td>
<td>Samsung</td>
<td>Brand</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>8</td>
<td>Finger Path.</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>9</td>
<td>Touchscreen</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>

The Meta-Star approach

- Implementation of a Meta-Star: the roll-up table
  - One row for each arc in the transitive closure of the hierarchy

<table>
<thead>
<tr>
<th>ChildId</th>
<th>RollUpSignature</th>
<th>FatherId</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>000000</td>
<td>1</td>
</tr>
<tr>
<td>2</td>
<td>000000</td>
<td>2</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>1</td>
<td>100000</td>
<td>2</td>
</tr>
<tr>
<td>2</td>
<td>010000</td>
<td>4</td>
</tr>
<tr>
<td>3</td>
<td>001000</td>
<td>7</td>
</tr>
<tr>
<td>4</td>
<td>000100</td>
<td>6</td>
</tr>
<tr>
<td>5</td>
<td>000001</td>
<td>9</td>
</tr>
<tr>
<td>6</td>
<td>000010</td>
<td>9</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>1</td>
<td>110000</td>
<td>4</td>
</tr>
<tr>
<td>2</td>
<td>110000</td>
<td>5</td>
</tr>
<tr>
<td>3</td>
<td>101000</td>
<td>7</td>
</tr>
<tr>
<td>4</td>
<td>100010</td>
<td>9</td>
</tr>
<tr>
<td>5</td>
<td>010100</td>
<td>6</td>
</tr>
<tr>
<td>6</td>
<td>010100</td>
<td>6</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>1</td>
<td>110100</td>
<td>6</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
</tbody>
</table>
**The Meta-Star approach**

- Implementation of a Meta-Star: the roll-up table
  - One row for each arc in the transitive closure of the hierarchy

<table>
<thead>
<tr>
<th>ChildId</th>
<th>RollUpSignature</th>
<th>FatherId</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>00000</td>
<td>1</td>
</tr>
<tr>
<td>2</td>
<td>00000</td>
<td>2</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>1</td>
<td>10000</td>
<td>3</td>
</tr>
<tr>
<td>2</td>
<td>01000</td>
<td>4</td>
</tr>
<tr>
<td>2</td>
<td>001000</td>
<td>7</td>
</tr>
<tr>
<td>4</td>
<td>000100</td>
<td>6</td>
</tr>
<tr>
<td>8</td>
<td>000001</td>
<td>9</td>
</tr>
<tr>
<td>2</td>
<td>0000010</td>
<td>9</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>1</td>
<td>110000</td>
<td>4</td>
</tr>
<tr>
<td>1</td>
<td>110000</td>
<td>5</td>
</tr>
<tr>
<td>1</td>
<td>101000</td>
<td>7</td>
</tr>
<tr>
<td>1</td>
<td>100010</td>
<td>9</td>
</tr>
<tr>
<td>2</td>
<td>010100</td>
<td>6</td>
</tr>
<tr>
<td>3</td>
<td>010100</td>
<td>6</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>5</td>
<td>110100</td>
<td>6</td>
</tr>
</tbody>
</table>

- Each bit of the roll-up signature corresponds to one roll-up semantics
- If the hierarchy includes a directed path from $t_1$ to $t_2$, the bits corresponding to the involved roll-up semantics are set to 1

**Slowly-changing Meta-Star**

- Meta-stars natively support data and schema changes, keeping track of the different version require further expedient
- Different techniques can be adopted. According to Kimball terminology
  - **Type-2 solution**: data versions are tracked by creating multiple tuples in the dimension table for the same natural key. No changes to the star schema is needed
  - **Full logging**: a couple of timestamps is added to dimension tables to explicitly model the temporal validity of each version so as to enable more expressive queries

- While handling different data versions is essentially a technical problem, dealing with changes in the schema of hierarchies is still a research issue, with only a few proposed solutions in the literature (e.g., Golfarelli, 06).
Slowly-changing Meta-Star

- Type-2 solution does not impact on the meta-star schema and is implemented by properly setting the ETL process only
- Full logging impacts on the meta-star schema
  - Tracking changes in the roll-up partial order requires timestamps in the roll-up table only
  - All the other operations also involve the topic table since a change in a topic/level must be reflected in all the related arcs (i.e., in the roll-up table)

<table>
<thead>
<tr>
<th>ID</th>
<th>Topic</th>
<th>Level</th>
<th>From</th>
<th>To</th>
<th>Master</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>8MP Camera</td>
<td>Component</td>
<td>Jan 01 2014</td>
<td>-</td>
<td>1</td>
</tr>
<tr>
<td>2</td>
<td>Galaxy III</td>
<td>Product</td>
<td>Jan 01 2014</td>
<td>-</td>
<td>2</td>
</tr>
<tr>
<td>3</td>
<td>Galaxy Tab</td>
<td>Product</td>
<td>Jan 01 2014</td>
<td>-</td>
<td>3</td>
</tr>
<tr>
<td>4</td>
<td>Smartphone</td>
<td>Type</td>
<td>Jan 01 2014</td>
<td>-</td>
<td>4</td>
</tr>
<tr>
<td>5</td>
<td>Tablet</td>
<td>Type</td>
<td>Jan 01 2014</td>
<td>-</td>
<td>5</td>
</tr>
<tr>
<td>6</td>
<td>Mobile Tech</td>
<td>Category</td>
<td>Jan 01 2014</td>
<td>-</td>
<td>6</td>
</tr>
<tr>
<td>7</td>
<td>Samsung</td>
<td>Brand</td>
<td>Jan 01 2014</td>
<td>-</td>
<td>7</td>
</tr>
<tr>
<td>8</td>
<td>Finger Path.</td>
<td>-</td>
<td>Jan 01 2014</td>
<td>-</td>
<td>8</td>
</tr>
<tr>
<td>9</td>
<td>Touchscreen</td>
<td>-</td>
<td>Jan 01 2014</td>
<td>-</td>
<td>9</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
</tbody>
</table>

Slowly-changing Meta-Star

<table>
<thead>
<tr>
<th>ChildId</th>
<th>RollUpSig</th>
<th>FatherId</th>
<th>From</th>
<th>To</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>00000000</td>
<td>1</td>
<td>Jan 01 14</td>
<td>-</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>1</td>
<td>1000000</td>
<td>2</td>
<td>Jan 01 14</td>
<td>Jan 31 14</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>2</td>
<td>0100000</td>
<td>4</td>
<td>Jan 01 14</td>
<td>Jan 31 14</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>3</td>
<td>0001000</td>
<td>6</td>
<td>Jan 01 14</td>
<td>Jan 31 14</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>4</td>
<td>0000001</td>
<td>6</td>
<td>Feb 01 14</td>
<td>-</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>10</td>
<td>0100000</td>
<td>10</td>
<td>Feb 01 14</td>
<td>-</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>ChildId</th>
<th>RollUpSig</th>
<th>FatherId</th>
<th>From</th>
<th>To</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0000000</td>
<td>1</td>
<td>Jan 01 14</td>
<td>-</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>1</td>
<td>10000000</td>
<td>2</td>
<td>Jan 01 14</td>
<td>Jan 31 14</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>2</td>
<td>01000000</td>
<td>4</td>
<td>Jan 01 14</td>
<td>Jan 31 14</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>3</td>
<td>00010000</td>
<td>6</td>
<td>Jan 01 14</td>
<td>Jan 31 14</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>4</td>
<td>00000010</td>
<td>6</td>
<td>Feb 01 14</td>
<td>-</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>10</td>
<td>01000000</td>
<td>10</td>
<td>Feb 01 14</td>
<td>-</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
</tbody>
</table>
Querying Meta-Stars

- The query class we adopt is the typical GPSJ:
  - Generalized Projection (group by)
  - Selection (where)
  - Join (from)

- The total number of positive occurrences in clips written in Italian during each month of 2014 and for each Media Type can be represented in a traditional cube as
  - Generalized Projection Media Type, Month
  - Selection Year = '2014' AND Language = 'Italian'
  - Join OCCURRENCE, CLIP

- While the query semantic is clear when standard hierarchies are involved, things become more complex when the full expressivity of topic hierarchy is involved…

---

Querying Meta-Stars

- Question: what is the current average sentiment over smartphones?
  - Facts can be associated to non-leaf topics
  - Result’s meaning is highly influenced by the involved semantics
Querying Meta-Stars

- Question: what is the current average sentiment over smartphones?
  - Facts can be associated to non-leaf topics
  - Result’s meaning is highly influenced by the involved semantics
Querying Meta-Stars

Question: what is the current average sentiment over smartphones?

- Facts can be associated to non-leaf topics
- Result’s meaning is highly influenced by the involved semantics
- Not all the topics could be associated to a level

We distinguish two main type of queries

- **Schema-free topic query**: the GP component is defined by topics
  - For each of the given topics a group is defined
- **Schema-aware topic query**: the GP component is defined by levels
  - As in traditional OLAP query a group is created for each topic of the given level
Querying Meta-Stars

We distinguish two main type of queries
- **Schema-free topic query**: the GP component is defined by topics
  - For each of the given topics a group is defined
- **Schema-aware topic query**: the GP component is defined by levels
  - As in traditional OLAP query a group is created for each topic of the given level

The semantic filters determines the composition of each group
- **Queries without topic aggregation**: only facts related to the specific topic are considered

---

**Smartphone**
- **Galaxy III Lumia 920**
- **Samsung**
- **Nokia**

**Product**
- **8MP Camera**
- **4.8in Display**

**Brand**
- **Mobile Tech**

**Type**
- **Tablet E5**
- **Tablet**

**Category**
- **Brand**
- **Type**
- **Product**
- **Component**
We distinguish two main type of queries:

- **Schema-free topic query**: the GP component is defined by topics
  - For each of the given topics a group is defined

- **Schema-aware topic query**: the GP component is defined by levels
  - As in traditional OLAP query a group is created for each topic of the given level

The semantic filters determines the composition of each group

- **Queries with full topic aggregation**: no filter on semantics is applied
  - It is the standard OLAP semantic

**Semantic Topic Aggregation**: semantic filter is user-defined

In the diagram:

- **Component**: 4.8in Display, 8MP Camera
- **Brand**: Lumia 920, Galaxy III
- **Type**: Smartphone, Tablet
- **Category**: Mobile Tech
- **Product**: E5, Samsun, Nokia
Querying Meta-Stars

- **Question:** what is the current average sentiment over smartphones?
  - Example of query with full-topic aggregation

![Diagram](image-url)

FT

<table>
<thead>
<tr>
<th>AvgSentim</th>
<th>IdT</th>
</tr>
</thead>
<tbody>
<tr>
<td>+5</td>
<td>1</td>
</tr>
<tr>
<td>-3</td>
<td>1</td>
</tr>
<tr>
<td>+2</td>
<td>1</td>
</tr>
<tr>
<td>+15</td>
<td>2</td>
</tr>
<tr>
<td>-20</td>
<td>2</td>
</tr>
<tr>
<td>...</td>
<td>3</td>
</tr>
<tr>
<td>+15</td>
<td>4</td>
</tr>
<tr>
<td>+20</td>
<td>4</td>
</tr>
<tr>
<td>...</td>
<td>5</td>
</tr>
<tr>
<td>...</td>
<td>6</td>
</tr>
<tr>
<td>...</td>
<td>7</td>
</tr>
<tr>
<td>...</td>
<td>8</td>
</tr>
</tbody>
</table>

ROLLUP T

<table>
<thead>
<tr>
<th>ChildId</th>
<th>RU_Sign</th>
<th>FatherId</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>000000</td>
<td>1</td>
</tr>
<tr>
<td>2</td>
<td>000000</td>
<td>2</td>
</tr>
<tr>
<td>4</td>
<td>000000</td>
<td>4</td>
</tr>
<tr>
<td>7</td>
<td>000000</td>
<td>7</td>
</tr>
<tr>
<td>1</td>
<td>100000</td>
<td>2</td>
</tr>
<tr>
<td>1</td>
<td>100000</td>
<td>3</td>
</tr>
<tr>
<td>2</td>
<td>010000</td>
<td>4</td>
</tr>
<tr>
<td>2</td>
<td>010000</td>
<td>7</td>
</tr>
<tr>
<td>4</td>
<td>001000</td>
<td>6</td>
</tr>
<tr>
<td>8</td>
<td>000001</td>
<td>9</td>
</tr>
<tr>
<td>2</td>
<td>000100</td>
<td>9</td>
</tr>
<tr>
<td>1</td>
<td>110000</td>
<td>4</td>
</tr>
<tr>
<td>1</td>
<td>110000</td>
<td>5</td>
</tr>
<tr>
<td>3</td>
<td>010000</td>
<td>7</td>
</tr>
<tr>
<td>1</td>
<td>100100</td>
<td>9</td>
</tr>
<tr>
<td>2</td>
<td>010100</td>
<td>6</td>
</tr>
<tr>
<td>3</td>
<td>010100</td>
<td>6</td>
</tr>
<tr>
<td>1</td>
<td>110100</td>
<td>6</td>
</tr>
</tbody>
</table>

BITAND ( RU_Sign, 010000 ) = RU_Sign

---

Querying Meta-Stars

- **Question:** what is the current average sentiment over smartphones?
  - Example of query with semantic topic aggregation

BITAND ( RU_Sign, 010000 ) = RU_Sign

![Diagram](image-url)

FT

<table>
<thead>
<tr>
<th>AvgSentim</th>
<th>IdT</th>
</tr>
</thead>
<tbody>
<tr>
<td>+5</td>
<td>1</td>
</tr>
<tr>
<td>-3</td>
<td>1</td>
</tr>
<tr>
<td>+2</td>
<td>1</td>
</tr>
<tr>
<td>+15</td>
<td>2</td>
</tr>
<tr>
<td>-20</td>
<td>2</td>
</tr>
<tr>
<td>...</td>
<td>3</td>
</tr>
<tr>
<td>+15</td>
<td>4</td>
</tr>
<tr>
<td>+20</td>
<td>4</td>
</tr>
<tr>
<td>...</td>
<td>5</td>
</tr>
<tr>
<td>...</td>
<td>6</td>
</tr>
<tr>
<td>...</td>
<td>7</td>
</tr>
<tr>
<td>...</td>
<td>8</td>
</tr>
</tbody>
</table>

ROLLUP T

<table>
<thead>
<tr>
<th>ChildId</th>
<th>RU_Sign</th>
<th>FatherId</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>000000</td>
<td>1</td>
</tr>
<tr>
<td>2</td>
<td>000000</td>
<td>2</td>
</tr>
<tr>
<td>4</td>
<td>000000</td>
<td>4</td>
</tr>
<tr>
<td>7</td>
<td>000000</td>
<td>7</td>
</tr>
<tr>
<td>1</td>
<td>100000</td>
<td>2</td>
</tr>
<tr>
<td>1</td>
<td>100000</td>
<td>3</td>
</tr>
<tr>
<td>2</td>
<td>010000</td>
<td>4</td>
</tr>
<tr>
<td>2</td>
<td>010000</td>
<td>7</td>
</tr>
<tr>
<td>4</td>
<td>001000</td>
<td>6</td>
</tr>
<tr>
<td>8</td>
<td>000001</td>
<td>9</td>
</tr>
<tr>
<td>2</td>
<td>000100</td>
<td>9</td>
</tr>
<tr>
<td>1</td>
<td>110000</td>
<td>4</td>
</tr>
<tr>
<td>1</td>
<td>110000</td>
<td>5</td>
</tr>
<tr>
<td>3</td>
<td>010000</td>
<td>7</td>
</tr>
<tr>
<td>1</td>
<td>100100</td>
<td>9</td>
</tr>
<tr>
<td>2</td>
<td>010100</td>
<td>6</td>
</tr>
<tr>
<td>3</td>
<td>010100</td>
<td>6</td>
</tr>
<tr>
<td>1</td>
<td>110100</td>
<td>6</td>
</tr>
</tbody>
</table>
Querying Meta-Stars

- Question: what is the current average sentiment over smartphones?
  - Example of query with full topic aggregation using static levels

<table>
<thead>
<tr>
<th>FT</th>
<th>AvgSentiment</th>
<th>IdT</th>
</tr>
</thead>
<tbody>
<tr>
<td>+5</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>-3</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>+2</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>+15</td>
<td>2</td>
<td></td>
</tr>
<tr>
<td>20</td>
<td>2</td>
<td></td>
</tr>
<tr>
<td>...</td>
<td>3</td>
<td></td>
</tr>
<tr>
<td>-15</td>
<td>4</td>
<td></td>
</tr>
<tr>
<td>+20</td>
<td>4</td>
<td></td>
</tr>
<tr>
<td>...</td>
<td>5</td>
<td></td>
</tr>
<tr>
<td>...</td>
<td>6</td>
<td></td>
</tr>
<tr>
<td>...</td>
<td>7</td>
<td></td>
</tr>
<tr>
<td>...</td>
<td>7</td>
<td></td>
</tr>
<tr>
<td>...</td>
<td>8</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>IdT</th>
<th>Topic</th>
<th>Level</th>
<th>Type</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>BMP Camera</td>
<td>Component</td>
<td>-</td>
</tr>
<tr>
<td>2</td>
<td>Galaxy III</td>
<td>Product</td>
<td>Smartphone</td>
</tr>
<tr>
<td>3</td>
<td>Galaxy Tab</td>
<td>Product</td>
<td>Tablet</td>
</tr>
<tr>
<td>4</td>
<td>Smartphone</td>
<td>Type</td>
<td>Smartphone</td>
</tr>
<tr>
<td>5</td>
<td>Tablet</td>
<td>Type</td>
<td>Tablet</td>
</tr>
<tr>
<td>6</td>
<td>Mobile Tech</td>
<td>Category</td>
<td>-</td>
</tr>
<tr>
<td>7</td>
<td>Samsung</td>
<td>Brand</td>
<td>-</td>
</tr>
<tr>
<td>8</td>
<td>Finger Path.</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>9</td>
<td>Touchscreen</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>

SQL Translations

**Schema-free query without topic aggregation**

*Total number of occurrences for two topics on June 22 2013*

```
SELECT TOPIC_T.Topic, SUM(FT.totalOcc) FROM TOPIC_T, DTCLIP, FT WHERE FT.IdT = TOPIC_T.IdT AND FT.IdC = DTCLIP.IdC AND T.Topic IN ('Touchscreen', 'Finger Pathologies') AND DTCLIP.Date = '06/22/2013' GROUP BY TOPIC_T.Topic;
```

**Schema-aware query without topic aggregation**

*Total number of occurrences for each brand on June 22 2013*

```
SELECT TOPIC_T.Topic, SUM(FT.totalOcc) FROM TOPIC_T, DTCLIP, FT WHERE FT.IdT = TOPIC_T.IdT AND FT.IdC = DTCLIP.IdC AND TOPIC_T.Level = 'Brand' AND DTCLIP.Date = '06/22/2013' GROUP BY TOPIC_T.Topic;
```
SQL Translations

Schema-aware query with semantic topic aggregation

Total number of occurrences for each brand considering only semantic hasBrand

```
SELECT TOPIC_T.Topic, SUM(FT.positiveOcc), SUM(FT.negativeOcc)
FROM TOPIC_T, ROLLUP_T, FT
WHERE FT.IdT = ROLLUP_T.ChildId AND ROLLUP_T.FatherId = TOPIC_T.IdT AND
TOPIC_T.Level = 'Brand' AND
BITAND(ROLLUP_T.RollUpSignature,001000) = ROLLUP_T.RollUpSignature
GROUP BY TOPIC_T.Topic;
```

Schema-aware query with full-topic aggregation and selection

Average sentiment for each type of category "Mobile Tech"

```
SELECT T2.Topic, AVG(FT.avgSentiment)
FROM TOPIC_T T1, TOPIC_T T2, ROLLUP_T R, FT
WHERE FT.IdT = R.ChildId AND T1.IdT = R.FatherId AND R.ChildId = T2.IdT AND
T1.Topic = 'Mobile Tech' AND T2.Level = 'Type'
GROUP BY T2.Topic;
```

Meta-star vs Star access plans

Meta-star vs Star access plans

- Meta-star
  - Schema-aware
    - [Diagram]
  - Schema-free
    - [Diagram]

- Star Schema
  - [Diagram]
Performances of Meta-Star are compared with traditional star schemata using queries with semantic topic aggregation.

<table>
<thead>
<tr>
<th>Topic hier. (Group-by)</th>
<th>FT1</th>
<th>FT2</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Meta-star</td>
<td>Star s.</td>
</tr>
<tr>
<td>H1</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>H2</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>H3</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>2</td>
</tr>
</tbody>
</table>

FT1 : 1M facts
FT2 : 10M facts
Performances of Meta-Star are compared with traditional star schemata using queries with semantic topic aggregation.

### Evaluation

<table>
<thead>
<tr>
<th>Topic hier.</th>
<th>Group-by</th>
<th>FT1</th>
<th>FT2</th>
</tr>
</thead>
<tbody>
<tr>
<td>H1 0</td>
<td>Meta-star</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>H1 1</td>
<td>Meta-star</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>H1 2</td>
<td>Meta-star</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>H2 0</td>
<td>Meta-star</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>H2 1</td>
<td>Meta-star</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>H2 2</td>
<td>Meta-star</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>H3 0</td>
<td>Meta-star</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>H3 1</td>
<td>Meta-star</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>H3 2</td>
<td>Meta-star</td>
<td>2</td>
<td>2</td>
</tr>
</tbody>
</table>

### Evaluation

Performances of Meta-Star are compared with traditional star schemata using queries with semantic topic aggregation.

**Number of levels in the group-by predicate**

<table>
<thead>
<tr>
<th>Topic hier.</th>
<th>Group-by</th>
<th>FT1</th>
<th>FT2</th>
</tr>
</thead>
<tbody>
<tr>
<td>H1 0</td>
<td>Meta-star</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>H1 1</td>
<td>Meta-star</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>H1 2</td>
<td>Meta-star</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>H2 0</td>
<td>Meta-star</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>H2 1</td>
<td>Meta-star</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>H2 2</td>
<td>Meta-star</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>H3 0</td>
<td>Meta-star</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>H3 1</td>
<td>Meta-star</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>H3 2</td>
<td>Meta-star</td>
<td>2</td>
<td>2</td>
</tr>
</tbody>
</table>
Performances of Meta-Star are compared with traditional star schemata using queries with semantic topic aggregation

- Tests run using the Oracle 11g RDBMS on a quad-core machine
- Each execution time (in seconds) is the average time of 3 different queries with different selection predicates

<table>
<thead>
<tr>
<th>Topic hier.</th>
<th>Group-by</th>
<th>PT1</th>
<th>PT2</th>
</tr>
</thead>
<tbody>
<tr>
<td>H1</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0</td>
<td></td>
<td>13.8</td>
<td>14.0</td>
</tr>
<tr>
<td>1</td>
<td></td>
<td>16.0</td>
<td>17.4</td>
</tr>
<tr>
<td>2</td>
<td></td>
<td>16.6</td>
<td>16.2</td>
</tr>
<tr>
<td>H2</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0</td>
<td></td>
<td>13.5</td>
<td>13.5</td>
</tr>
<tr>
<td>1</td>
<td></td>
<td>16.7</td>
<td>17.9</td>
</tr>
<tr>
<td>2</td>
<td></td>
<td>17.0</td>
<td>17.5</td>
</tr>
<tr>
<td>H3</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0</td>
<td></td>
<td>12.2</td>
<td>13.1</td>
</tr>
<tr>
<td>1</td>
<td></td>
<td>15.9</td>
<td>14.7</td>
</tr>
<tr>
<td>2</td>
<td></td>
<td>35.1</td>
<td>18.7</td>
</tr>
</tbody>
</table>

In most cases star schemata outperform meta-stars, but the gap is quite limited and perfectly acceptable
Performances of Meta-Star are compared with traditional star schemata using queries with Topic Aggregation

- Tests run using the Oracle 11g RDBMS on a quad-core machine
- Each execution time (in seconds) is the average time of 3 different queries with different selection predicates

<table>
<thead>
<tr>
<th>Topic hier.</th>
<th>Group-by</th>
<th>FT1</th>
<th>FT2</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Meta-star</td>
<td>Star s.</td>
</tr>
<tr>
<td>H1</td>
<td>0</td>
<td>13.8</td>
<td>12.7</td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>16.0</td>
<td>5.8</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>16.6</td>
<td>14.6</td>
</tr>
<tr>
<td>H2</td>
<td>0</td>
<td>13.5</td>
<td>13.0</td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>16.7</td>
<td>5.6</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>17.0</td>
<td>16.2</td>
</tr>
<tr>
<td>H3</td>
<td>0</td>
<td>12.2</td>
<td>9.0</td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>15.9</td>
<td>14.1</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>35.1</td>
<td>16.9</td>
</tr>
</tbody>
</table>

The execution time is mostly spent on the fact table, as the increase of execution time is proportional to the increase of the fact table size.

Performances of Meta-Star are compared with traditional star schemata using queries with Topic Aggregation

- Tests run using the Oracle 11g RDBMS on a quad-core machine
- Each execution time (in seconds) is the average time of 3 different queries with different selection predicates

<table>
<thead>
<tr>
<th>Topic hier.</th>
<th>Group-by</th>
<th>FT1</th>
<th>FT2</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Meta-star</td>
<td>Star s.</td>
</tr>
<tr>
<td>H1</td>
<td>0</td>
<td>13.8</td>
<td>12.7</td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>16.0</td>
<td>5.8</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>16.6</td>
<td>14.6</td>
</tr>
<tr>
<td>H2</td>
<td>0</td>
<td>13.5</td>
<td>13.0</td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>16.7</td>
<td>5.6</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>17.0</td>
<td>16.2</td>
</tr>
<tr>
<td>H3</td>
<td>0</td>
<td>12.2</td>
<td>9.0</td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>15.9</td>
<td>14.1</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>35.1</td>
<td>16.9</td>
</tr>
</tbody>
</table>

Execution times on the meta-star increase smoothly for group-by’s with increasing number of levels.
Evaluation

- Performances of Meta-Star are compared with traditional star schemata using queries with Topic Aggregation
  - Tests run using the Oracle 11g RDBMS on a quad-core machine
  - Each execution time (in seconds) is the average time of 3 different queries with different selection predicates

<table>
<thead>
<tr>
<th>Topic hier.</th>
<th>(Group-by)</th>
<th>PT1</th>
<th>PT2</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Meta-star</td>
<td>Star s.</td>
<td>Meta-star</td>
</tr>
<tr>
<td>H1</td>
<td>0</td>
<td>13.8</td>
<td>12.7</td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>16.0</td>
<td>5.8</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>16.6</td>
<td>14.6</td>
</tr>
<tr>
<td>H2</td>
<td>0</td>
<td>13.8</td>
<td>13.0</td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>16.7</td>
<td>5.6</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>17.0</td>
<td>16.2</td>
</tr>
<tr>
<td>H3</td>
<td>0</td>
<td>12.2</td>
<td>9.0</td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>15.9</td>
<td>14.1</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>35.1</td>
<td>16.9</td>
</tr>
</tbody>
</table>

Execution times on the meta-star increase slowly for topic and roll-up tables with increasing cardinality.

Evaluation

- Meta-star allows performances comparable (in many cases better) with implementations based on traditional star schema
  - The main cost remains accessing the fact table
  - Execution plans are similar
  - Roll-up table indexing makes its size not relevant

- The worst case for meta-star is when
  - Fine group by
  - High number of topics (1M of topic, 14M or arcs) in the roll-up table
  - Selection applied on topic

![Graph of execution time vs. selectivity for Meta-star and Star schema]
Evaluation

- Meta-star allows performances comparable (in many cases better) with implementations based on traditional star schema
  - The main cost remains accessing the fact table
  - Execution plans are similar
  - Roll-up table indexing makes its size not relevant
- The worst case for meta-star is when
  - Fine group by
  - High number of topics (1M of topic, 14M or arcs) in the roll-up table
  - Selection applied on topic

The execution time drops when the optimizer moves from full scan to indexed access of the roll-up table.
Evaluation

- Meta-star performance is paid in terms of roll-up table size
  - The size of the fact table is always predominant
  - The number of topic monitored in real SBI applications is far from reaching critical sizes
  - Sizes are reported for 10M of facts and an increasing number of topics organized on a 6 levels hierarchy

![Disk space occupation graph]

Future Works

- Coupling SQL and OWL
  Study the possibility of using the OWL language to directly query the topic hierarchy

- Summarizability for N-M relationships
  Study which summarization rationales are valid and can be adopted to produce interesting results

- OLAP front-end
  Investigate how commercial OLAP front-ends can be extended to efficiently support meta-stars
A Case Study on 2014 European Election

Our projects

- We collaborated with the following projects (carried out without adopting an ad-hoc methodology)
  - DOXA – the widest Italian Market Analysis Company
  - Amadori - the italian leader in the poultry industry
  - City mood - Bologna
- We are now running a large Social BI project within the WebPoIEU – FIRB project (http://webpoleu.altervista.org)
  - The project aims at studying the nexus between politics and social media in comparative perspective from the viewpoint of both citizens and political actors
- We monitor the European Election 2014 over three different countries (Italy, England, Germany)
  - 2 months of listening
  - 10 millions of raw clips
  - 60,000 web sources
  - 3 different languages
  - 2 full-time + 2 half-time + 5 end user
Our architecture

- We adopted an “end-to-end” solution
We adopted an “end-to-end” solution

Brandwatch + GNIP

Crawling

Semantic Enrichment

DM

OLAP & Dashb.

ETL

Crawling

ETL & Data cleaning

Semantic Enrichment

DM

OLAP & Dashb.

ETL
We adopted an “end-to-end” solution
We adopted an “end-to-end” solution.

Syntheta 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.

OPEN NLP for German

Talend + PL/SQL

Brandwatch + GNIP

8 CPUs

64 GB RAM

1 TB RAID 10 (15,000RPM)
Demo time!

Behind the scenes

- 4 cubes
- 2 enrichment techniques
  - Information Retrieval – IR scans the clips searching for topics and alias using a brute force technique
    - IR does not carry out sentiment analysis and exploit the crawlers’s polarization
  - Natural Language Processing – NLP: Exploit Syn a commercial system that carry out a deep analysis of the text
    - Morphological Analysis
    - Lexical Analysis
    - Syntax Analysis
    - Semantic Analysis
Each cell models a clip. The cubes report the measure available at the clip level.

Each cell models the occurrences of a topic within a clip.
NLP Topic Occurrence

- Each cell models the occurrence of a PosEntity within a clip
- A PosEntity is the lemma of a word + its Part of Speech (noun, verb, adverb, etc.)

NLP Topic CoOccurrence

- Each cell models the co-occurrence of two topics within the same sentence
- Co-occurrence is defined as the presence of the two topics in the same sentence even if the NLP engine did not detect a semantic relationship between them.
NLP Semantic Relationship

- Each cell models the presence of a semantic relationship between two topics or a topic and an entity in a clip.

Fact Cardinality

<table>
<thead>
<tr>
<th>Cardinality</th>
<th>ITA</th>
<th>ENG</th>
<th>DEU</th>
</tr>
</thead>
<tbody>
<tr>
<td>Topic</td>
<td>464</td>
<td>432</td>
<td>513</td>
</tr>
<tr>
<td>Alias</td>
<td>895</td>
<td>694</td>
<td>870</td>
</tr>
<tr>
<td>pos entity</td>
<td>1,262,790</td>
<td>2,902,942</td>
<td>6,093,724</td>
</tr>
<tr>
<td>entity</td>
<td>1,242,402</td>
<td>2,867,726</td>
<td>5,145,714</td>
</tr>
<tr>
<td>IR Clip</td>
<td>2,393,568</td>
<td>3,275,193</td>
<td>933,438</td>
</tr>
<tr>
<td>IR Topic Occurrence</td>
<td>15,400,783</td>
<td>25,005,664</td>
<td>16,569,387</td>
</tr>
<tr>
<td>NLP Entity Occurrence</td>
<td>226,953,012</td>
<td>519,446,526</td>
<td>524,784,320</td>
</tr>
<tr>
<td>NLP Topic Occurrence</td>
<td>14,214,686</td>
<td>23,398,601</td>
<td>7,504,815</td>
</tr>
<tr>
<td>NLP Semantic Relationship</td>
<td>17,364,421</td>
<td>23,837,474</td>
<td>38,231,162</td>
</tr>
</tbody>
</table>
Meta-star in action!

ETL flows

- Processing time (secs.) for 10 K clips.
  - Green is for NLP specific process
  - Blue is for IR specific process
- Deep semantic analysis deeply impacts on processing time and data size. This extra effort must be carefully evaluated.
Count vs sentiment

The two techniques largely agree when counting topic occurrences

<table>
<thead>
<tr>
<th>Fact</th>
<th>ITA</th>
<th>ENG</th>
</tr>
</thead>
<tbody>
<tr>
<td>IR Topic Occurrence</td>
<td>14,215 K</td>
<td>23,399 K</td>
</tr>
<tr>
<td>NLP Topic Occurrence</td>
<td>15,401 K</td>
<td>25,006 K</td>
</tr>
<tr>
<td>Shared Occurrence %</td>
<td>83.9%</td>
<td>86.0%</td>
</tr>
</tbody>
</table>

Count vs sentiment

The two techniques disagree in associating a polarization to the clips.

- Brandwatch is more conservative in assigning a non-neutral polarization
- Most of the clips that are found to be negative by Brandwatch are negative for Synthema too.

<table>
<thead>
<tr>
<th>Sentiment</th>
<th>ITA</th>
<th>ENG</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>NLP</td>
<td>IR</td>
</tr>
<tr>
<td>Positive</td>
<td>566 K</td>
<td>36 K</td>
</tr>
<tr>
<td>Neutral</td>
<td>893 K</td>
<td>2,340 K</td>
</tr>
<tr>
<td>Negative</td>
<td>934 K</td>
<td>17 K</td>
</tr>
</tbody>
</table>

We are manually assigning a polarization to the clips in order to define an effectiveness score

- Surprisingly, a draft result seems to show that the two techniques are quite close
Count vs sentiment

- The Sympton's paradox takes place! (Wagner, 1982)
  - A trend that appears in different group of data disappears or reverses when these groups are combined

- The percentage of sentences that have a sentiment different from the clip's one is 76% and 70% for English and Italian respectively

- The percentage is strictly related to the number of sentence per clip that is 20.8 in the average.
  - The disagreement drops to 15% for Twitter whose clips are composed by 1.6 sentences in the average.

- This encourages the brand reputation solutions strictly based on Twitter.

Some comments

- Twitter is largely the main (viral) clip source

- The project scope determines the quantity of data to be handled
  - In many cases storing can be handled with traditional technologies but in many others a Big data approach must be followed
  - 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
  - It is not clear (at least to me) if this extra cost has a value for customers
  - It will have one when the quality level of NLP/text mining will be high enough
  - The polarization correctness has still a statistic value
    - The polarization correctness is typically less than 70% when web/social sources are involved
    - May be higher than 90% on very specific sources, topics and vocabulary (e.g. a survey submitted to bank customers about strong and weak points of the bank)
Some comments

- Twitter is largely the main (viral) clip source
- The project scope determines the quantity of data to be handled
  - In many cases storing can be handled with traditional technologies but in many others a Big data approach must be followed
  - 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
  - It is not clear (at least to me) if this extra cost has a value for customers
  - It will be when the quality level of NLP/text mining will be high enough
  - The polarization correctness has still a statistic value
  - The polarization correctness is typically less than 70% when web/social sources are involved
  - Maybe higher than 90% on very specific sources, topics and vocabulary (e.g. a survey submitted to bank customers about strong and weak points of the bank)
- We are working towards turning our dataset in a public benchmark
  - Multi lingual (Italian, English, German)
  - Crawled using the same keywords
  - Enriched with a manually-defined ontology of topics
  - Rich of manual annotation (the top 1000 sources have been labeled and grouped according to their type type - blog, news, etc - political area)

A methodology for SBI
Social BI Projects

Social BI projects are characterized by:
- Quickly changing requirements due to topic dynamicty
- Data sources are not known a priori and keyword query are a rough tool for their selection
- Cubes schema is project independent, modeling mainly involves topic hierarchy/ontology
- Meta-star makes a change in the hierarchy not affecting the physical schema
- Project complexity depends on the type of project adopted

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. def.</td>
<td>ETL design and impl.</td>
</tr>
<tr>
<td>Level 2: end-to-end</td>
<td>source selection, query design, content rel. analysis</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>

The methodology: Macro Analysis

During this activity, users are interviewed to define the project scope and the set of inquiries the system will answer to.
- Project scope is the domain of interest for the users
  - Italian national politics
- An inquiry captures an informative need of a user
  - what? the Prime Minister
  - how? top related topics
  - where? the Wall Street Journal website

Inquiries drive the definition of:
- Themes defines and roughly classify which information is to be collected
  - Politicians & Parties
  - Policies
- Topics, their early definition is useful as
  - a foundation for designing a core taxonomy
  - a crawling query draft
The methodology: Ontology Design

- The domain ontology describes the project scope. It includes the list of topic and their relationship.
- The domain ontology becomes a key input for almost all process phases:
  - Semantic enrichment relies on it to better understand UGC meaning.
  - Crawling design benefits from topics in the ontology to develop better crawling queries and establish the content relevance.
  - ETL and OLAP design heavily uses the ontology to develop more expressive, comprehensive, and intuitive dashboards.
- The main task of this activity consists in detecting as many domain-relevant topics, alias and themes as possible and organizing them into a classification hierarchy.
  - Depending on the adopted model the classification hierarchy may have a fixed or dynamic number of levels.

The methodology: Ontology Design

- **Aliases** are alternative terms used for defining the same concept (i.e. synonymous) or slightly different concepts that we want to bring back to the same one:
  - They are part of the ontology.
  - They are not included in the topic hierarchy.
  - They are used in crawling queries.
  - During the ETL process references to aliases are linked to the corresponding topic.
- For example possible alias:
  - *Border* is an alias for *Frontier*.
  - *Lib-Dem, libdemocratic* are aliases for *Liberal Democratic*. 

Aliases are alternative terms used for defining the same concept (i.e. synonymous) or slightly different concepts that we want to bring back to the same one:
- They are part of the ontology.
- They are not included in the topic hierarchy.
- They are used in crawling queries.
- During the ETL process references to aliases are linked to the corresponding topic.
- For example possible alias:
  - *Border* is an alias for *Frontier*.
  - *Lib-Dem, libdemocratic* are aliases for *Liberal Democratic*.
The methodology: Coverage Analysis

- An effective way to measure the ontology maturity level is to use as an indicator the coverage.
  - The ontology coverage is the percentage of clips that include at least one ontology topic.

- Note that assessing the ontology coverage is made harder by off-topic clips that negatively impact on the coverage; this induces a strong connection between ontology design and crawling design.

The methodology: Source Selection

- Source selection is aimed at identifying as many web domains as possible for crawling.
- Sources can be split in two families:
  - **Primary sources:**
    - All the corporate communication channels
    - Generalist sources, such as the online version of the major publications
  - **Minor sources:**
    - Lots of small platforms which produce valuable information with high informative value because of their major focus on themes related to the subject area.
**The methodology: Source Selection**

- There are several ways for identifying the set of potentially relevant sources:
  - Conducting interviews with domain experts
  - Analyzing back-links and third-party references to the corporate communication channels
  - Searching the web using themes and topics as keywords, which can be done through search engines
  - Considering all the local editions of major newspapers
- Deciding which of them are to be actually crawled is the result of a trade-off between achieving a satisfying coverage, and optimizing the effort for analyzing the retrieved clips

**The methodology: Crawling Design**

- Crawling design aims at retrieving in-topic clips by filtering off-topic clips out.
  - **Template design** consists in an analysis of the code structure of the source website to enable the crawler to detect and extract only the informative UGC
  - Only in Level 1 projects
  - **Query design** develops a set of queries to extract the relevant clips.
  - **Content relevance analysis** aims at evaluating the effectiveness of crawling by measuring the percentage of in-topic clips.
The methodology: Crawling Design

- Filtering off-topic clips at crawling time could be difficult due to the limitations of the crawling language, and also risky because the in-topic perimeter could change during the analysis process.
- For these reasons, we can choose to release some constraints and let a wider set of clips “slip through the net”, and only filter them at a later stage.

The methodology: Semantic Enrichment Design

- Involves several tasks whose purpose is to increase the accuracy of text analytics so as to maximize the process effectiveness in terms of extracted named-entities and sentiment assigned to clips.
- Dictionary enrichment
  - Adding entity to the dictionary
  - Adding alias to existing entities
  - Changing entity polarization
- Inter-word relation definition establishes or modifies the existing semantic, and sometimes also syntactic, relations between words.
The methodology: Semantic Enrichment Design

- Modifications in the linguistic resources may produce undesired side effects
- Correctness analysis should be executed aimed at measuring the actual improvements introduced and the overall ability of the process in understanding a text and assigning the right sentiment to it
- This is normally done, using regressive test techniques, by manually tagging an incrementally-built sample set of clips with a sentiment

<table>
<thead>
<tr>
<th></th>
<th>Project 1</th>
<th>Project 2</th>
<th>Improvement</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Non-tuned</td>
<td>Tuned</td>
<td>Improvement</td>
</tr>
<tr>
<td>Total</td>
<td>54.0%</td>
<td>57.4%</td>
<td>3.4%</td>
</tr>
<tr>
<td>Social</td>
<td>52.5%</td>
<td>55.9%</td>
<td>3.4%</td>
</tr>
<tr>
<td>Qualified</td>
<td>55.0%</td>
<td>58.3%</td>
<td>3.3%</td>
</tr>
<tr>
<td>Hard</td>
<td>34.3%</td>
<td>37.2%</td>
<td>2.9%</td>
</tr>
<tr>
<td>Standard</td>
<td>67.3%</td>
<td>71.1%</td>
<td>3.8%</td>
</tr>
<tr>
<td>Negative</td>
<td>46.6%</td>
<td>46.6%</td>
<td>0.0%</td>
</tr>
<tr>
<td>Neutral</td>
<td>45.6%</td>
<td>49.1%</td>
<td>3.5%</td>
</tr>
<tr>
<td>Positive</td>
<td>69.5%</td>
<td>76.3%</td>
<td>6.8%</td>
</tr>
</tbody>
</table>

The methodology: ETL & OLAP Design

- ETL design and implementation, strongly depends on
  - features of the semantic engine
  - richness of the meta-data retrieved by the crawler (e.g., URLs, author, source type, platform type),
  - presence of specific data acquisition channels like CRM, enterprise databases, etc.
- KPI design
  - depends on which kinds of meta-data the crawler fetches
    - Clip relevance
    - Alerting
    - Author/source ranking
- Dashboard design
The methodology: Execution & Testing

- Testing has a basic role in the methodology
  - Crawling queries are executed
  - The resulting clips are processed
  - Reports are launched over the enriched clips

- The specific tests related to each single activity, can be executed separately though they are inter-related

- The most critical activity is crawling query improvements since a wrong query may lead to losing relevant clips that could be hardly retrieved later
  - Many iterations are typically required not only for tuning the queries but also for handling topic dynamicity
  - Checking query coverage is a daily task

- Improving semantic enrichment module and the domain ontology are also two time-consuming phases too

Social BI Roles

<table>
<thead>
<tr>
<th>Phase</th>
<th>Task</th>
<th>Programmer</th>
<th>Designer</th>
<th>User</th>
</tr>
</thead>
<tbody>
<tr>
<td>Crawling</td>
<td>template design</td>
<td>Exec</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>source selection</td>
<td>Exec</td>
<td>Exec</td>
<td>Exec</td>
</tr>
<tr>
<td></td>
<td>query design</td>
<td>Exec</td>
<td>Exec</td>
<td>Exec</td>
</tr>
<tr>
<td></td>
<td>content rel. analysis</td>
<td>Exec</td>
<td></td>
<td>Exec</td>
</tr>
<tr>
<td></td>
<td>macro-analysis</td>
<td>Exec</td>
<td>Exec</td>
<td>Exec</td>
</tr>
<tr>
<td>Semantic Enrichment</td>
<td>dictionary enrichment</td>
<td>Partic</td>
<td>Exec</td>
<td>Exec</td>
</tr>
<tr>
<td></td>
<td>inter-word relat. def.</td>
<td>Partic</td>
<td>Exec</td>
<td>Exec</td>
</tr>
<tr>
<td></td>
<td>polarization</td>
<td>Exec</td>
<td>Exec</td>
<td></td>
</tr>
<tr>
<td></td>
<td>correctness analysis</td>
<td>Exec</td>
<td></td>
<td>Exec</td>
</tr>
<tr>
<td></td>
<td>ontology coverage analysis</td>
<td>Exec</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Storing &amp; Analysis</td>
<td>macro-analysis</td>
<td>Exec</td>
<td>Exec</td>
<td>Exec</td>
</tr>
<tr>
<td></td>
<td>ontology design</td>
<td>Exec</td>
<td>Exec</td>
<td>Exec</td>
</tr>
<tr>
<td></td>
<td>KPI &amp; dashboard design</td>
<td>Exec</td>
<td>Partic</td>
<td>Partic</td>
</tr>
<tr>
<td></td>
<td>ETL design and impl.</td>
<td>Exec</td>
<td>Partic</td>
<td></td>
</tr>
</tbody>
</table>
# Social BI Roles

<table>
<thead>
<tr>
<th>Phase</th>
<th>Task</th>
<th>Programmer</th>
<th>Designer</th>
<th>User</th>
</tr>
</thead>
<tbody>
<tr>
<td>Crawling</td>
<td>template design</td>
<td>Exec</td>
<td>Exec</td>
<td>Exec</td>
</tr>
<tr>
<td></td>
<td>source selection</td>
<td>Exec</td>
<td>Exec</td>
<td>Exec</td>
</tr>
<tr>
<td></td>
<td>query design</td>
<td>Exec</td>
<td>Exec</td>
<td>Exec</td>
</tr>
<tr>
<td></td>
<td>content rel. analysis</td>
<td>Exec</td>
<td>Exec</td>
<td>Exec</td>
</tr>
<tr>
<td></td>
<td>macro-analysis</td>
<td>Exec</td>
<td>Exec</td>
<td>Exec</td>
</tr>
<tr>
<td></td>
<td>ontology coverage analysis</td>
<td>Exec</td>
<td>Exec</td>
<td>Exec</td>
</tr>
<tr>
<td></td>
<td>dictionary enrichment</td>
<td>Exec</td>
<td>Exec</td>
<td>Exec</td>
</tr>
<tr>
<td></td>
<td>inter-word relat. def.</td>
<td>Exec</td>
<td>Exec</td>
<td>Exec</td>
</tr>
<tr>
<td></td>
<td>polarization</td>
<td>Exec</td>
<td>Exec</td>
<td>Exec</td>
</tr>
<tr>
<td></td>
<td>correctness analysis</td>
<td>Exec</td>
<td>Exec</td>
<td>Exec</td>
</tr>
<tr>
<td></td>
<td>ontology design</td>
<td>Exec</td>
<td>Exec</td>
<td>Exec</td>
</tr>
<tr>
<td></td>
<td>KPI &amp; dashboard design</td>
<td>Exec</td>
<td>Partic</td>
<td>Partic</td>
</tr>
<tr>
<td></td>
<td>ETL design and impl.</td>
<td>Exec</td>
<td>Partic</td>
<td>Partic</td>
</tr>
</tbody>
</table>

### The programmer
The programmer, besides traditional BI skills, needs competences in the Information Retrieval, Text Mining, and NLP areas.

### The designer
The designer is a real Social BI expert and must be able to drive the customer in all the project specific choices that range from properly choosing the crawling keywords to correctly organizing the topic ontology.
Outcomes

- Responsiveness in an SBI project is not a choice but rather a necessity, since the frequency of changes requires a tight involvement of domain experts to detect these changes and rapid iterations to keep the process well-tuned.
  - With reference to traditional BI projects a radical change in the project management approach is needed as well as a huge effort to both end users and developers (about one full-time person in both our projects)
- 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
- 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 discussed in this paper for brevity) to record and share information between activities.
  - A set of tests to be applied. The definition of each test includes the testing method and the indicators that measure the test results, for instance in terms of correctness of a process phase, as well as how these results have improved over the previous iteration.

Thank you for your attention

Matteo Golfarelli
matteo.golfarelli@unibo.it
Bibliography


