Social Business Intelligence

5th European Business Intelligence Summer School
July 8, 2015

Matteo Golfarelli

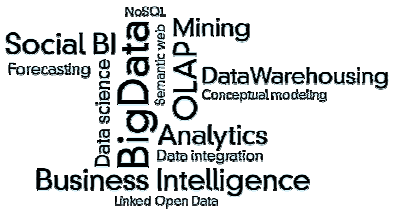
University of Bologna - Italy

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



Business Intelligence Group

Introduction to SBI

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 [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

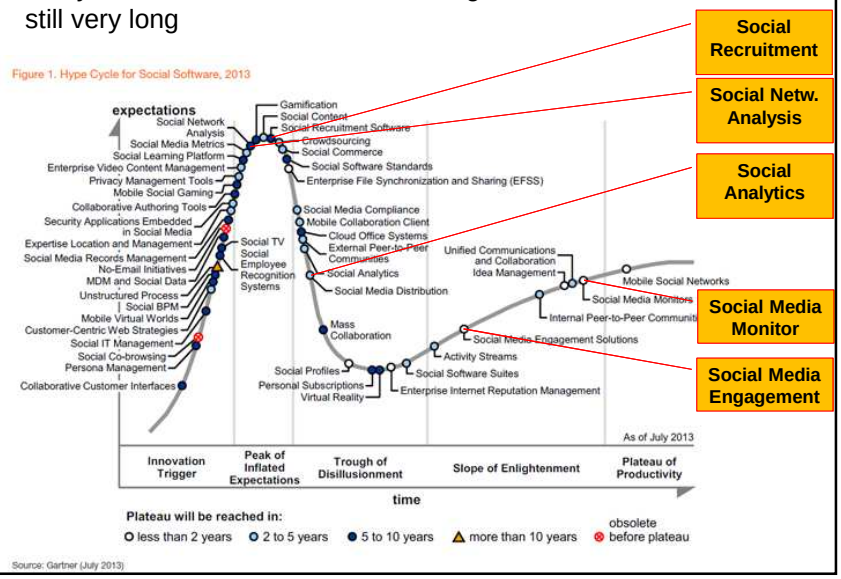
- UGCs do not carry only news, preferences, opinions, etc. but they are also rich of meta-data such as:
  - ✓ Geo-localization
  - ✓ Information about the authors
  - ✓ Information about the media

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.

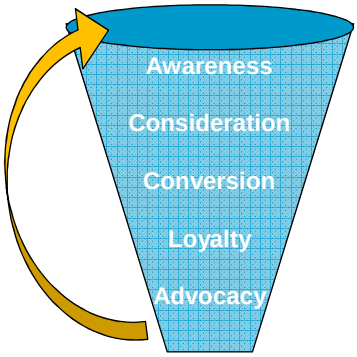
UGC Exploitation

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



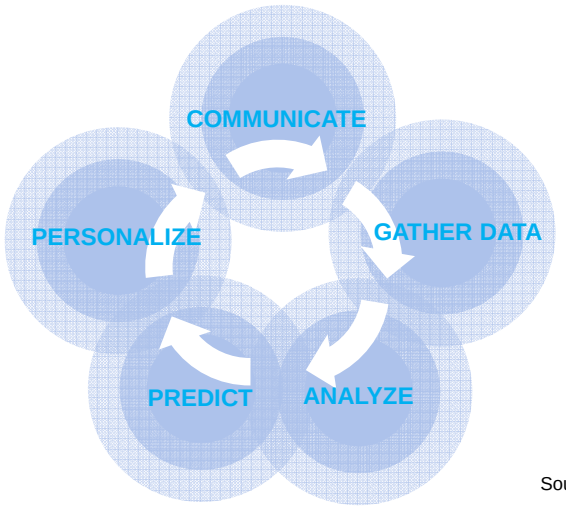
UGC Exploitation: Digital Marketing

- 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



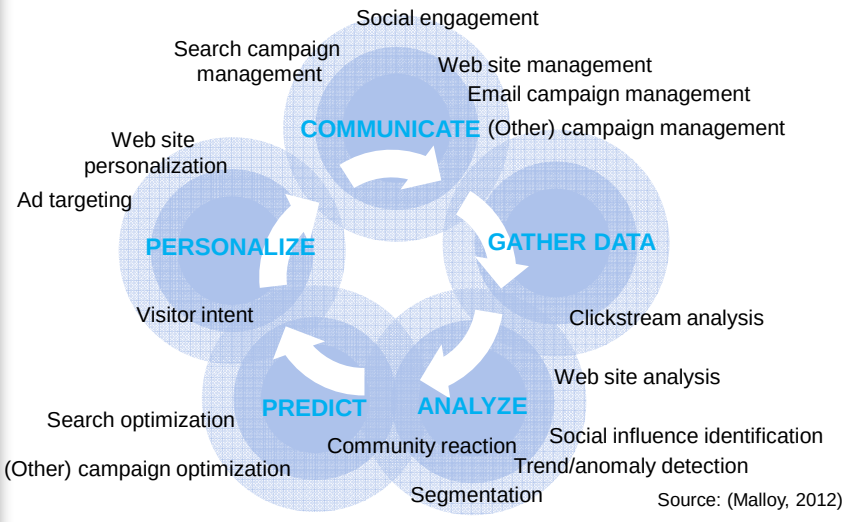
Source: (Malloy, 2012)

Digital Marketing: activities breakdown



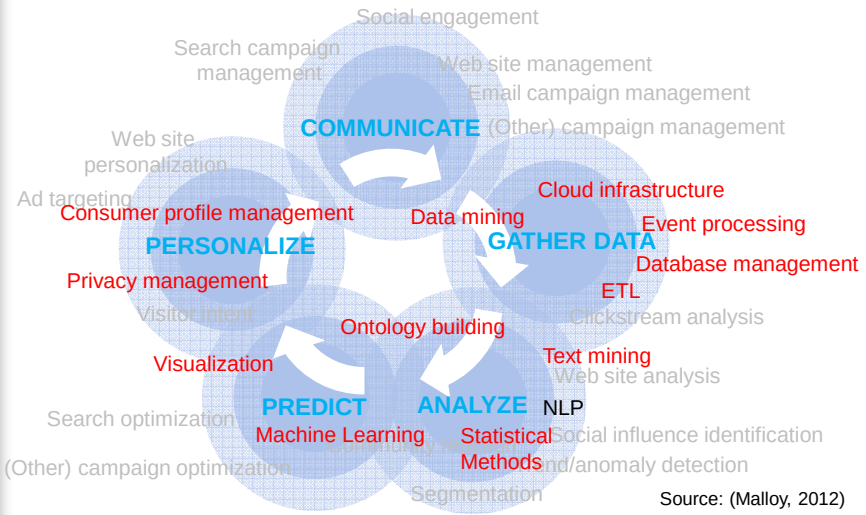
Source: (Malloy, 2012)

Digital Marketing: Activities Breakdown



Source: (Malloy, 2012)

Digital Marketing: Technology Building Blocks



Source: (Malloy, 2012)

Social-Media Monitoring tools

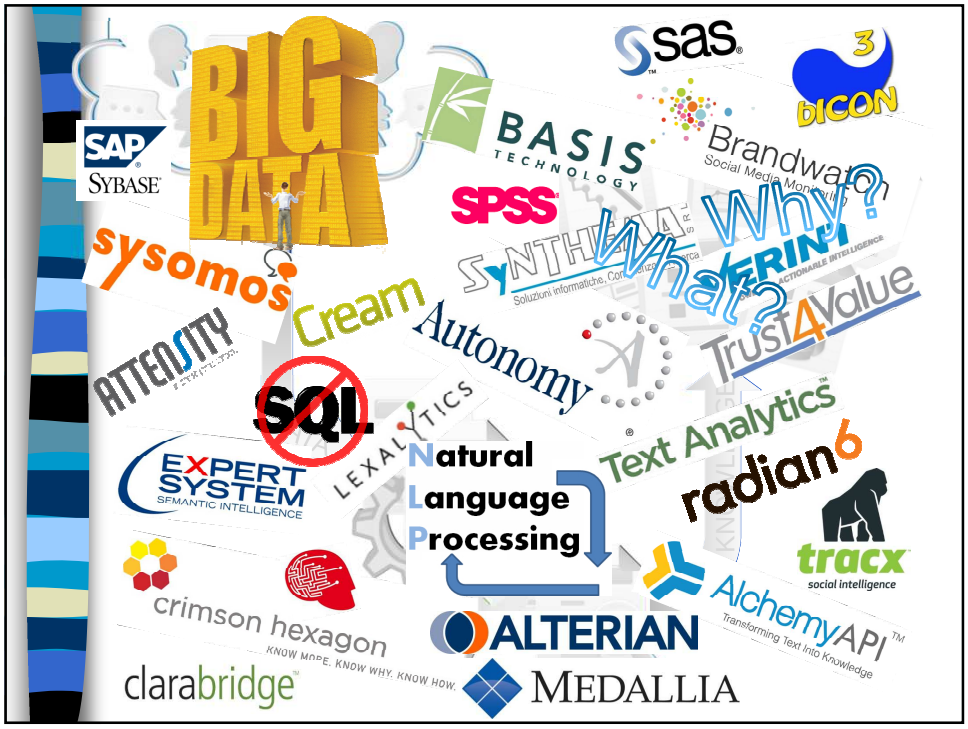
- Many commercial tools and platforms are available for analyzing the UGC
  - ✓ Brandwatch
  - ✓ Tracx
  - ✓ Clarabridge
- They typically rely on a large but fix set of glitzy dashboards that analyze the data from set of points of view…
  - ✓ Topic usage
  - ✓ Topic correlation
  - ✓ Brand reputation
- … 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

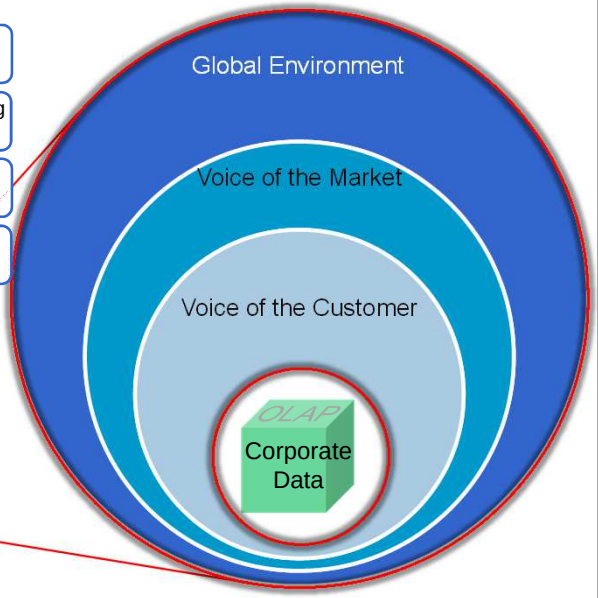


Social BI: a Definition

- **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

The Trend

- 2005: • Business Intelligence • Owned Data
- 2010: • Web/Social Media Monitoring • User Generated Content
- 2015: • Social Business Intelligence • Owned Data + Social Data
- 2020: • ?



Verticalization

- 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

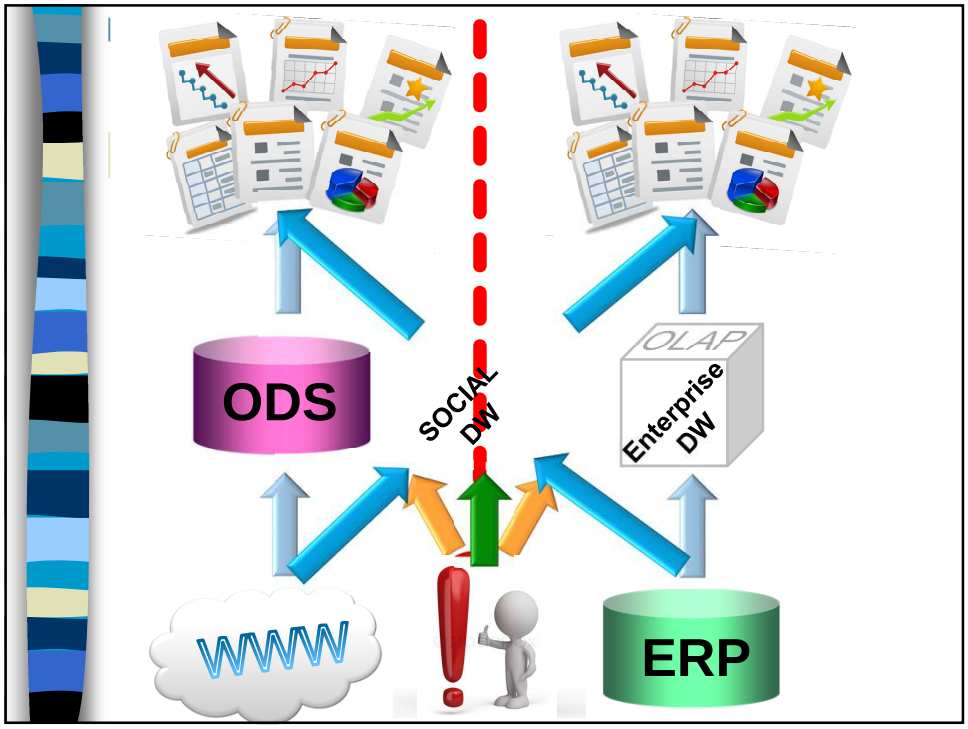
Integration with Enterprise Data

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

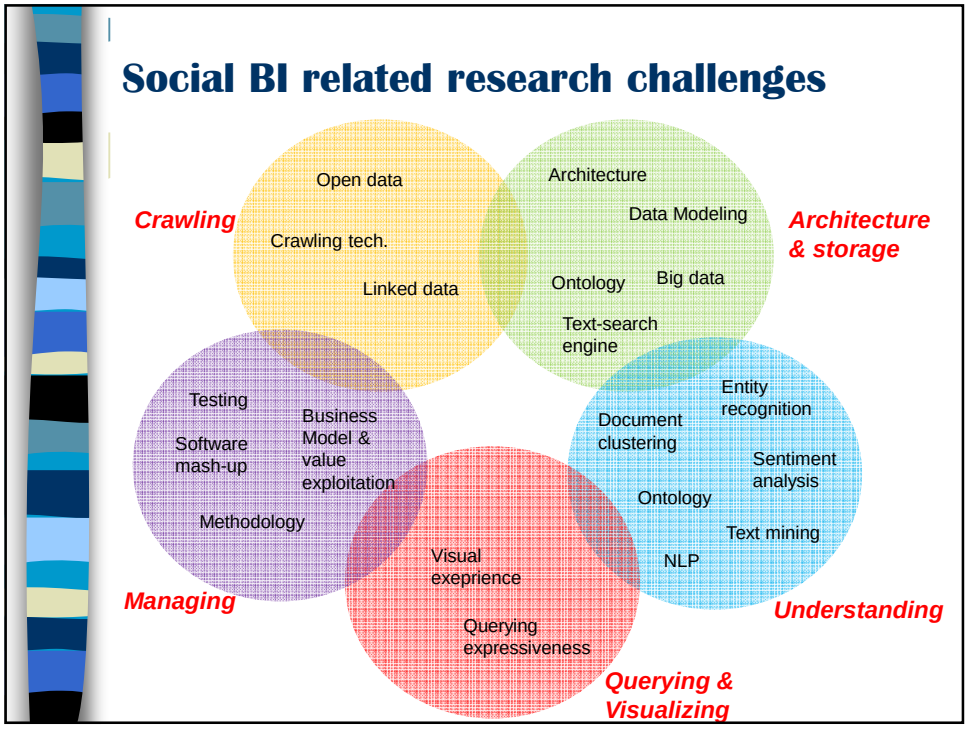
Integration with Enterprise Data

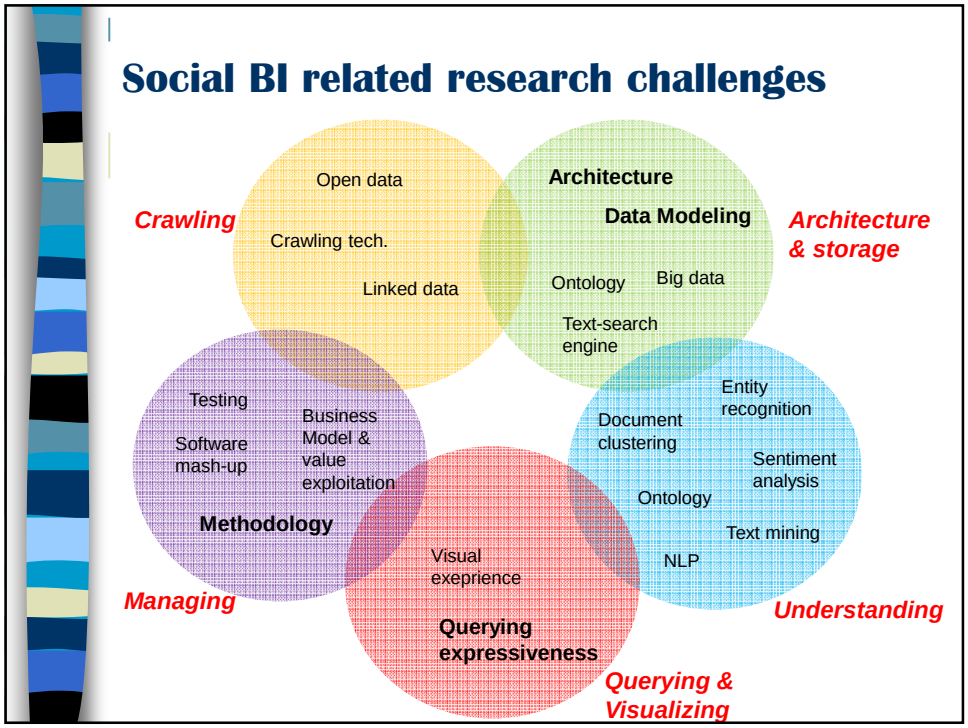
- Carrying out cross-analysis between enterprise and social data is fundamental to properly understand the impact of social events on the enterprise
  - ✓ Ex-post analysis
    - Coupling the trend of a product sentiment with its sales
    - Coupling customer complaints with the churn rate
    - Coupling 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

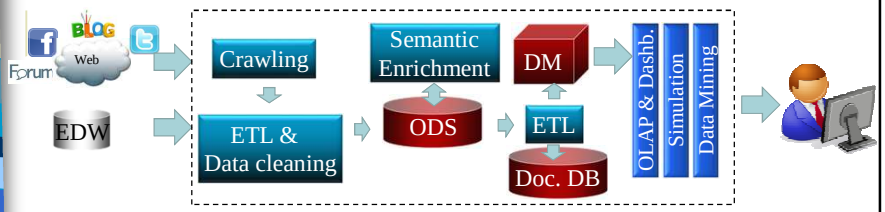
- The interest around Social-Media Monitoring and Social BI is high since companies cannot ignore UGC-derived information
- Nowadays project costs are largely determined by the licenses of commercial tools
  - ✓ Crawling engines for gathering data from the web
  - ✓ Semantic engines for data enrichment
- 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
Ontology
Big data
Text-search engine
Architecture & storage
Testing
Software mash-up
Business Model & value exploitation
Methodology
Managing
Visual exeprience
Querying expressiveness
Querying & Visualizing
Entity recognition
Document clustering
Sentiment analysis
Ontology
Text mining
NLP
Understanding


Social BI related research challenges
Crawling
Open data
Crawling tech.
Linked data
Architecture
Data Modeling
Ontology
Big data
Text-search engine
Architecture & storage
Testing
Software mash-up
Business Model & value exploitation
Methodology
Managing
Visual exeprience
Querying expressiveness
Querying & Visualizing
Entity recognition
Document clustering
Sentiment analysis
Ontology
Text mining
NLP
Understanding

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

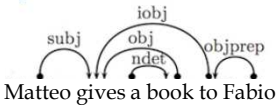
- This module is in charge of capturing UGCs through a set of keyword-based queries
- While this task is quite easy when the UGC sources are controlled by the enterprise (e.g. Enterprise CRM), it becomes very hard when the listening domain is the web
  - ✓ Parsing XML or JSON data
  - ✓ Split content 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.

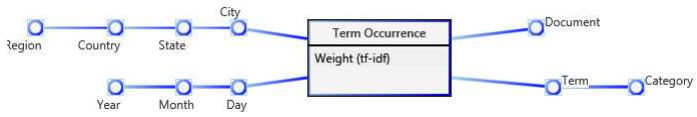
  - ✓ Polysemy

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

Data Modeling in SBI

Analysis of textual UGC

- 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.

Analysis of textual UGC

- 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*

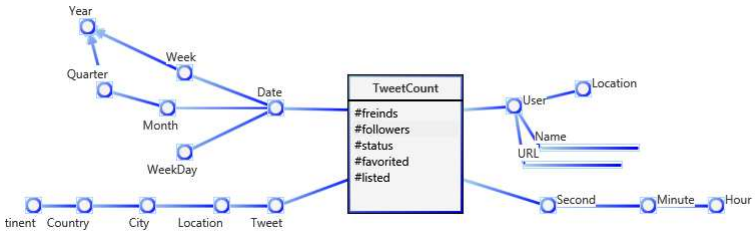
$$w_{ij} = tf_{ij}(t) \times idf_{ij}(t)$$

$$tf_{ij}(t) = \frac{n_{ij}(t)}{n_{ij}} \qquad idf_{ij}(t) = \log \frac{d_{ij}+1}{d_{ij}(t)}$$

    - 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)

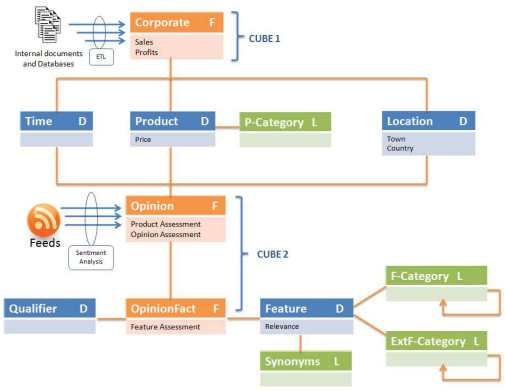
Analysis of textual UGC

- 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.



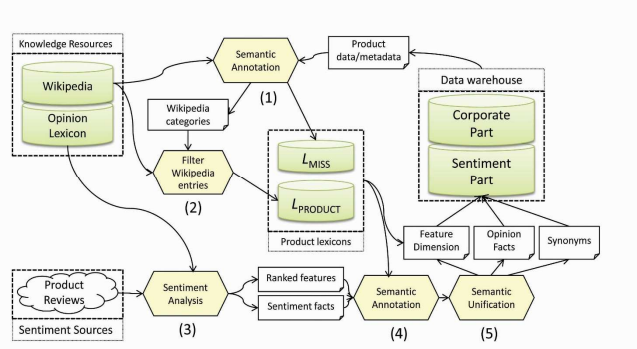
Analysis of textual UGC

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



Analysis of textual UGC

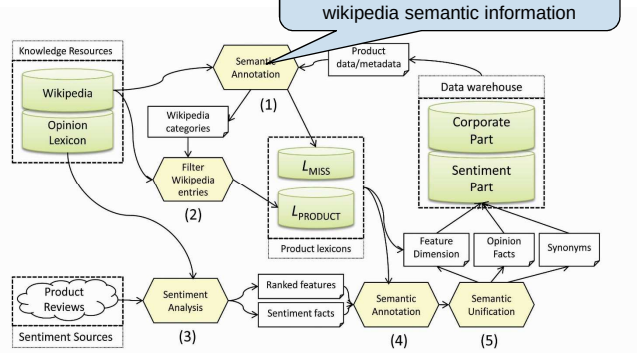
- In (Garcia-Moya 2013) the authors propose a complete architecture for Social BI
  - ✓ Corporate and social cubes are integrated through conformed hierarchies
  - ✓ Automatic Semantic Annotation is carried out in order to semantically enrich both the corporate and social data



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 socal data

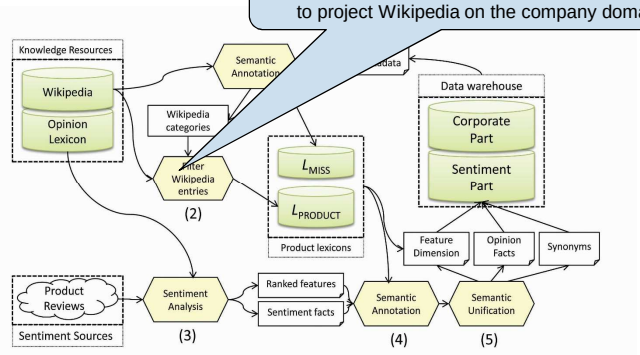
Corporate data and meta-data related to Product are annotated with wikipedia semantic information



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 socal data

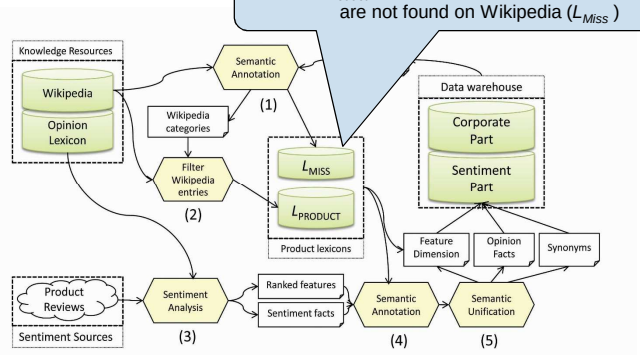
The main Wikipedia categories related to the Corporate products are manually selected in order to project Wikipedia on the company domain



Analysis of textual UGC

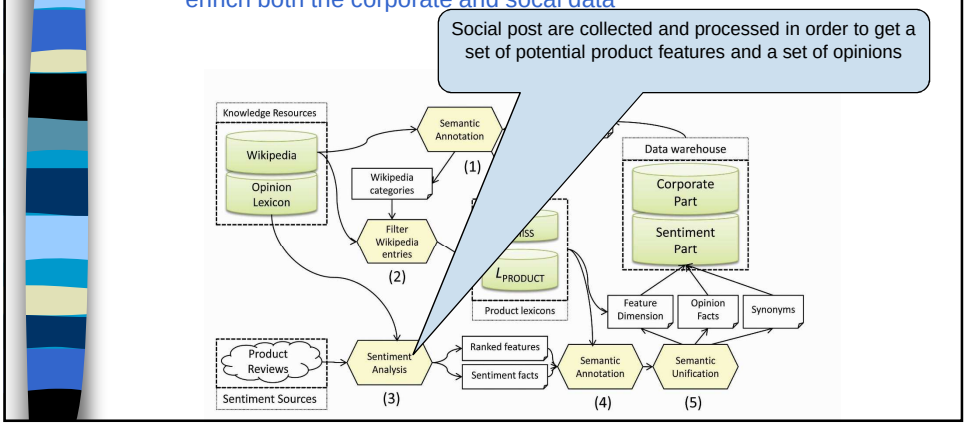
- In (Garcia-Moya 2013) the authors propose a complete architecture for Social BI
  - ✓ Corporate and social cubes are integrated through conformed hierarchies
  - ✓ Automatic Semantic Annotation is carried out in order to semantically enrich both the corporate and socal data

The lexicon related to semantically annotated concepts is stored ($L_{Product}$) together with those corporate terms that are not found on Wikipedia (L_{Miss})



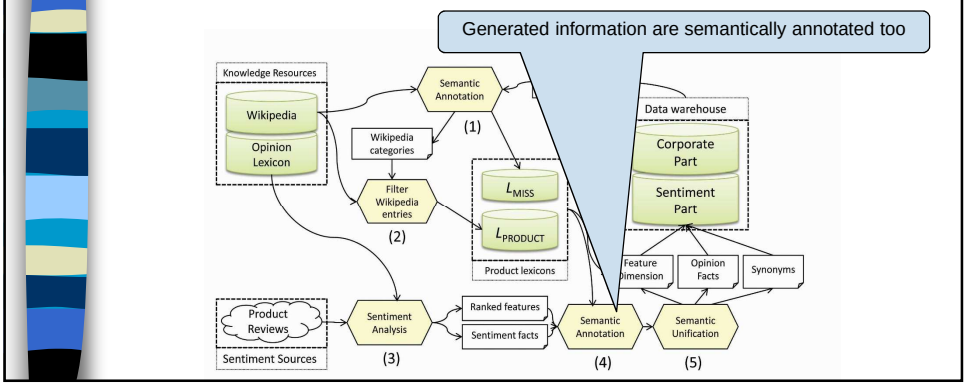
Analysis of textual UGC

- 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 socal data



Analysis of textual UGC

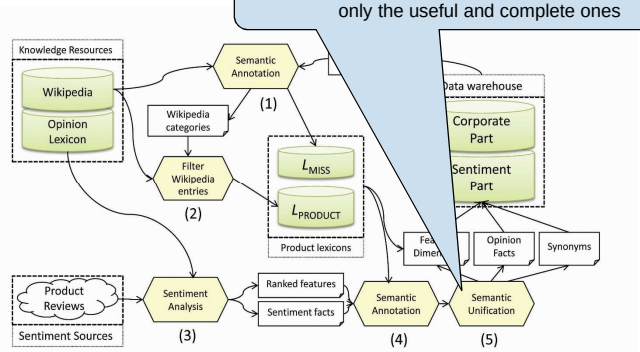
- In (Garcia-Moya 2013) the authors propose a complete architecture for Social BI
  - ✓ Corporate and social cubes are integrated through conformed hierarchies
  - ✓ Automatic Semantic Annotation is carried out in order to semantically enrich both the corporate and socal data



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 socal data

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



Analysis of textual UGC

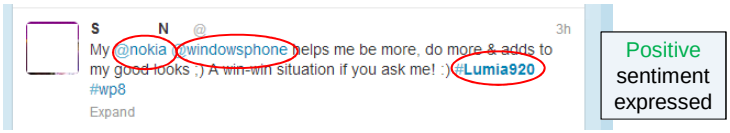
- In (Dayal 2012) the authors propose a complete architecture for Social BI
  - ✓ A first advanced solution for modeling the topic hierarchy is proposed
  - ✓ The authors stress the need to couple enterprise data with social one in order to achieve situational awareness for enterprise related events.
  - ✓ 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

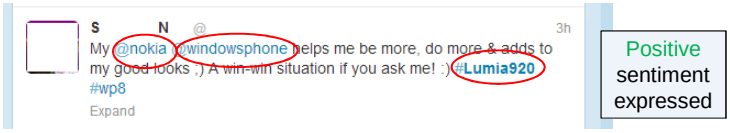
Analysis of textual UGC through relevant topics

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

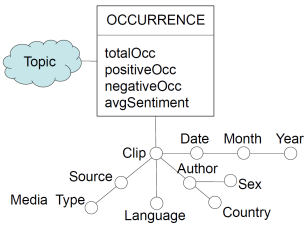


Analysis of textual UGC through relevant topics

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

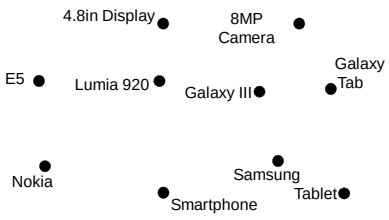


- Topics are an obvious candidate dimension of the cubes for Social BI, but:
  - ✓ Trending topic are heterogeneous and change quickly over time
  - ✓ A classical dimension table with a static hierarchy is not suitable



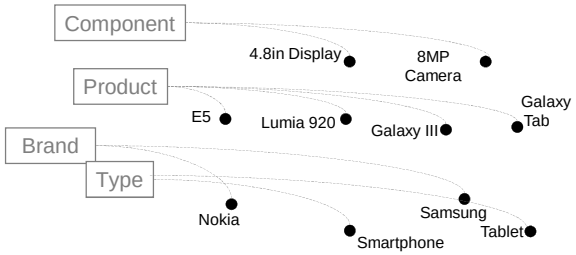
Topic hierarchy schema

- Consider a mobile-oriented scenario



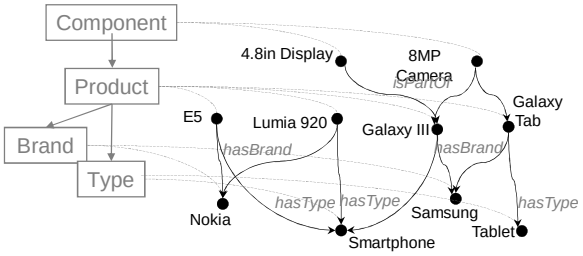
Topic hierarchy schema

- Consider a mobile-oriented scenario
  - ✓ Most topics can be classified into levels, that correspond to aggregation levels in traditional hierarchies



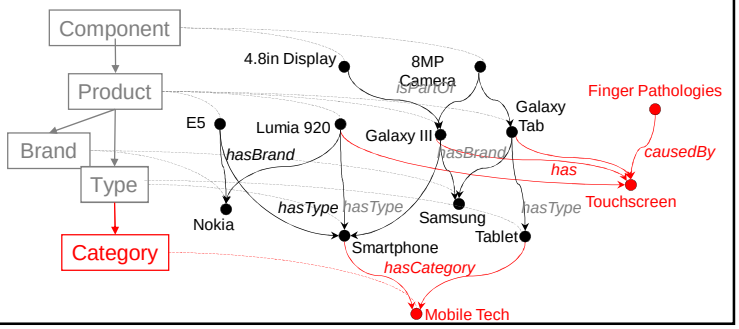
Topic hierarchy schema

- 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



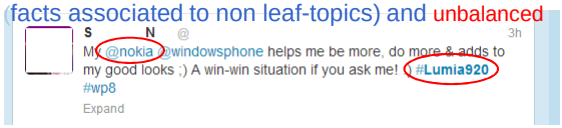
Topic hierarchy schema

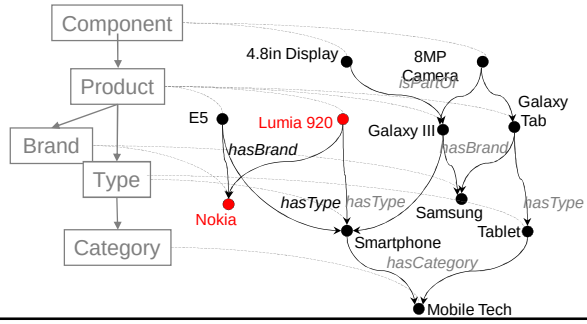
- How is a topic hierarchy different from a traditional hierarchy?
  1. Dynamicity: new topics, relationships and aggregation levels might be added at any time



Topic hierarchy schema

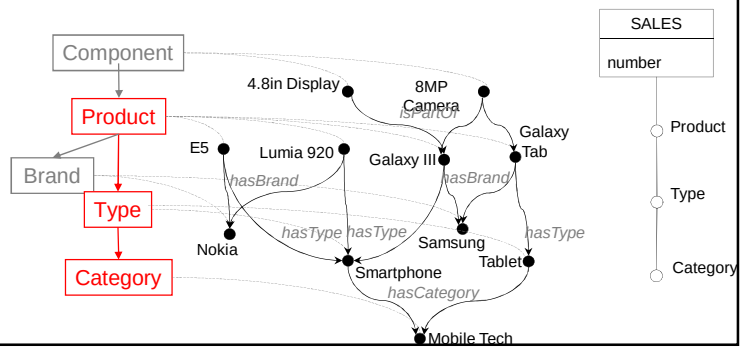
- How is a topic hierarchy different from a traditional hierarchy?
  1. Dynamicity: new topics, relationships and aggregation levels might be added at any time
  2. Mixed granularity (facts associated to non leaf-topics) and unbalanced hierarchies





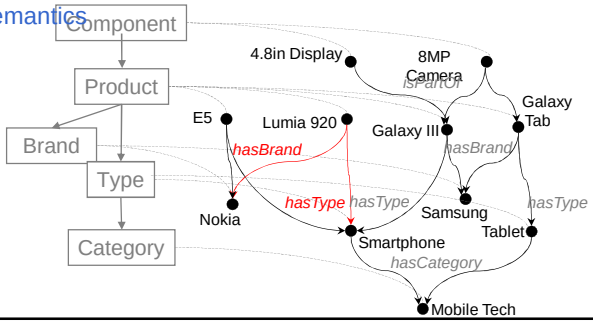
Topic hierarchy schema

- 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
  3. Integration: some topics are also part of business hierarchies and require a direct connection with the enterprise cube



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
  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

TOPIC_T

IdT	Topic	Level
1	8MP Camera	Component
2	Galaxy III	Product
3	Galaxy Tab	Product
4	Smartphone	Type
5	Tablet	Type
6	Mobile Tech	Category
7	Samsung	Brand
8	Finger Path.	-
9	Touchscreen	-
...	...	...

The Meta-Star approach

- Implementation of a Meta-Star: the **topic table**
  - ✓ One row for each topic

TOPIC_T

IdT	Topic	Level
1	8MP Camera	Component
2	Galaxy III	**Product**
3	Galaxy Tab	**Product**
4	Smartphone	**Type**
5	Tablet	**Type**
6	Mobile Tech	**Category**
7	Samsung	Brand
8	Finger Path.	-
9	Touchscreen	-
...	...	...

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

TOPIC_T

IdT	Topic	Level	**Product**	**Type**	**Category**
1	8MP Camera	Component	-	-	-
2	Galaxy III	Product	Galaxy III	Smartphone	Mobile Tech
3	Galaxy Tab	Product	Galaxy Tab	Tablet	Mobile Tech
4	Smartphone	Type	-	Smartphone	Mobile Tech
5	Tablet	Type	-	Tablet	Mobile Tech
6	Mobile Tech	Category	-	-	Mobile Tech
7	Samsung	Brand	-	-	-
8	Finger Path.	-	-	-	-
9	Touchscreen	-	-	-	-
…	…	…	…	…	…

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

ROLLUP_T

ChildId	RollUpSignature	FatherId
1	000000	1
2	000000	2
…	000000	…
1	100000	2
1	100000	3
2	010000	4
2	001000	7
4	000100	6
8	000001	9
2	000010	9
…	…	…
1	110000	4
1	110000	5
1	101000	7
1	100010	9
2	010100	6
3	010100	6
…	…	…
1	110100	6
…	…	…

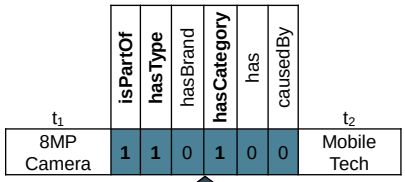
The Meta-Star approach

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

ROLLUP_T

ChildId	RollUpSignature	FatherId
1	000000	1
2	000000	2
…	000000	…
1	100000	2
1	100000	3
2	010000	4
2	001000	7
4	000100	6
8	000001	9
2	000010	9
…	…	…
1	110000	4
1	110000	5
1	101000	7
1	100010	9
2	010100	6
3	010100	6
…	…	…
1	110100	6
…	…	…

- 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

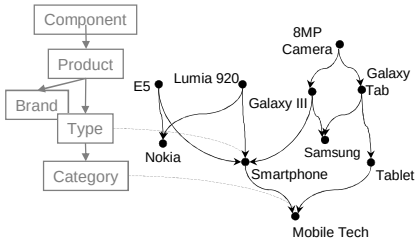
- Meta-stars natively support data and schema changes, *keeping track* of the different version require further expedient
- Different techniques can be adopted. According to Kimball terminology
  - ✓ **Type-2 solution:** data versions are tracked by creating multiple tuples in the dimension table for 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

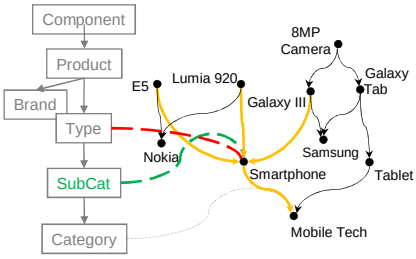
- Type-2 solution does not impact on the meta-star schema and is implemented by properly setting the ETL process only
- Full logging impacts on the meta-star schema
  - Tracking changes in the roll-up partial order requires timestamps in the roll-up table only
  - All the other operations also involve 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)

TOPIC_T

IdT	Topic	Level	From	To	Master
1	8MP Camera	Component	Jan 01 2014	-	1
2	Galaxy III	Product	Jan 01 2014	-	2
3	Galaxy Tab	Product	Jan 01 2014	-	3
4	Smartphone	Type	Jan 01 2014	-	4
5	Tablet	Type	Jan 01 2014	-	5
6	Mobile Tech	Category	Jan 01 2014	-	6
7	Samsung	Brand	Jan 01 2014	-	7
8	Finger Path.	-	Jan 01 2014	-	8
9	Touchscreen	-	Jan 01 2014	-	9
...	...	...		-	



Slowly-changing Meta-Star



TOPIC_T

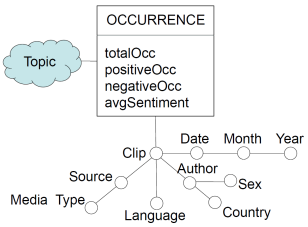
IdT	Topic	Level	From	To	Master
1	8MP Camera	Component	Jan 01 2014	-	1
2	Galaxy III	Product	Jan 01 2014	-	2
3	Galaxy Tab	Product	Jan 01 2014	-	3
4	**Smartphone**	**Type**	**Jan 01 2014**	**Jan 31 2014**	**4**
5	Tablet	Type	Jan 01 2014	-	5
6	Mobile Tech	Category	Jan 01 2014	-	6
7	Samsung	Brand	Jan 01 2014	-	7
8	Finger Path.	-	Jan 01 2014	-	8
9	Touchscreen	-	Jan 01 2014	-	9
10	**Smartphone**	**SubCat**	**Feb 01 2014**		**4**
...	...	...	...	-	

ROLLUP_T

ChildId	RollUpSig	FatherId	From	To
1	000000	1	Jan 01 14	-
...	000000	...	...	...
1	100000	2	Jan 01 14	Jan 31 14
2	**010000**	**4**	**Jan 01 14**	**Jan 31 14**
4	**000100**	**6**	**Jan 01 14**	**Jan 31 14**
...	...	...	...	...
10	**000001**	**6**	**Feb 01 14**	**-**
2	**010000**	**10**	**Feb 01 14**	**-**
...	...	...	...	...

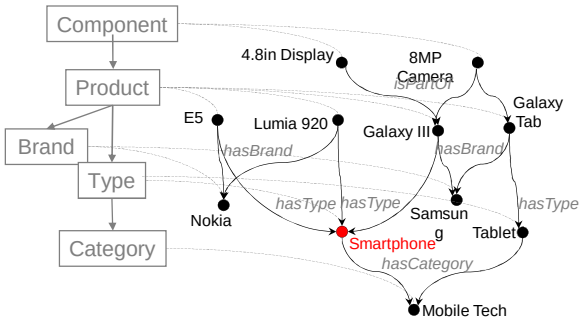
Querying Meta-Stars

- 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…



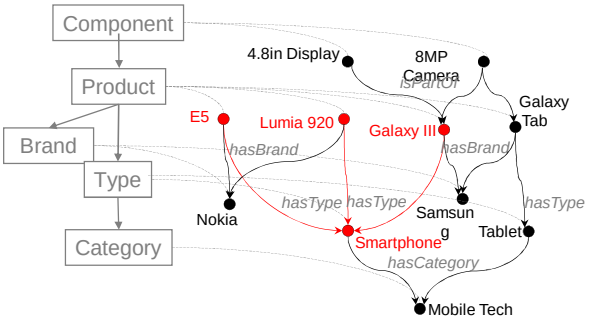
Querying Meta-Stars

- 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



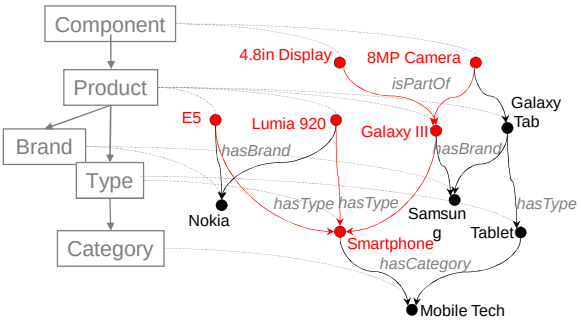
Querying Meta-Stars

- 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



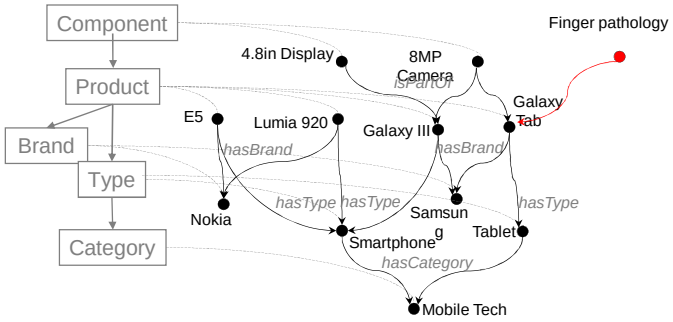
Querying Meta-Stars

- 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



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

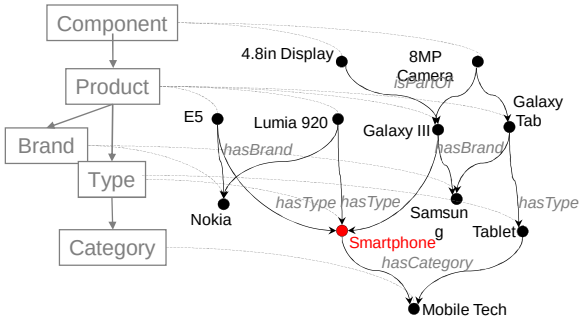


Querying Meta-Stars

- 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

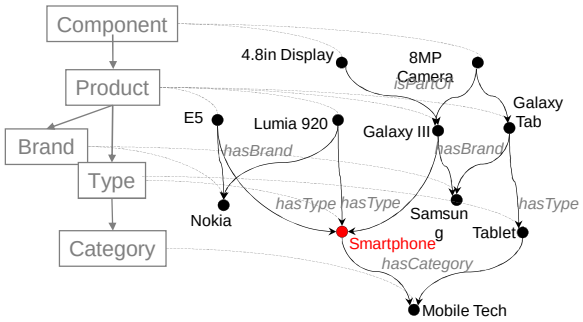
Querying Meta-Stars

- 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



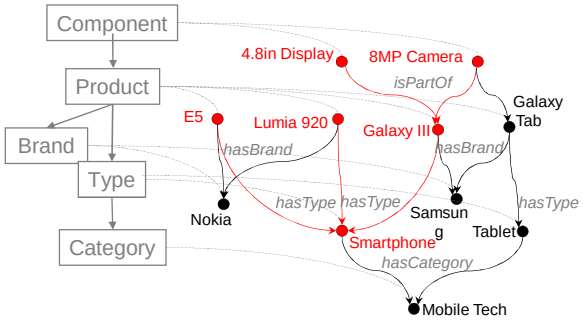
Querying Meta-Stars

- 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



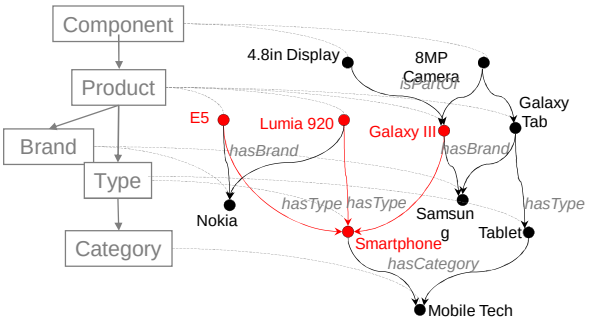
Querying Meta-Stars

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



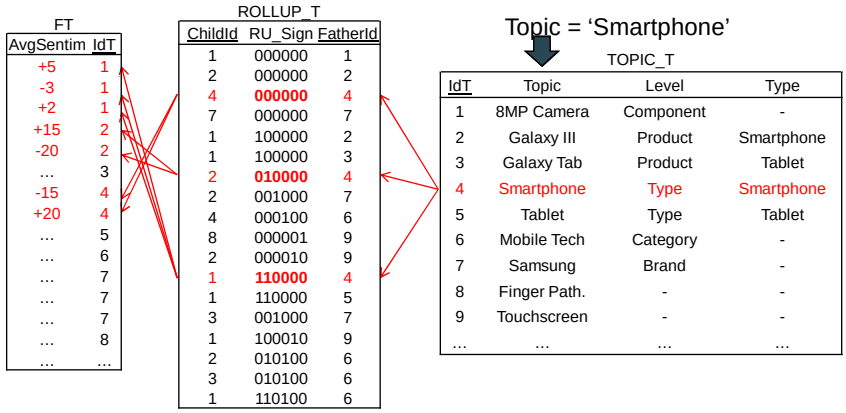
Querying Meta-Stars

- 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 with semantic topic aggregation**: semantic filter is user-defined



Querying Meta-Stars

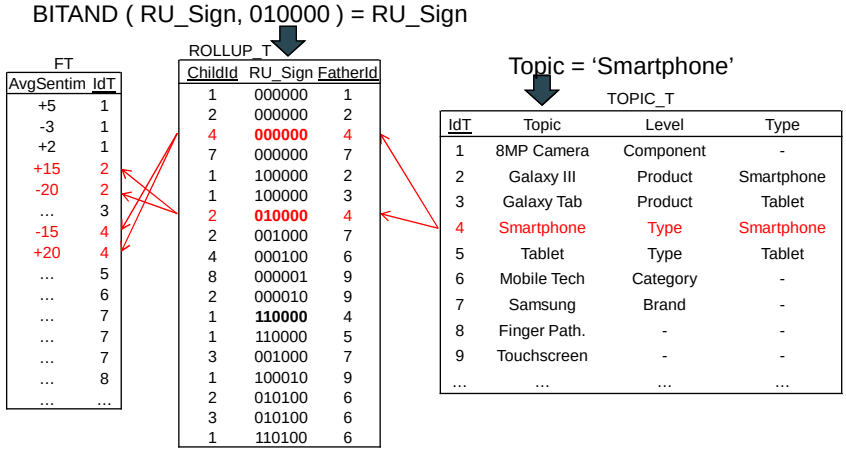
- Question: what is the current average sentiment over smartphones?
  - Example of query with full-topic aggregation



Querying Meta-Stars

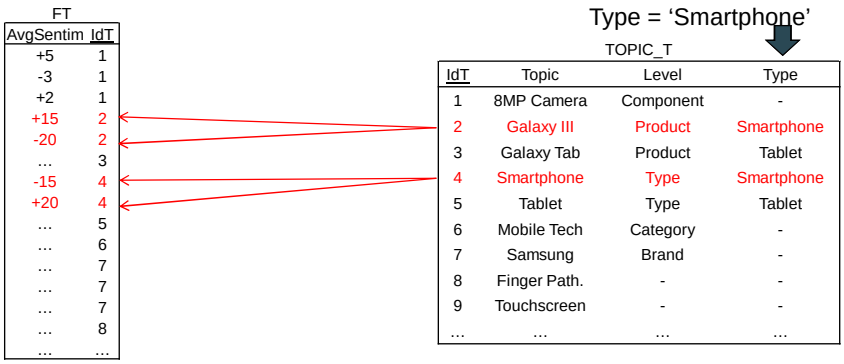
- Question: what is the current average sentiment over smartphones?
  - Example of query with semantic topic aggregation

BITAND (RU_Sign, 010000) = RU_Sign



Querying Meta-Stars

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



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;
```

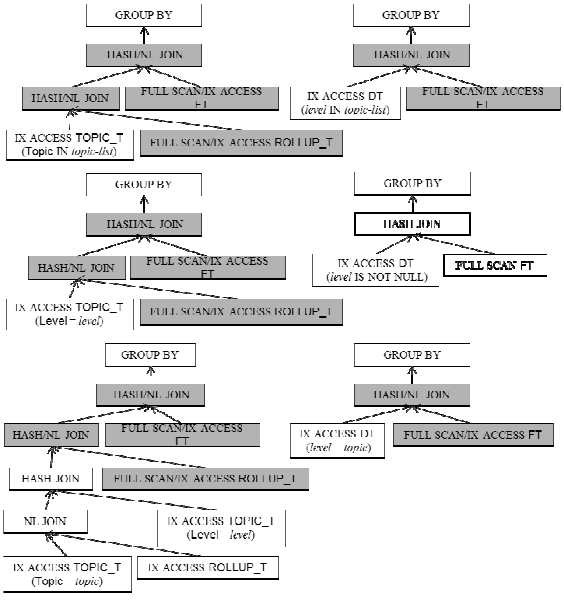
Meta-star vs Star access plans

Meta-star

Schema-free

Schema-aware

Schema-aware with sel.



Star Schema

Evaluation

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

Topic hier.	|Group-by|	FT1		FT2	
		Meta-star	Star s.	Meta-star	Star s.
H1	0				
	1				
	2				
H2	0				
	1				
	2				
H3	0				
	1				
	2				

Evaluation

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

FT1 → 1M facts
FT2 → 10M facts

Topic hier.	|Group-by|	**FT1**		**FT2**	
		Meta-star	Star s.	Meta-star	Star s.
H1	0				
	1				
	2				
H2	0				
	1				
	2				
H3	0				
	1				
	2				

Evaluation

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

FT1 → 1M facts
FT2 → 10M facts

Topic hier.	|Group-by|	FT1 Meta-star	FT1 Star s.	FT2 Meta-star	FT2 Star s.
H1	0				
	1				
	2				
H2	0				
	1				
	2				
H3	0				
	1				
	2				

Topic hier.	| TOPIC_T |	| ROLLUP_T |	fan-out	tree-height
H1	106	626	4	4
H2	658	4514	8	4
H3	27,306	334,962	4	8

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

FT1 → 1M facts
FT2 → 10M facts

Topic hier.	**|Group-by|**	FT1 Meta-star	FT1 Star s.	FT2 Meta-star	FT2 Star s.
H1	0				
	1				
	2				
H2	0				
	1				
	2				
H3	0				
	1				
	2				

Topic hier.	| TOPIC_T |	| ROLLUP_T |	fan-out	tree-height
H1	106	626	4	4
H2	658	4514	8	4
H3	27,306	334,962	4	8

Evaluation

- 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

Topic hier.	|Group-by|	FT1		FT2	
		Meta-star	Star s.	Meta-star	Star s.
H1	0	13.8	12.7	140.0	137.2
	1	16.0	5.8	174.6	64.3
	2	16.6	14.6	162.4	162.1
H2	0	13.6	13.0	136.0	133.6
	1	16.7	5.6	179.5	179.4
	2	17.0	16.2	175.8	162.2
H3	0	12.2	9.0	139.1	126.6
	1	15.9	14.1	147.3	172.1
	2	35.1	16.9	187.1	144.2

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

Topic hier.	|Group-by|	FT1		FT2	
		Meta-star	**Star s.**	**Meta-star**	**Star s.**
H1	0	13.8	12.7	140.0	137.2
	1	16.0	5.8	174.6	64.3
	2	16.6	14.6	162.4	162.1
H2	0	13.6	13.0	136.0	133.6
	1	16.7	5.6	179.5	179.4
	2	17.0	16.2	175.8	162.2
H3	0	12.2	9.0	139.1	126.6
	1	15.9	14.1	147.3	172.1
	2	35.1	16.9	187.1	144.2

In most cases star schemata outperform meta-stars, but the gap is quite limited and perfectly acceptable

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

Topic hier.	|Group-by|	FT1		FT2	
		Meta-star	Star s.	Meta-star	Star s.
H1	0	13.8	12.7	140.0	137.2
	1	16.0	5.8	174.6	64.3
	2	16.6	14.6	162.4	162.1
H2	0	13.6	13.0	136.0	133.6
	1	16.7	5.6	179.5	179.4
	2	17.0	16.2	175.8	162.2
H3	0	12.2	9.0	139.1	126.6
	1	15.9	14.1	147.3	172.1
	2	35.1	16.9	187.1	144.2

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

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

Topic hier.	|Group-by|	FT1		FT2	
		Meta-star	Star s.	Meta-star	Star s.
H1	**0**	13.8	12.7	140.0	137.2
	1	16.0	5.8	174.6	64.3
	2	16.6	14.6	162.4	162.1
H2	**0**	13.6	13.0	136.0	133.6
	1	16.7	5.6	179.5	179.4
	2	17.0	16.2	175.8	162.2
H3	**0**	12.2	9.0	139.1	126.6
	1	15.9	14.1	147.3	172.1
	2	35.1	16.9	187.1	144.2

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

Evaluation

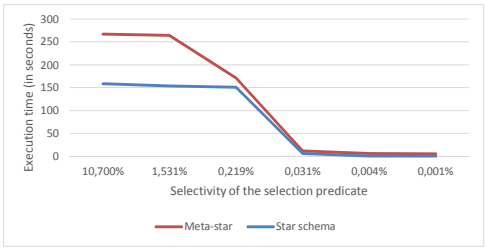
- Performances of Meta-Star are compared with traditional star schemata using queries with Topic Aggregation
  - ✓ Tests run using the Oracle 11g RDBMS on a quad-core machine
  - ✓ Each execution time (in seconds) is the average time of 3 different queries with different selection predicates

Topic hier.	|Group-by|	FT1		FT2	
		Meta-star	Star s.	Meta-star	Star s.
H1	0	13.8	12.7	140.0	137.2
	1	16.0	5.8	174.6	64.3
	2	16.6	14.6	162.4	162.1
H2	0	13.6	13.0	136.0	133.6
	1	16.7	5.6	179.5	179.4
	2	17.0	16.2	175.8	162.2
H3	0	12.2	9.0	139.1	126.6
	1	15.9	14.1	147.3	172.1
	2	35.1	16.9	187.1	144.2

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

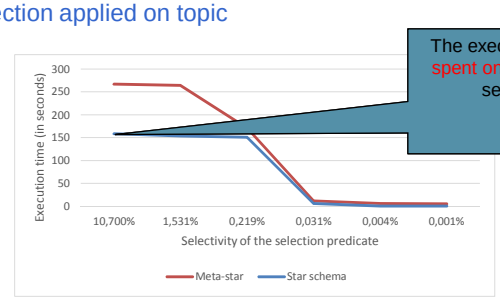
Evaluation

- 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



Evaluation

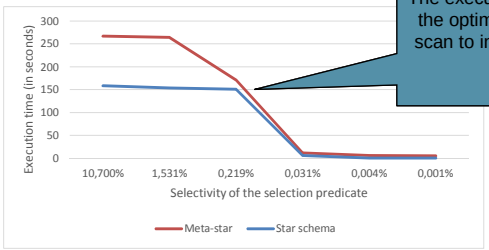
- Meta-star allows performances comparable (in many cases better) with implementations based on traditional star schema
  - ✓ The main cost remains accessing the fact table
  - ✓ Execution plans are similar
  - ✓ Roll-up table indexing makes its size not relevant
- The worst case for meta-star is when
  - ✓ 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 is mostly spent on the fact table when selectivity is low

Evaluation

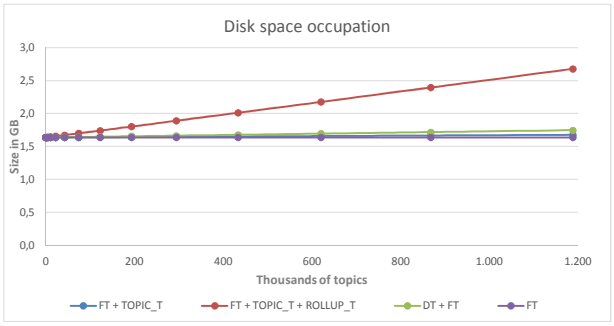
- Meta-star allows performances comparable (in many cases better) with implementations based on traditional star schema
  - ✓ The main cost remains accessing the fact table
  - ✓ Execution plans are similar
  - ✓ Roll-up table indexing makes its size not relevant
- The worst case for meta-star is when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



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

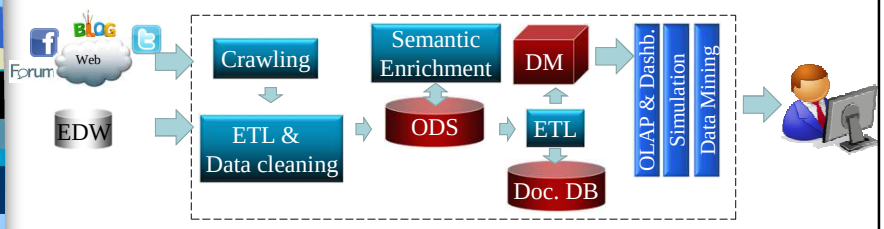
A Case Study on 2014 European Election

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 WebPolEU – 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 milions of raw clips
  - 60,000 web sources
  - 3 different languages
  - 2 full-time + 2 half-time + 5 *end user*

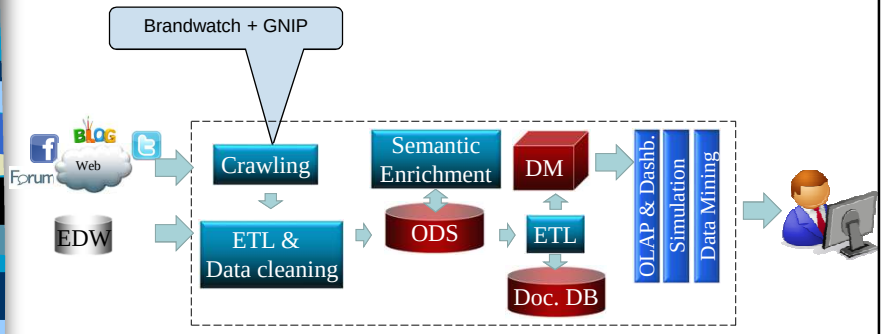
Our architecture

- We adopted an “end-to-end” solution



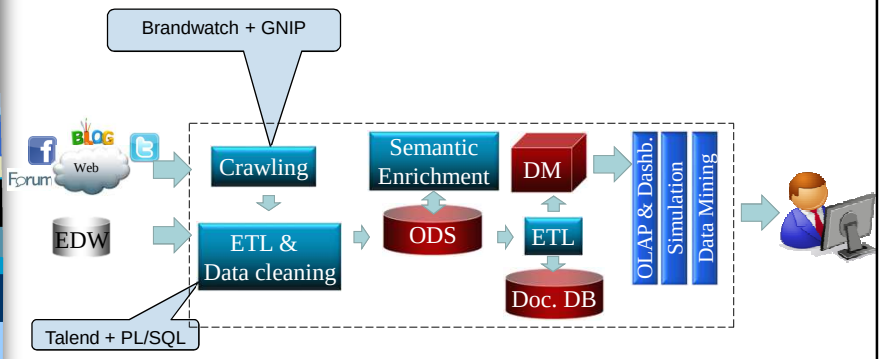
Our architecture

- We adopted an “end-to-end” solution



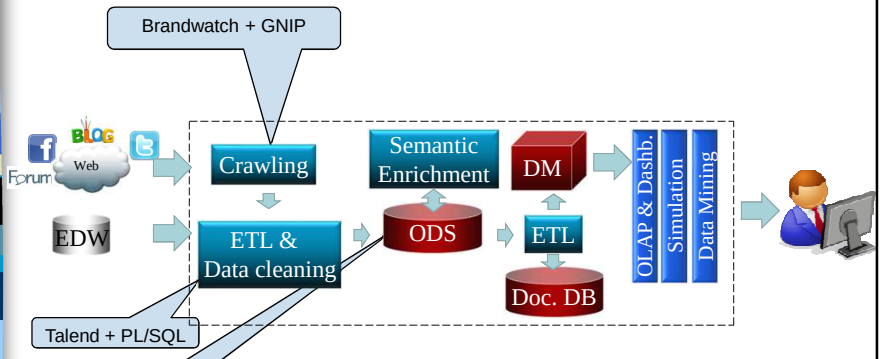
Our architecture

- We adopted an “end-to-end” solution



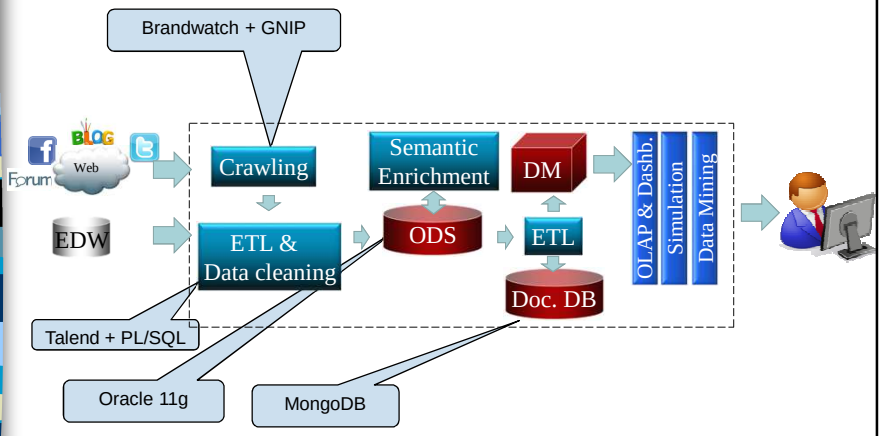
Our architecture

- We adopted an “end-to-end” solution



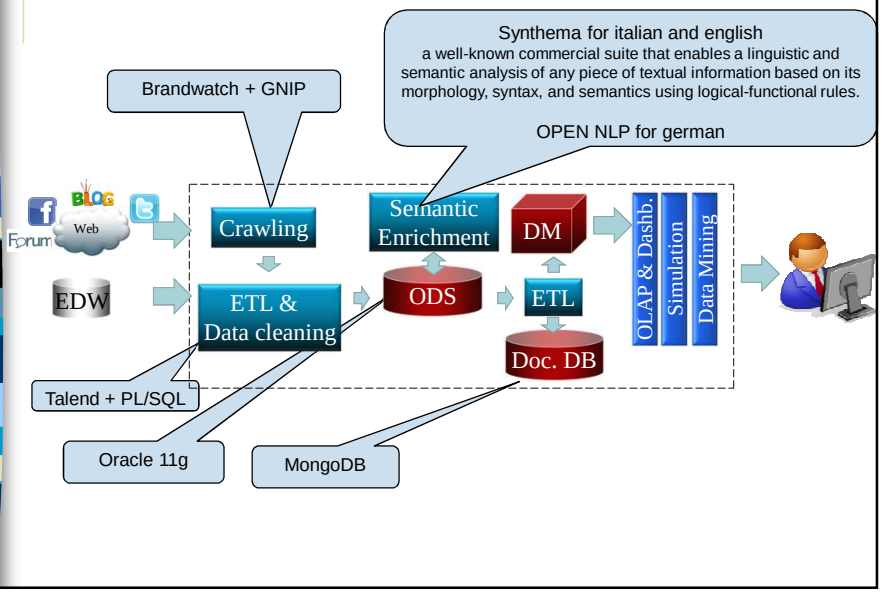
Our architecture

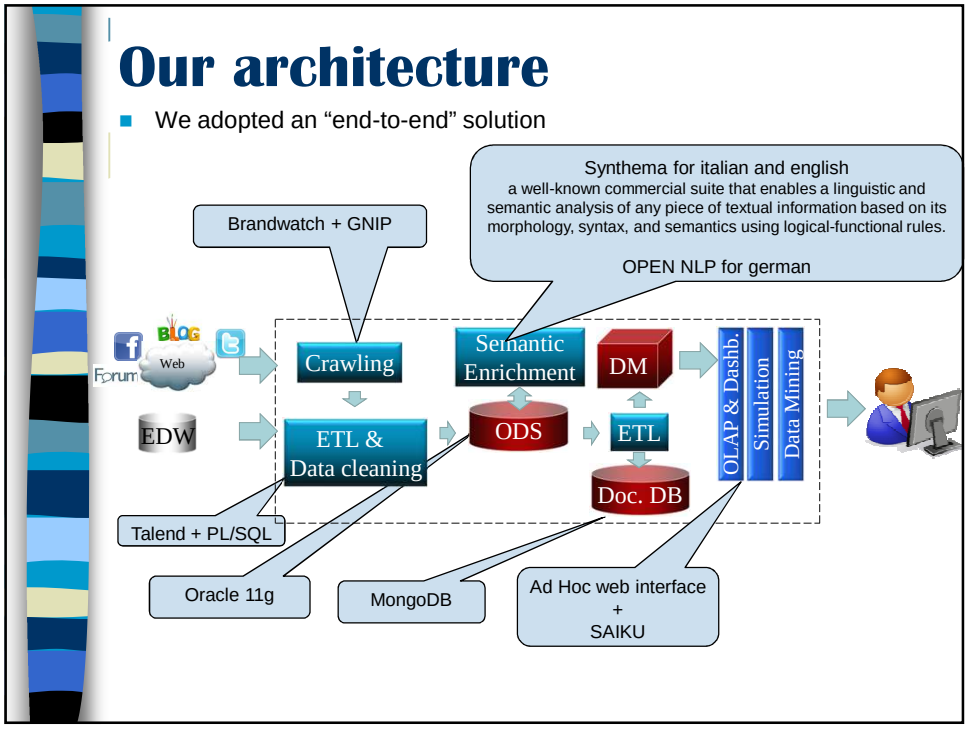
- We adopted an “end-to-end” solution

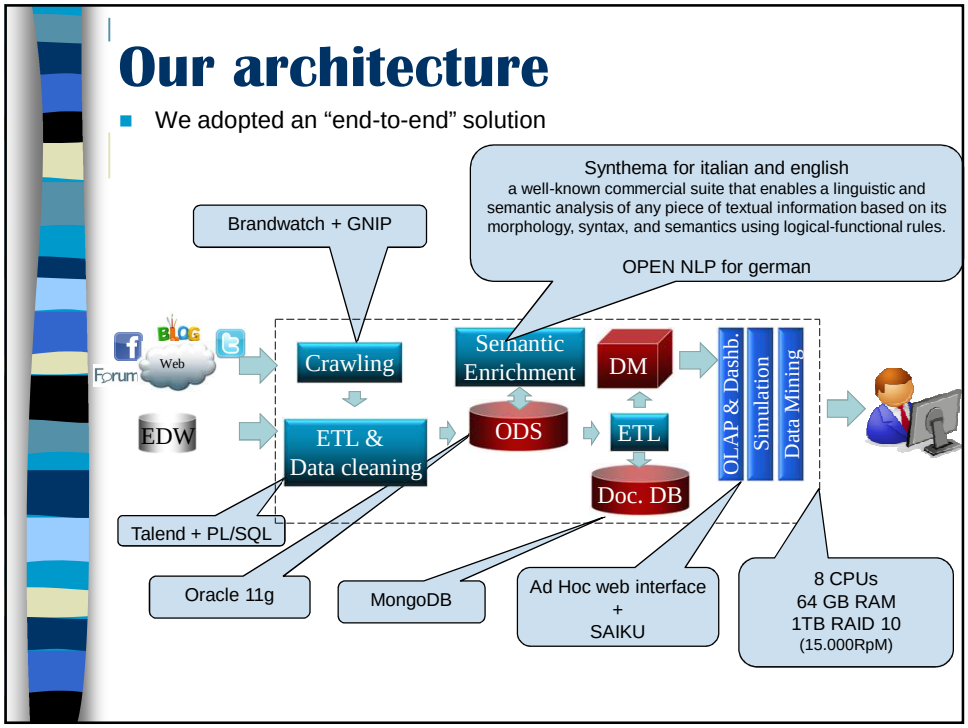


Our architecture

- We adopted an “end-to-end” solution




Our architecture
We adopted an “end-to-end” solution
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.
OPEN NLP for german
Brandwatch + GNIP
Web
Forum
BLOG
EDW
Crawling
ETL & Data cleaning
Semantic Enrichment
ODS
DM
ETL
Doc. DB
OLAP & Dashb.
Simulation
Data Mining
Talend + PL/SQL
Oracle 11g
MongoDB
Ad Hoc web interface + SAIKU


Our architecture
We adopted an “end-to-end” solution
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.
OPEN NLP for german
Brandwatch + GNIP
Web
Forum
BLOG
EDW
Crawling
ETL & Data cleaning
Semantic Enrichment
ODS
DM
ETL
Doc. DB
OLAP & Dashb.
Simulation
Data Mining
Talend + PL/SQL
Oracle 11g
MongoDB
Ad Hoc web interface + SAIKU
8 CPUs
64 GB RAM
1TB RAID 10
(15.000RpM)

Demo time!

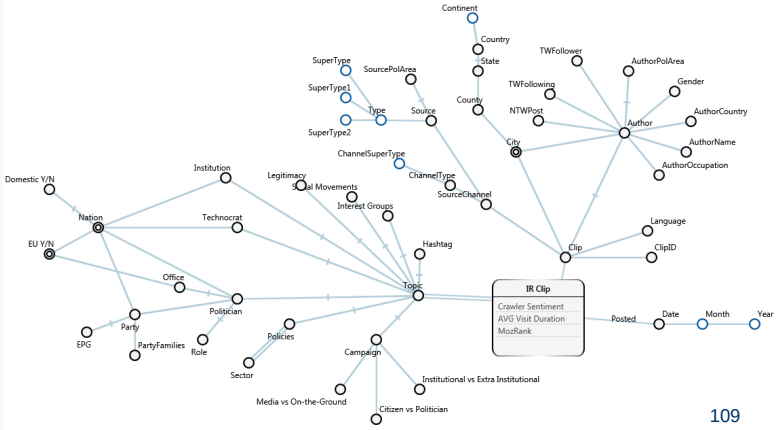
Behind the scenes

- 4 cubes
- 2 enrichment techniques
  - ✓ Information Retrival – 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 analys of the text
    - Morphological Analysis
    - Lexical Analysis
    - Syntax Analysis
    - Semantic Analysis

108

IR Clip

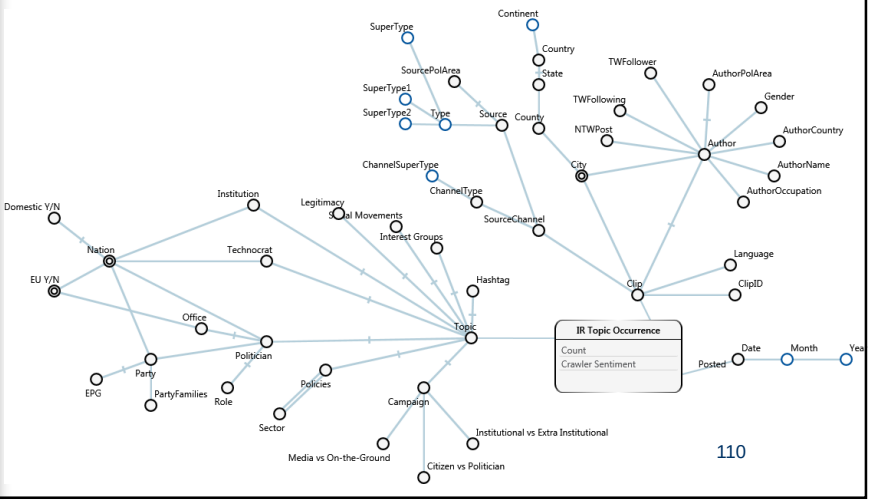
- Each cell models a clip
- The cubes reports the measure available at the clip level



109

IR Topic Occurrence

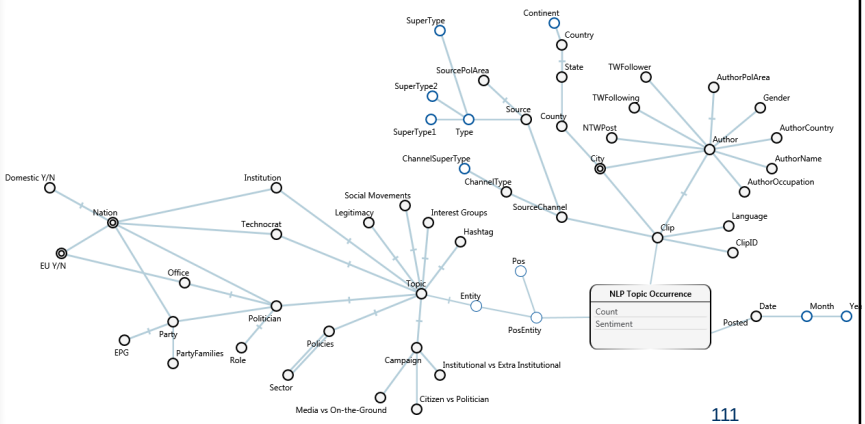
- Each cell models the occurrences of a **topic** within a clip



110

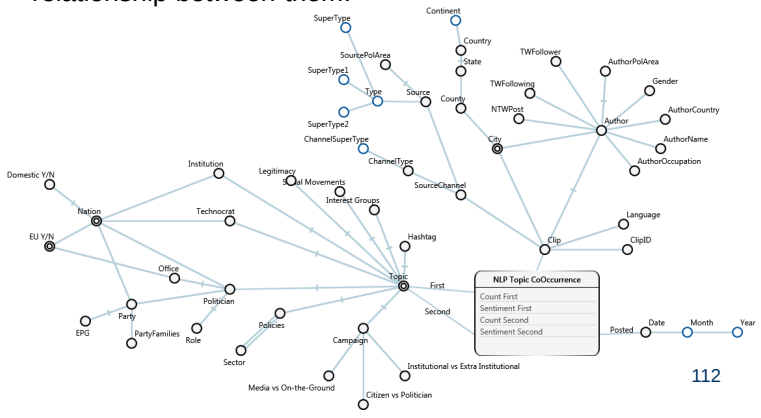
NLP Topic Occurrence

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



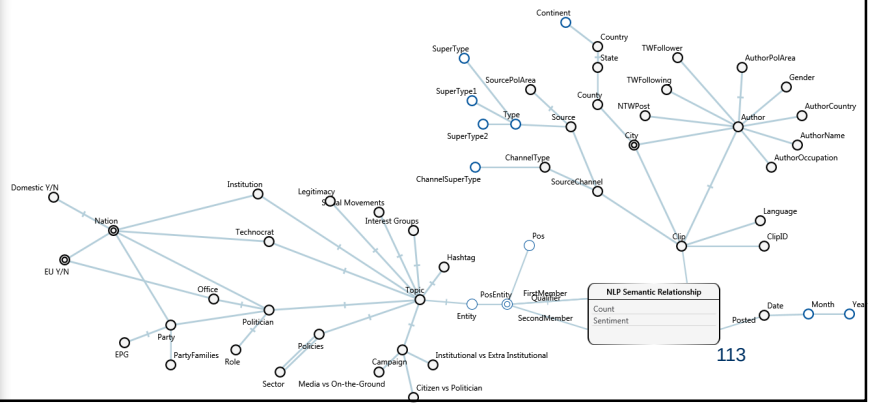
NLP Topic CoOccurrence

- 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 e 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



113

Fact Cardinality

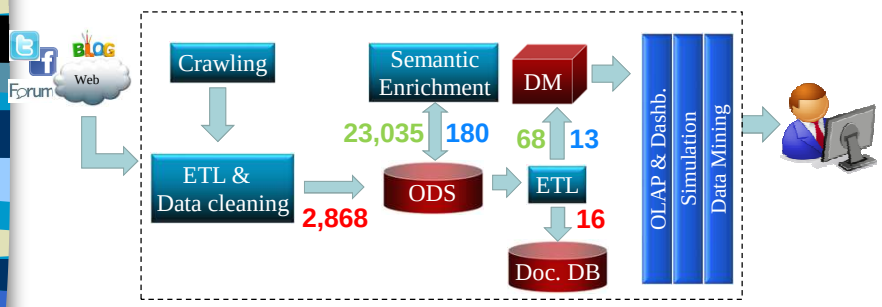
Cardinality	ITA	ENG	DEU
Topic	464	432	513
Alias	895	694	870
pos entity	1.262.790	2.902.942	6.093.724
entity	1.242.402	2.867.726	5.145.714
IR Clip	2.393.568	3.275.193	933.438
IR Topic Occurrence	15.400.783	25.005.664	16.569.387
NLP Entity Occurrence	226.953.012	519.446.526	524.784.320
NLP Topic Occurrence	14.214.686	23.398.601	7.504.815
NLP Semantic Relationship	17.364.421	23.837.474	38.231.162

114

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

Fact	ITA	ENG
IR Topic Occurrence	14,215 K	23,399 K
NLP Topic Occurrence	15,401 K	25,006 K
Shared Occurrence %	***83.9%***	***86.0%***

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.

	ITA			ENG		
Sentiment	**NLP**	**IR**	**Shared %**	**NLP**	**IR**	**Shared %**
Positive	566 K	36 K	***52.6%***	1,090 K	142 K	***75.5 %***
Neutral	893 K	2,340 K	***38.0%***	1,368 K	2,973 K	***45.0 %***
Negative	934 K	17 K	***82.3%***	817 K	159 K	***70.2 %***

- 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 Sympson'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 useing 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

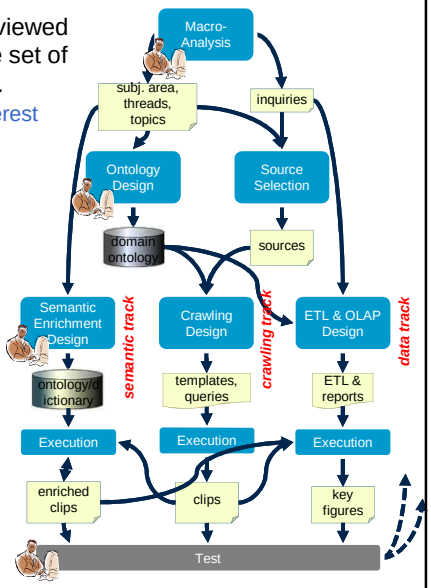
- Social BI projects are characterized by:
  - ✓ Quickly changing requirements due to topic dynamicity
  - ✓ Data sources are not known a priori and keyword query are a rough tool for their selection
  - ✓ Cubes schema is project independent, modeling mainly involves topic hierarchy/ontology
    - ✓ 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

Project Type	Crawling	Semantic Enrichment	Storing & Analysis
Level 1: Best-of-Breed	template design	dictionary enrichment, inter-word relat. def.	ETL design and impl.
Level 2: end-to-end	source selection, query design, content rel. analysis	polarization, correctness analysis, ontology coverage analysis	ontology design, KPI & dashboard design
Level 3:Off-the-Shelf	macro-analysis	macro-analysis	macro-analysis

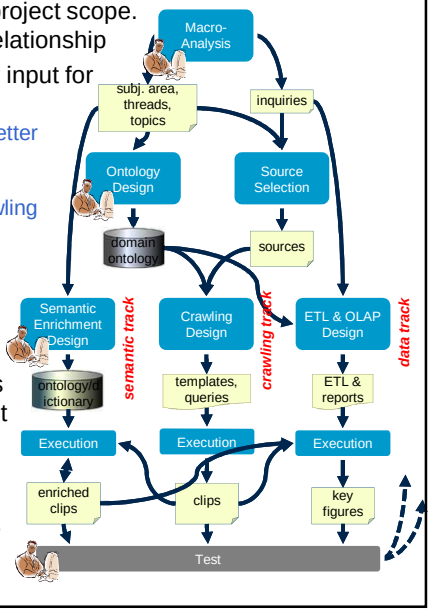
The methodology: Macro Analysis

- 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
    - ✓ Politiacian & Parties
    - ✓ Policies
  - ✓ Topics, their early definition is useful as
    - ✓ a foundation for designing a core taxonomy
    - ✓ a crawling query draft



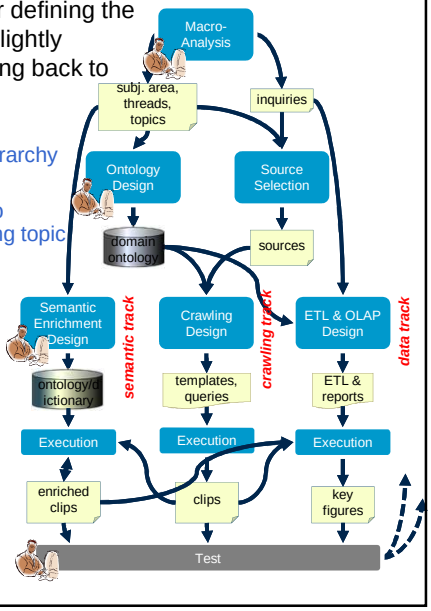
The methodology: Ontology Design

- 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



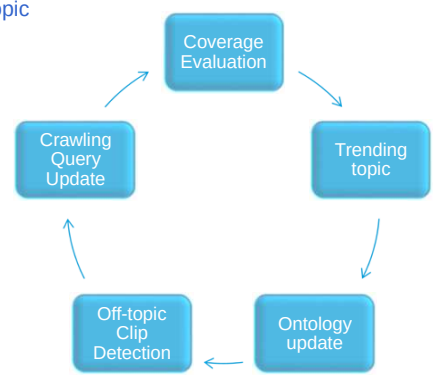
The methodology: Ontology Design

- 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*



The methodology: Coverage Analysis

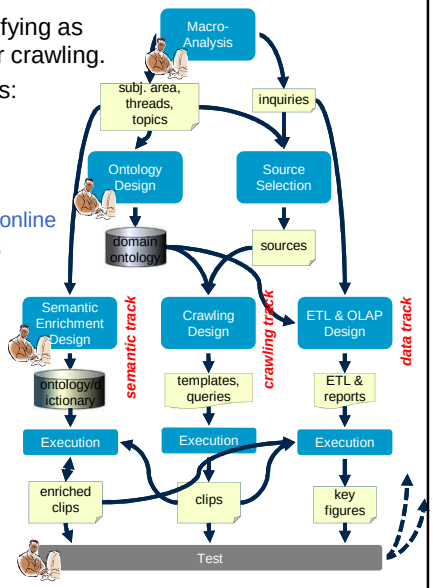
- An effective way to measure the ontology maturity level is to use as an indicator the coverage
  - ✓ The ontology coverage is the percentage of clips that include at least one ontology topic



- 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.

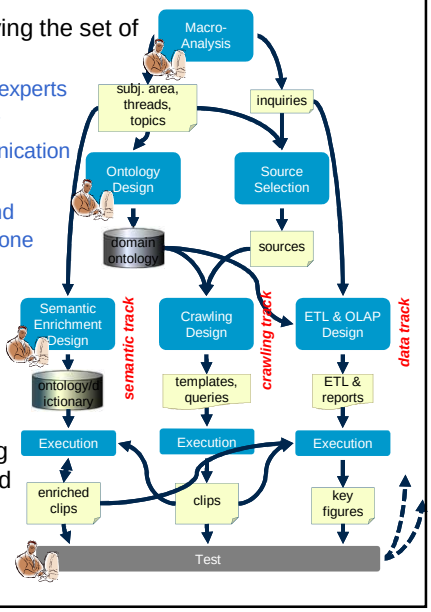
The methodology: Source Selection

- 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



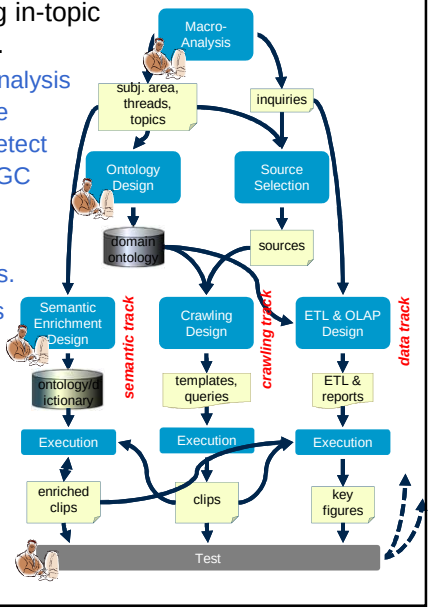
The methodology: Source Selection

- 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



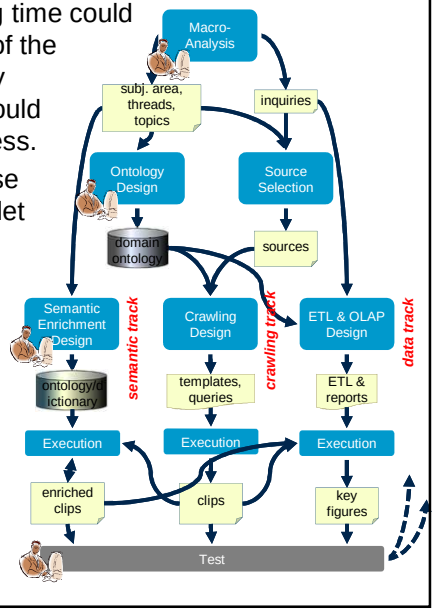
The methodology: Crawling Design

- 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

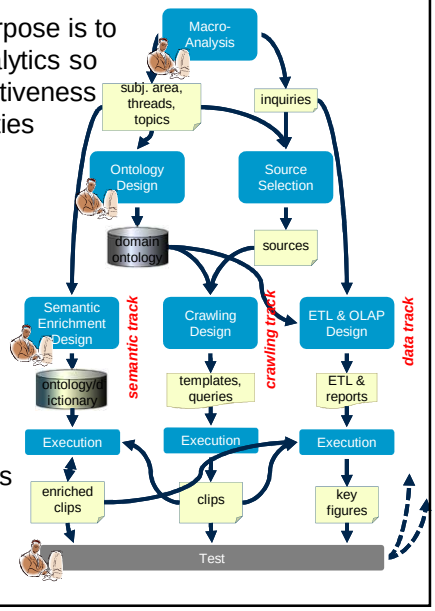


The methodology: Semantic Enrichment Design

- 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.



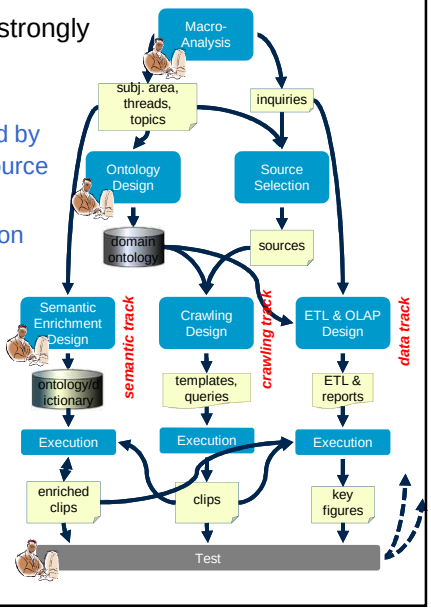
The methodology: Semantic Enrichment Design

- 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

	Project 1			Project 2		
	Non-tuned	*Tuned*	*Improvement*	*Non-tuned*	*Tuned*	*Improvement*
Total	54.0%	57.4%	3.4%	51.8%	60.3%	8.5%
Social	52.5%	55.9%	3.4%	46.1%	58.1%	12.0%
Qualified	55.0%	58.3%	3.3%	54.6%	61.4%	6.8%
Hard	34.3%	37.2%	2.9%	35.0%	47.0%	12.0%
Standard	67.3%	71.1%	3.8%	61.4%	68.1%	6.7%
Negative	46.6%	46.6%	0.0%	50.5%	59.7%	9.2%
Neutral	45.6%	49.1%	3.5%	62.0%	71.3%	9.3%
Positive	69.5%	76.3%	6.8%	47.8%	52.4%	4.6%

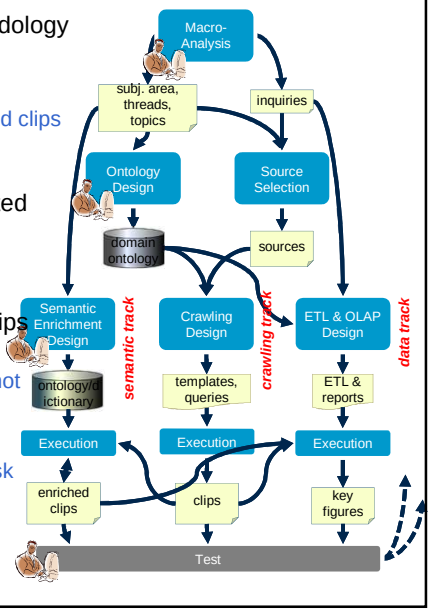
The methodology: ETL & OLAP Design

- 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 loosing 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–spending phases too



Social BI Roles

Phase	Task	Programmer	Designer	User
Crawling	template design	Exec		
	source selection		Exec	Exec
	query design		Exec	Exec
	content rel. analysis	Exec		Exec
	macro-analysis		Exec	Exec
Semantic Enrichment	dictionary enrichment	Partic	Exec	Exec
	inter-word relat. def.	Partic	Exec	Exec
	polarization		Exec	Exec
	correctness analysis	Exec		Exec
	ontology coverage analysis	Exec		
	macro-analysis		Exec	Exec
Storing & Analysis	macro-analysis		Exec	Exec
	ontology design		Exec	Exec
	KPI & dashboard design	Exec	Partic	Partic
	ETL design and impl.	Exec	Partic	

Social BI Roles

Phase	Task	Programmer	Designer	User
Crawling	template design	Exec		
	source selection		Exec	Exec
	query design		Exec	Exec
	content rel. analysis	Exec		Exec
			Exec	Exec
Semant Enrichment		Partic	Exec	Exec
		Partic	Exec	Exec
			Exec	Exec
	correctness analysis	Exec		Exec
	ontology coverage analysis	Exec		
	macro-analysis		Exec	Exec
Storing & Analysis	macro-analysis		Exec	Exec
	ontology design		Exec	Exec
	KPI & dashboard design	Exec	Partic	Partic
	ETL design and impl.	Exec	Partic	

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

Social BI Roles

Phase	Task	Programmer	Designer	User
Crawling	template design	Exec		
	source selection		Exec	Exec
	query design		Exec	Exec
	content rel. analysis	Exec		Exec
			Exec	Exec
Semantic Enrichment			Exec	Exec
			Exec	Exec
			Exec	Exec
	correctness analysis			Exec
	ontology coverage analysis	Exec		
	macro-analysis		Exec	Exec
Storing & Analysis	macro-analysis		Exec	Exec
	ontology design		Exec	Exec
	KPI & dashboard design	Exec	Partic	Partic
	ETL design and impl.	Exec	Partic	

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

(Francia, 2014) M. Francia, M. Golfarelli, S. Rizzi. A Methodology for Social BI. In *Proc. IDEAS2014*, Porto, Portugal, 2014.

(Gallinucci, 2013) E. Gallinucci, M. Golfarelli, S. Rizzi. Meta-Stars: Multidimensional Modeling for Social Business Intelligence. In *Proc. DOLAP 2013*, San Francisco, USA, 2013.

(Gallinucci, 2015) E. Gallinucci, M. Golfarelli, S. Rizzi. Advanced Topic Modeling for Social Business Intelligence. In *Information Systems*, vol 53, pp. 87-106, 2015

(García-Moya, 2013) L. García-Moya, S. Kudama,M. J. Aramburu, R. Berlanga. *Storing and analyzing voice of the market data in the corporate data warehouse.* In Information Systems Frontier2012. Vol. 15(3), pp. 331-349, 2013.

(Golfarelli, 2006), J.Lechtenborger, S.Rizzi, G.Vossen. Schema versioning in data warehouses: Enabling cross-version querying via schema augmentationIn *Data Knowl. Eng* vol. 59, pp. 435-459, 2006.

(Grimes, 2014) Seth Grimes. *Sentiment Analysis and Business Sense*. Retrieved on 30th April 2014 from clarabridge.com.

(Lee, 2000) J. Lee, D. Grossman, O. Frieder, M.C. McCabe. *Integrating structured data and text: a multi-dimensional approach*, in Int. Conf. on Information Technology: Coding and Computing, Las Vegas, 2000.

(Liu, 2012) Bing Liu. Sentiment Analysis and Opinion Mining. Morgan & Claypool, 2012.

(Malloy, 2012) Tom Malloy. *Revolutionizing Digital Marketing with Big Data*, In CIKM 2012. Hawaii (USA), 2012.

(Ravat, 2008) F. Ravat, O. Teste, R. Tournier, G. Zurfluh *Top_Keyword: an Aggregation Function for Textual Document OLAP*. In DaWaK 2008 Turin, Italy, 2008.

(Rehman, 2012) N. Rehman, S. Mansmann, A. Weiler, M.H. Scholl. *Building a Data Warehouse for Twitter Stream Exploration*. In Int.Conf. on Advances in Social Networks Analysis and Mining (ASONAM), Istanbul, Turkey, 2012

(Wagner, 1982) C.H. Wagner. Simpson's paradox in real life. The American Statistician, 36(1):46-48, 1982.