Making Sense of Social Media Data

Dr. Asuman Suenbuel,
Senior Director of Development
Office of the CTO
SAP Labs, Palo Alto
asuman.suenbuel@sap.com

Overview

1. Social media data processing challenges
   - From where to get the data? Where are the real customers? Not necessarily on twitter/facebook/social x
   - “Garbage in garbage out”
   - Volume of data
2. Demo: connecting social media on the fly with your ERP
3. Example how customers are using social media data
4. What to look for in the 100+ social media tools (only US!)

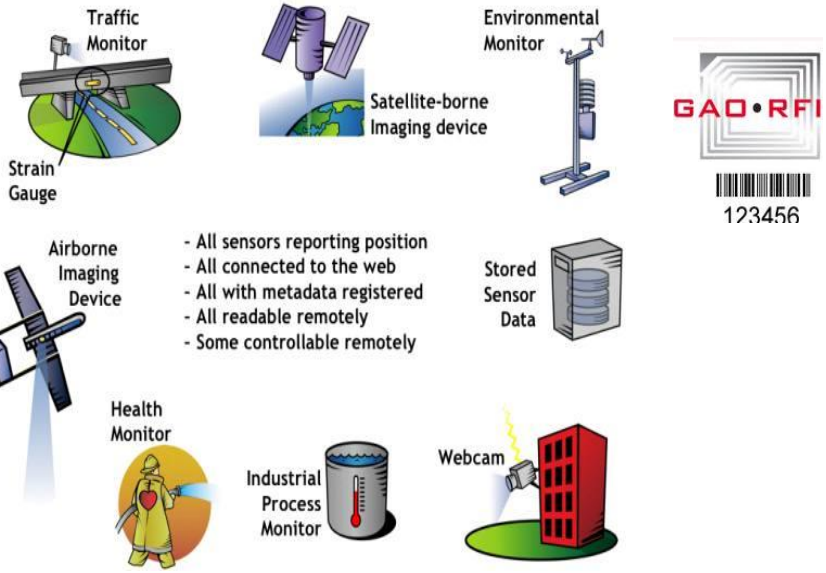
➢ Question & Discussion

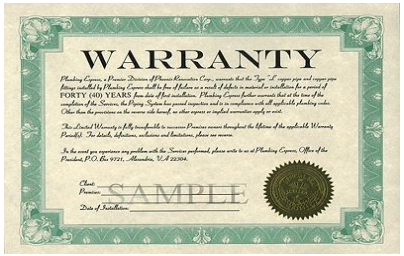
Technical Classification: Source of Unstructured Data

deserves special attention!

1. Social Media

2. Sensors/RFID and Barcode





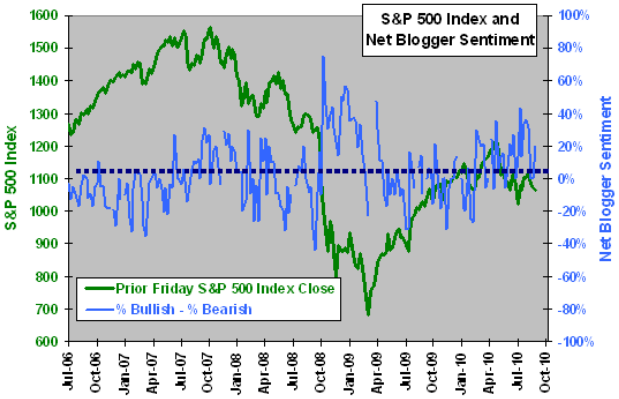
social media will not be the largest source of information in the long run, **sensor data** will create a lot more unstructured data

3. Other Unstructured/ Semi Structured Business Data

1. The Power of Social Media

News Sentiment

Actual market price



- Social networks sites are used by millions of users,
- Some of these postings are about businesses, products, operations and services.
- Example:
  - Bosch excellent customer engagement!
  - Kryptonite
- *Extracting information and transforming it into actionable knowledge requires business intelligence, and analytics capabilities."*
- **Tools to** uncover + integrate data-driven insights from blogs, social networks, groups, boards and other consumer generated platforms to your business.
- What are the technical challenges?

Challenge: Building a System of Records

- Extract and build a system of records and perform the same analytics on unstructured data that is currently possible on structured data
- Combine structured and unstructured data for seamless analysis
- Analyze content from many different facets on the fly
- Automatically identify and alert any unusual relations between data that might require your attention

Challenges

- Unlike carefully authored news text and regular web context, social media streams pose a number of new challenges, due to their large scale, short, noisy, context dependent, and dynamic nature.
- *Twitter: Feuerdrache: "Heute frueher Schluss!…Kommt doch rueber ! Geil, 1:7! Waz los Brazil?*"
- **Short Message:** Most Facebook + Twitter messages are very short. Semantic based methods supplement these with extra information
- **Noisy content:** unusual spelling, irregular capitalization (all upper or lower case), location based linguistic variances. Emoticons are used as sentiment indicators
- **Multilingual:** Social media content is strongly multilingual. Automatic language detection is a prerequisite
- **User generated content** is relatively small, corpus based statistical methods cannot be applied successfully
- **Social context** is crucial for the correct interpretation of social media content

Some Techniques Applied

- Natural language processing
- **Semantic annotations**: tying semantic models and natural language
- Opinion mining
- Dynamic creation of interrelationship between ontologies
- Information Extraction (IE): a form of natural language analysis, it is becoming the central technology in bridging the gap between structured and unstructured text and formal knowledge expressed in ontologies.
- Also crucial: the underlying database

What to look for in state of the art solutions?

- Most vendors implementation is based on sentiment classification that is keyword-based.
- In this approach, terms, mainly adjectives (e.g. awesome, awful, good, bad, love, hate) and fixed expressions (e.g. police state, on cloud nine), are used as sentiment indicators.
- The list of indicators can be prepared manually (the most common approach), composed semi- automatically, or acquired by machine learning algorithms that infer the best indicators from tagged samples in the domain of interest.
- the language taxonomies that are supported with natural language processing (NLP) provide the heart and soul to their solutions.
- We encourage looking beyond the pretty charts, plotting the sentiment over time to overlaying the actual comments (black/white buzz) and look for the inter-relationships. Don't use a simple comment count with a negative sentiment score as the metric for black buzz.

Examples for Applications, once data processing is done:

- SAP Trend Intelligence allows brand monitoring as well as voice of the customer, as well as correlation to the intranet
- Most state of the art tools are industry specific applications, e.g only for hospitals, or a particular other domain
- Trend Intelligence has a core NPL engine, data loader, data storage and analytical components, it provides the core technology to build all these applications and more

SAP Trend Intelligence
DEMO

Demo Focus:

Connecting Structured & Unstructured Data in MRP Domain

Conclusion: Benefits Are There, Challenges Remain!

Planning and focusing

effective tool for auditing an organization, products, services and its environment.

It is the first stage of planning and helps marketers to focus on key issues.

On the fly information:

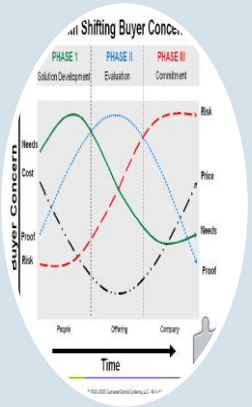
WIGO reveals the strengths, weaknesses, opportunities, and threats of a company/products.

helpful in matching the firm's resources + capabilities to the competitive environment in which it operates.

Opportunities

are critical external factors,

→WIGO can help to make these factors explicit, not based on a handful of opinions, but based on the voice of customers.



Competitive Advantage

Wigo can help to determine the strength parameter which is used as a basis for developing a competitive advantage

Strength can be the patents, strong brand names, good reputation among customers, specialist-marketing expertise, a new and innovative product or service, the location of the business, the quality processes and procedures, or any other aspect of the business that adds value to his product or service.

.

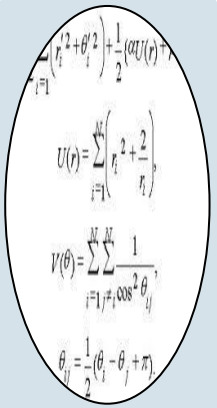
Market Entry and Threat Detection

It is moving into new market segments,

or new international markets that offer improved profits, mergers, joint ventures or strategic alliances,

even a market vacated by an ineffective competitor

In contrast to opportthe company from carrying unity: threat may prevent out a new policy.



Potential

Discover weakness of a business:

lack of marketing expertise,

undifferentiated products/ services, inconvenient location of business,

poor quality goods/ services, damaged reputation

Question & Discussion

- Panacea or pain:
  1. Social media is a panacea for my business
  2. The challenges, investment does not really pay off for my business

- Source is crucial for the output/
  1. **Source is crucial**: Random web crawling for social media may not be really beneficial, “garbage in, garbage out”. A more proper approach would be hand selected sources based on customer profiling, “where are my customers”, what are relevant sources
  2. **Not relevant:** This is contrary to many social media tools to capture sentiments & opinions. You might miss certain sources.

Language Support in Text Analysis XI 3.0

	Text Analysis Language Specific Processing*					Tools and Applications**			
Language	Pre-defined entity extraction	Concept extraction	Custom Catalog / Rules	Categorization	Summarization	Processing Manager***	ThingFinder Workbench	Categorizer Workbench	Annotation Manager
Arabic	✓	✓	✓	✓	✓	✓	✓	✓	✓
Catalan	X	✓	✓	X	✓	X	✓	X	X
Chinese (Simplified)	✓	✓	✓	✓	✓	✓	✓	✓	✓
Chinese (Traditional)	X	✓	✓	✓	✓	✓	✓	✓	✓
Croatian	X	✓	✓	X	X	X	X	X	X
Czech	X	✓	✓	✓	X	✓	✓	✓	✓
Danish	X	✓	✓	X	✓	X	✓	X	X
Dutch	X	✓	✓	✓	✓	✓	✓	✓	✓
English	✓	✓	✓	✓	✓	✓	✓	✓	✓
Farsi	✓	✓	✓	✓	✓	✓	✓	✓	✓
Finnish	X	✓	✓	X	✓	X	✓	X	X
French	✓	✓	✓	✓	✓	✓	✓	✓	✓
German	✓	✓	✓	✓	✓	✓	✓	✓	✓
Greek	X	X	✓	X	X	X	X	X	X
Hebrew	X	X	✓	X	X	X	X	X	X
Hungarian	X	X	✓	X	X	X	X	X	X
Italian	X	✓	✓	✓	✓	✓	✓	✓	✓
Japanese	X	✓	✓	✓	✓	✓	✓	✓	✓
Korean	✓	✓	✓	✓	✓	✓	✓	✓	✓
Norwegian (Bokmal)	X	✓	✓	X	✓	X	✓	X	X
Norwegian (Nynorsk)	X	✓	✓	X	✓	X	✓	X	X
Polish	X	X	✓	X	X	X	X	X	X
Portuguese	X	✓	✓	✓	✓	✓	✓	✓	✓
Romanian	X	X	✓	X	X	X	X	X	X
Russian	✓	✓	✓	✓	✓	✓	✓	✓	✓
Serbian	X	✓	✓	X	X	X	X	X	X
Slovak	X	✓	✓	X	X	X	X	X	X
Slovenian	X	✓	✓	X	X	X	X	X	X
Spanish	✓	✓	✓	✓	✓	✓	✓	✓	✓
Swedish	X	✓	✓	✓	✓	✓	✓	✓	✓
Thai	X	X	✓	X	X	X	X	X	X
Turkish	X	X	✓	X	X	X	X	X	X

Valid as of March 2009. Some features may vary per language; please consult the product documentation for full details.

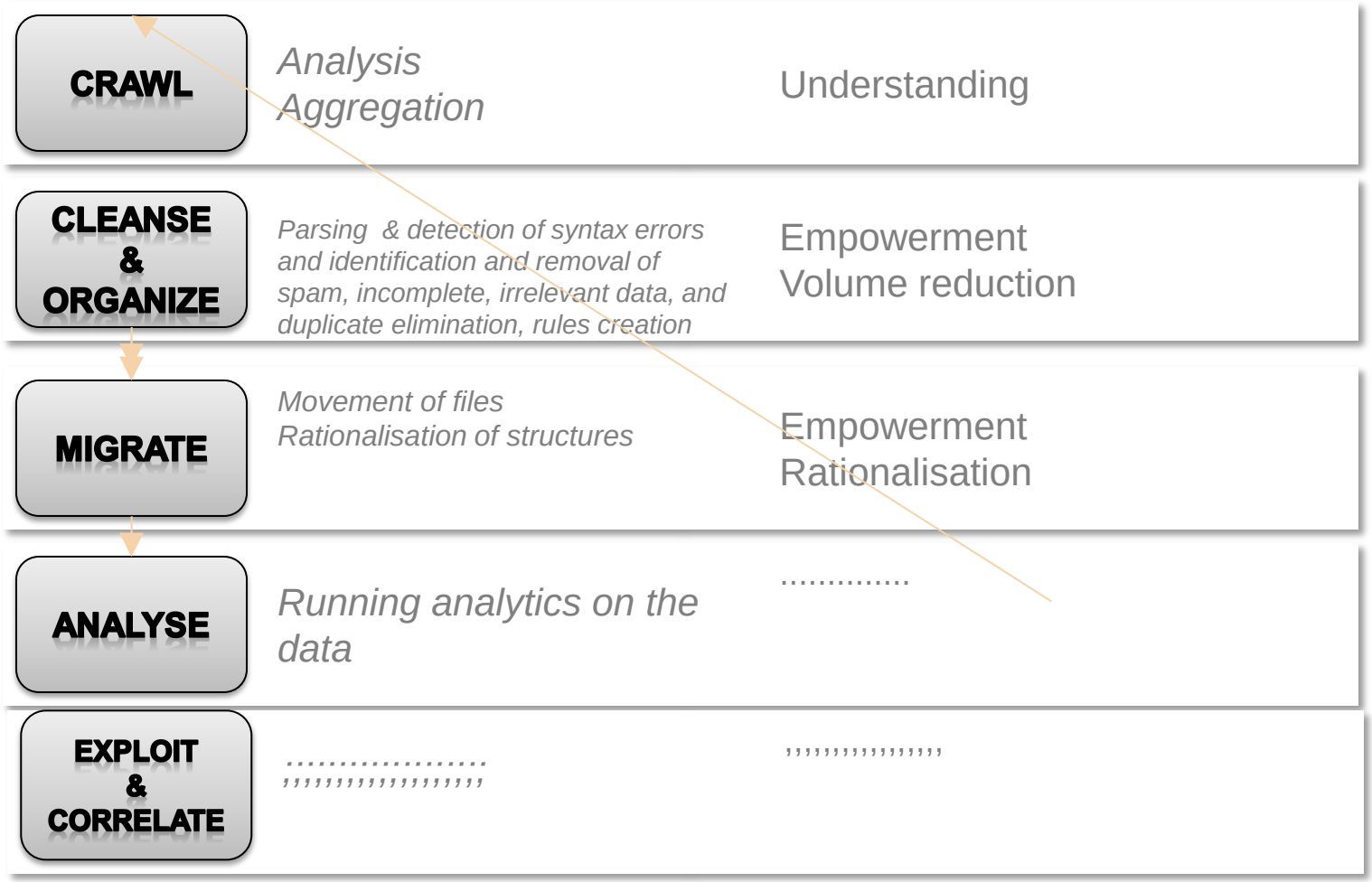
* This refers to the ability to process text in the given languages, performing language-aware analytics.

** These tools have been certified for use with specific languages, character sets, etc. It does not apply to the user interface or localization of any other part of the product, such as documentation

*** Processing Manager is certified for extraction and categorization only in Czech, and it is certied for categorization only in Swedish.

✓ Certification is coming in Text Analysis XI 3.0 SP2.

WIGO Overview of methodology -- tbd



Topic-driven data loading