

Web Intelligence for Improved Decision Making (WISDOM)

Final Presentation – January 22, 2014



Agenda

- 1. Introduction
- 2. Key technologies
- 3. Project highlights & publications

Key Technologies (1/3)

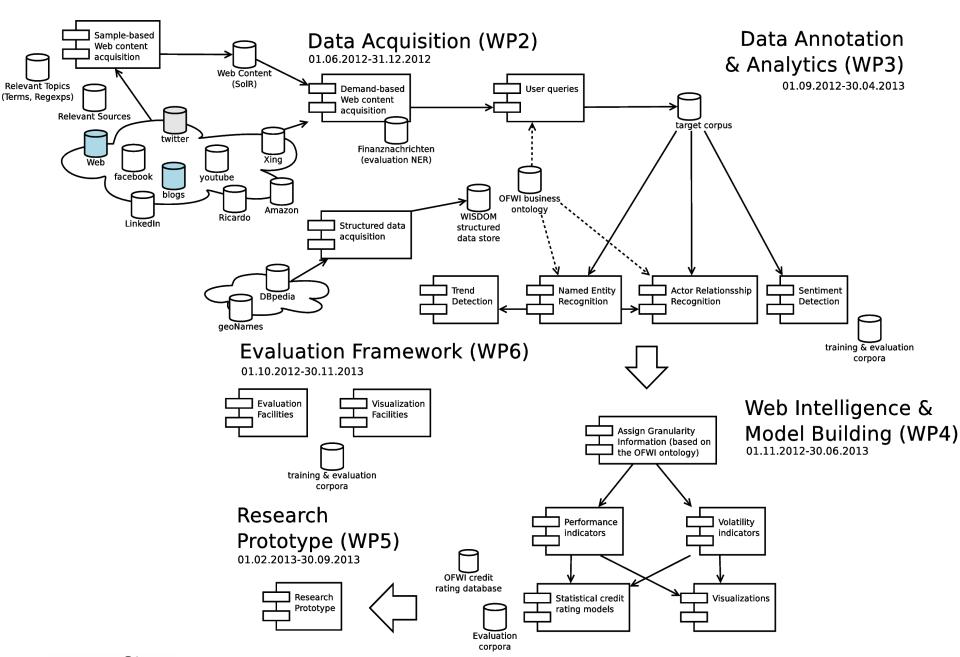
Key Technology	Application areas	Maturity
Linked Enterprise Data	Data integration	~ ~
WISDOM document repository	Web intelligence	✓ ✓
Multilingual context aware sentiment analysis	Automatically detect the sentiment polarity of Web articles.	De (
	Automatically identify and annotate named entities in Web documents.	Locations: (✓ ✓)
	Data quality and consistency checking; automatic suggestion	Companies: (✓ ✓)
	of invalid and outdated entities.	People: (✓)

Key Technologies (2/3)

Key Technology	Application areas	Maturity
Actor relationship assignment	Automatically identify relationships between key players. Identify clusters of companies and stakeholders. Data quality and consistency	Relation detection (✓) Assign entity classes (✓)
	checking; suggestion of missing relations.	Assign types (△)
	Automatically assign values (such as revenues, stock ticker symbols, growth) to entities.	Value assignment (△△)
Frequency and volatility based Web intelligence metrics	Web intelligence – assess the market volatility and media coverage.	✓ ✓

Key Technologies (3/3)

Key Technology	Application areas	Maturity
Network-based Web intelligence metrics / Spreading activation	Web intelligence – simulate how economic events affect interconnected company networks.	•
Visualization of Web intelligence metrics	Quickly assess a company's performance.	✓



Key Technologies

- 1. French sentiment analysis (Daniel)
- 2. Named entity resolution (Daniel)
- 3. Actor relationship detection & visualization (Norman)
- 4. Web intelligence and model building (Albert)
 - → frequency-based Web intelligence metrics
 - → network-based Web Intelligence metrics
- Prototype (Thomas)

- Sentiment analysis: identifying and aggregating polar opinions – i.e., positive or negative statements about facts
- Extend the existing framework to support French among English and German
- Tasks
 - 1. Evaluate a text processing framework
 - 2. Acquire suitable polarity lexicons
 - 3. Negation detection
 - 4. Evaluation
 - 5. Adaptation to the business domain

- Text processing
 - Text → Sentences → Tokens and word forms (POS)
 - Special characters and sequences
 - Word forms from an annotated corpus
- Stanford NLP
 - continuous ongoing development process
 - documented support for English and German
 - availability of a French tokenizer

The quick brown fox jumps over the lazy dog

<u>Token</u>	<u>Tag</u>	<u>Description</u>
The	DT	Determiner
quick	JJ	Adjective
brown	JJ	Adjective
fox	NN	Noun, singular or mass
jumps	VBZ	Verb, 3rd person singular present
over	IN	Preposition or subordinating conjunction
the	DT	Determiner
lazy	JJ	Adjective
dog	NN	Noun, singular or mass

Victor jagt zwölf Boxkämpfer quer über den großen Sylter Deich

<u>Token</u>	<u>Tag</u>	<u>Description</u>
Vicor	NE	Eigennamen
jagt	VVFIN	finites Verb, voll
zwölf	CARD	Kardinalzahl
Boxkämpfer	NN	normales Nomen
quer	ADJD	adverbiales oder prädikatives Adjektiv
über	APPR	Präposition; Zirkumposition links
den	ART	bestimmter oder unbestimmter Artikel
großen	ADJA	attributives Adjektiv
Sylter	NN	normales Nomen
Deich	NE	Eigennamen

Portez ce vieux whisky au juge blond qui fume

<u>Token</u>	<u>Tag</u>	Description
Portez	V	verb
ce	D	determiner
vieux	Α	adjective
whisky	N	noun
au	Р	preposition
juge	N	noun
blond	Α	adjective
qui	PRO	strong pronoun
fume	V	verb

Polarity lexicons

- Word lists with sentiment
- Resources
 - Amazon Reviews
 - General Inguirer Augmented Spreadsheet
 - UHZ SNF project "Bi-directional Sentiment Composition"

French Amazon customer reviews

- Approx. 25000 reviews with 4 or 5 stars (positive)
 Robuste, souple et agréable à toucher.
- Approx. 25000 reviews with 1 or 2 stars (negative)
 Inutilisable dans ces conditions.
- Naïve Bayes classifier
 - Convert reviews to feature sets
 - Train
 - Extract most informative features

French Amazon customer reviews: Evaluation

- Accuracy 0.87
- Precision+ 0.89
- Recall+ 0.85
- F-Score+ 0.87
- Precision- 0.86
- Recall- 0.89
- F-Score- 0.86

French Amazon customer reviews: most informative features

→ ~220 positive and ~190 negative terms

```
NEGATI : POSITI =
                                                             173.1:
   V(reçu) = 'reçu'
1.0
    V(décu) = 'décu'
                                    NEGATI : POSITI =
                                                         142.4 :
1.0
    V(d\hat{u}) = 'd\hat{u}'
                                    NEGATI : POSITI =
                                                             94.8:
1.0
    N(qoût) = 'qoût'
                                    NEGATI : POSITI =
                                                             89.4:
1.0
    N(\hat{a}me) = '\hat{a}me'
                                    NEGATI : POSITI =
                                                             65.3 :
1.0
    N(modération) = 'modération' POSITI : NEGATI =
                                                             56.7:
1.0
    N(rôle) = 'rôle'
                                    NEGATI : POSITI =
                                                             46.8:
1.0
HTWN(ใกษายี่ไ) = 'noël'
                                                             45.2:
                                    NEGATI : POSITI =
1 0
```

General Inguirer Augmented Spreadsheet

- 1. ignore ambiguous words
- 2. translate the words into German and French
- 3. keep triples consisting of three distinct words
- 4. remove triples which contain a french translation containing spaces
- 5. remove duplicate entries in French
- 6. eliminate misspelled tuples by applying Hunspell
 - → 1'194 words remain, 504 with positive, 687 with negative sentiment

UHZ SNF project "Bi-directional Sentiment Composition"

- 7'108 entries
- Positive, negative and ambiguous
 - → 1'926 positive and 3'348 negative terms

Word list evaluation

classify ~ 50'000 Amazon reviews

Word list	<u>Pos</u>	<u>Neg</u>	<u>Total</u>	<u>P+</u>	<u>R+</u>	<u>F+</u>	<u>P-</u>	<u>R-</u>	<u>F-</u>
Amazon	6'029	12'233	18'262	0.93	0.22	0.35	0.93	0.44	0.61
Inquirer	19'147	12'841	31'988	0.60	0.45	0.51	0.64	0.32	0.43
Sentimental.li	33'291	13'812	47'103	0.59	0.77	0.66	0.68	0.37	0.48
All lists combined	32'487	15'103	47'590	0.62	0.79	0.70	0.74	0.44	0.55

Negation detection

- the sentiment of words after a negation trigger is negated (default)
 - Je n'aime pas comme il joue.
 - Je ne veux pas de beurre.
 - Personne n'est venu.
- French negation trigger
- Improvement: invert the sentiment of the subsequent x words (window)

French negation trigger

Negation trigger	<u>Examples</u>	English translation
n'	Je n'aime pas comme il joue	I don't like how he plays
ne	Je ne veux pas de beurre	I don't want butter
non	Pourquoi non?	Why not?
pas	Je n'ai pas d' argent	I don't have money
plus	Je n'ai plus de monnaie	I don't have money anymore
guère	Je ne ris guère	I don't laugh often
jamais	Je ne pleure jamais	I never cry
rien	II n'a rien vu	He didn't see anything

HTW Chur

Evaluation

 classify ~ 50'000 Amazon reviews with Inquirer and sentimental.li lists

<u>Variant</u>	<u>P+</u>	<u>R+</u>	<u>F+</u>	<u>P-</u>	<u>R-</u>	<u>F-</u>
Default	0.59	0.78	0.67	0.70	0.38	0.50
Window 2	0.60	0.77	0.68	0.70	0.40	0.51
Window 3	0.60	0.77	0.68	0.70	0.41	0.52
Window 4	0.60	0.76	0.67	0.70	0.42	0.52

Adapting the sentiment lexicons to the business domain

- Combine the three lists
 - 1'096 entries from the Inquirer list
 - 5'274 entries from the Sentimental.li list
 - 417 entries from the Amazon list
- Classify 130'000 French AWP messages
- Use messages with a polarity of +/-0.25 to train a Naïve Bayes classifier
- Extract new most informative features
 - → ~140 new positive/negative terms each

- Recognyze component (Java)
 - Identify:
 - Locations
 - People
 - Organizations
 - Assign entites to Linked Open Data (LOD) resources
- Architecture
- Workflow
- Algorithms
- Evaluation

Architecture

- Linked open/enterprise data repository
- Configuration
- Recognyze profile (Lexicon, disambiguation and search)
- REST api

Workflow

- Indexing
- Search

Linked open/enterprise data repository

- URI
- Names
- Context information (Text, Turnover)

Configuration

- Repository to query
- Sparql query
- ResultHandler (Lexicon type, Indexing, Disambiguation)
- Stopwords, Filters, Entity type

Recognyze profile

- Lexicon (Geo, Person or Organization)
- Disambiguation (Geo, Person, Disambiguation w/o context)
- Search (close to O(1))

REST api

- Add, list, de-/serialize, remove profiles
- Search (text/XML, serial/parallel, output format, combined search)
- Various actions to inspect the component and profiles

1. Indexing

- Query the repository
- Process the retrieved data
- Build the search

2. Search

- Find matching entities
- Disambiguate and score
- Return results (entities, text positions and confidence) in the desired output format (standard, minimal and annie)

Indexing algorithm for organizations

- Generate short names (suffixes and affixes)
- Generate non-composite variants (case, umlauts)
- Use as name (length, stopwords, firstname/lastname filter)
- Generates names, ambiguous names and alternative names

Search algorithms

- Geo: Relevance matrix, focus location, population, tree length
- Person and Location:
 - Lucene similarity: combination of
 - retrieval (boolean model) and
 - weighting (Vector Space Model)
 - Boosting for terms (occurrences) and fields (names, keywords
 - Ground on names, separation
- Rescore

Evaluation

- Regarding throughput, memory consumption, precision and recall
- Iterative improvements:
 - Lowercase/capitalize
 - Dictionaries
 - Context
 - Name normalization
 - Un-/ambiguous needles

Relation Extraction | Introduction

Goal:

Detect relations between business entities in text documents. Classify (or group) the named entities.

Business Entity: Company, Person

Relation: Different types of relations

Relation Extraction | Example

Example Docment:

Richemont will den britischen Onlinehändler NET-A-PORTER Limited übernehmen.

Detection:

Richemont related to NET-A-PORTER.

Relation Extraction | Classes

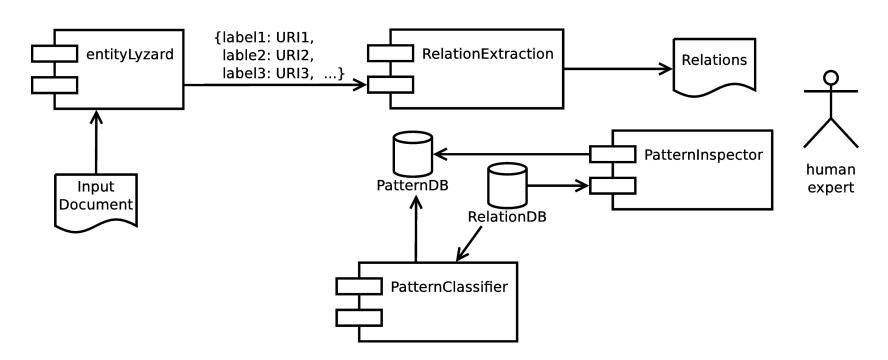
Example Document (frequent in the awp news): Schwergewichte im Aktionariat, Compagnie Financière Rupert, 50,02% Public Investment Corp. Ltd (PIC), (ZA), Verwaltungsrat, Johann Peter Rupert, Präsident, Richard Lepeu, Franco Cologni, Ruggero Magnoni, Dominique Perrin...

Goal:

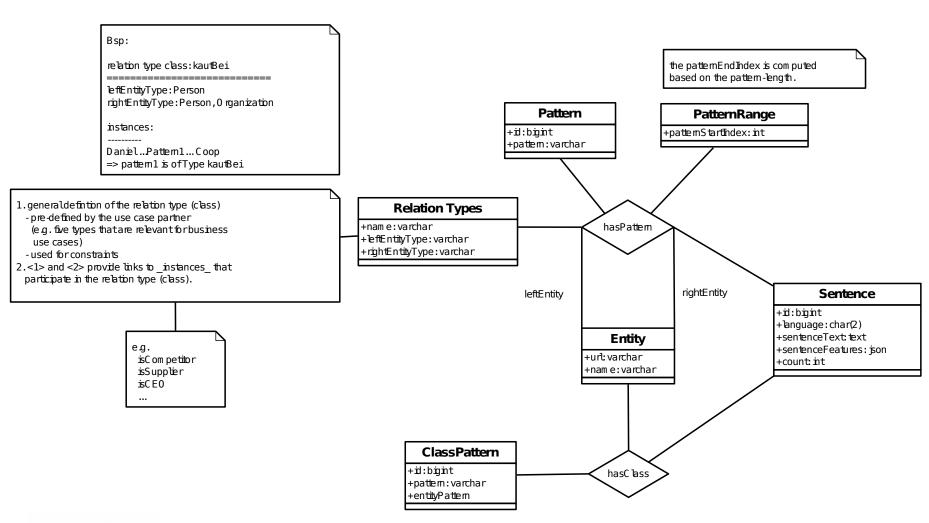
Assign all named entities to the same class

Relation Extraction | System Architecture

ch.htwchur.wisdom.relative



Relation Extraction | UML



Relation Extraction | Technical Concepts

- Relation patterns Canonical Form
 - Term > Stem > POS > Person > Organization > Datatype>
- Two-level matching (Regular Expression)

- Enumeration handling (Detected with Regular Expressions):
 - Das Mangement der Liwet Holding AG und Bergean Holding AG unter Vladimir Kuznetov, Präsident Kurt Hausheer, Urs Meyer, entschied
 - ORG1 und ORG2, ... IN PERS1, PERS2, PERS3 VRB \rightarrow ORG1 IN PERS1 VRB , ORG2 IN PERS3 VRB

Reports | Relation Extraction

- Based on Open Information Extraction
 - extraction patterns → entities → extraction patterns
- Assign patterns and entities to relation types (isCompetitor, isCEO, ...)
 - Rule based approach
 - Machine learning

Relation Type	Pattern	Entity 1	Entity 2	Sentence Snippet
isCompetitor	ORG werben ORG NN ab.	CS	UBS	CS wirbt UBS Kunden ab
isCompetitor	ORG schlagen ORG	UBS	CS	UBS schlägt CS mit
isCEO	PER leiten ORG	David	UBS	David leitet UBS
isCEO	ORG werben PERS ab	CS	David	CS wirbt David ab

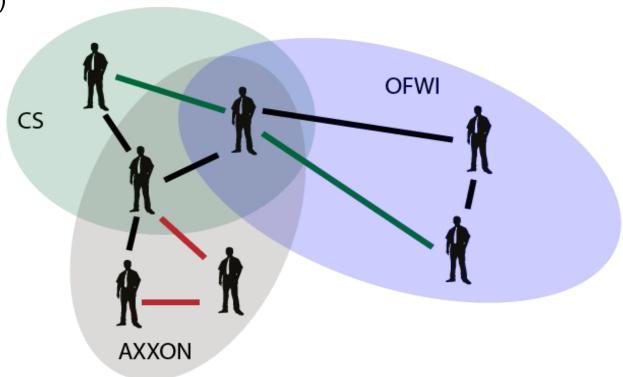
Actor relationship visualization

- Evaluation of different visualization approaches and visualizations
- Develop a visualization concept
- Implementation of the concept using the D3-Framework and GraphML
- Data is provided by the WISDOM relation extraction component (Relative)

Relation visualization | Dimensions

Three visualization dimensions

- Entities (nodes)
- Relations (edges)
- Classes (cloud)



Relation visualization | GraphML

Example Data delivered by Relative Rest Interface:

```
<node id="http://www.semanticlab.net/proj/wisdom/ofwi/person/Jürg Präsident (018245)</pre>
) ">
<data key="entityClasses"></data>
<data key="name">J&#252;rg Pr&#228;sident</data>
<data key="entityType">Person</data>
</node>
<node
id="http://www.semanticlab.net/proj/wisdom/ofwi/person/Sergio Marchionne (024586)">
<data key="entityClasses"></data>
<data key="name">Sergio Marchionne</data>
<data key="entityType">Person</data>
</node>
<edge directed="false" source="</pre>
http://www.semanticlab.net/proj/wisdom/ofwi/person/Jürg Präsident (018245"
target="http://www.semanticlab.net/proj/wisdom/ofwi/person/Sergio Marchionne (024586
)"/>
```

Relation visualization | Technologies

JQuery

- Straightforward implementation of asynchronous Javascript (AJAX)
 Clean Code
- Cross Browser Compatibility

Visualization Framework: D3

- Javascript Visualization Framework, Force Layout "out of the box", highly customizable
- Based on Standarts (HTML, Javascript, SVG)

Relation visualization | Concepts

Link distance between nodes:

Distance between nodes with same classes: short Distance between nodes with different classes: long

Link weight:

Number of equal relations (count relations with the same entitys from the result dataset) Saturate the link color and link thickness according to the link weight

Group nodes from according to their classes:

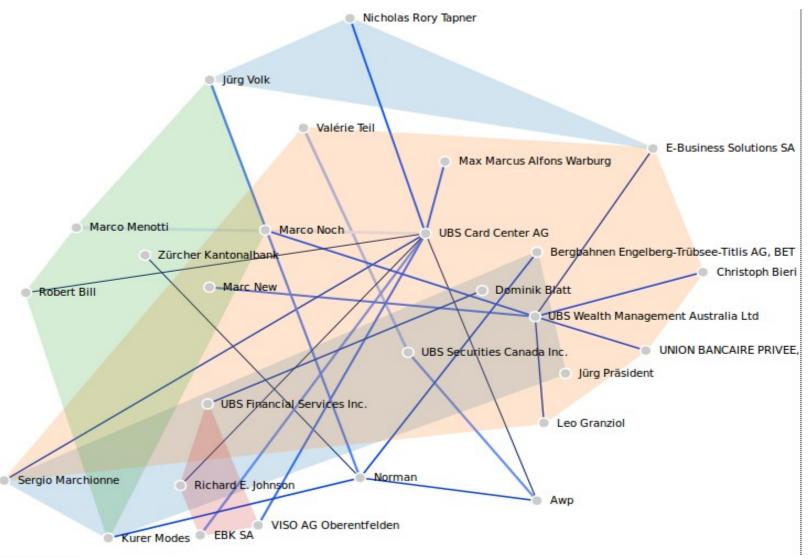
Cluster center: Node with highest link weight (for every class)

Challenge:

Multiple forces on the graph at the same time \rightarrow best compromise?

HTW Chur

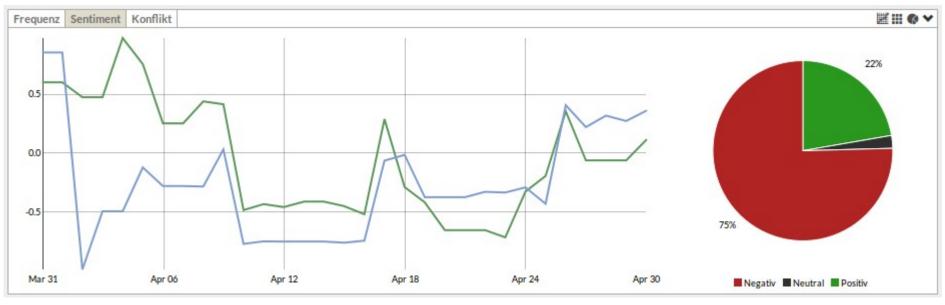
Relation visualization | Prototype



Web intelligence metrics

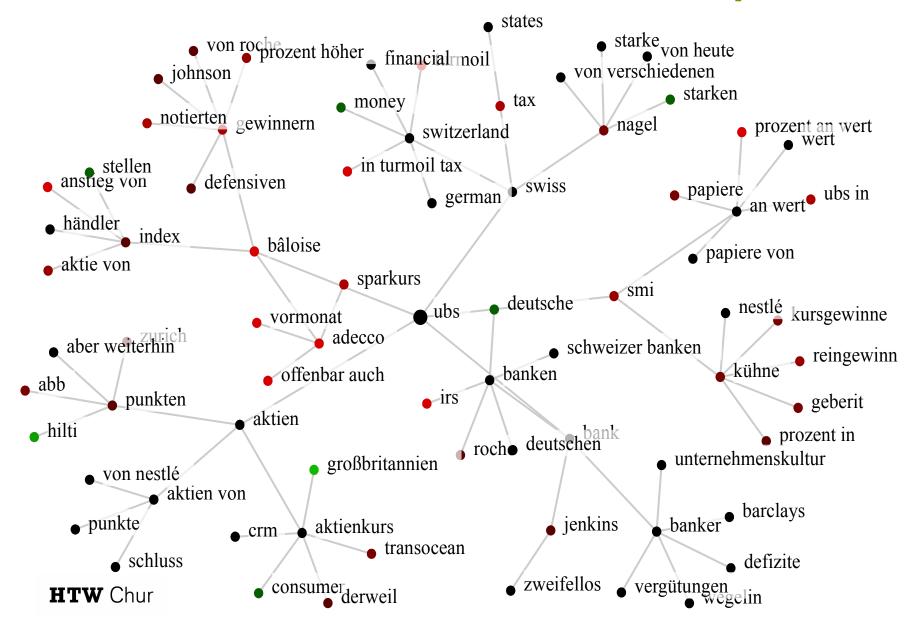
- Frequency based metrics
 - volatility (terms, named entities)
 - sentiment (polarity, deviation)
 - associations
 - trends
- Network based metrics
 - standard network metrics and analytics
 - Influence networks

Sentiment and Conflict Indicator for UBS

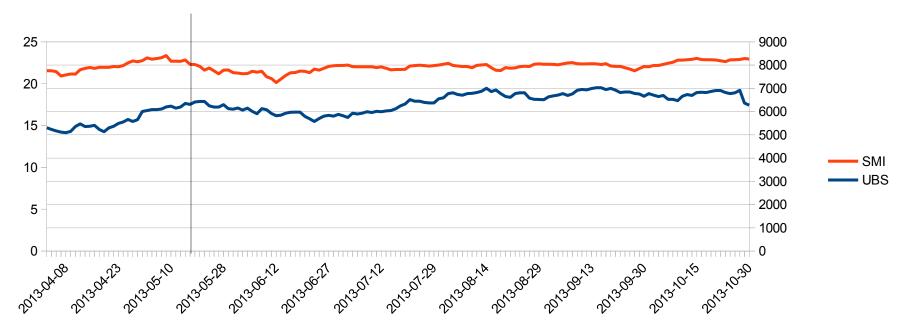


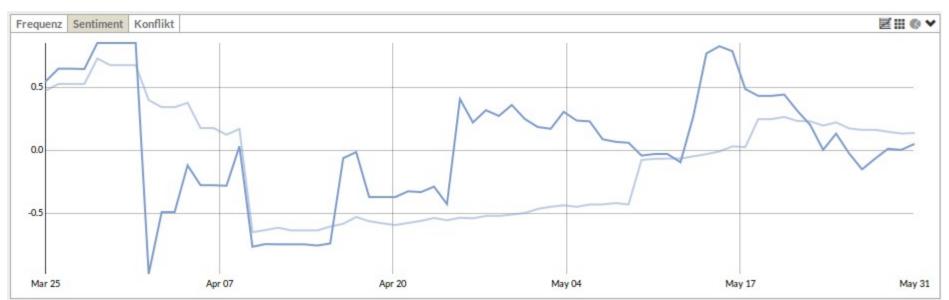


Associations and Relevant Concepts



Comparison with Real-World Indicators





Web intelligence metrics

- Network-based metrics
 - business relations (Lau, R. et al., 2012)
 - competitive advantage dimensions based on Porter's five forces model
 - bargaining power of customers / suppliers
 - thread of substitutes
 - intensity of rivalry
 - thread of new entrants
 - assess the relative credit rating, revenue, turnover within an industry
- Influence networks
 - How do strengths and weaknesses of companies impact related companies.

Web intelligence metrics

- Influence networks / Spreading activation
 - builds upon the actor relationship networks
 - parameters:
 - relations between actors (type)
 - relations strength (number of occurrences)
 - "firewall effects" (legal constructs, ...)
 - sources (revenue, turnover, contracts, ...)
 - sinks (liabilities, risks, ...)
- Translation: influence network → spreading activation network
 - perform simulations (similar to neural networks)

Usability Evaluation – Study Design

- Multi-level approach (expert and user-orientated methods)
 - Heuristic evaluation (ISO-Norm 9241-110 and 9241-12, Nielsen's 10 usability heuristics, Shneiderman's eight golden rules of interface design)
 - Formative usability test (including thinking aloud and eye tracking)



Heuristic Evaluation

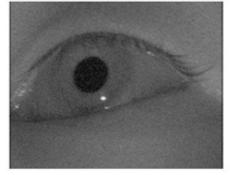
 Assessment of the user interface and judgement about its compliance with recognized usability principles ("heuristics")

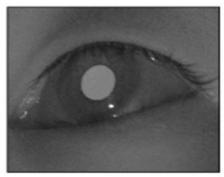
ISO 9241-110	ISO 9241-12
Suitability for the task	Clarity
Suitability for learning	Discriminability
Suitability for individualization	Conciseness
Conformity with user expectations	Consistency
Self descriptiveness	Detectability
Controllability	Legibility
Error tolerance	Comprehensibility

Formative usability test

- Observation of users while working on predefined tasks (screen capturing user interaction, web cam facial expressions)
- Probands simultaneously described their actions and thoughts (audio recording method of thinking aloud)
- Additionally recording of the gaze data of the test subjects (eye tracking)

Test subject	<u>Gender</u>	<u>Age</u>
Proband A	male	36
Proband B	female	49
Proband C	male	25
Proband D	female	26



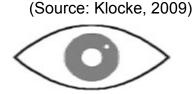








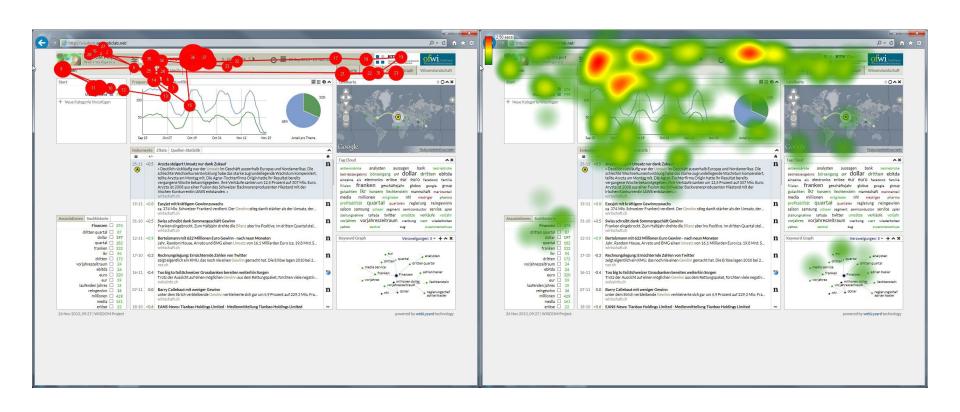
Looking under camera



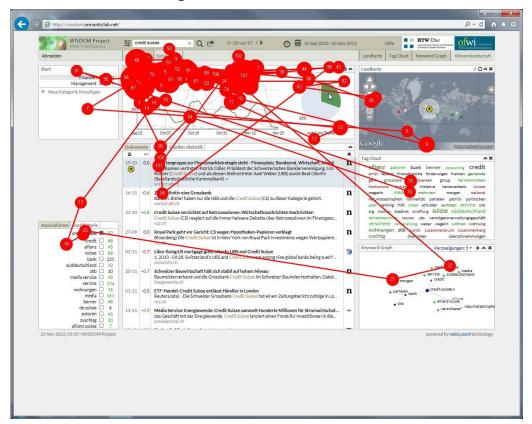
Looking under camera and to the right

HTW Chur

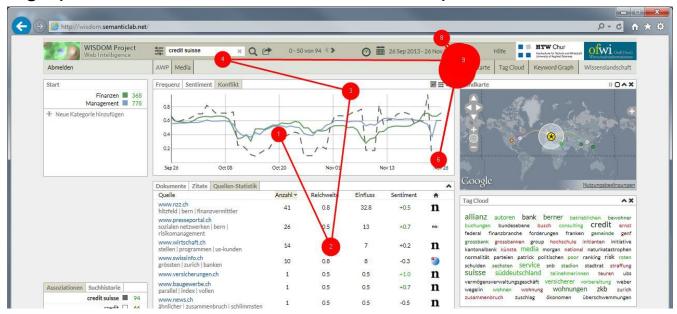
Eyetracking – Examples



- Probands had no problems in choosing the desired time period and performing searches
- Initially difficulties in finding the search results in the line chart

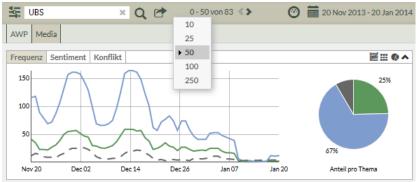


- No problems in switching between different views
- Tab concept is well suited and intuitive
- Also most icons are quite intuitive (except sentiment barometer)
- However, for some UI elements tool tips are missing
- Sorting options of the result list should be emphasized



- Users had problems in limiting the results to only positive or negative documents (icon not intuitive)
 - Test subjects tried to set limitations by clicking on the green/red areas of the pie chart
- Problems in increasing the number of simultaneously displayed results
- Export functions could be used easily by all probands





- Visual analytic tools could also be used without problems by most users
 - Maximizing the geographic map covers the other areas
 - Setup icon of the geographic map not consistent with the rest of UI
 - Problems in distinguishing the color shades, which are used for the visualization of the sentiment values
- Purpose of the topics area for novel users not immediately understandable
 - Possibility to combine several search terms by using categories was not recognized
 - Problems in editing topics (topic editor)

Conclusion

- No major usability issues could be identified
- Discovered shortcomings have no big impact on users
- WISDOM prototype is a powerful tool a certain level of complexity can not be avoided
- Evaluation results show that novel users can learn to use the tool rather quickly
 - At the beginning partially some problems with the real-time manipulation options
 - Users did not recognize immediately the impact of their actions
 - After a short learning phase this problem could be overcome

Published Peer-reviewed Journal and Conference Articles

- Weichselbraun, Albert, Gindl, Stefan and Scharl, Arno. (2013).
 Extracting and Grounding Context-Aware Sentiment Lexicons. IEEE
 Intelligent Systems 28 (2): 39-46
- Gindl, Stefan, Weichselbraun, Albert and Scharl, Arno. (2013).
 Rule-based Opinion Target and Aspect Extraction to Acquire
 Affective Knowledge. First WWW Workshop on Multidisciplinary
 Approaches to Big Social Data Analysis (MABSDA 2013), Rio de
 Janeiro, Brazil
- Weichselbraun, Albert, Scharl, Arno and Lang, Heinz-Peter. (2013).
 Knowledge Capture from Multiple Online Sources with the Extensible Web Retrieval Toolkit (eWRT). Seventh International Conference on Knowledge Capture (KCAP-2013), Banff, Canada

Articles under Review

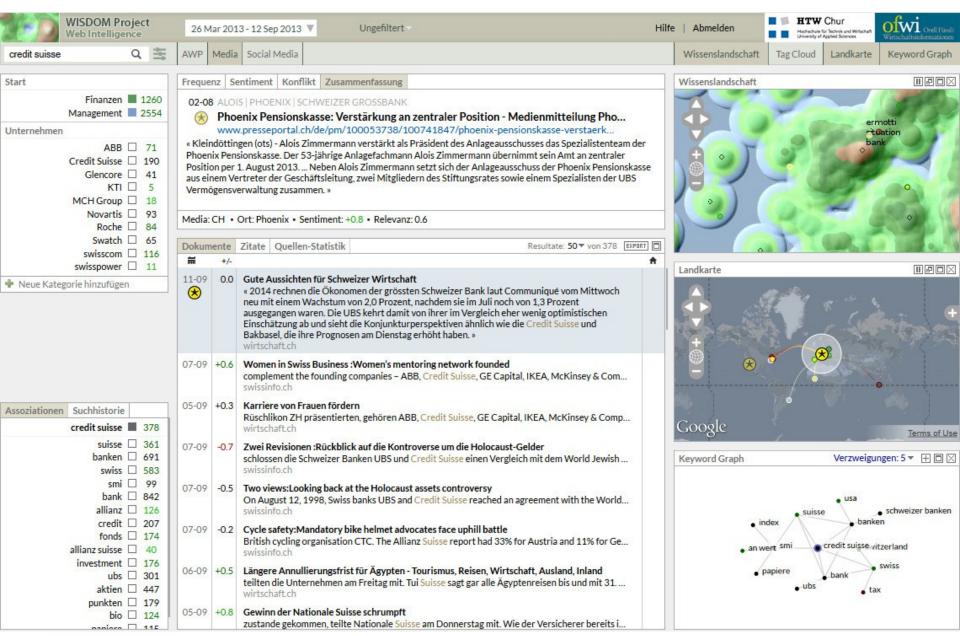
- Weichselbraun, Albert, Gindl, Stefan and Scharl, Arno.
 Enriching Semantic Knowledge Bases for Opinion Mining in Big Data Applications, submitted to Knowledge-Based Systems Special Issue on Big Data for Social Analysis
- Weichselbraun, Albert, Schreiff, Daniel and Scharl, Arno. Linked Enterprise Data for Fine Grained Named Entity Linking and Web Intelligence, submitted to the International Conference on Web Information Systems and Mining (WISM 2014).

Master and Bachelor Thesis

- Michael Aschwanden. (2013). Konzipierung eines Leitfadens zur Handhabung heterogener und dezentraler Datenquellen, Master thesis, University of Applied Sciences Chur
- Laurin Wegelin. (2013). Follow the best, Bachelor thesis,
 University of Applied Sciences Chur
- Franziska Walser. (2012). Named Entity Recognition in deutschsprachigen Texten, Bachelor thesis, University of Applied Sciences Chur

Speeches and Presentations

- Weichselbraun, Albert (2012). Coping with Evolving Knowledge - Dynamic Domain Ontologies for Web Intelligence, Invited Speech, 11th International Workshop on Web Semantics and Information (WebS 2012) in conjunction with the 23rd International Conference on Database and Expert Systems Applications (DEXA 2012)
- The Rector's Conference of the Swiss Universities of Applied Sciences (KFH) has selected the WISDOM project as one of the three research projects to be presented at the Swiss Innovation Forum in November 2013.



Miscellaneous