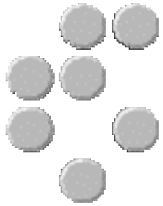


Text Mining and beyond

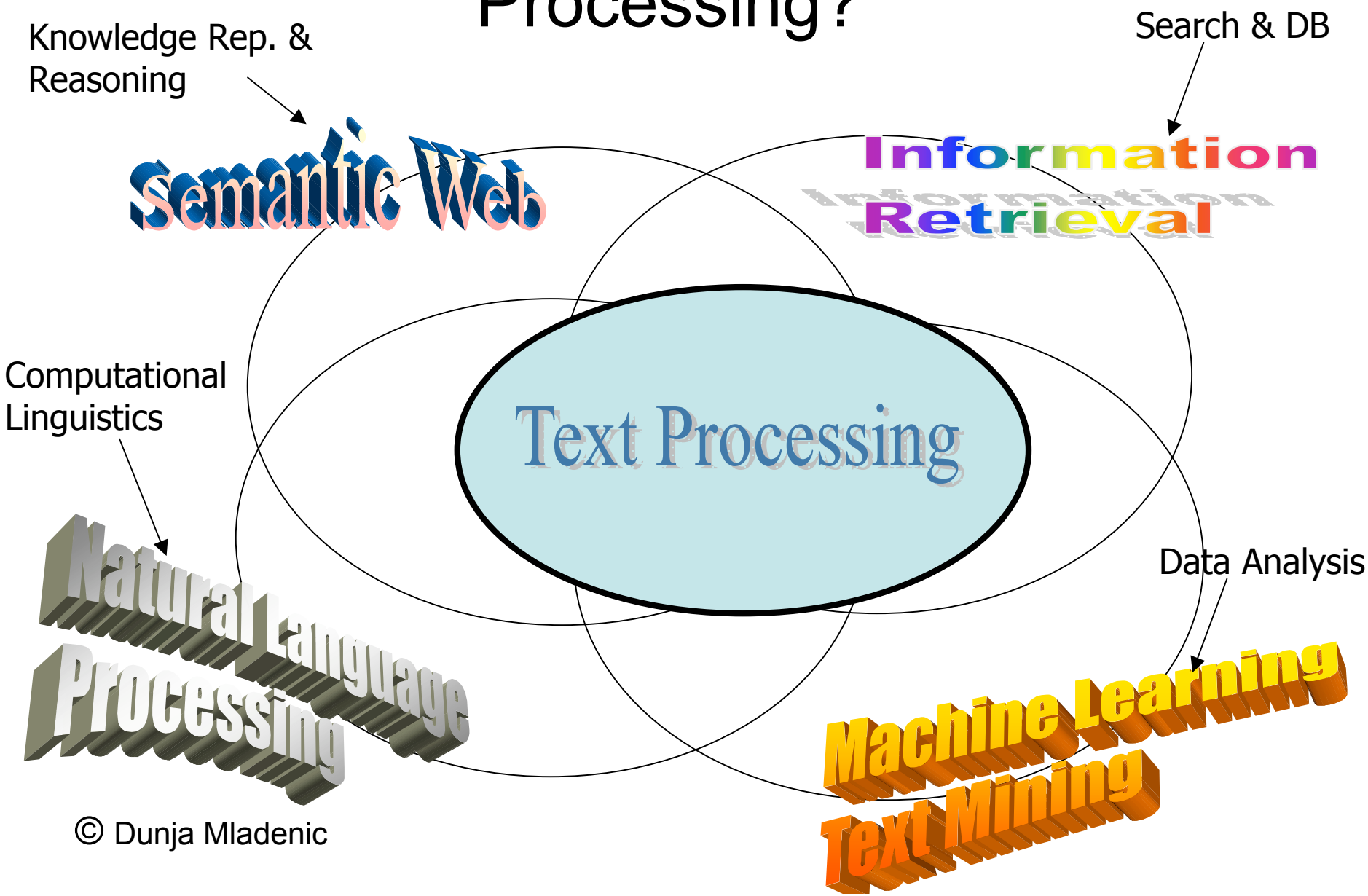
Dunja Mladeni_
J. Stefan Institute, Ljubljana, Slovenia
Dunja.Mladenic@ijs.si

What is Text-Mining?

- "…finding **interesting** regularities in large **textual** datasets…" (Usama Fayad, adapted)
  - …where **interesting** means: non-trivial, hidden, previously unknown and potentially useful
- "…finding semantic and abstract information from the surface form of textual data…"

© Dunja Mladenic

Which areas are active in Text Processing?

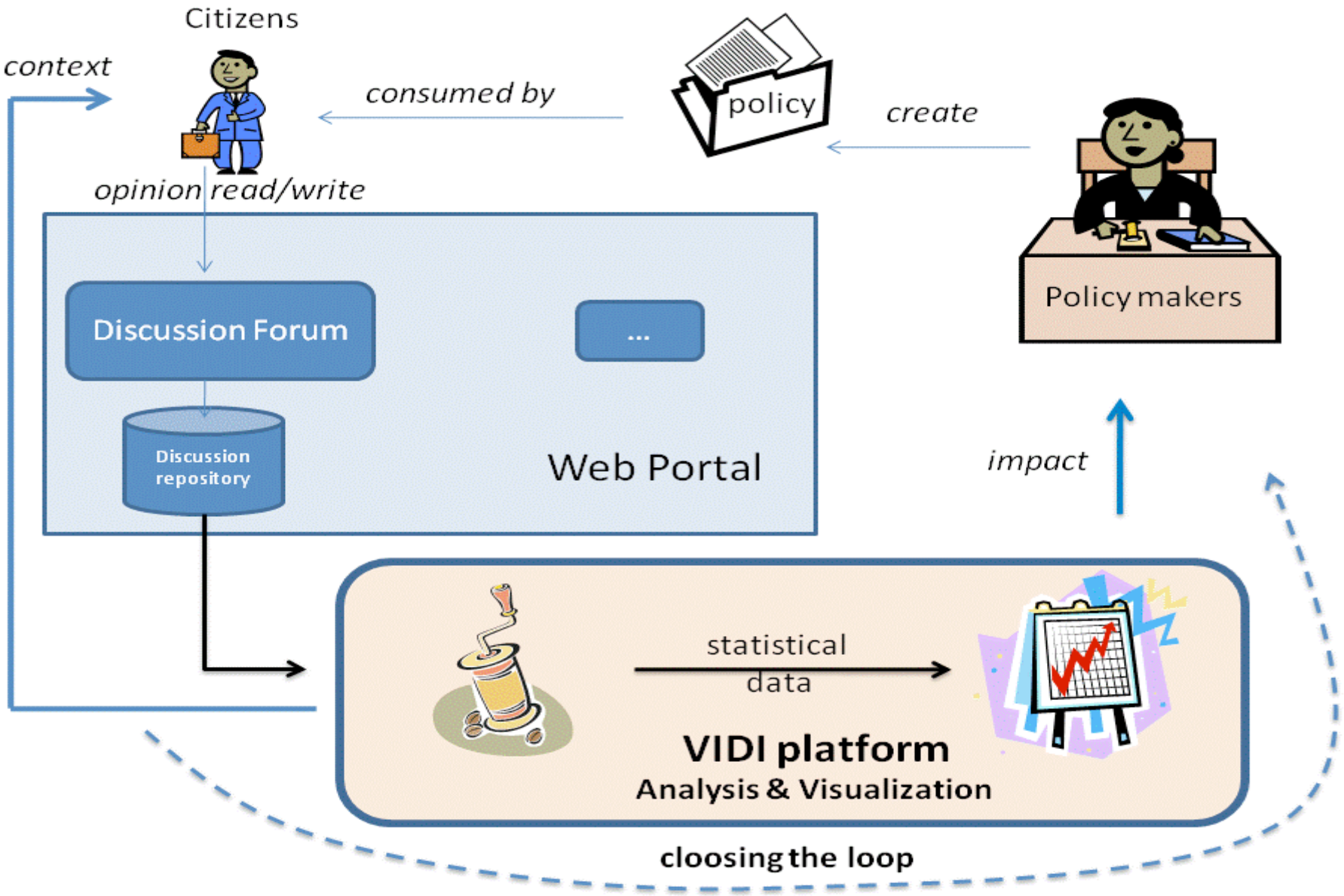


© Dunja Mladenic

Example scenarios

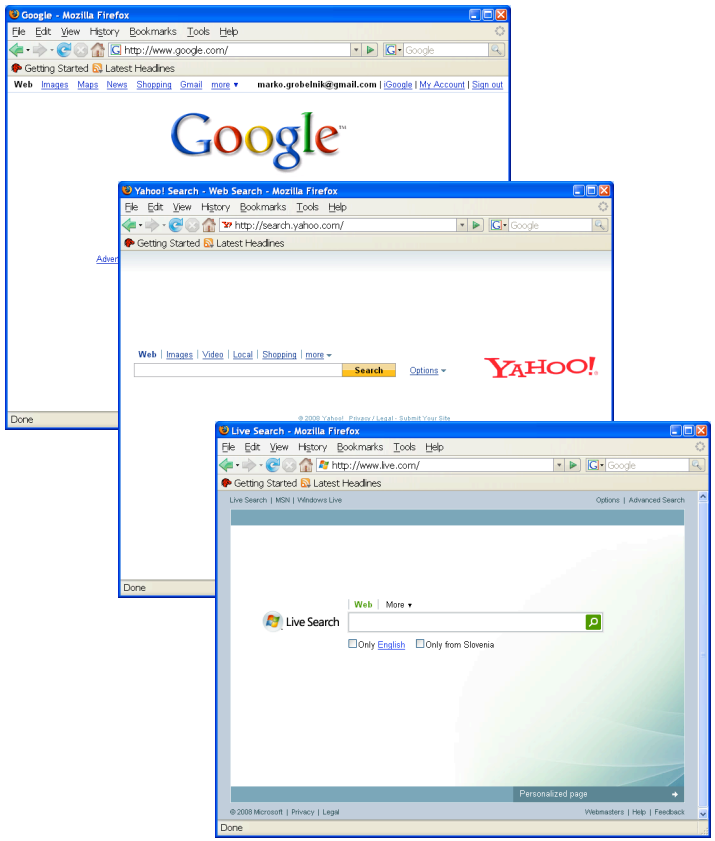
- Public opinion visualization
- Search on the Web
- Document collecton visualization
- Extracting structured data from text
- Analysis of collaborations on EU projects
- Real-Time information processing

VIDI


Citizens
context
consumed by
policy
create
opinion read/write
Policy makers
Discussion Forum
...
Discussion repository
Web Portal
impact
statistical data
VIDI platform
Analysis & Visualization
cloosing the loop

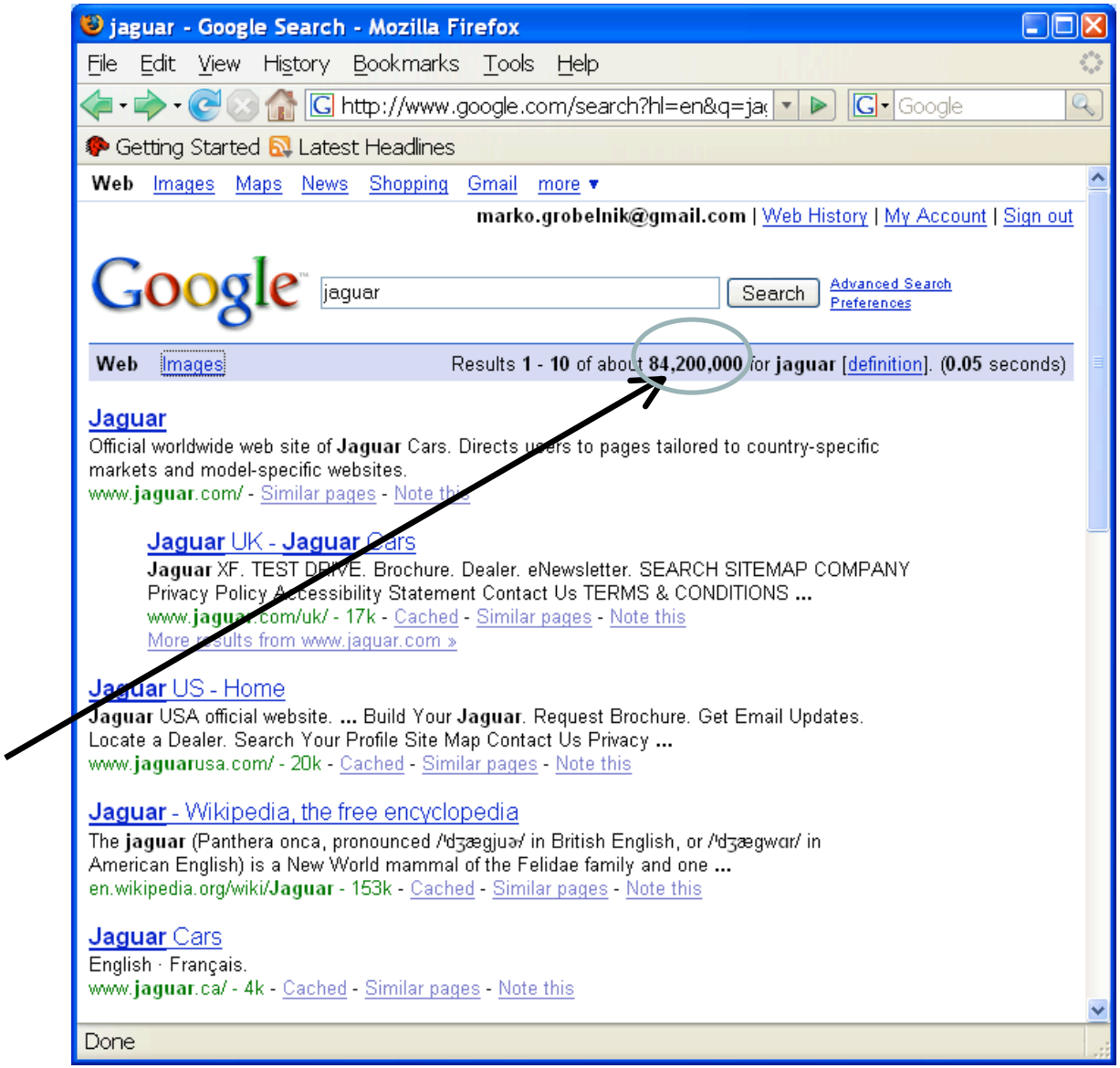
Semantic Web – search

- What are the most common tasks where we manipulate texts in everyday life?
  - "Internet search"!
- …but – how smart is search technology today?
  - …not too smart!
  - It is sophisticated, but not smart



Example – Searching for “jaguar”

- Query “**jaguar**” has many meanings…
- …but the first page of search engines doesn’t provide us with many answers
- …there are 84M more results



Context sensitive search

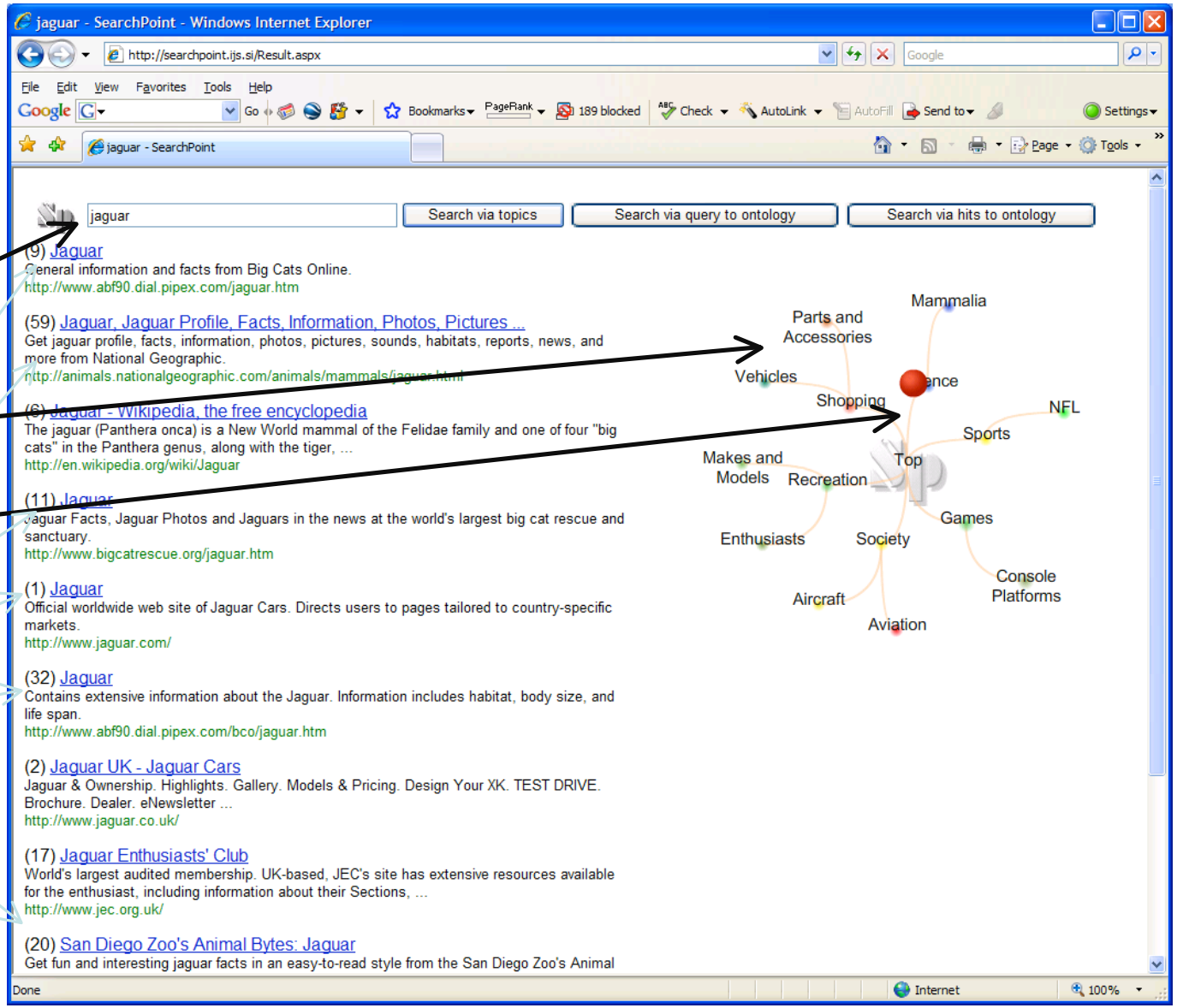
http://searchpoint.ijs.si

Query

Conceptual map

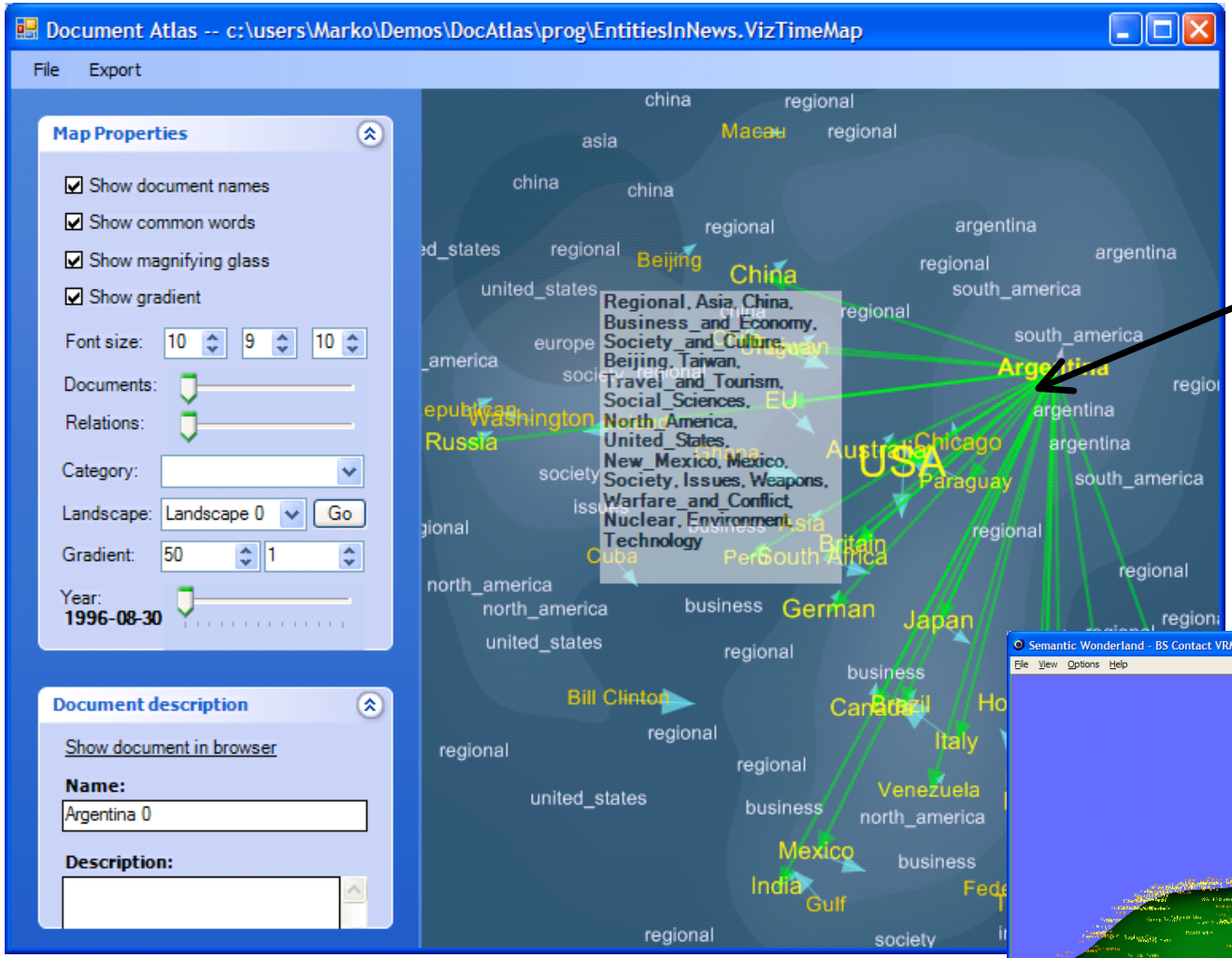
Search Point

Dynamic contextual ranking based on the search point



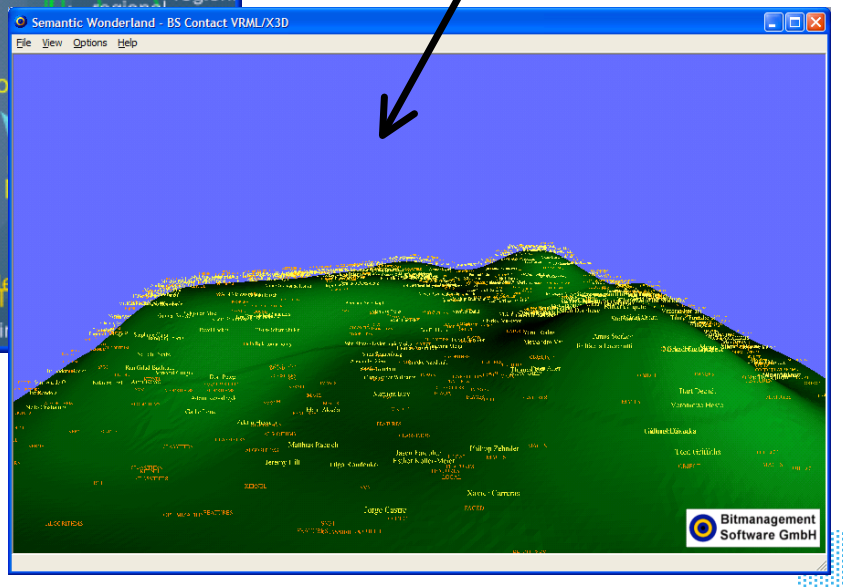
DEPARTMENT OF KNOWLEDGE TECHNOLOGIES
Jožef Stefan Institute

Document collection visualization



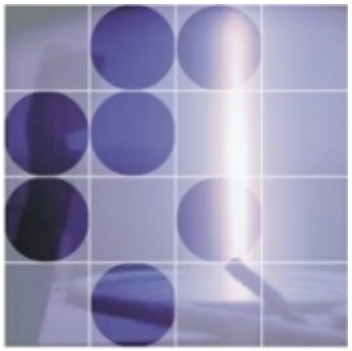
Relationships between entities in the news

3D version



http://docatlas.ijs.si

DEPARTMENT OF KNOWLEDGE TECHNOLOGIES
Jožef Stefan Institute



Blaz Fortuna, Delia Rusu, Lorand Dali, Ruben Sipo_, Marko Grobelnik , Dunja Mladeni_

Jo_ef Stefan Institute, Slovenia (http://www.ijs.si/)

EXTRACTING STRUCTURED DATA

Congress approves budget plans, long road ahead

Fri Apr 3, 2009 3:12am EDT

Email | Print | Share | Reprints | Single Page [-] Text [+]

1 of 1 Full Size

By Jeremy Pelofsky and Richard Cowan

WASHINGTON (Reuters) - The Democratic-controlled U.S. Congress on Thursday approved budget blueprints embracing President Barack Obama's agenda but leaving many hard choices until later and a government deeply in the red.

With no Republican support, the House of Representatives and Senate approved slightly different, less expensive versions of Obama's $3.55 trillion budget plan for fiscal 2010, which begins on October 1. The differences will be worked out over the next few weeks.

Obama, who took office in January after eight years of the Republican Bush presidency, has said the Democrats' budget is critical to turning around the recession-hit U.S. economy and paving the way for sweeping healthcare, climate change and education reforms he hopes to push through Congress this year.

Obama, traveling in Europe, issued a statement praising the votes as "an important step toward rebuilding our struggling economy." Vice President Joe Biden, who serves as president of the Senate, presided over that chamber's vote.

Democrats in both chambers voted down Republican alternatives that focused on slashing massive deficits with large cuts to domestic social spending but also offered hefty tax breaks for corporations and individuals.

ongress

es
_voting
val-noun-4
bEbGdrcN5Y2

budget
dbpedia:Budget

pencyc:Mx4rvVjS0JwpEbGdrcN5Y2
9ycA

ves budget

Assertions (e.g. Cyc)

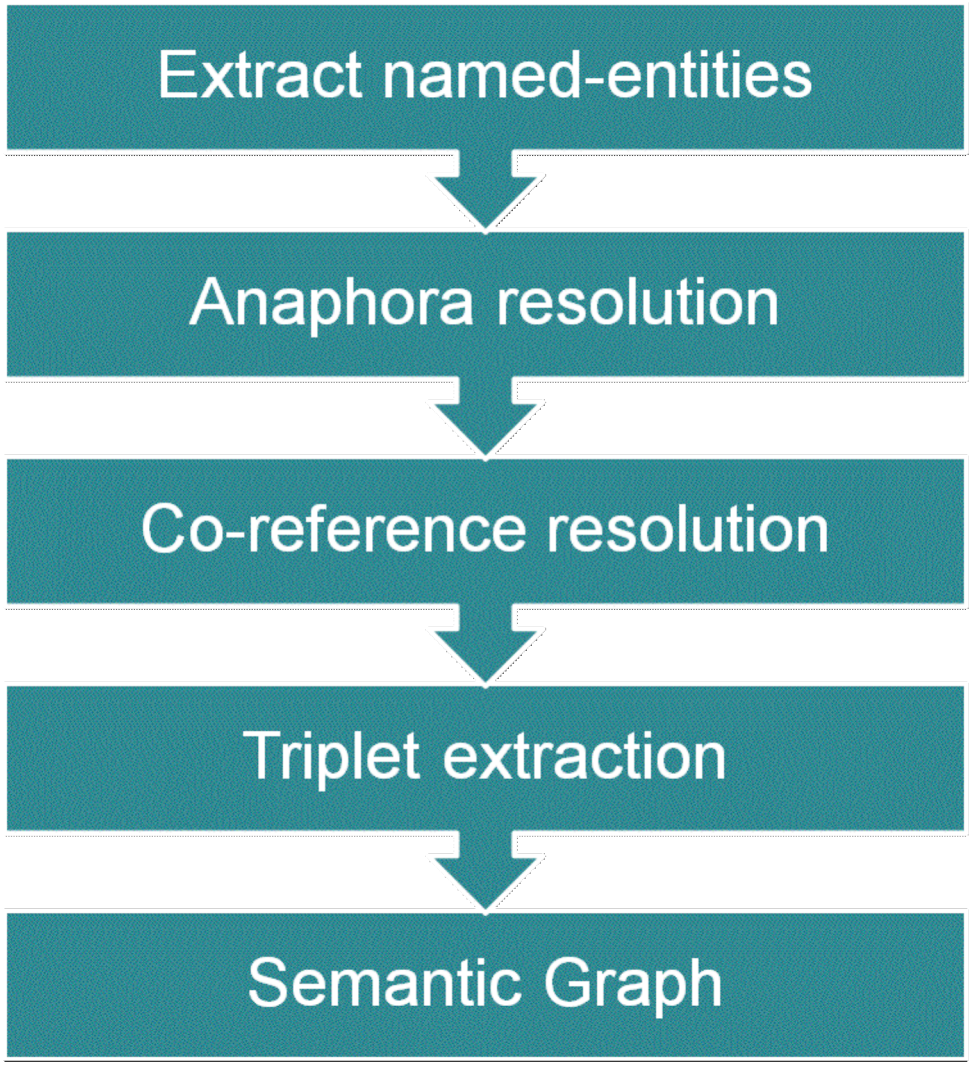
concepts

text

Expressivity

As of Feb (#$February). 24 (24), Air Force (#$UnitedStatesAirForce) officials (#$PublicOfficial #$OrganizationRepresentative) reported (#$RegisteringAComplaint #$Reporting) that personnel (#$Employee) in the area (#$Area 0 #$FieldOfStudy #$Region-Underspecified) numbered (#$Counting) close to 8,000 (8000). The 100 (100) aircraft (#$AirTransportationDevice) based (#$Base-Support #$MilitaryBase-Grounds #$BaseOfLandProtrusion #$NitrogenBase #$ChemicallyBasicSubstance) in Saudi Arabia (#$SaudiArabia) for patrols (#$Patrolling) over southern Iraq ((#$SouthernRegionFn #$Iraq)) has (#$possesses) seen (#$VisualPerception #$MeetingSomeone #$sees) the addition (#$DoingAddition) of two (2) dozen (12) F-15 (#$FighterPlane-F15) and F-16 fighter jets (#$FighterPlane-F16) to Bahrain (#$Bahrain-TheIsland #$Bahrain (#$CityNamedFn Bahrain #$Bahrain)). The Air Force (#$UnitedStatesAirForce) has (#$possesses) also authorized (#$GrantingPermission) the dispatch (#$SendingSomething) of 12 (12) F-117 (#$FighterPlane-F117) stealth (#$DodgeStealthCar) fighter jets (#$JetOfFluid #$JetPropelledAircraft) to Kuwait (#$CityOfKuwaitKuwait (#$ProperSubcollectionNamedFn-Ternary kuwait #$Individual 34057665-f4ed-11d9-9bea-0002b3a85b0b) #$Kuwait), three (3) B-1 bombers (#$B-1-Bomber) to Bahrain (#$Bahrain-TheIsland #$Bahrain (#$CityNamedFn Bahrain #$Bahrain)) and 14 (14) B-52 (#$B-52-Bomber) bombers (#$SubmarineSandwich #$BomberPlane #$Bomber) to the island (#$Island) of Diego Garcia. It also has (#$possesses) diverted (#$AmusingSomeone #$DivertingSomething) dozens (#$Dozens-Quant 12) of support (#$SupportingSomething #$ShowingSupportForSomeone (#$SubcollectionOfWithRelationFromTypeFn #$PartiallyTangible #$supportingObject #$SupportingSomething)) aircraft (#$AirTransportationDevice) to the region (#$TheRegion) for refueling (#$Refueling (#$MakingAvailableFn #$CombustibleFuelSubstance)),

Extraction Pipeline



Example

Volkswagen may overtake Toyota as No.1 in Q1

Fri Apr 17, 2009 12:59am EDT

Email | Print | Share | Reprints | Single Page [-] Text [+]

1 of 1 Full Size

RELATED NEWS

- Nippon Steel, Toyota agree steel price cut: source
  12:59am EDT

By Chang-Ran Kim and Christiaan Hetzner

TOKYO/FRANKFURT (Reuters) - Volkswagen AG may have passed Toyota Motor Corp as the world's top selling automaker in the first quarter, helped by robust demand in its main markets, while its Japanese rival suffered sharp declines, partial company data suggests.

The German automaker, with its nine car and truck brands including Audi, Skoda, Seat and Skania, has set a goal of overtaking Toyota and General Motors Corp to be the world's No.1 seller by 2018 -- a target that was initially met with skepticism.

But a deepening recession and credit crisis have crippled demand in Toyota's top markets, with U.S. sales falling 38 percent and Japan sliding 24 percent in January-March.

Volkswagen, meanwhile, is benefiting from government stimulus plans that have boosted sales in China, Germany and Brazil, which together accounted for 44 percent of group sales last year, making it more likely that it beat Toyota or at least came close.

In the first quarter of last year, the German group delivered 1.57 million vehicles, a third less than Toyota's 2.41 million, which included sales at minivehicle and truck units Daihatsu Motor Co and Hino Motors Ltd.

Toyota has given no forecast for retail sales, but its latest estimate for shipments for the 2009 first quarter is 1.23 million vehicles, down 47 percent from a year earlier.

Its first-quarter U.S. sales fell 36 percent, while sales in Japan for the core Toyota brand plummeted 31 percent. The two markets account for just under half of Toyota's global sales.

Volkswagen AG may have passed Toyota Motor Corp as the world's top selling automaker in the first quarter, helped by robust demand in its main markets, while its Japanese rival suffered sharp declines, partial company data suggests.

The German automaker, with its nine car and truck brands including Audi, Skoda, Seat and Skania, has set a goal of overtaking Toyota and General Motors Corp to be the world's No.1 seller by 2018 -- a target that was initially met with skepticism.

...

http://www.reuters.com/article/GCA-autos/idUSTRE53G0L420090417

Name Entities

Volkswagen AG may have passed **Toyota Motor Corp** as the world's top selling automaker in the first quarter, helped by robust demand in its main markets, while its Japanese rival suffered sharp declines, partial company data suggests.

The German automaker, with its nine car and truck brands including Audi, **Skoda**, **Seat** and **Skania**, has set a goal of overtaking **Toyota** and **General Motors** Corp to be the world's No. seller by 2018 -- a target that was initially met with skepticism.

…

Anaphora resolution

Volkswagen AG may have passed Toyota Motor Corp as the world's top selling automaker in the first quarter, helped by robust demand in ***its*** main markets, while ***its*** Japanese rival suffered sharp declines, partial company data suggests.

…

Co-reference resolution

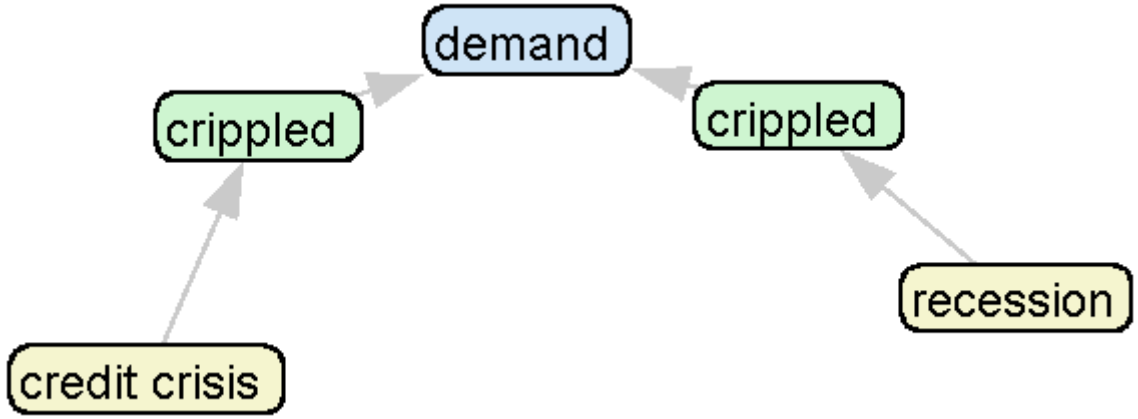
Volkswagen AG may have passed **Toyota Motor Corp** as the world's top selling automaker in the first quarter, helped by robust demand in its main markets, while its Japanese rival suffered sharp declines, partial company data suggests.

The German automaker, with its nine car and truck brands including Audi, Skoda, Seat and Skania, has set a goal of overtaking **Toyota** and General Motors Corp to be the world's No. seller by 2018 -- a target that was initially met with skepticism.

…

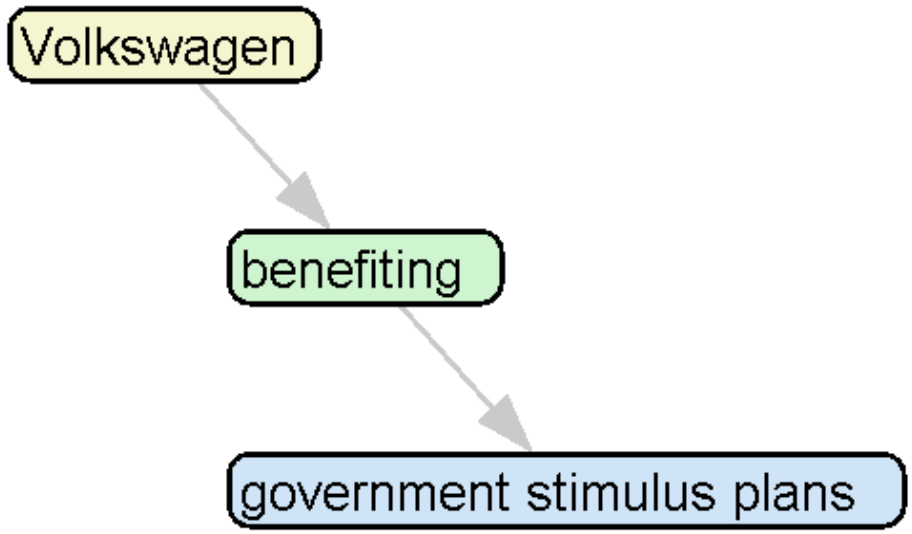
Extract triples

But a deepening **recession** and **credit crisis** have **crippled** **demand.**

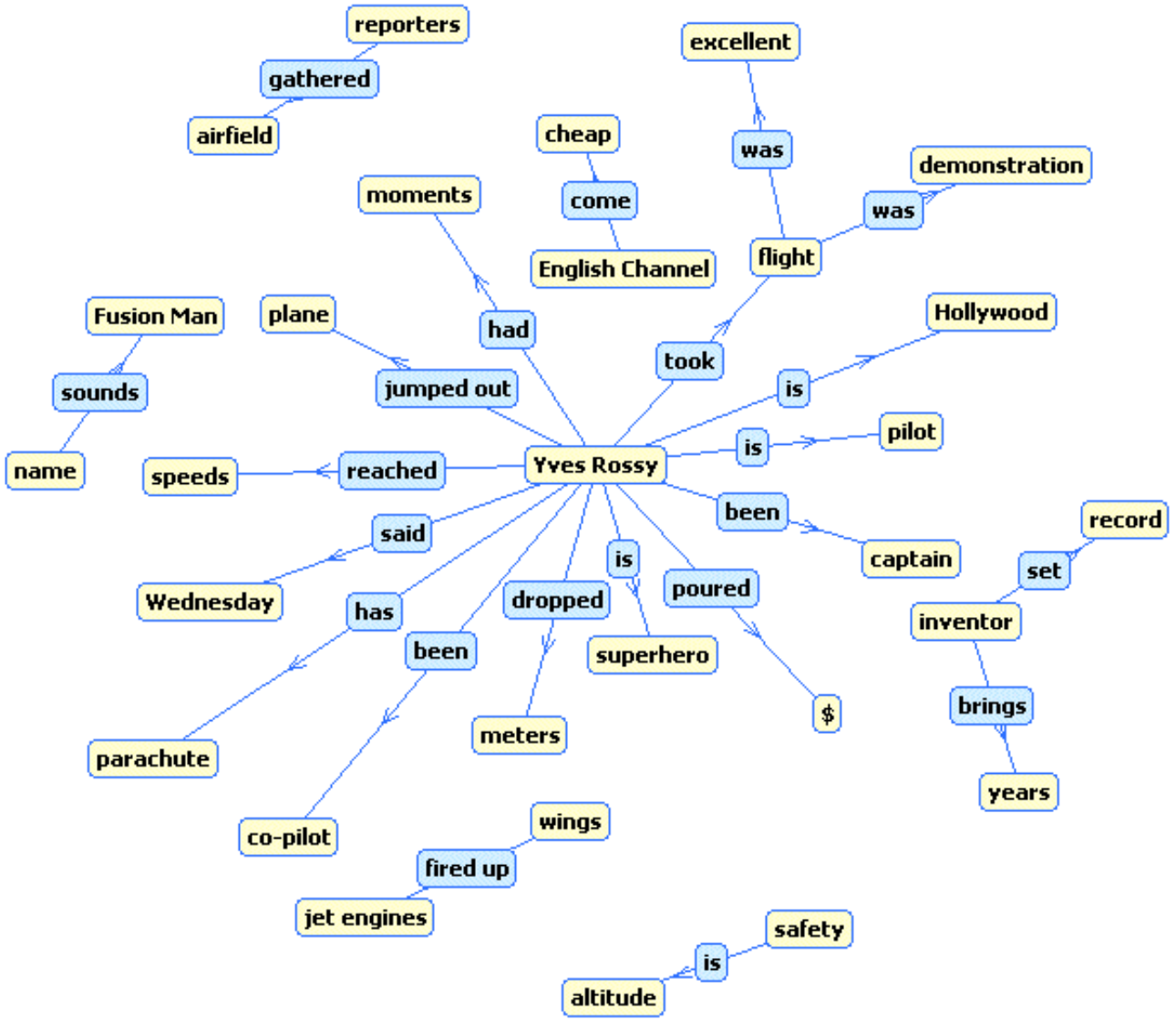


Extract triples

Volkswagen, meanwhile, is
from **government stimulus plans**.



Semantic graph



Do

Bavaria de

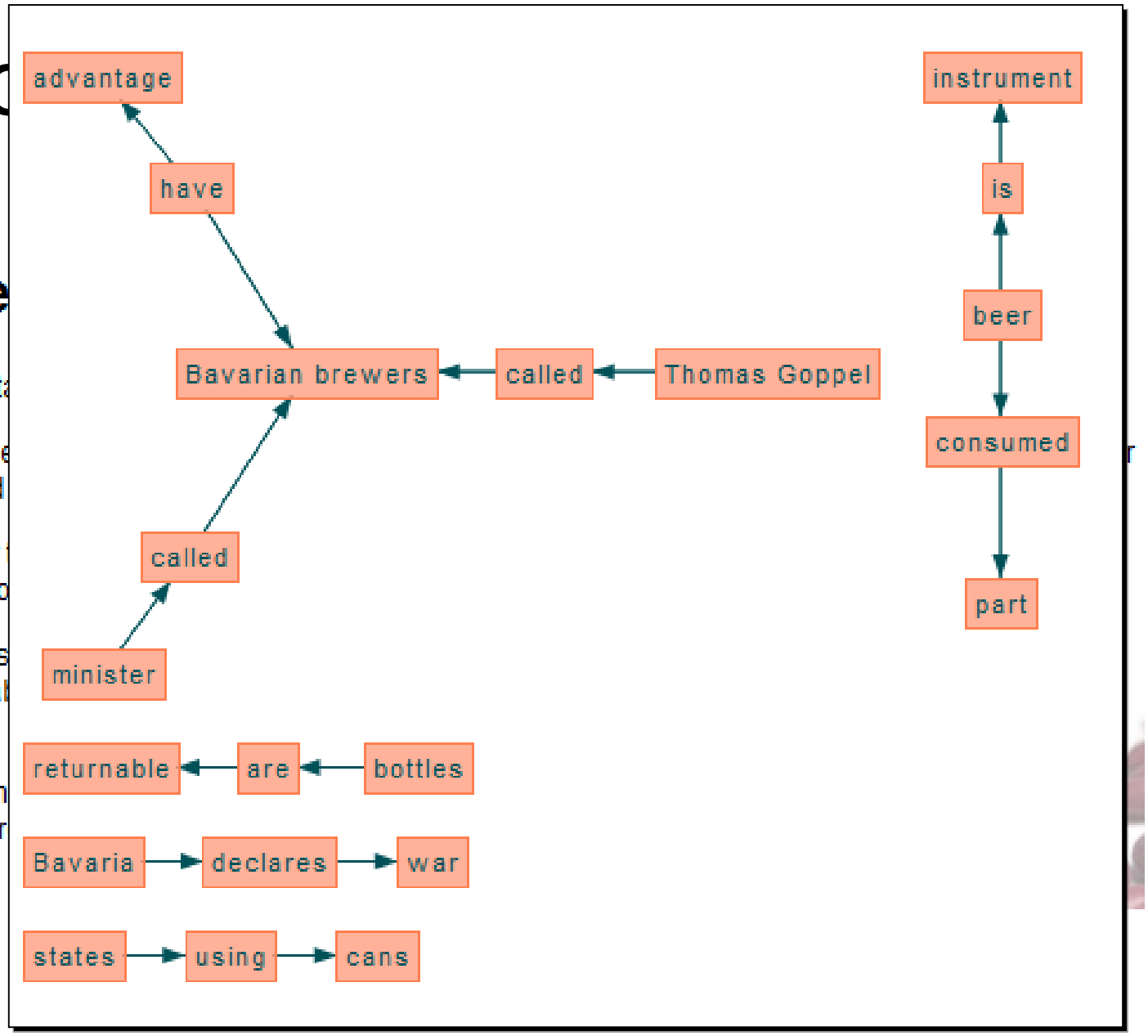
Germany's beer capita

The government in the
part of Germany, said

Bavaria's minister for
cans for their beer pro

Because most bottles
German states and a
market share.

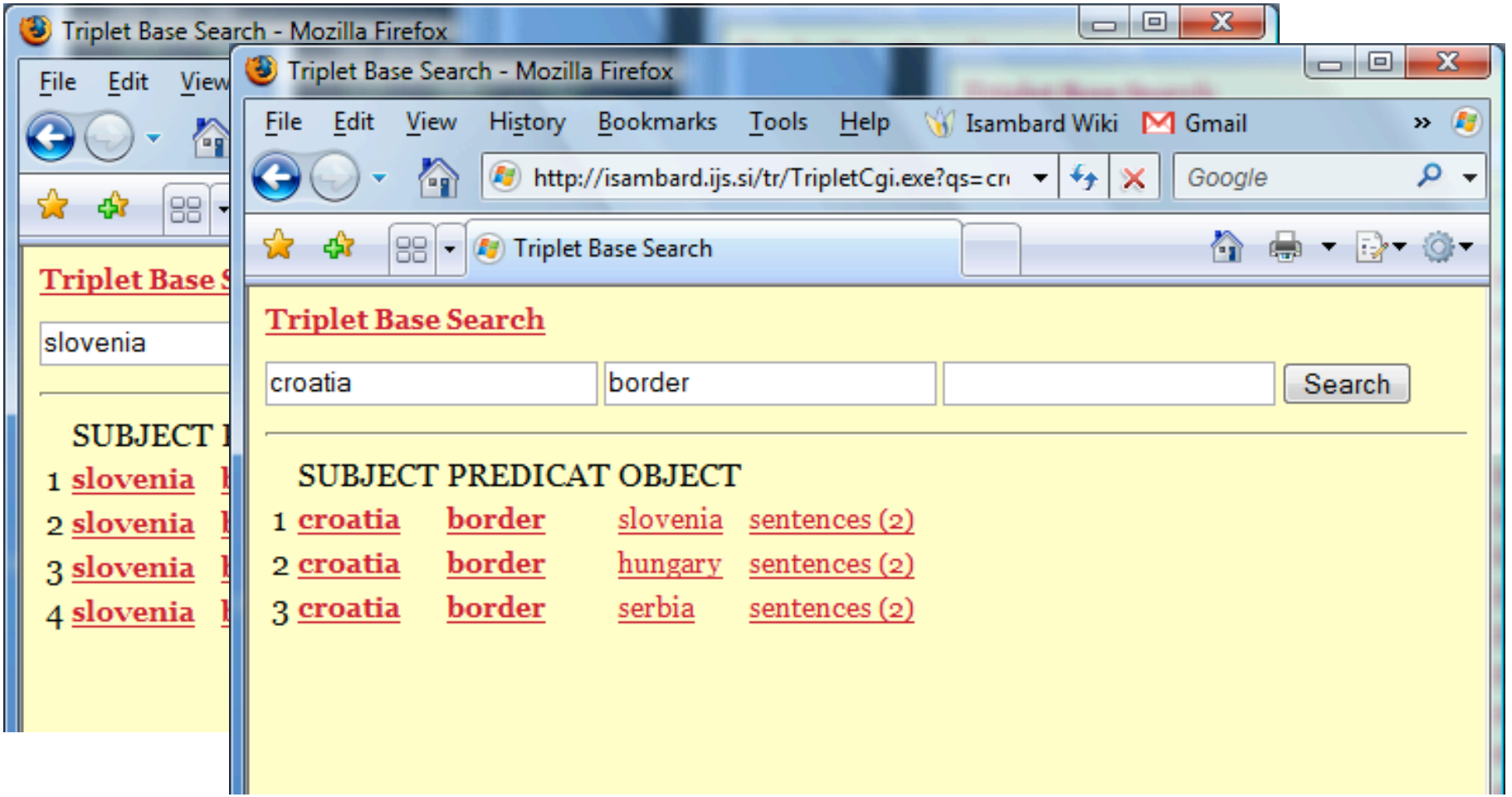
"Canned beer is an in
said. "For ecological r
from canned beer."



http://answerart.net/docview/?id=19961114\189541newsML.xml

Search over triplets

- Pipeline ran over Reuters corpus
  - 800k news articles from 1996 to 1997



http://isambard.ijs.si/triplet/search/

Question Answering

answerArt

where do

We foun

tigers

Siberian t

tigers

Chinese me
Indonesia w
conservatior

where do birds fly

Ask

We found that

birds	fly	the following
bird	flew	engines, rotor, rotor
birds	flying	kilometre, kilometre
Elmo doll	flew	store shelves
U.S. shuttle	flies	satellite
northeast shuttle		Washington
shuttle	flown	spacecraft
U.S. shuttle	fly	satellite

Related documents

engines **USA: American Air jet turned back by bird in engine.** An American Airlines flight was forced to turn around shortly after take-off on Friday when a **bird flew** into one of the **engines**, officials said.

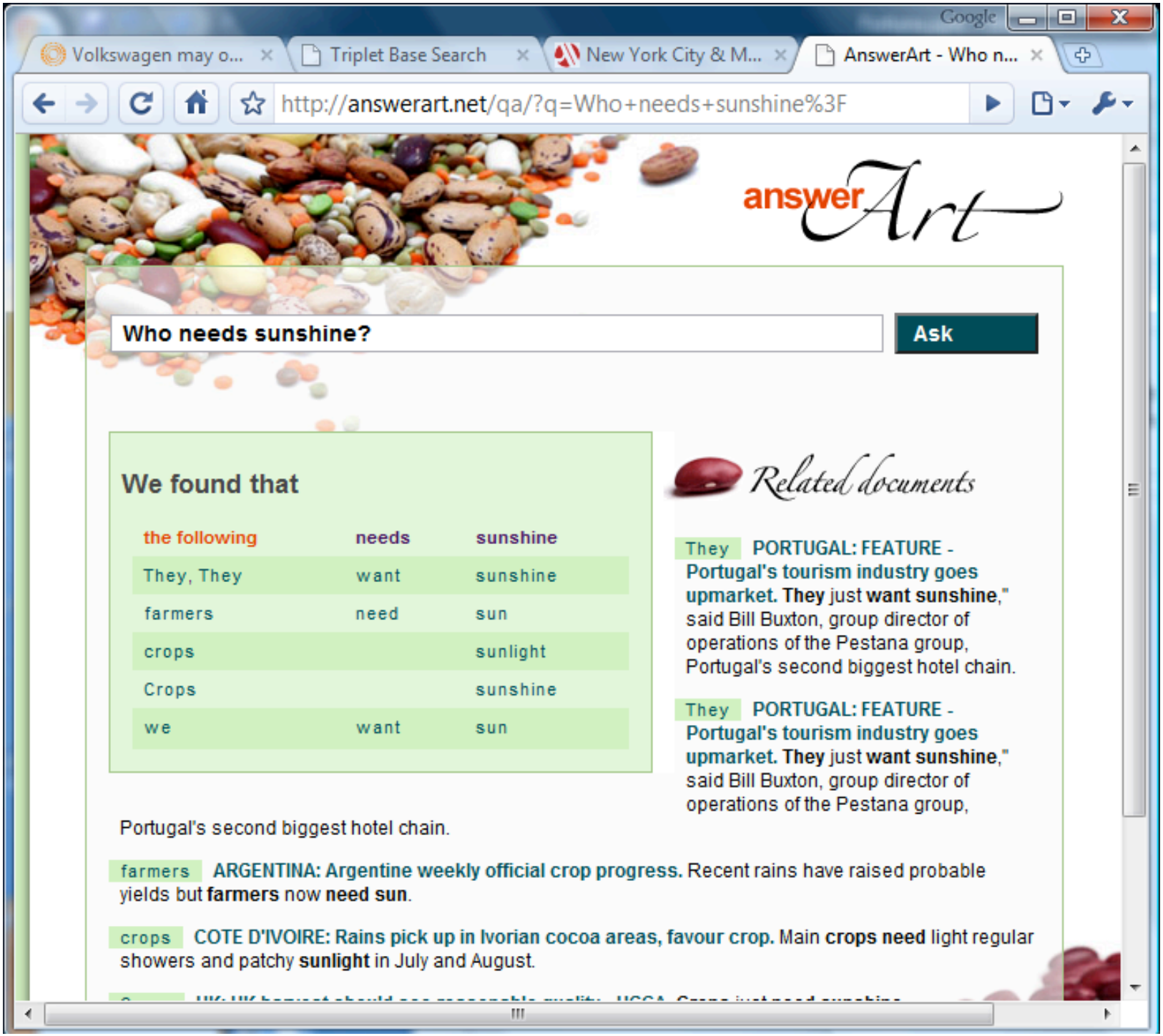
rotor **UK: Five hurt in Guinness helicopter crash-landing.** "It would appear that a **bird flew** into the rear **rotor** of the helicopter, and the helicopter came down."

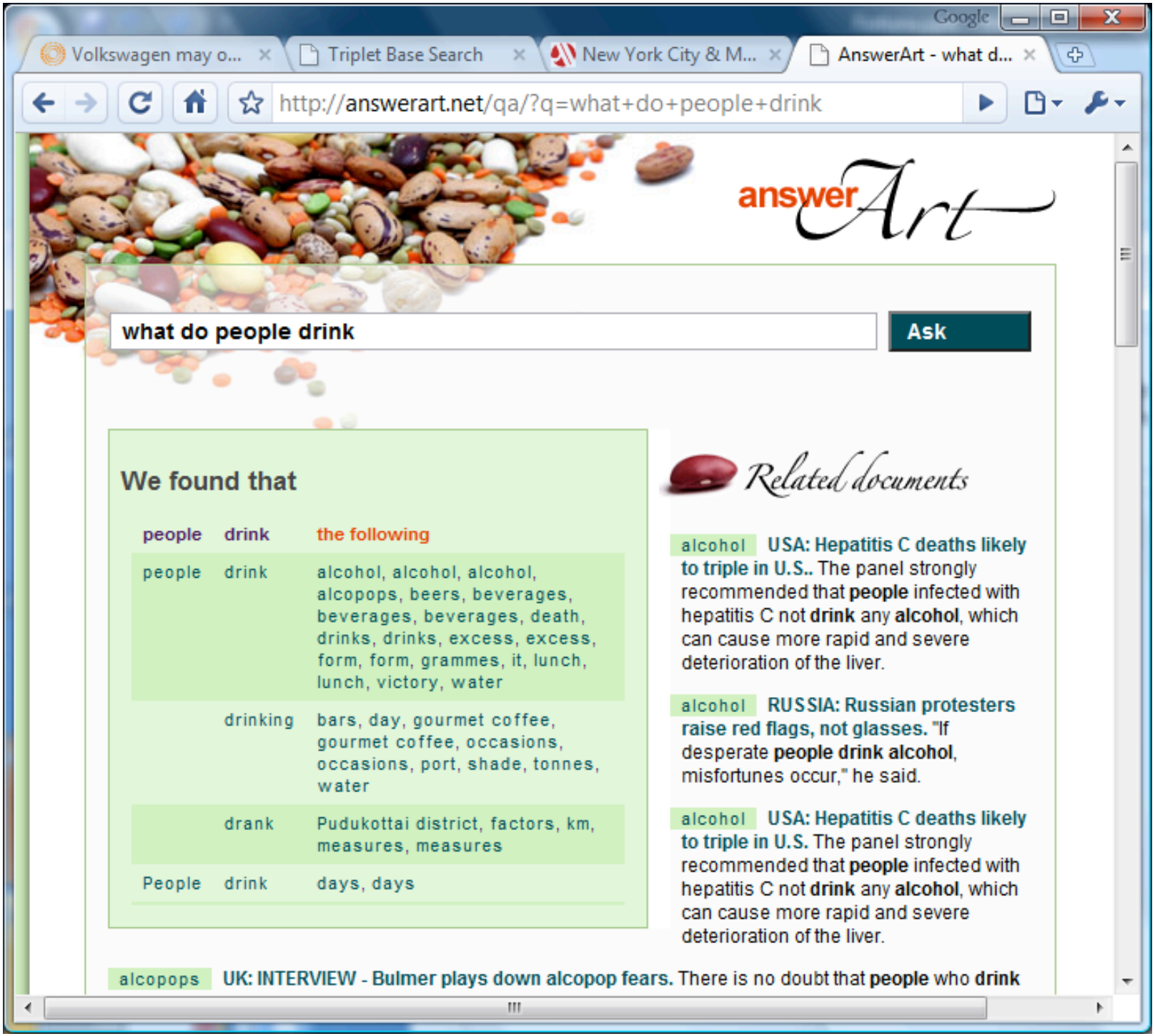
rotor **UK: Five hurt in Guinness helicopter crash landing.** "It would appear that a **bird flew** into the rear **rotor** of the helicopter, and the helicopter came down."

kilometre **PERU: Explosion heard near Peru siege site.** "We heard a very loud boom, and as I looked back, I saw **birds flying** up and away about one **kilometre** (half-mile) away," said Reuters photographer Gregg Newton at the residence.

kilometre **PERU: Peru rebels say no talks without prisoners' issue.** "We heard a very loud boom, and as I looked back, I saw **birds flying** up and away about one **kilometre** (half-mile) away," said Reuters photographer Gregg Newton at the residence.

http://answerart.net/


Google
Volkswagen may o...
Triplet Base Search
New York City & M...
AnswerArt - Who n...
http://answerart.net/qa/?q=Who+needs+sunshine%3F
answerArt
Who needs sunshine?
Ask
We found that
the following needs sunshine
They, They want sunshine
farmers need sun
crops sunlight
Crops sunshine
we want sun
Related documents
They PORTUGAL: FEATURE - Portugal's tourism industry goes upmarket. They just want sunshine," said Bill Buxton, group director of operations of the Pestana group, Portugal's second biggest hotel chain.
They PORTUGAL: FEATURE - Portugal's tourism industry goes upmarket. They just want sunshine," said Bill Buxton, group director of operations of the Pestana group,
Portugal's second biggest hotel chain.
farmers ARGENTINA: Argentine weekly official crop progress. Recent rains have raised probable yields but farmers now need sun.
crops COTE D'IVOIRE: Rains pick up in Ivorian cocoa areas, favour crop. Main crops need light regular showers and patchy sunlight in July and August.


Google
Volkswagen may o...
Triplet Base Search
New York City & M...
AnswerArt - what d...
http://answerart.net/qa/?q=what+do+people+drink
answerArt
what do people drink
Ask
We found that
people drink the following
people drink alcohol, alcohol, alcohol, alcopops, beers, beverages, beverages, beverages, death, drinks, drinks, excess, excess, form, form, grammes, it, lunch, lunch, victory, water
drinking bars, day, gourmet coffee, gourmet coffee, occasions, occasions, port, shade, tonnes, water
drank Pudukottai district, factors, km, measures, measures
People drink days, days
Related documents
alcohol USA: Hepatitis C deaths likely to triple in U.S.. The panel strongly recommended that people infected with hepatitis C not drink any alcohol, which can cause more rapid and severe deterioration of the liver.
alcohol RUSSIA: Russian protesters raise red flags, not glasses. "If desperate people drink alcohol, misfortunes occur," he said.
alcohol USA: Hepatitis C deaths likely to triple in U.S. The panel strongly recommended that people infected with hepatitis C not drink any alcohol, which can cause more rapid and severe deterioration of the liver.
alcopops UK: INTERVIEW - Bulmer plays down alcopop fears. There is no doubt that people who drink

Volkswagen may overtake Toyota as No.1 in Q1

Volkswagen AG may have passed Toyota Motor Corp as the world's top selling automaker in the first quarter, helped by robust demand in its main markets, while its Japanese rival suffered sharp declines, partial company data suggests.

The German automaker, with its nine car and truck brands including Audi, Skoda, Seat and Skania, has set a goal of overtaking Toyota and General Motors Corp to be the world's No.1 seller by 2018 -- a target that was initially met with skepticism.

But a deepening recession and credit crisishave crippled demand in Toyota's top markets, with U.S. sales falling 38 percent and Japan sliding 24 percent in January-March.

Volkswagen, meanwhile, is benefiting from government stimulus plans that have boosted sales in China, Germany and Brazil, which together accounted for 44 percent of group sales last year, making it more likely that it beat Toyota or at least came close.

In the first quarter of last year, the German group delivered 1.57 million vehicles, a third less than Toyota's 2.41 million, which included sales at minivehicle and truck units Daihatsu Motor Co and Hino Motors Ltd.

Toyota has given no forecast for retail sales, but its latest estimate for shipments for the 2009 first quarter is 1.23 million vehicles, down 47 percent from a year earlier.

Its first-quarter U.S. sales fell 36 percent, while sales in Japan for the core Toyota brand plummeted 31 percent. The two markets account for just under half of Toyota's global sales.

Volkswagen had projected a 10 percent decline in its global sales for 2009 back in January, but the sharp reversal in trends in Germany and China could alter that outcome.

"Volkswagen is a big competitor for Toyota," said Koji Endo, auto analyst at Credit Suisse in Tokyo. "Audi is strong, Volkswagen is strong, and they're making good use of their small cars."

The automakers are expected to disclose their worldwide first-quarter vehicle sales over the next week.

The ranking could easily change in subsequent quarters.

Toyota is counting on a third-generation Prius hybrid car due for roll-out next month to jump-start sales as more countries offer consumers incentives to buy energy-efficient cars. It will launch 16 new models in Europe this year following a product drought in 2008.

Volkswagen, for its part, will have a full year of contribution from the remodeled Golf, a perennial best-seller, and the relaunch of its popular Polo compact car.

Volkswagen has also moved up in stock value ranking, grabbing the No.2 spot behind Toyota, whose market capitalization of $133 billion still dwarfs the German carmaker's $100 billion.

Market research company R.L. Polk Germany predicted this month that Volkswagen would overtake GM as the world's second-largest automaker as

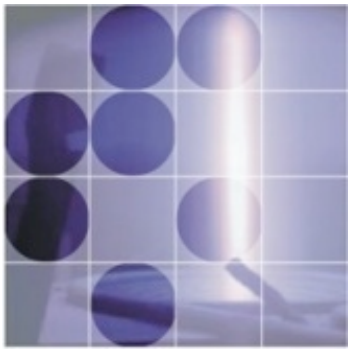
Summarization

Volkswagen, meanwhile, is benefiting from government stimulus plans that have boosted sales in China, Germany and Brazil, which together accounted for 44 percent of group sales last year, making it more likely that it beat Toyota or at least came close.

In the first quarter of last year, the German group delivered 1.57 million vehicles, a third less than Toyota's 2.41 million, which included sales at minivehicle and truck units Daihatsu Motor Co and Hino Motors Ltd.

Toyota has given no forecast for retail sales, but its latest estimate for shipments for the 2009 first quarter is 1.23 million vehicles, down 47 percent from a year earlier.

demo



Marko Grobelnik , Dunja Mladeni_
Jo_ef Stefan Institute, Slovenia (http://www.ijs.si/)

ANALYSIS OF COLLABORATIONS

Analysis of collaborations

Two sources of the data:

- Table of IST projects from internal EC database with fields:
  - Project Ref., Acronym, Key Action, Unit, Officer
  - Org. Name, Country, Org Type, Role in project
- List of IST project descriptions as 1-2 page text summaries from the Web (Cordis at http://dbs.cordis.lu/fep/FP5/FP5_PROJI_search.html)

IST 5FP has 2786 projects in which participate 7886 organizations

© Dunja Mladenic

SolEuNet data mining & decision support

Example of data for Sol-Eu-Net (1)

Table of all IST projects – for each project list of partners

	A	B	C	D	E	F	G	H	I	J
1	Project Ref	Acronym	Domain / k	Unit	PO	Legal Name	Legal Country	Type of organisation	Participant role	
2	IST-1999-11495	SOL-EU-NET	KA2	C2	HANSEN RALF	ALARIX, D.O.O.	SLOVENIA	Private non research org.	CR	
3	IST-1999-11495	SOL-EU-NET	KA2	C2	HANSEN RALF	AUSTRIAN RESEARCH INS	AUSTRIA	Research centres	CR	
4	IST-1999-11495	SOL-EU-NET	KA2	C2	HANSEN RALF	CZECH TECHNICAL UNIVEF	CZECH REPUE	Higher education	CR	
5	IST-1999-11495	SOL-EU-NET	KA2	C2	HANSEN RALF	DIALOGIS SOFTWARE & S	GERMANY	Private non research org.	CR	
6	IST-1999-11495	SOL-EU-NET	KA2	C2	HANSEN RALF	FACHHOCHSCHULE BONN	GERMANY	Higher education	CR	
7	IST-1999-11495	SOL-EU-NET	KA2	C2	HANSEN RALF	FRAUNHOFER GESELLSCI	GERMANY	Research centres	CO	
8	IST-1999-11495	SOL-EU-NET	KA2	C2	HANSEN RALF	GMD - FORSCHUNGSZENT	GERMANY	Research centres	CR	
9	IST-1999-11495	SOL-EU-NET	KA2	C2	HANSEN RALF	INSTITUT JOZEF STEFAN	SLOVENIA	Research centres	CR	
10	IST-1999-11495	SOL-EU-NET	KA2	C2	HANSEN RALF	KATHOLIEKE UNIVERSITEI	BELGIUM	Higher education	CR	
11	IST-1999-11495	SOL-EU-NET	KA2	C2	HANSEN RALF	STUDIO PHI D.O.O., COMM	SLOVENIA	Private non research org.	AC	
12	IST-1999-11495	SOL-EU-NET	KA2	C2	HANSEN RALF	TEMIDA D.O.O., COMPANY	SLOVENIA	Private non research org.	CR	
13	IST-1999-11495	SOL-EU-NET	KA2	C2	HANSEN RALF	THE CHANCELLOR, MASTE	UNITED KINGD	Higher education	CR	
14	IST-1999-11495	SOL-EU-NET	KA2	C2	HANSEN RALF	UNIVERSIDADE DO PORTC	PORTUGAL	Private non research org.	CR	
15	IST-1999-11495	SOL-EU-NET	KA2	C2	HANSEN RALF	UNIVERSITY OF BRISTOL	UNITED KINGD	Higher education	CR	

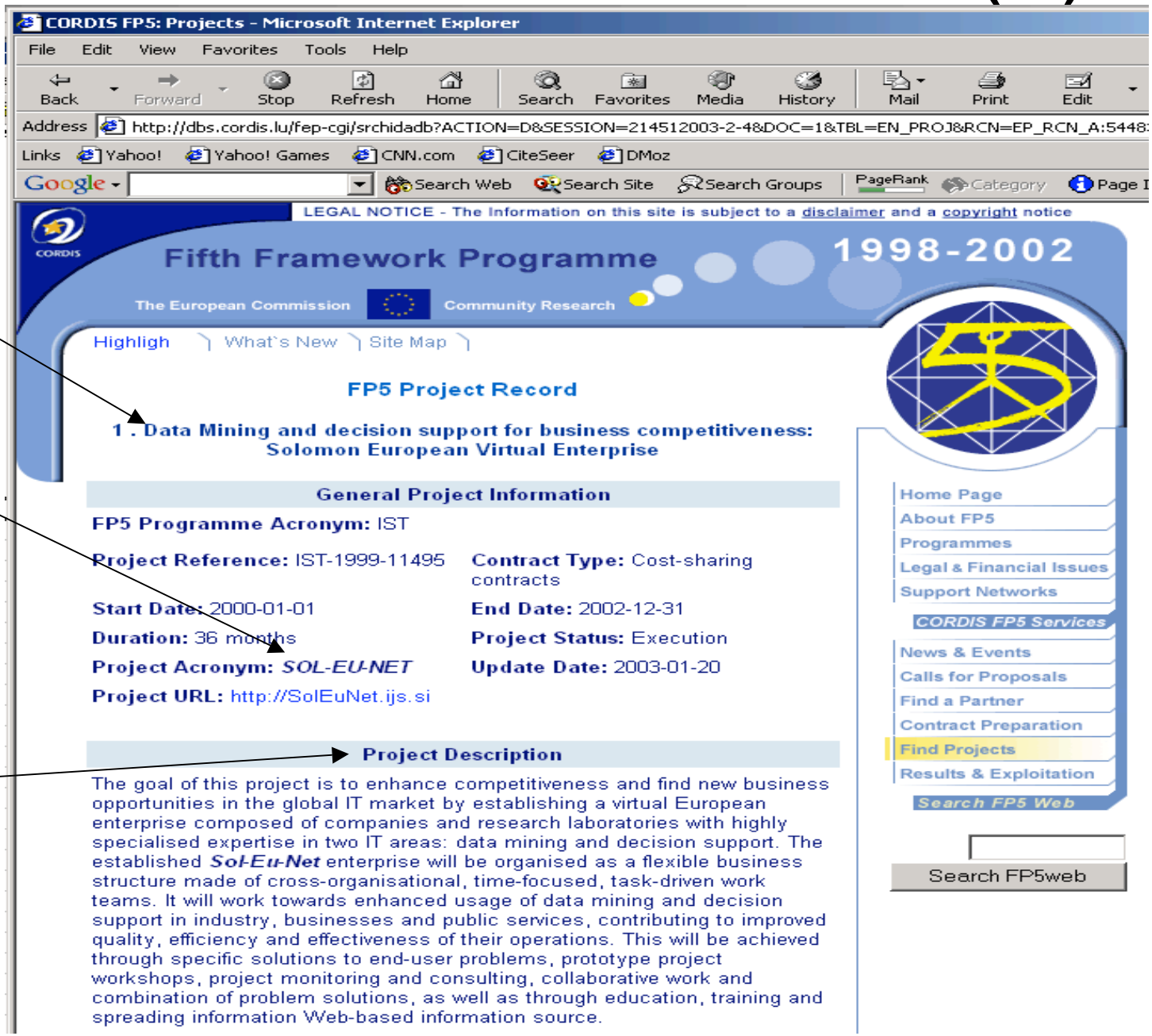
© Dunja Mladenic

Example of data for Sol-Eu-Net (2)

Project Title

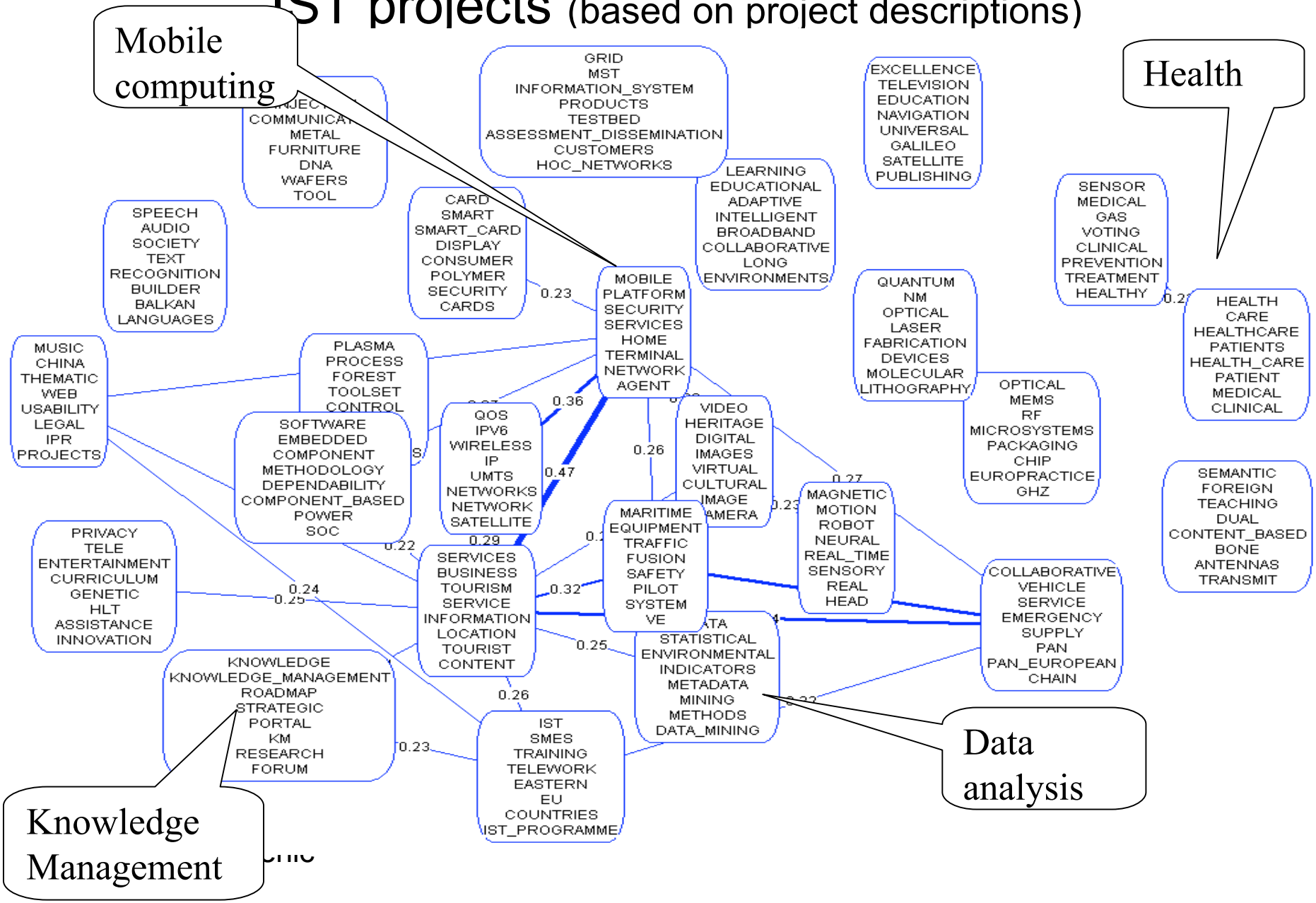
Project Acronym

Project Description



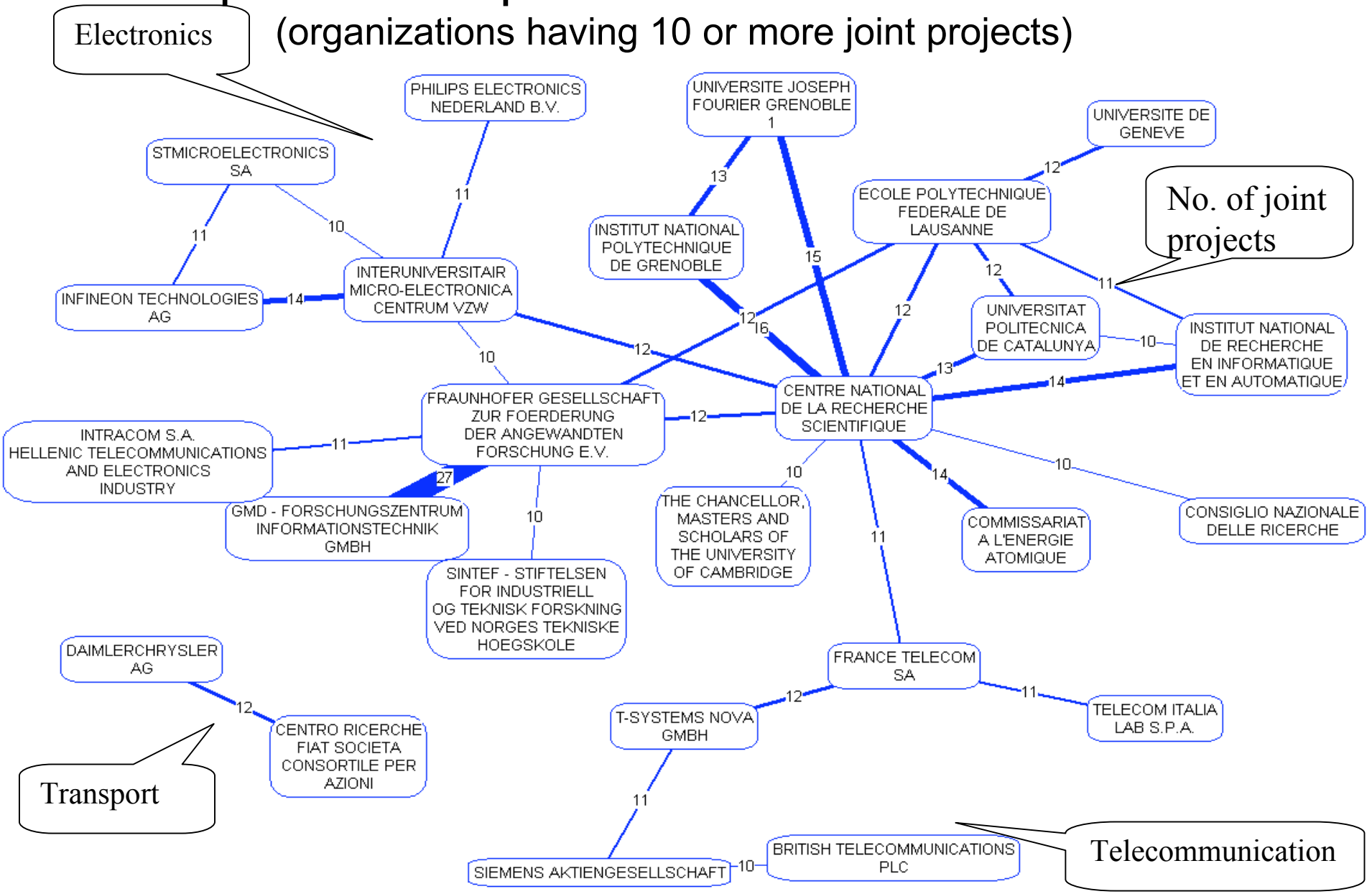
© Dunja Mladenic

Knowledge map: Visualization into 25 groups of 2786 IST projects (based on project descriptions)



Competence map: Institutional Backbone of IST

(organizations having 10 or more joint projects)



© Dunja Mladenic

Community identification
(based on project partnership)

Organizations “more connected” between each other than to the rest of “the world”

Example of a **star-shaped** cooperation (around Fraunhofer):

- 'FRAUNHOFER GESELLSCHAFT ZUR FOERDERUNG DER ANGEWANDTEN FORSCHUNG':0.758
- 'UNIVERSITAET STUTTGART':0.177
- 'THALES BROADCAST MULTIMEDIA':0.155
- 'STAEDTISCHE KLINIKEN OFFENBACH':0.129
- 'AVATARME':0.107
- 'NTEC MEDIA ADVANCED DIGITAL MOTION PICTURE SOLUTIONS':0.089
- 'FOERSAEKRINGSAKTIEBOLAGET SKANDIA PUBL':0.085
- 'EXODUS':0.085
- …

© Dunja Mladenic

Community identification
(based on project partnership)

- Example of a **cycle-shaped** (clique) cooperation (mainly Greece, some Germany and Portugal,...):
  - 'NATIONAL TECHNICAL UNIVERSITY ATHENS':0.548
  - 'INTRACOM HELLENIC TELECOMMUNICATIONS ELECTRONICS INDUSTRY':0.412
  - 'ATHENS UNIVERSITY ECONOMICS BUSINESS':0.351
  - 'NOKIA CORPORATION':0.229
  - 'POULIADIS ASSOCIATES CORP':0.153
  - 'NATIONAL KAPODISTRIAN UNIVERSITY ATHENS':0.139
  - 'LAMBRAKIS RESEARCH FOUNDATION':0.129
  - 'PORTUGAL TELECOM INOVACAO':0.116
  - 'INTRASOFT INTERNATIONAL':0.106
  - 'SEMA GROUP':0.102
  - 'SIEMENS INFORMATION COMMUNICATION NETWORKS':0.097
  - 'UNIVERSITAET ZU KOELN':0.083
  - 'HELLENIC BROADCASTING CORPORATION':0.083
  - 'STADT KOELN':0.081
  - 'HELLENIC TELECOMMUNICATIONS ORGANIZATION':0.081

© Dunja Mladenic

Identifying thematic consortia given a set of keywords

- The task is to list relevant institutions for the given set of keywords
- This can be seen as generating a knowledge map
- The set of institutions can be understood as proposed consortium for a given thematic area

© Dunja Mladenic

Thematic consortia identification

Example of possible Data Mining consortium:

Top 20 institutions for the set of “data-mining” related keywords: “knowledge discovery text mining classification machine learning data mining data analysis

1. (1.537) FRAUNHOFER GESELLSCHAFT ZUR FOERDERUNG DER ANGEWANDTEN FORSCHUNG - [KDNE
2. (1.305) GMD FORSCHUNGSZENTRUM INFORMATIONSTECHNIK - [SPIN!, SOL-EU-NET, XML-KM, ITCOLE]
3. (1.120) UNIVERSITAET DORTMUND - [KDNET, MINING MART, DREAM, INTERMON]
4. (0.939) RESEARCH ACADEMIC COMPUTER TECHNOLOGY INSTITUTE - [NEMIS]
5. (0.817) CZECH TECHNICAL UNIVERSITY PRAGUE - [KDNET, SOL-EU-NET, CLOCKWORK, EUTIST-IMV]
6. (0.727) UNIVERSITA DEGLI STUDI DI BARI - [KDNET, SPIN!, ASSO]
7. (0.725) INSTITUT JOZEF STEFAN - [KDNET, SOL-EU-NET, ELENA]
8. (0.705) UNIVERSITY BRISTOL - [KDNET, SOL-EU-NET, TRUST]
9. (0.696) VYSOKA SKOLA EKONOMICKA PRAZE - [KDNET, MINING MART]
10. (0.696) PEROT SYSTEMS NEDERLAND - [KDNET, MINING MART]
11. (0.678) UNIVERSITY MANCHESTER - [PARMENIDES, E-UTILITIES]
12. (0.668) EUROPEAN COMMISSION JOINT RESEARCH CENTRE - [KDNET, MINEO, EDEN-IW, DISMAR]
13. (0.659) KATHOLIEKE UNIVERSITEIT LEUVEN - [KDNET, SOL-EU-NET]
14. (0.638) QUANTOS - [NEMIS, X-STATIS]
15. (0.620) UNIVERSITAT POLITECNICA DE CATALUNYA - [NEMIS, ESIS, INTERFACE, ALCOM-FT]
16. (0.587) ROYAL HOLLOWAY BEDFORD COLLEGE - [KDNET, KERMIT]
17. (0.567) TEKNILLINEN KORKEAKOULU - [KDNET, E-SHARING, OR-WORLD, NOMAD]
18. (0.557) DIALOGIS SOFTWARE SERVICES - [SPIN!, SOL-EU-NET]
19. (0.552) ATKOSOFT - [X-STATIS, VITAMIN S]
20. (0.543) PIXELPARK - [KDNET, CERENA]
21. (0.530) UNIVERSITEIT VAN AMSTERDAM - [KDNET, ITCOLE, CODEX-IP, COMMORG]
22. (0.524) UNIVERSITA DEGLI STUDI DI ROMA LA SAPIENZA - [NEMIS, ITCOLE]
23. (0.516) ECOLE POLYTECHNIQUE FEDERALE DE LAUSANNE - [NEMIS, INTERFACE]
24. (0.482) UNIVERSITEIT UTRECHT - [KDNET, ITCOLE, ALCOM-FT]
25. (0.470) KUNGLIGA TEKNISKA HOEGSKOLAN - [KDNET, WEBLABS]

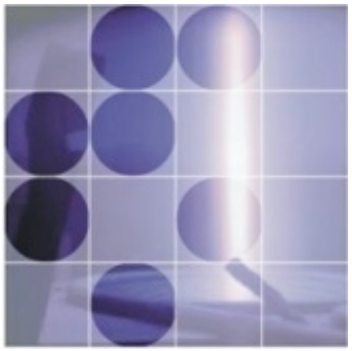
© Dunja Mladenic

Project Intelligence Web site

- All demos, reports and results available at the web at http://pi.ijs.si/

© Dunja Mladenic

demo



Marko Grobelnik (marko.grobelnik@ijs.si)
Dunja Mladeni_ (dunja.mladenic@ijs.si)
Jo_ef Stefan Institute, Slovenia (http://www.ijs.si/)

REAL-TIME INFORMATION PROCESSING

Motivation

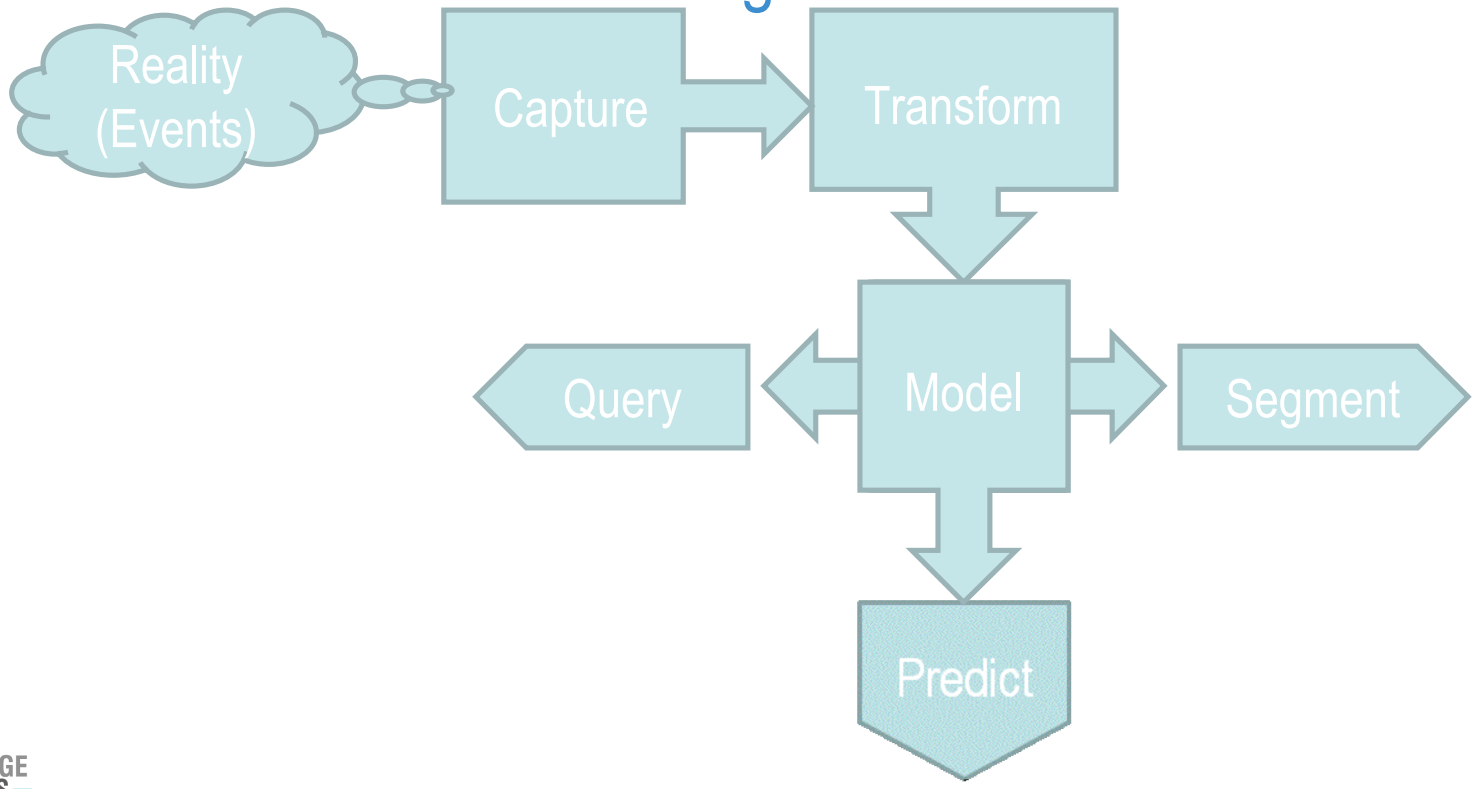
- Why one would need (near) real-time information processing?
  - …because **Time** and **Reaction Speed** correlate with many target quantities – e.g.:
    - …on stock exchange with **Earnings**
    - …in controlling with **Quality of Service**
    - …in fraud detection with **Safety**, etc.
  - Generally, we can say: **Reaction Speed == Value**
    - …if our systems react fast, we create new value!

Introduction – Who?

- Who works with real time data processing?
  - "**Stream Mining**" (subfield of "**Data Mining**") dealing with mining data streams in different scenarios in relation with machine learning and data bases
    - http://en.wikipedia.org/wiki/Data_stream_mining
  - "**Complex Event Processing**" is a research area discovering complex events from simple ones by inference, statistics etc.
    - http://en.wikipedia.org/wiki/Complex_Event_Processing

Introduction – What?

- What is Real-Time information processing?
  - It is defined by a set of approaches enabling operations on the observed incoming stream of data:



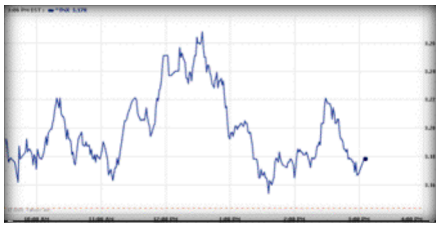
Approaches

- When dealing with streams is really a problem?
  - …when we have an **intensive** data stream and **complex** operations on data are required!
- In such situations usually…
  - …the volume of data is **too big** to be stored
  - …the data can be scanned thoroughly **only once**
  - …the data is highly non-stationary (changes properties through time), therefore approximation and adaptation are key to success
- Therefore, a typical solution is…
  - …**not** to store observed data **explicitly,** but rather in the **aggregate form** which allows execution of required operations

Applications

- All typical applications are “mission critical”
  - …they have intensive streams and complex queries

- Example applications:
  - Dynamic tracking of stock fluctuations
  - Surveillance for frauds and money laundering
  - Network traffic monitoring
  - Sensor network data analysis
  - Web click stream mining
  - Power consumption measurement



- Next slides show some concrete applications…

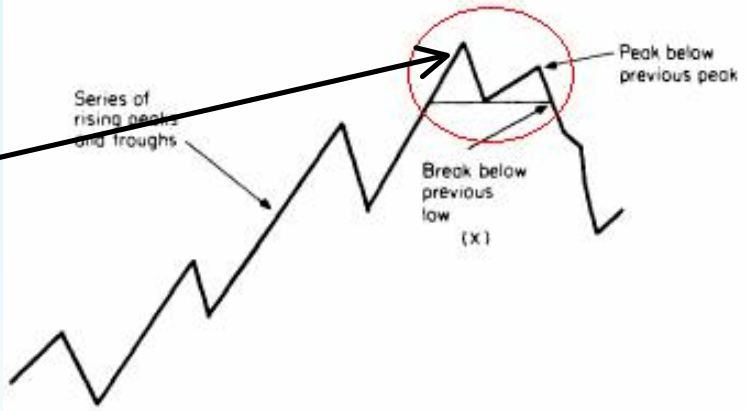
Application: Stock Monitoring

Stock monitoring

- Stream of price and sales volume of stocks over time
- Technical analysis/charting for stock investors
- Support trading decisions

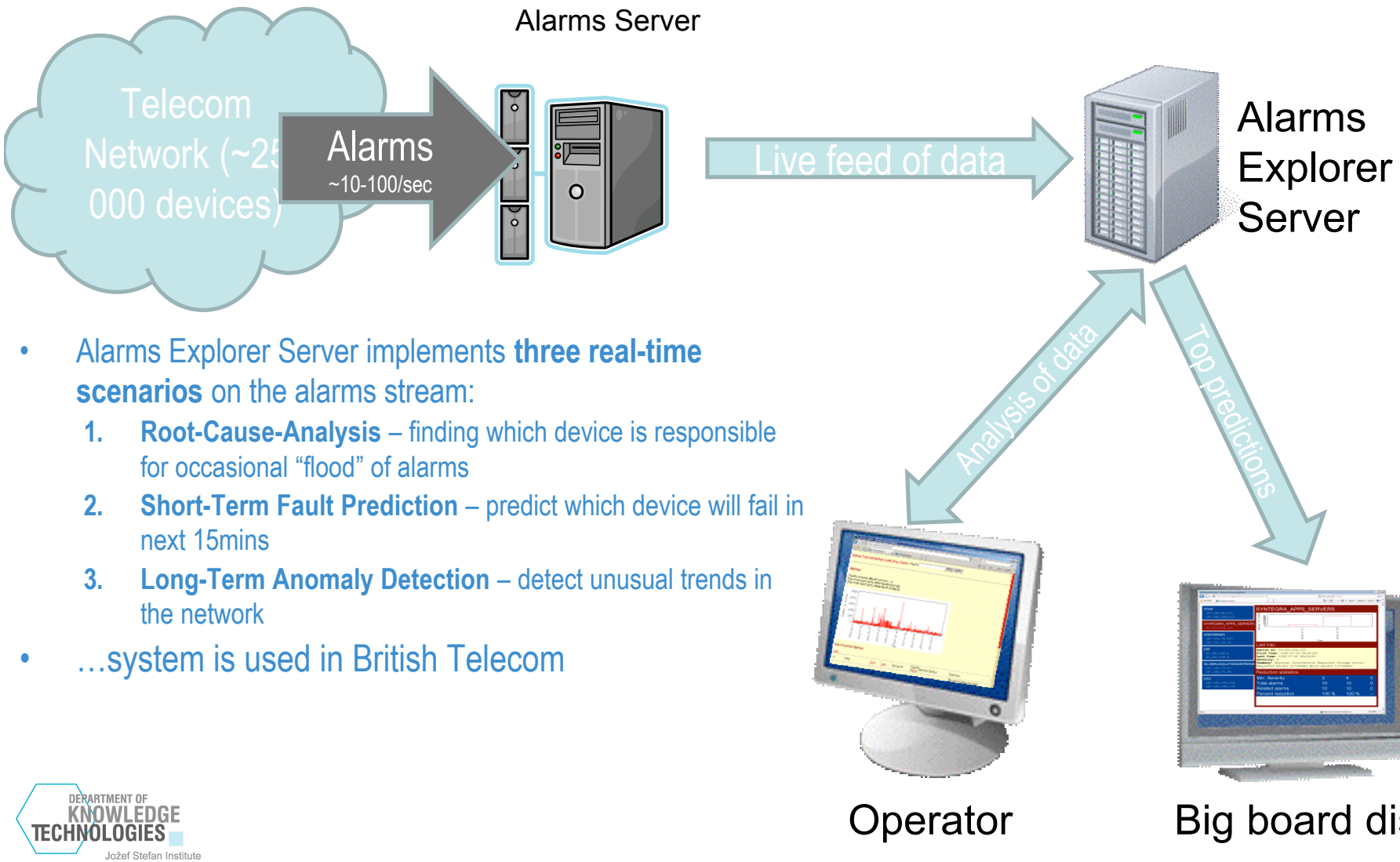
Example Queries (Stream Triggers):

- *Notify me when the price of IBM is above $83, and the first MSFT price afterwards is below $27.*
- *Notify me when some stock goes up by at least 5% from one transaction to the next.*
- *Notify me when the price of any stock increases monotonically for ≥30 min.*
- *Notify me whenever there is **double top formation** in the price chart of any stock*
- *Notify me when the difference between the current price of a stock and its 10 day moving average is greater than some threshold value*



Source: Gehrke 07 and Cayuga application scenarios (Cornell University)

Applications: Telecommunication Network Monitoring



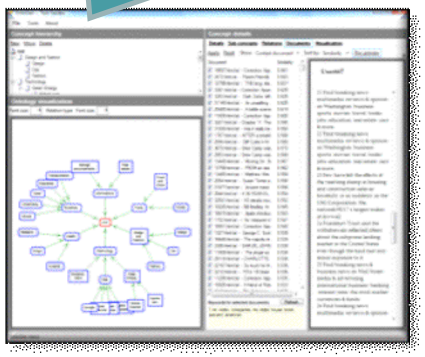
- Alarms Explorer Server implements **three real-time scenarios** on the alarms stream:
  1. **Root-Cause-Analysis** – finding which device is responsible for occasional "flood" of alarms
  2. **Short-Term Fault Prediction** – predict which device will fail in next 15mins
  3. **Long-Term Anomaly Detection** – detect unusual trends in the network
- …system is used in British Telecom

Application: Online Advertising for NYTimes (microtrends detection)

Log Files (~100M page clicks per day)

User profiles

Trend Detection System



NYT articles (~830K news)

Trends and updated segments

Segment	Keywords
Stock Market	Stock Market, mortgage, banking, investors, Wall Street, turmoil, New York Stock Exchange
Health	diabetes, heart disease, disease, heart, illness
Green Energy	Hybrid cars, energy, power, model, carbonated, fuel, bulbs,
Hybrid cars	Hybrid cars, vehicles, model, engines, diesel
Travel	travel, wine, opening, tickets, hotel, sites, cars, search, restaurant
...	...

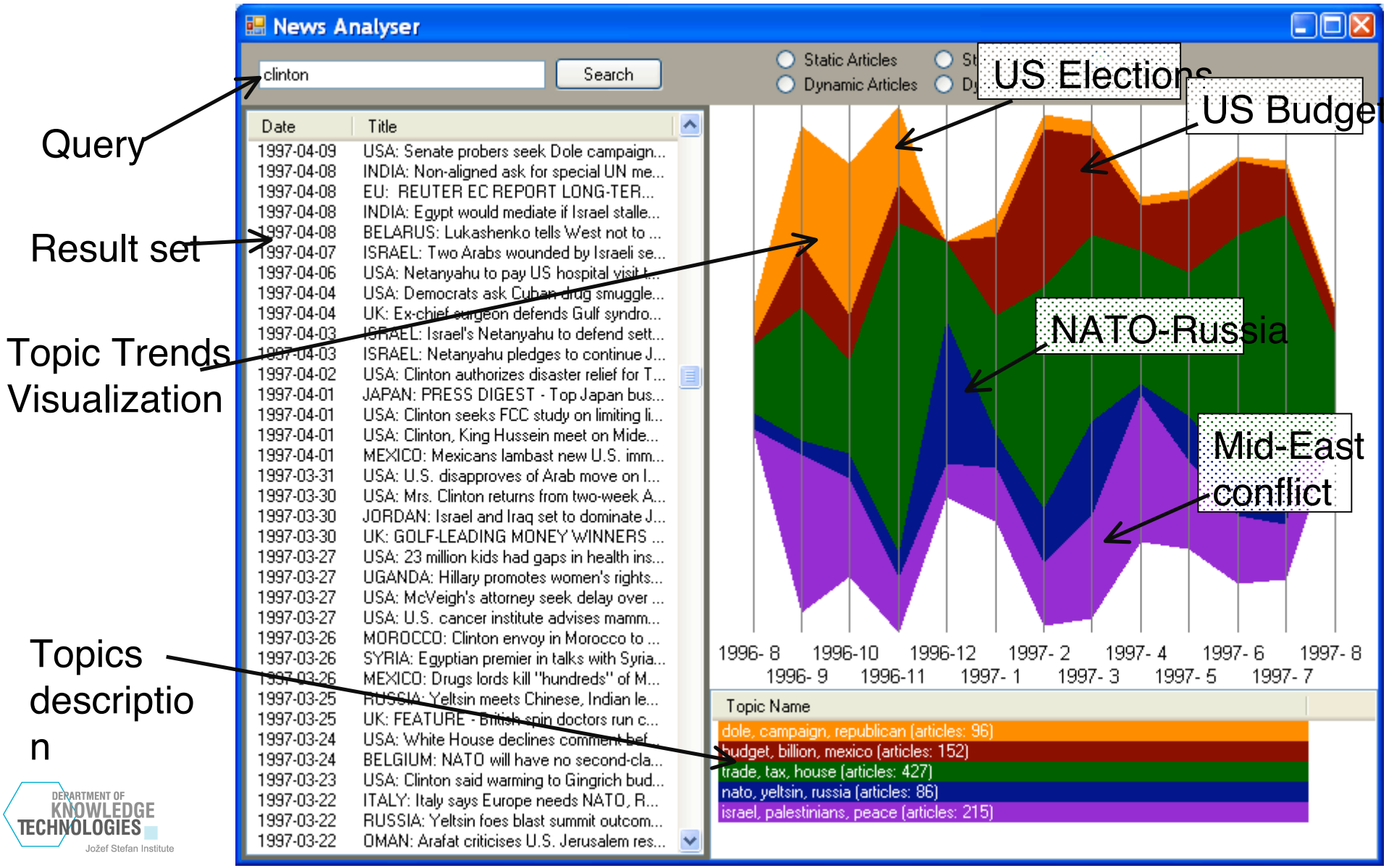
Sales

Campaign to sell segments

$

Advertisers

Application: Topic Tracking



DEPARTMENT OF KNOWLEDGE TECHNOLOGIES
Jožef Stefan Institute

Video Tutorials @ videolectures.net

- **Text Mining and Link Analysis:**
  - Marko Grobelnik, Dunja Mladenic, J. Stefan Institute
    http://videolectures.net/kdd07_grobelnik_tmala/
  - Thomas Hofmann, Brown University

  http://videolectures.net/mlss06au_hofmann_irtm/
- **State of the Art in Data Stream Mining**:
  - Joao Gama, University of Porto

  http://videolectures.net/ecml07_gama_sad/
- **Data stream management and mining:**
  - Georges Hebrail, Ecole Normale Superieure

  http://videolectures.net/mmdss07_hebrail_dsmm/