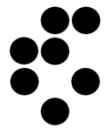
Media Streams

Blaž Fortuna

Artificial Intelligence Laboratory

Jožef Stefan Institute



Outline

- Introduction
- Analyzing document streams
- Web click stream mining

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?

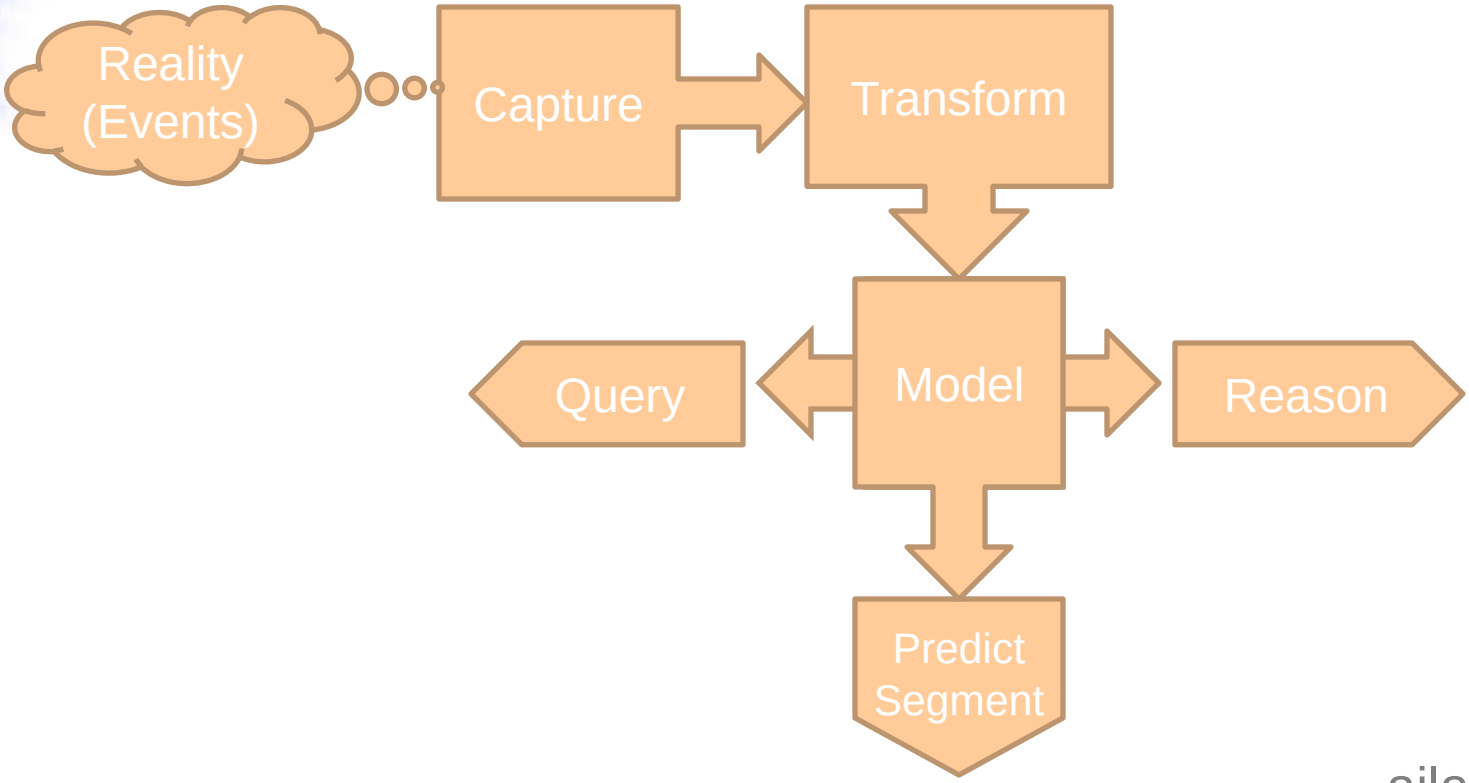
- Who works with real time data processing?
  - "**Stream Mining**" (subfield of "**Data Mining**") dealing with mining data streams in different scenarios in relation with machine learning and data bases
    - http://en.wikipedia.org/wiki/Data_stream_mining
  - "**Complex Event Processing**" is a research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

Approaches

- When dealing with streams is really a problem?
  - …dealing with streams can be often easy, but…
  - …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

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:



ANALYZING DOCUMENT STREAMS

Document Corpus and Relationship Networks

- What is different if we have many documents?
  - First, we have lots of data and linguistic structure doesn’t matter anymore
  - …usually we are interested in entities (nodes) and their relationships (context how they co-appear)

Document representation

- Bag of words:
  - Vocabulary: $\{w_i \mid i = 1, \ldots, N\}$
  - Documents are represented with vectors (word space):

$$D = \{\boldsymbol{x}_i \in \mathbb{R}^N; i = 1, \ldots, \ell\}$$

- Example:

Document set:

d_1 = "*Canonical Correlation Analysis*"
d_2 = "*Numerical Analysis*"
d_3 = *"Numerical Linear Algebra"*

Document vector representation:

x_1 = (1, 1, 1, 0, 0, 0)
x_2 = (0, 0, 1, 1, 0, 0)
x_3 = (0, 0, 0, 1, 1, 1,)

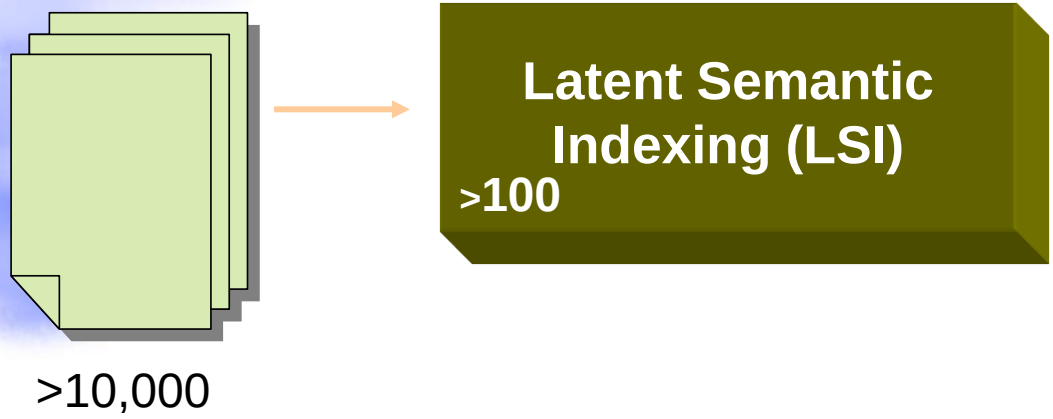
Vocabulary:
{*"Canonical "*, *"Correlation "*, *"Analysis"*, *"Numerical "*, *"Linear "*, *"Algebra"*}

Visualization

- Documents in bag-of-words representation live in a very high dimensional space – usually >10,000 dims!

- For visualisation the number of dimensions must be reduced to just 2!

The Big Picture



Latent Semantics Indexing

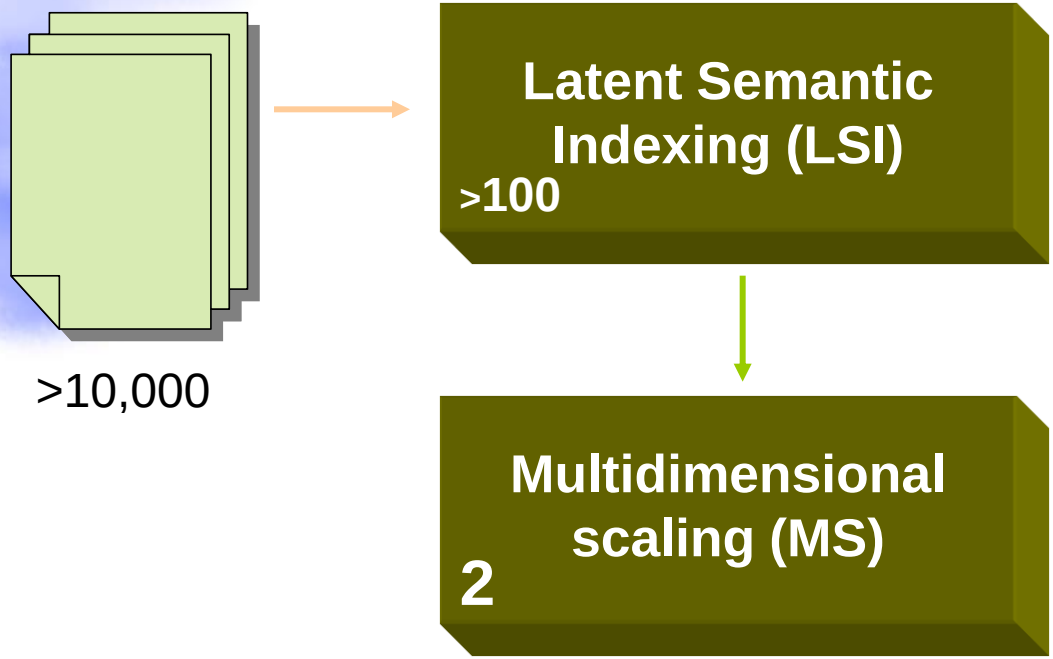
What is LSI?

- A linear technique for finding words with similar meaning based on concurrences in the documents
- Similar words are grouped into latent variables (*concepts),* one word can appear in more concepts
- Documents are described by these concepts instead of words (== much lower dimension).

Background

- Uses Singular Value Decomposition (SVD) to find the best low-dimensional approximation of the documents.
- Latent variables are the basis vectors of this low-dimensional subspace

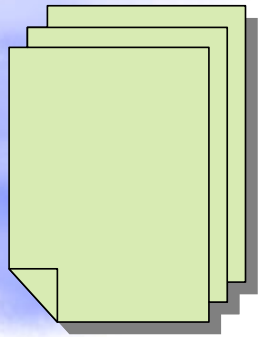
The Big Picture



Multidimensional scaling

- Non-linear technique for dimensionality reduction
- Finds a position of points in lower dimension space so that the Euclidian distances best match original distances
- Iterative gradient descent algorithm
- We use it to position documents into two dimensional plane

The Big Picture



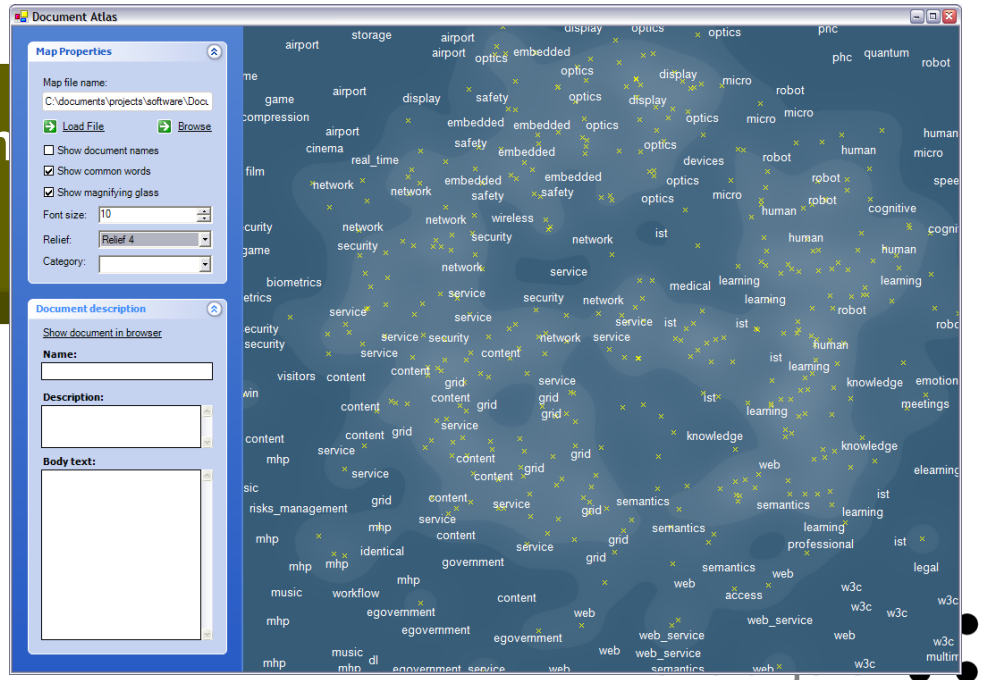
>10,000

Latent Semantic Indexing

>100

Multidimension scaling

2

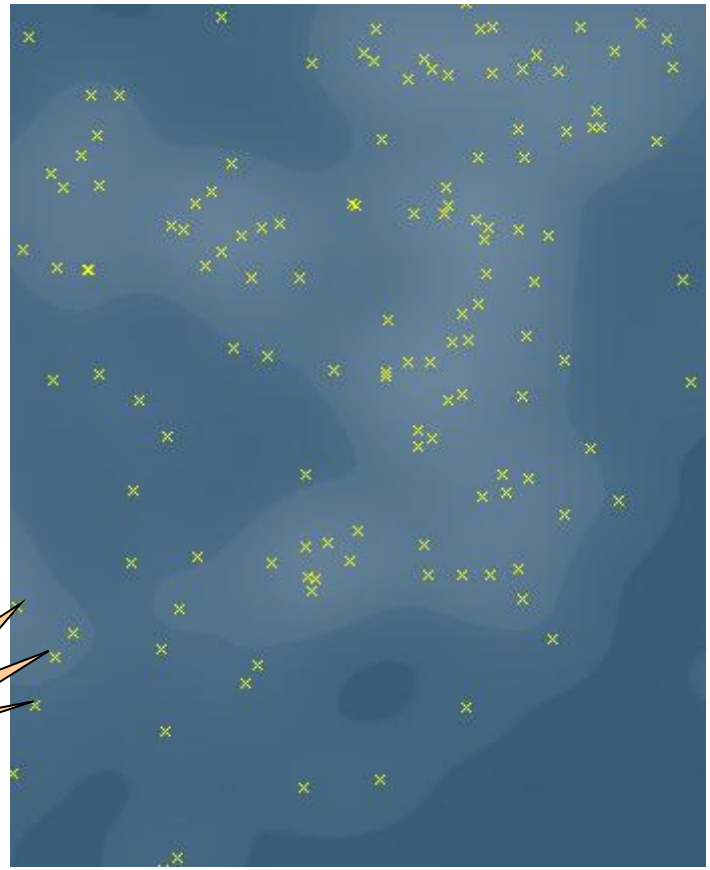


Landscape generation

Density of points is used to generate a landscape.

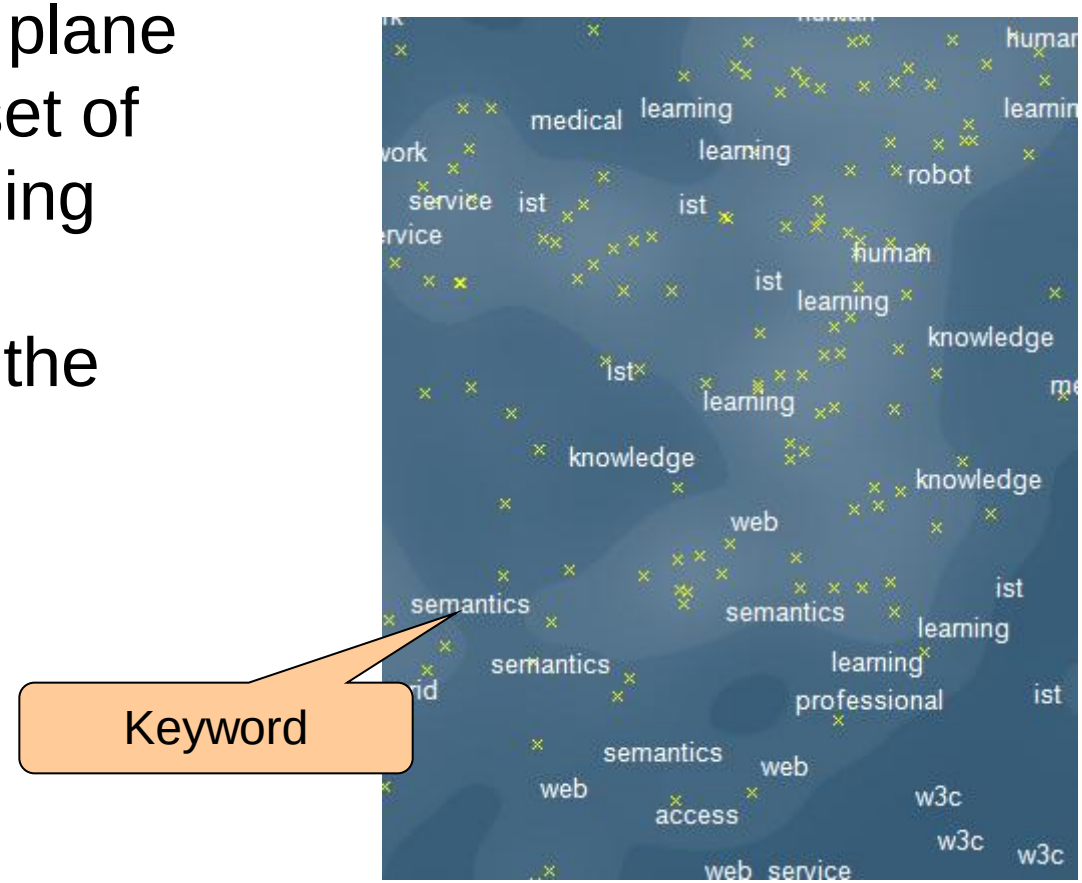
Landscape is used as a background – lighter is higher.

Clusters of high density can be emphasized by drawing contour lines.



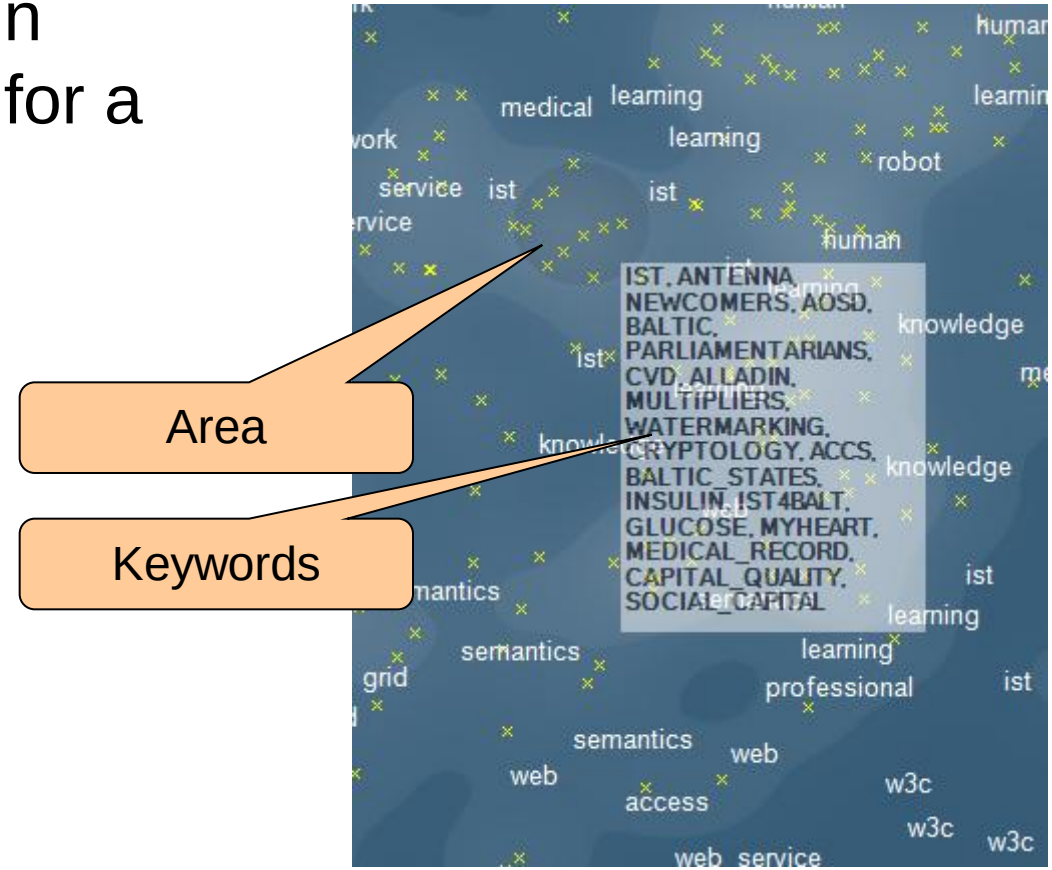
Keywords

Each point from the plane can be assigned a set of keywords by averaging TFIDF vectors of documents close to the point.



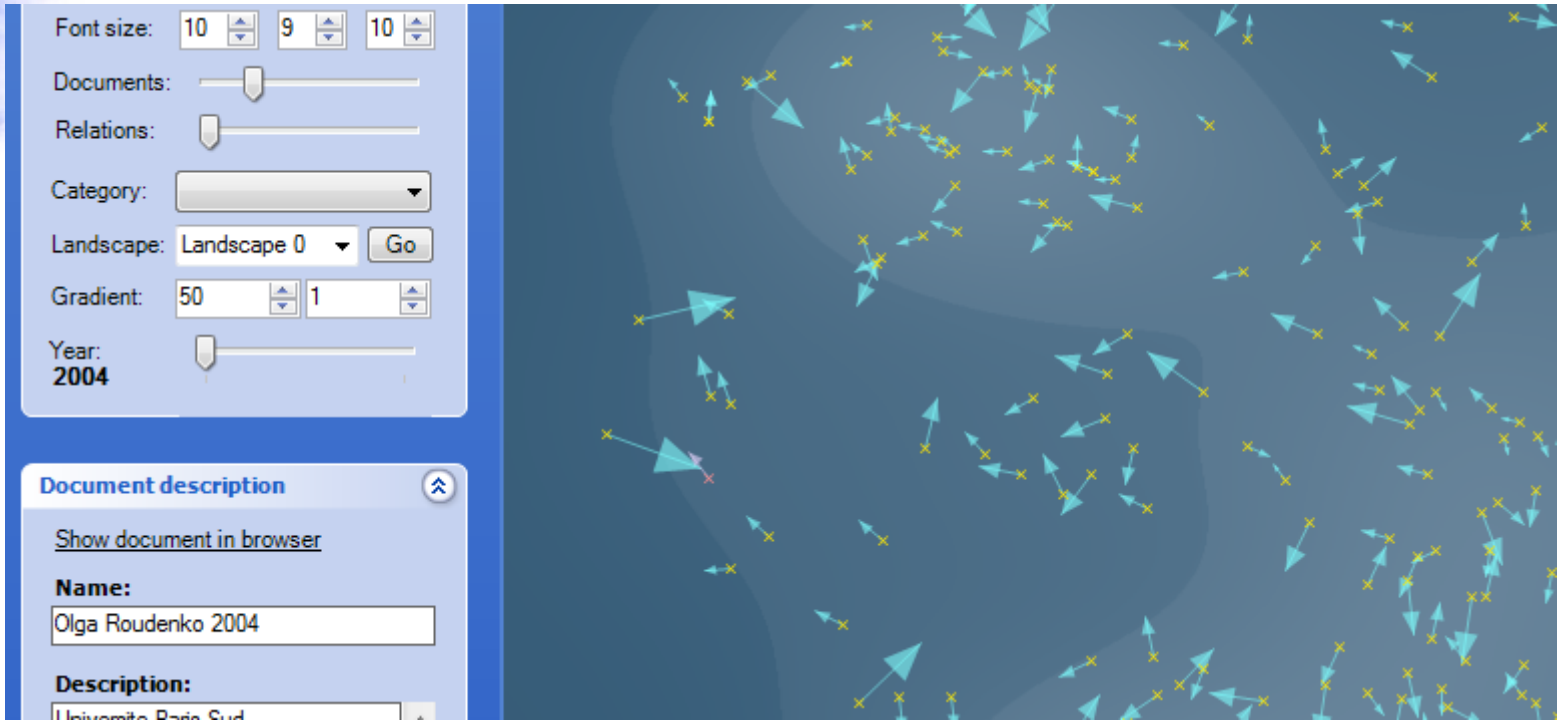
Keywords

User can also zoom in and check keywords for a specific area.



…adding time into the picture

- Often corpora include documents happening in time, so we need to address time dimension
  - …resulting into a dynamic network



DEMO – PASCAL NETWORK

Extracting triples from text and linking to LOD (DBpedia, OpenCyc, Yago) with Enrycher (http://enrycher.ijs.si/)

Plain text

Slovenia's dramatic w… Russia Wednesday, and to a lesser extent Ireland's narrow loss to France, capped off a grueling two-year qualifying period that saw some of the smallest countries in the world kick some of soccer's biggest names in the teeth. After a century of near-domination from the likes of Brazil, Italy and Germany, internation… entering the era of the Cinderella. It may not happen t… but given the increasing flow of talent, training and inf… borders, it's almost certain that a small upstart nation b… athletes and better luck will make a legitimate run at t… coveted trophy.

Russia's Yuri Zhirkov, right, fights for the ball with Slovenia's Valter Birsa Wednesday.

Text Enrichment

Extracted graph of triples from text

entities

- Brazil
- Italy
- Germany
- Cinderella
- Paris
- John O'Shea
- Manchester United
- Robbie Keane
- Shay Given
- Greece
- Portugal
- Bosnia-Herzegovina
- Cristiano Ronaldo
- Uruguay

keywords

Sports, Soccer, CONCACAF, Competitions, United States, Sports and Hobbies, Kids and Teens, World Cup, Women,

categories

- Top/Kids_and_Teens/Sports_and_Hobbies/Sports/Soccer
- Top/Sports/Soccer/Competitions
- Top/Sports/Soccer/Competitions/World_Cup
- Top/Sports/Soccer/CONCACAF

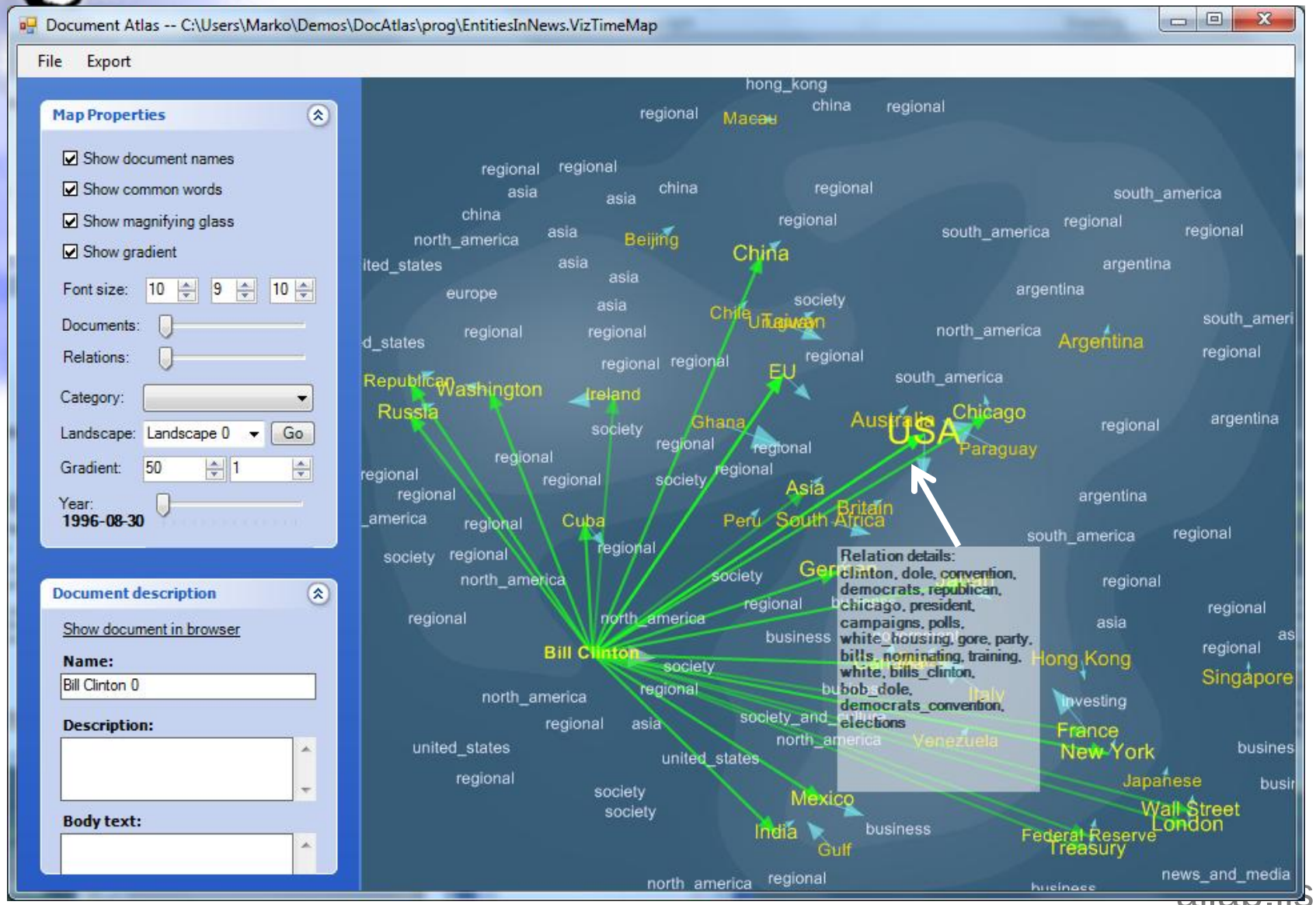
Diego Maradona Semantics:
owl:sameAs: http://dbpedia.org/resource/Diego_Maradona
owl:sameAs: http://sw.opencyc.org/concept/Mx4rvofERZwpEbGdrcN5Y29ycA
rdf:type: http://dbpedia.org/class/yago/ArgentinaInternationalFootballers
rdf:type: http://dbpedia.org/class/yago/ArgentineExpatriatesInItaly
rdf:type: http://dbpedia.org/class/yago/ArgentineFootballManagers
rdf:type: http://dbpedia.org/class/yago/ArgentineFootballers

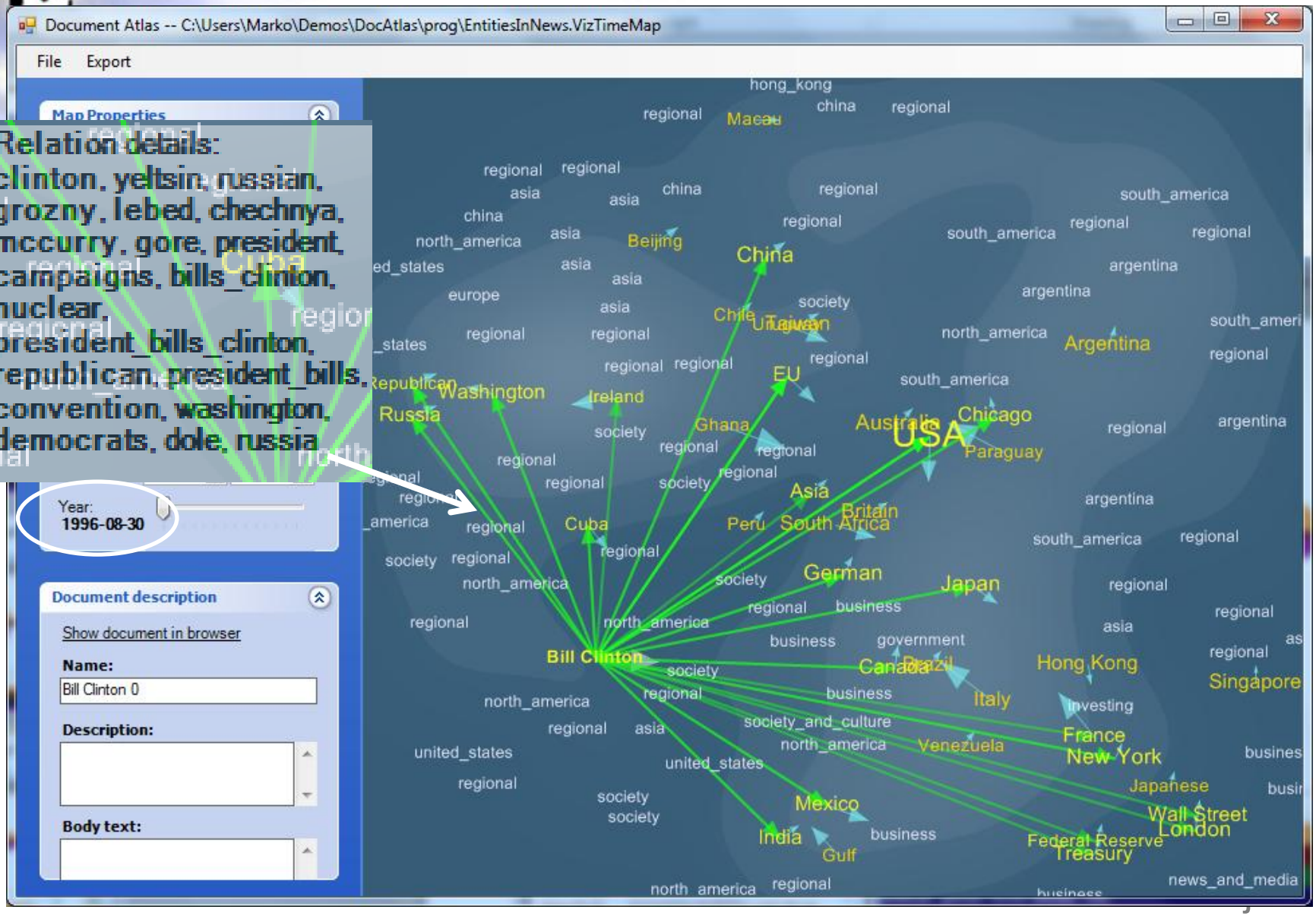
"Enrycher" is available as as a web-service generating Semantic Graph, LOD links, Entities, Keywords, Categories, Text Summarization

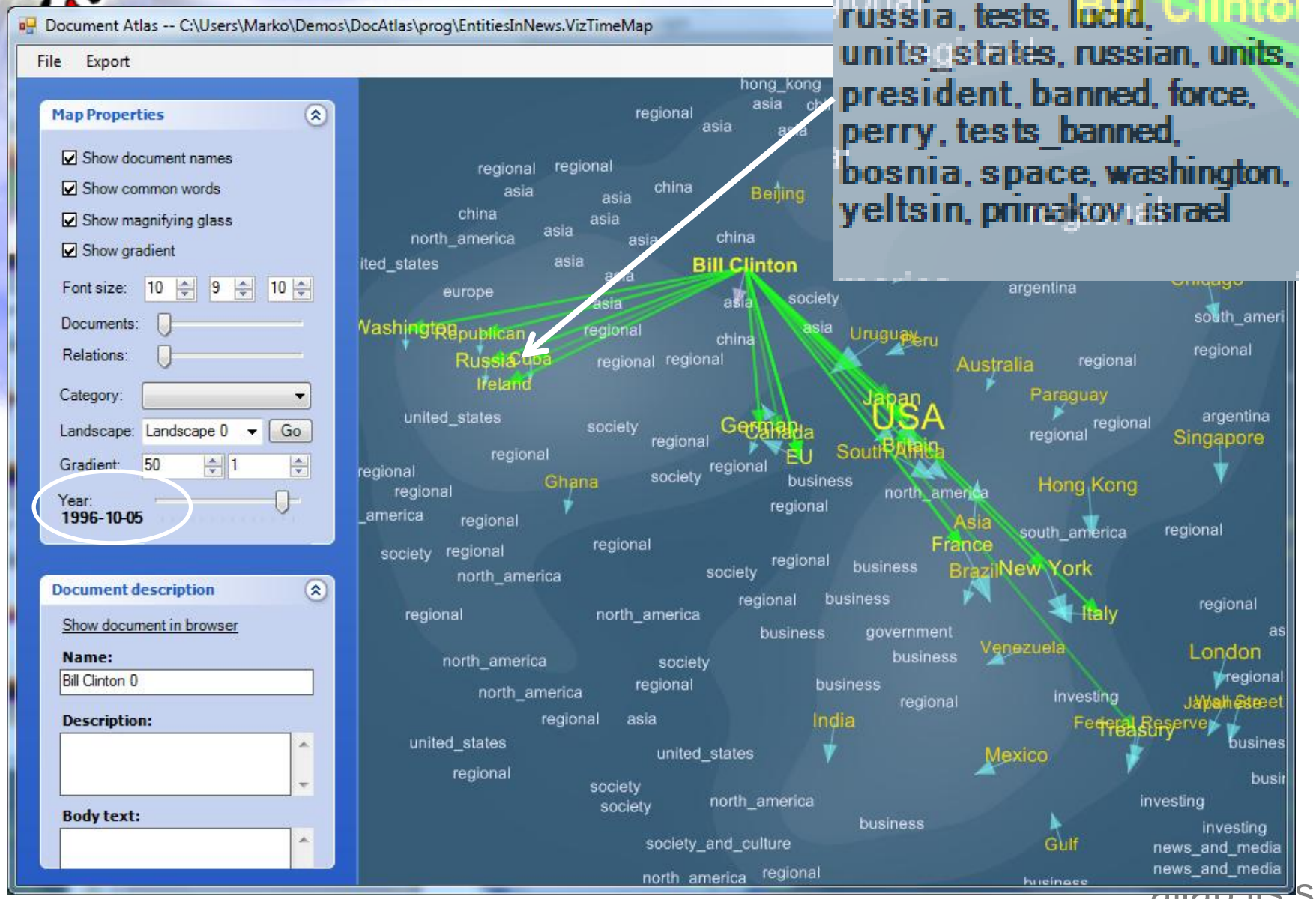
Robbie Keane Semantics:
owl:sameAs: http://dbpedia.org/resource/Robbie_Keane
rdf:type: http://dbpedia.org/class/yago/CoventryCityF.C.Players
rdf:type: http://dbpedia.org/class/yago/ExpatriateFootballPlayersInItaly
rdf:type: http://dbpedia.org/class/yago/F.C.InternazionaleMilanoPlayers

DEMO – ENRYCHER

Relating entities with a keyword context

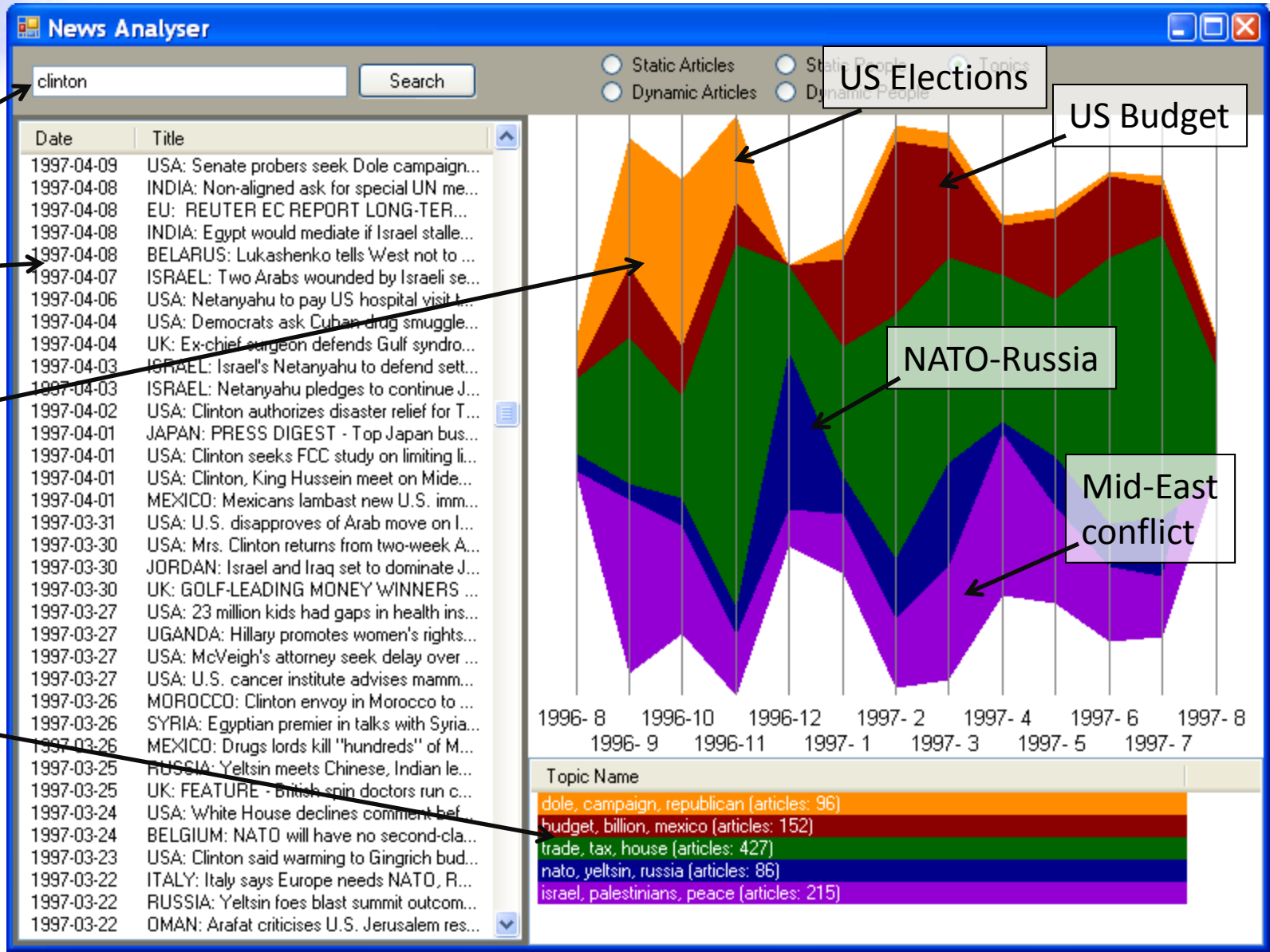



Document Atlas -- C:\Users\Marko\Demos\DocAtlas\prog\EntitiesInNews.VizTimeMap
File
Export
Map Properties
Relation details:
clinton, yeltsin, russian, grozny, lebed, chechnya, mccurry, gore, president, campaigns, bills_clinton, nuclear, president_bills_clinton, republican, president_bills, convention, washington, democrats, dole, russia
Year:
1996-08-30
Document description
Show document in browser
Name:
Bill Clinton 0
Description:
Body text:
Bill Clinton
USA
China
Beijing
Russia
Washington
Ireland
Cuba
Ghana
EU
Asia
Britain
South Africa
Peru
Chile
Taiwan
Argentina
Chicago
Paraguay
Australia
German
Japan
Canada
Brazil
Italy
Mexico
India
Gulf
Venezuela
France
New York
Hong Kong
Singapore
Japanese
Wall Street
London
Federal Reserve
Treasury
Macau
hong_kong
china
regional
asia
europe
north_america
south_america
argentina
society
business
government
investing
society_and_culture
united_states
news_and_media


Relation details:
clinton, treaty, nuclear, russia, tests, lucid, units_states, russian, units, president, banned, force, perry, tests_banned, bosnia, space, washington, yeltsin, primakov, israel
Document Atlas -- C:\Users\Marko\Demos\DocAtlas\prog\EntitiesInNews.VizTimeMap
File
Export
Map Properties
Show document names
Show common words
Show magnifying glass
Show gradient
Font size:
Documents:
Relations:
Category:
Landscape:
Landscape 0
Go
Gradient:
Year:
1996-10-05
Document description
Show document in browser
Name:
Bill Clinton 0
Description:
Body text:
Bill Clinton
Beijing
Washington
Republican
Russia
Ireland
Uruguay
Peru
Australia
Paraguay
Japan
USA
Britain
South Africa
Canada
EU
Ghana
Hong Kong
Singapore
Asia
France
Brazil
New York
Italy
Venezuela
London
India
Mexico
Federal Reserve
Treasury
Gulf
Wall Street
Japanese
hong_kong
china
europe
argentina
south_ameri
south_america
north_america
united_states
business
government
investing
news_and_media
society
society_and_culture
regional
asia

DEMO – ENTITY NETWORK

Topic Tracking from News or Social media



Query

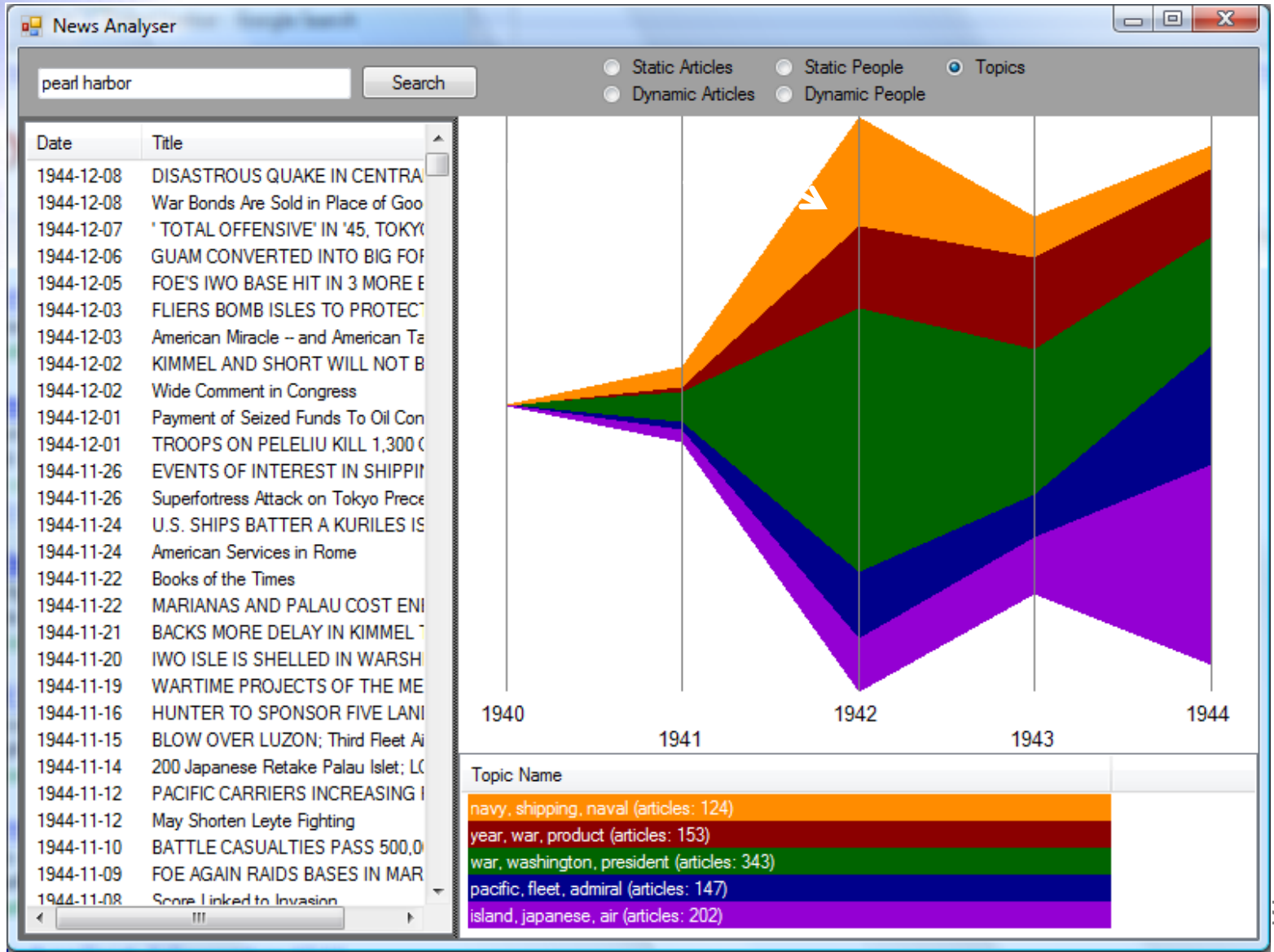
Result set

Topic Trends Visualization

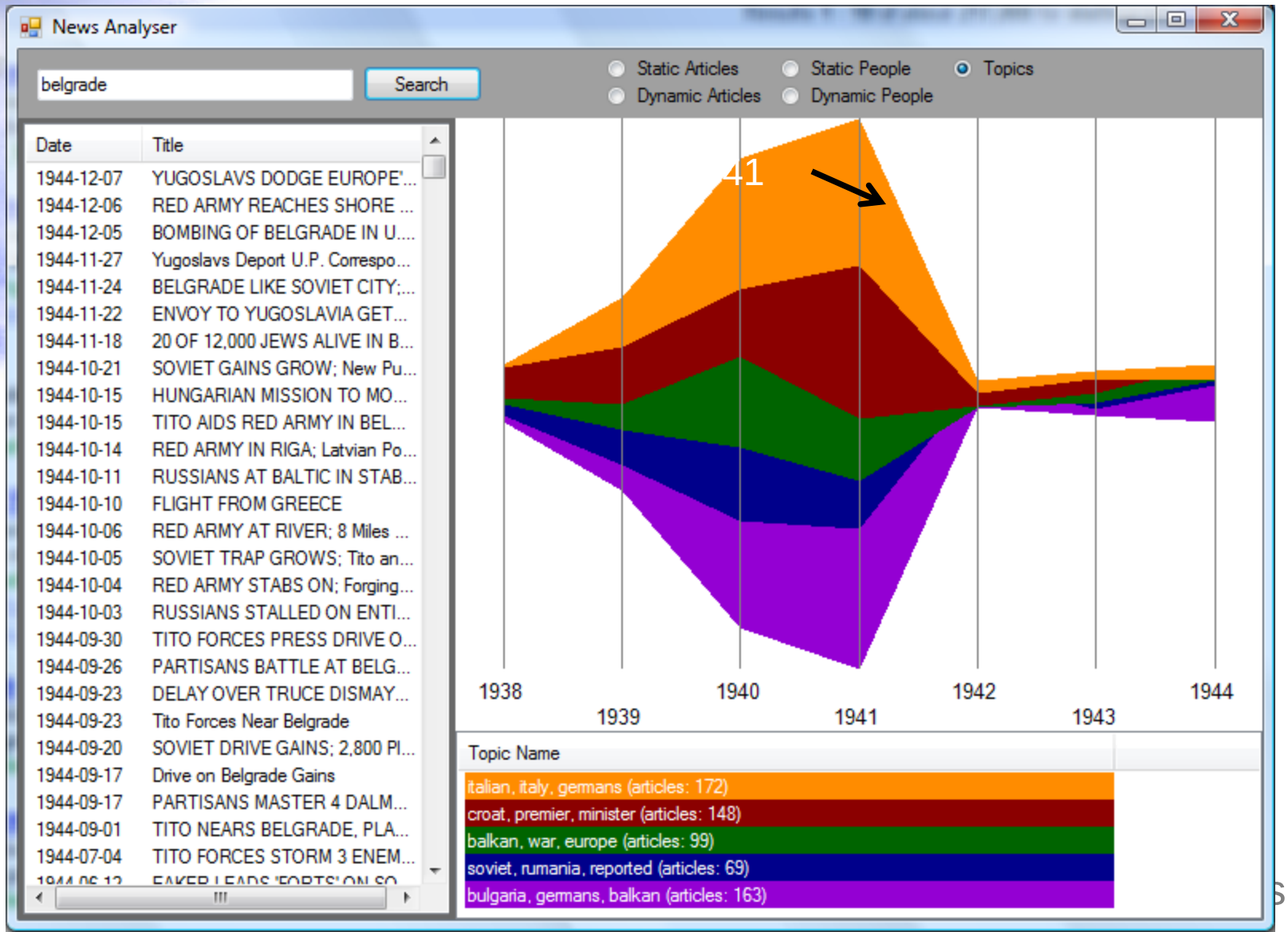
Topics description

DEMO – ARCHIVE EXPLORER

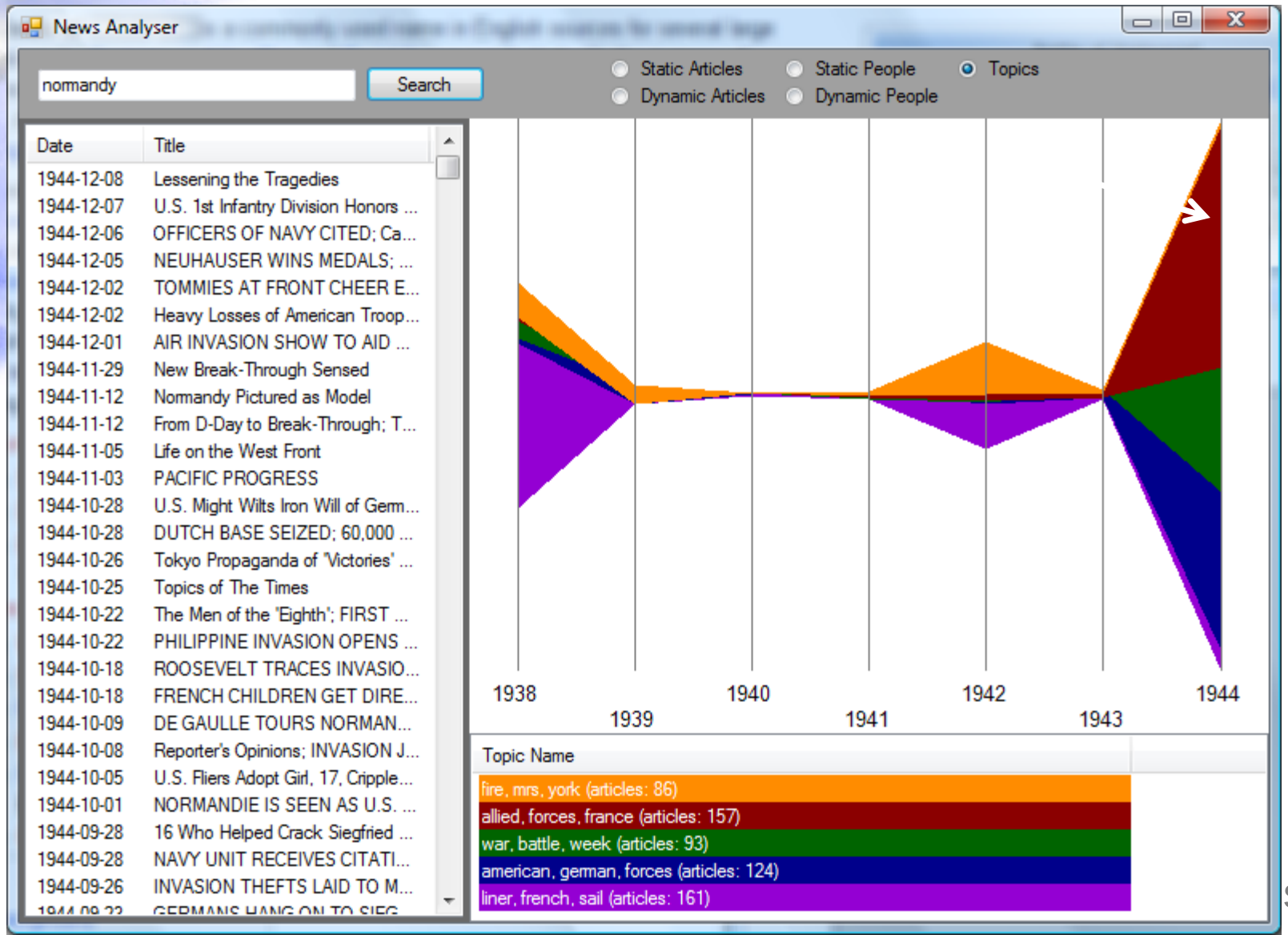
WW2 query “Pearl Harbor” into NYTimes archive



WW2 query “Belgrade” into NYTimes archive



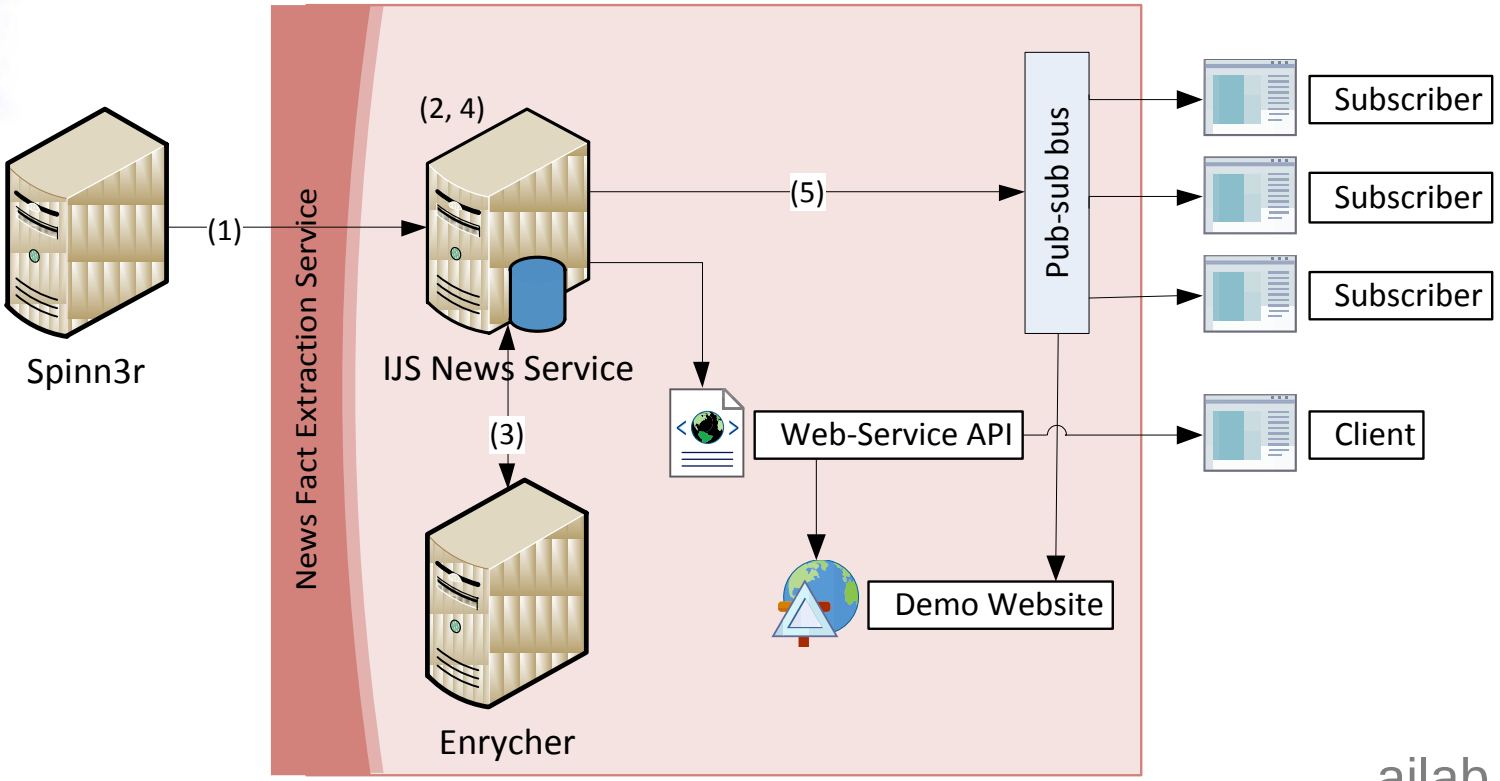
WW2 query “normandy” into NYTimes archive



Application: News & Social Media Fact Extraction Service

Spinn3r: 30 million documents per day

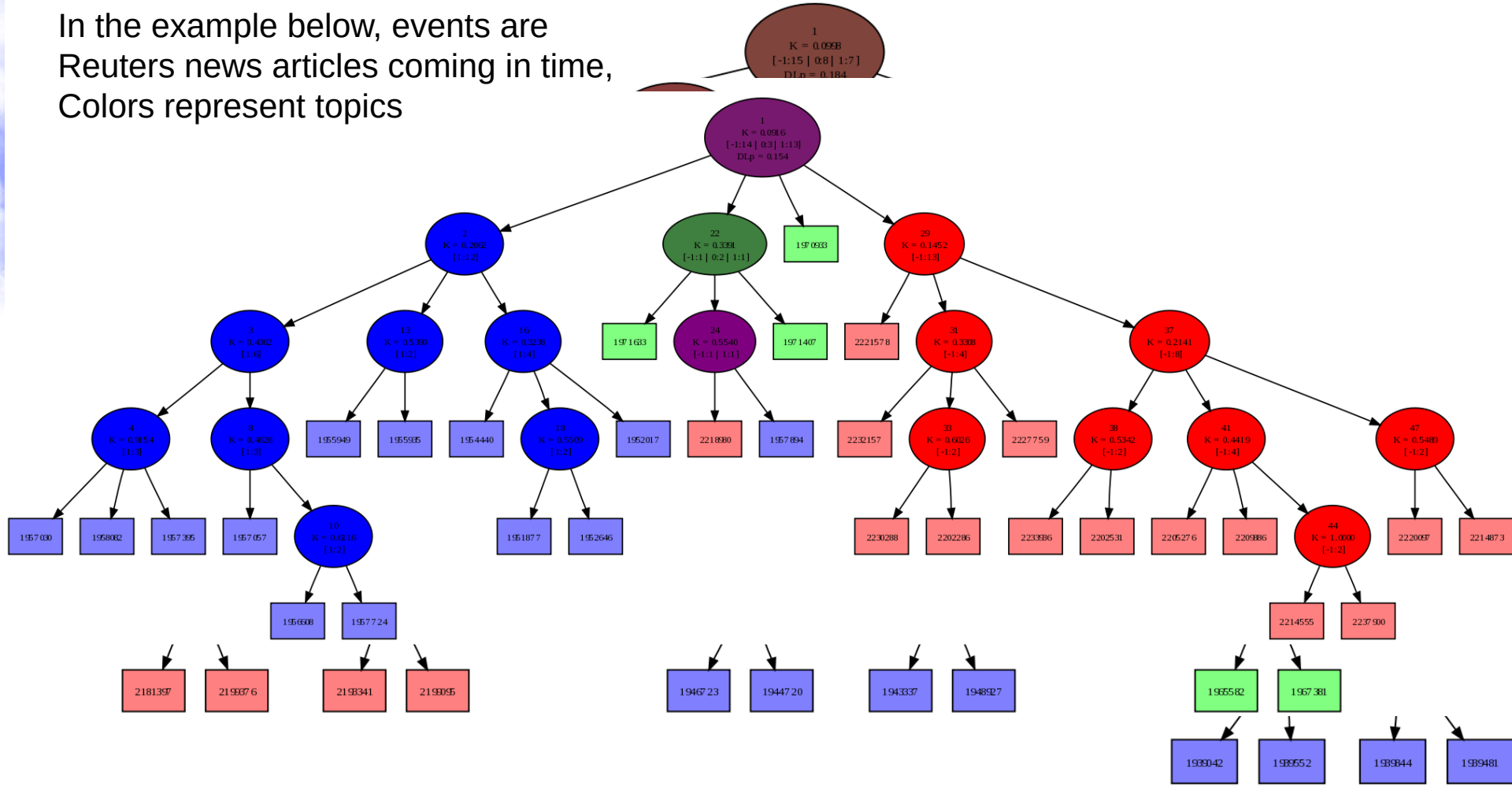
- mainstream news (10,000 monitored sources)
- Blog (40 million monitored blogs)
- micro blog (e.g. Twitter, Facebook status updates)



Example: Stream summarization by incremental hierarchical clustering

The goal is to maintain summary of data from stream in a form of a taxonomy of prototype clusters – each new events updates the taxonomy

In the example below, events are Reuters news articles coming in time, Colors represent topics



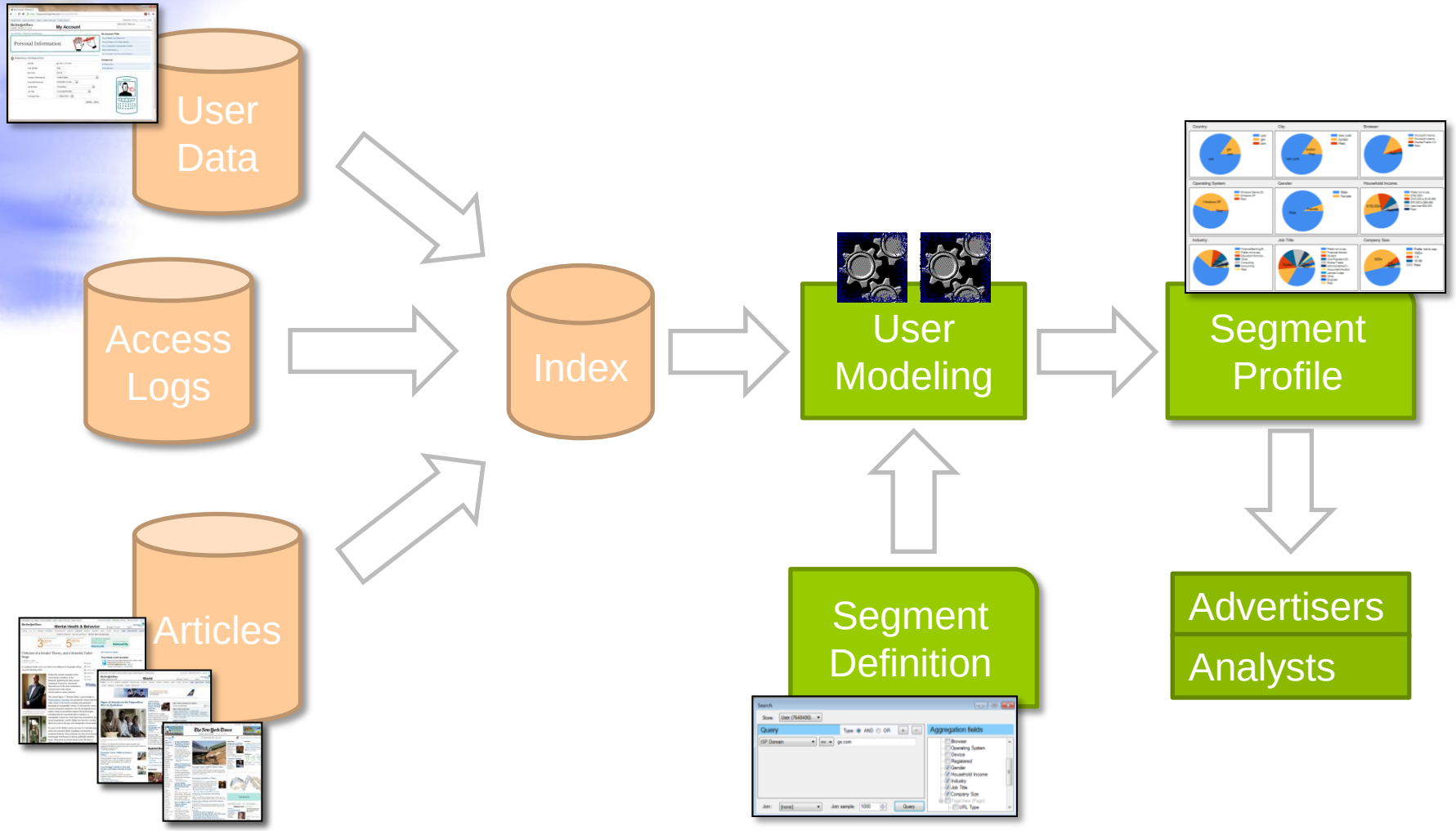
[Blaz Novak, 2008]

WEB CLICK STREAM MINING

Motivation

- Various streams available on the web:
  - Content streams
  - Click streams
  - User data
- Tasks:
  - Identifying and modeling user segments
  - Recommending content

System Overview



Access Logs

- User interactions with the website
- Each page-view described with:
  - **User ID**
  - **Date and Time**
  - **Location** (from IP address)
  - **Requested page**
  - **Referring page**
  - **Search query** (from Referring page)
  - **Browser, Operating System, Device** (from User agent)
- Users tracked using cookies
  - Tag with unique ID at the first visit

Example

User ID cookie: `1234567890`

IP: `123.123.123.123` (Beijing, China)

Requested URL:
http://www.nytimes.com/2009/08/23/weekinreview/23baker.html

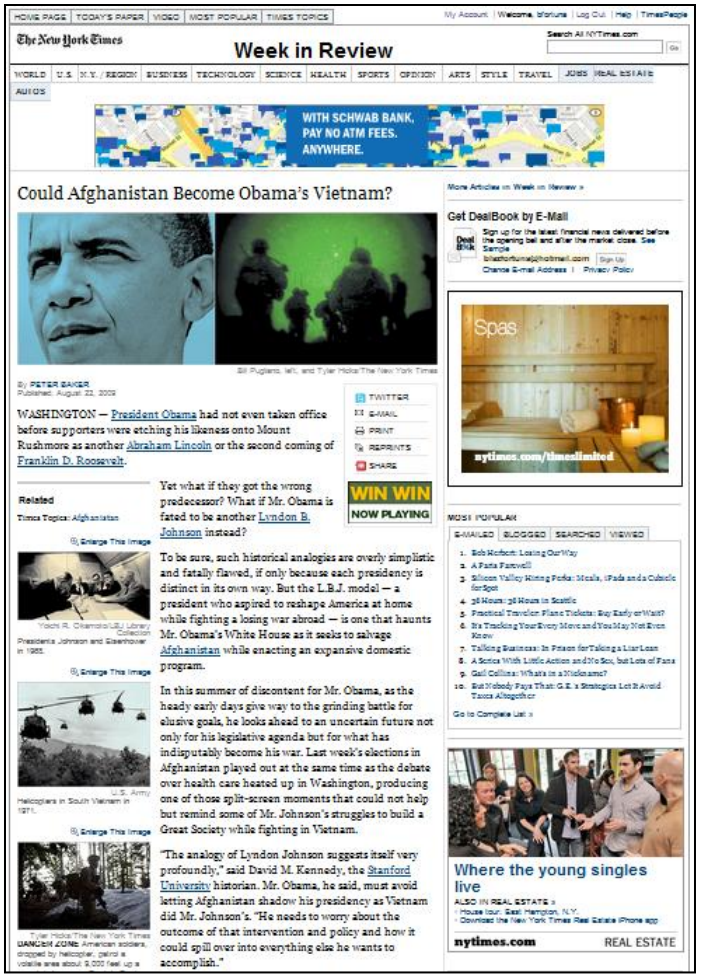
Referring URL:
http://query.nytimes.com/search/sitesearch?query=obama

Date and time: `2009-08-25 08:12:34`

User agent: `Mozilla/5.0 (Windows; U; Windows NT 5.1; en) AppleWebKit/526.9 (KHTML, like Gecko) Version/4.0dp1 Safari/526.8` (Safari, Windows, PC)

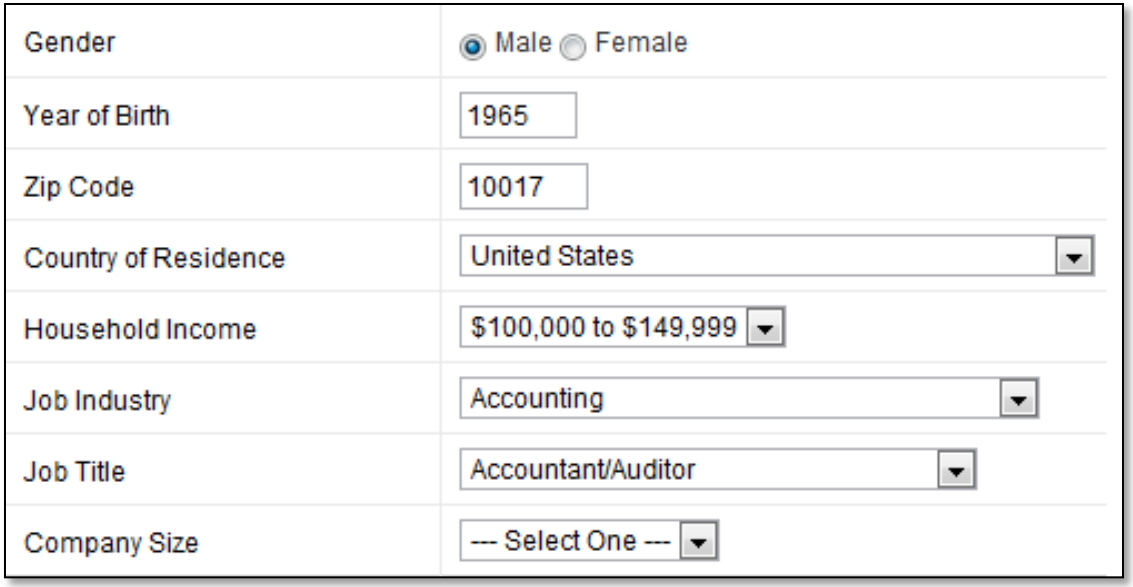
Articles

- Content and Semantics about requested pages
- Each page described with:
  - **Content**
  - **Annotations**
    - **Named Entities** (e.g. Obama, Mount Rushmore, Afghanistan, Vietnam)
    - **Topics** (e.g. politics, opinion, sports)
  - **Content meta-data** (e.g. author, publish date, editorial desk)
  - **Page meta-data** (e.g. article, home-page, section-front)



User Data

- Provided only for registered users
  - ~20% unique users in our case
  - Can generalize to all using machine learning
- Each registered users described with:
  - **Gender**
  - **Year of birth**
  - **Household income**
- Noisy



User Segment

- User segment:

> Subset of website visitors sharing some common characteristics

- Example:
  - [Gender = Male]
  - [Age ≥ 40]
  - [Referring domain = facebook.com]
  - [Requested page topic = Travel]
  - …

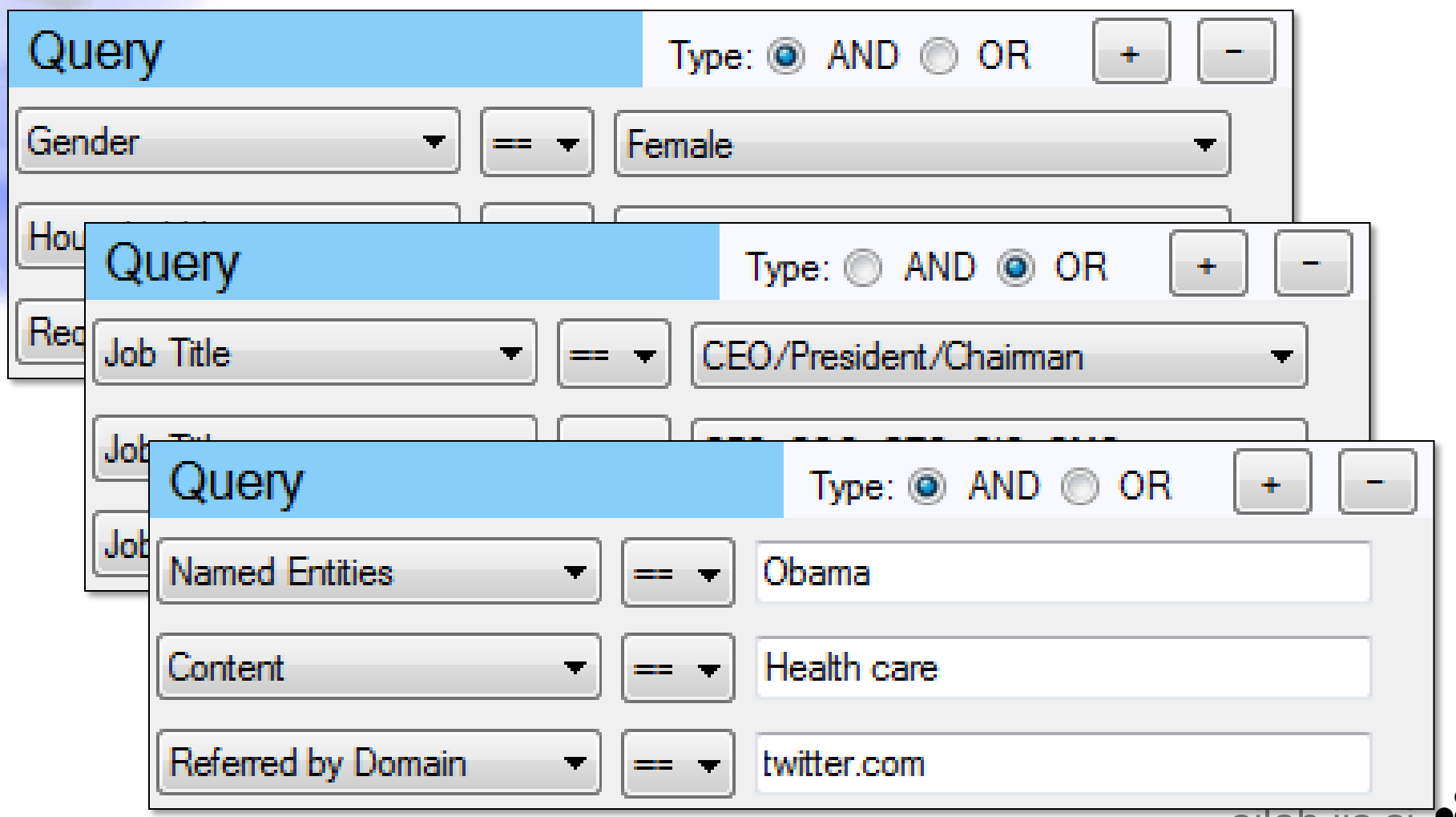
Defining Segments

- Must be simple enough so it can be used by domain experts
- Our solution
  - Index all users using inverted index
  - Segment definition equals faceted search query over users
  - Ad-hoc segment definitions

Indexed fields:

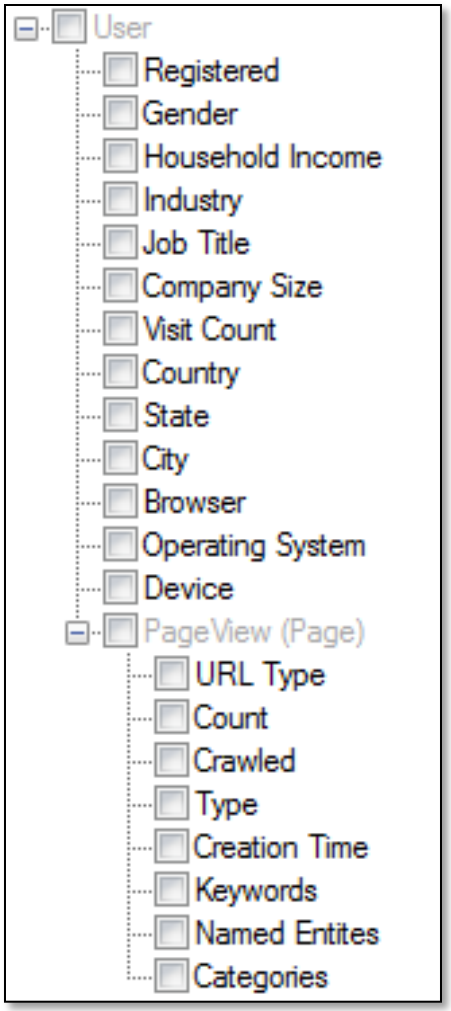
- Domain
- Sub-domain
- Page URL
- Page Meta Tags
- Page Title
- Page Content
- Named Entities
- Referring Search Term
- Referring Domain
- Referring URL
- Country (from IP)
- State (from IP)
- City (from IP)
- Date
- Day of the Week
- Hour of the day
- User Agent
- Income
- Age
- Gender

Example


Query
Type: AND OR
+
−
Gender
==
Female
Query
Type: AND OR
+
−
Job Title
==
CEO/President/Chairman
Query
Type: AND OR
+
−
Named Entities
==
Obama
Content
==
Health care
Referred by Domain
==
twitter.com

User Modeling

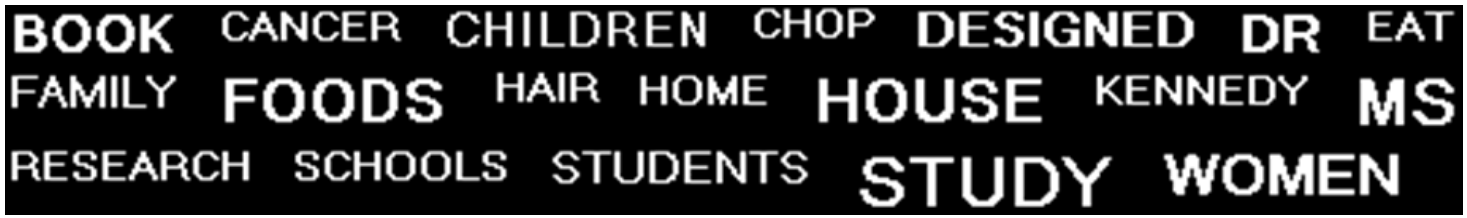
- Feature space
  - Extracted from subset of fields
  - Using vector space model
  - Vector elements for each field are normalized
- Training set
  - One visit = one vector
  - One user = a centroid of all his/her visits
  - Users from the segment form positive class
  - Sample of other users form negative class
- Classification algorithm
  - Support Vector Machine
  - Good dealing with high dimensional data
  - Linear kernel
  - Stochastic gradient descent
    - Good for sampling



Segment visualization

- Using SVM for feature selection
- Visualize a segment by displaying keywords significant for correct classification
- Useful information for the website editors

Gender = female
Income ≥ $100,000
Meta Data = Category Style



Experimental setting

- Real-world dataset from a major news publishing website
  - 5 million daily users, 1 million registered
- Tested prediction of three demographic dimensions:
  - Gender, Age, Income
- Three user groups based on the number of visits:
  - ≥2, ≥10, ≥50
- Evaluation:
  - Break Even Point (BEP)
  - 10-fold cross validation

Category	Size
Male	250,000
Female	250,000

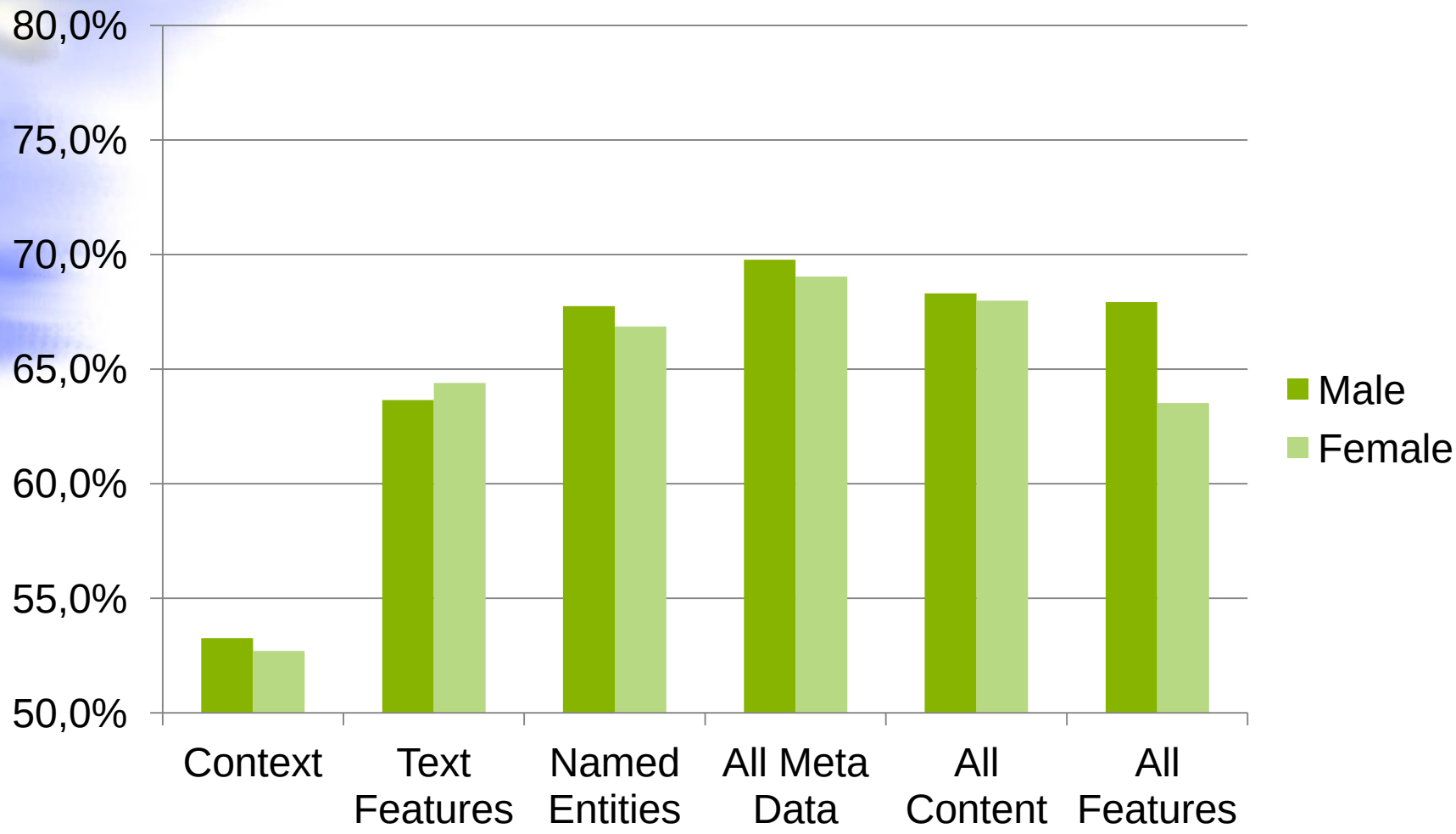
Category	Size
21-30	100,000
31-40	100,000
41-50	100,000
51-60	100,000
61-80	100,000

Category	Size
0-24k	50,000
25k-49k	50,000
50k-74k	50,000
75k-99k	50,000
100k-149k	50,000
150k-254k	50,000

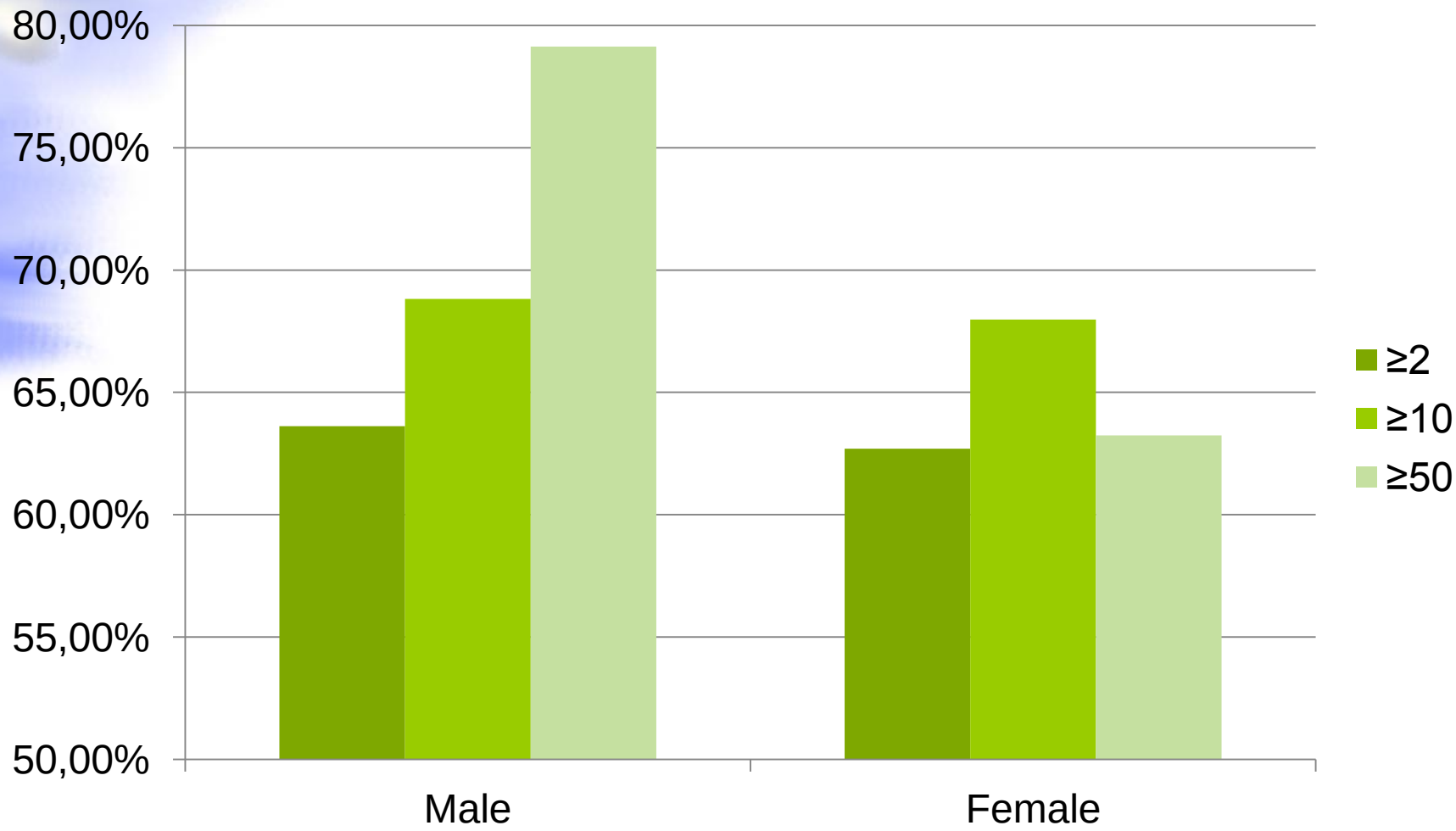
Combining Features

- **Context** – features that can be obtained from access logs, such as time, referring page, location and device.
- **Content features**:
  - **Text Features** – keywords extracted from the articles
  - **Named Entities** – automatically extracted named entities
  - **All Metadata** –assigned to the article by the authors and editors
    - byline; topics; main keywords; people, organization and countries mentioned in the article; publish date.
- **All Content** – combination of text features, named entities and metadata features.
- **All Features** – combination of all above features.

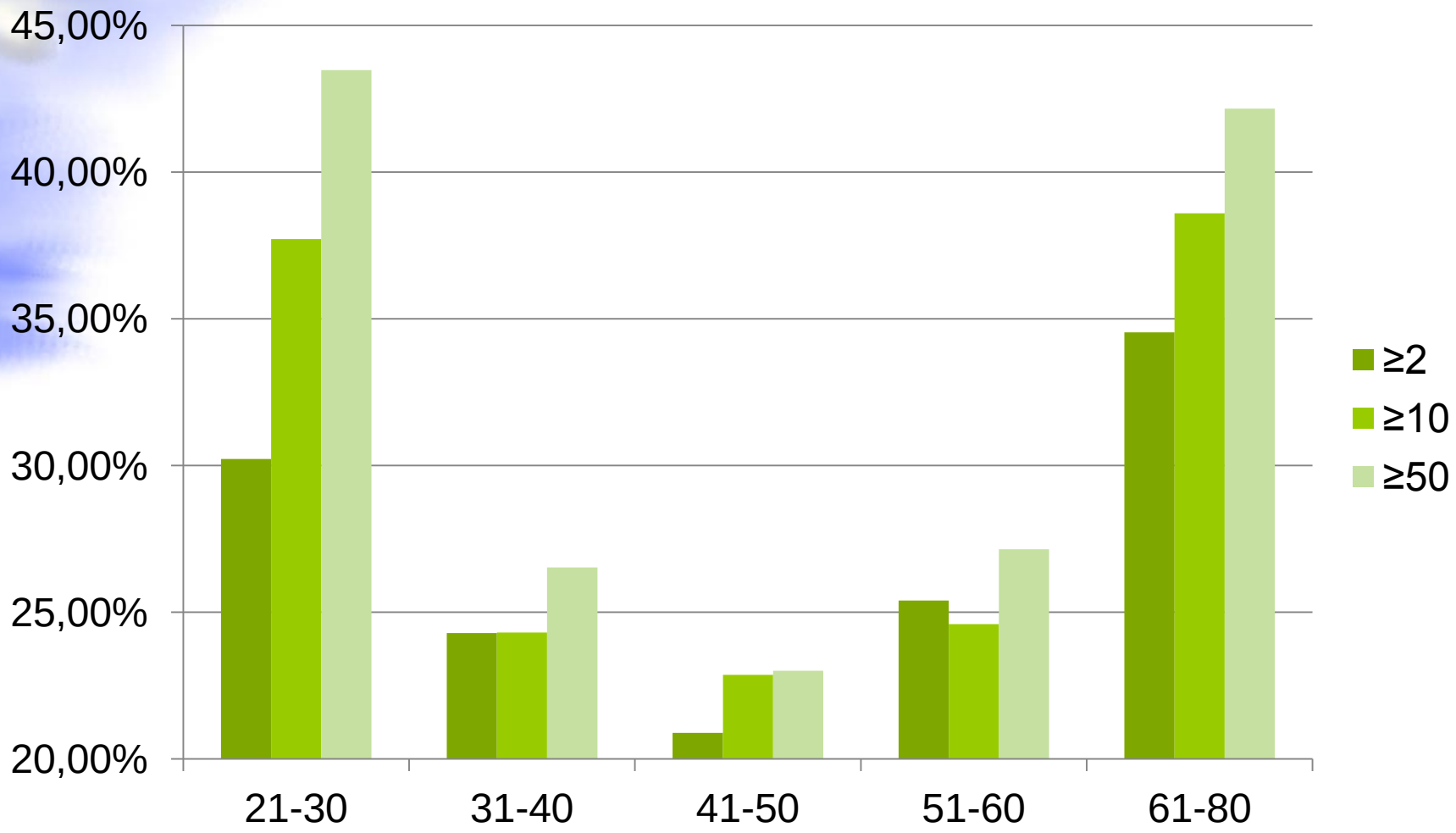
Gender (≥10 visits)



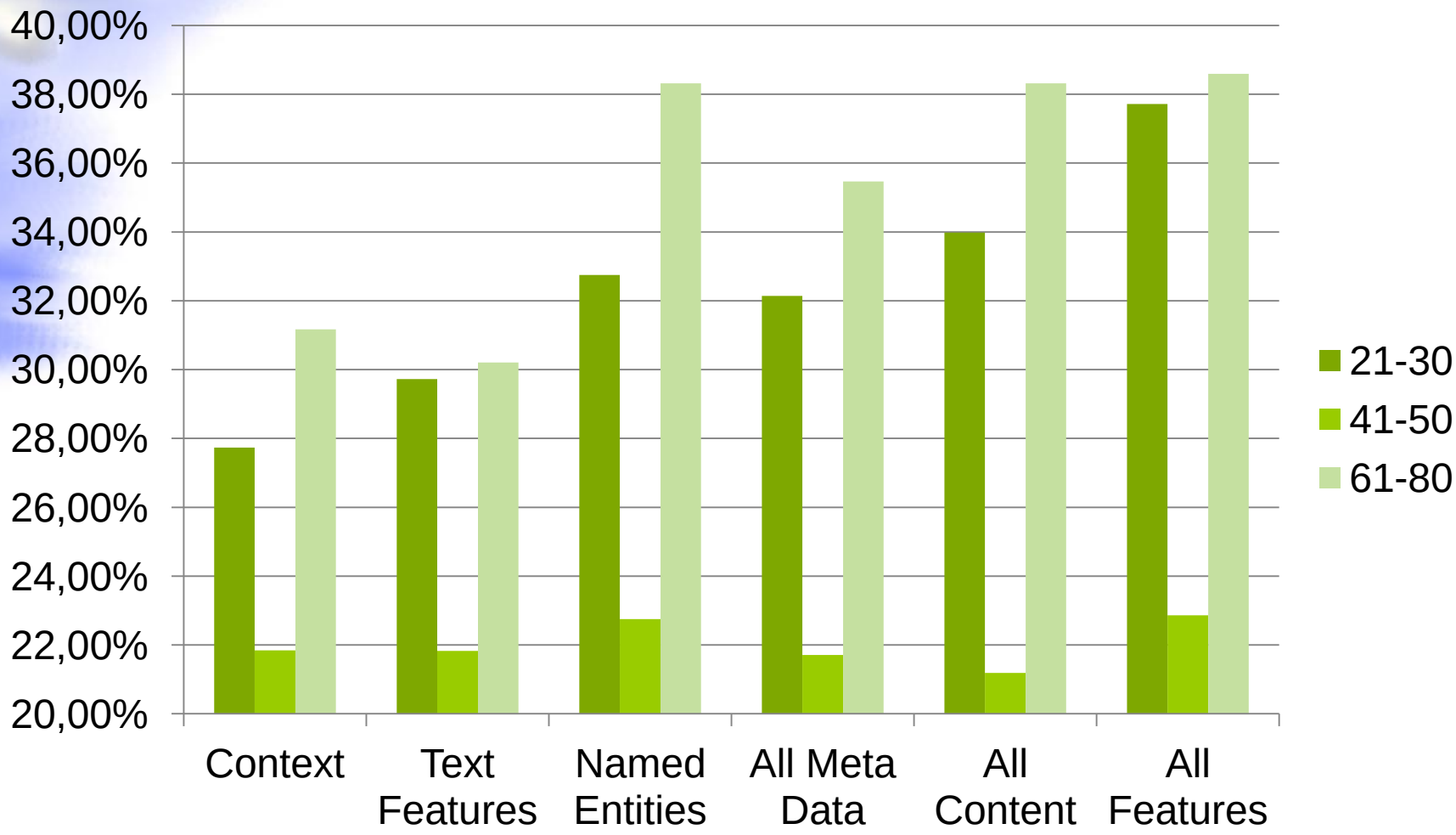
Gender



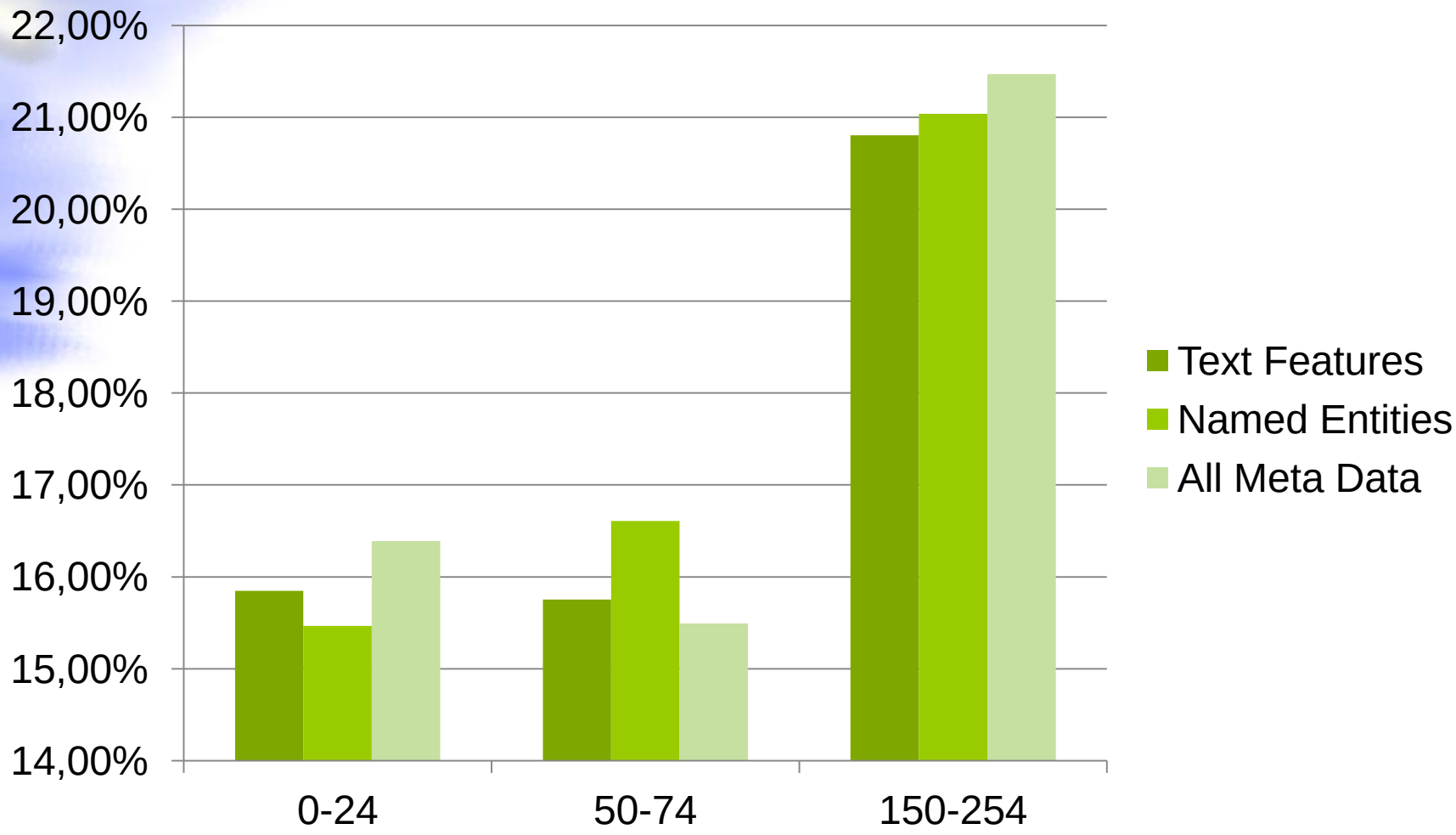
Age (all features)



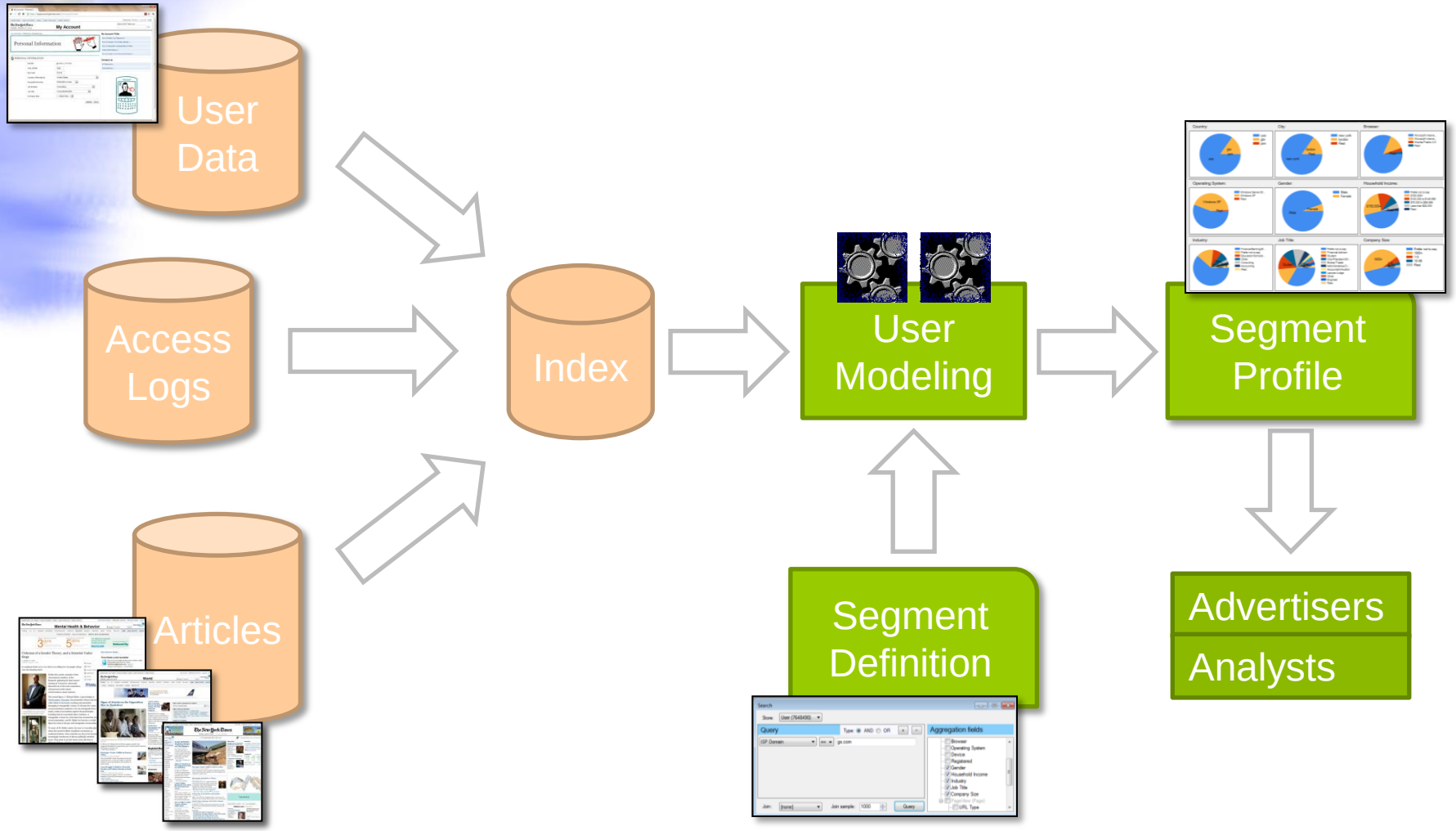
Age (≥10 visits)



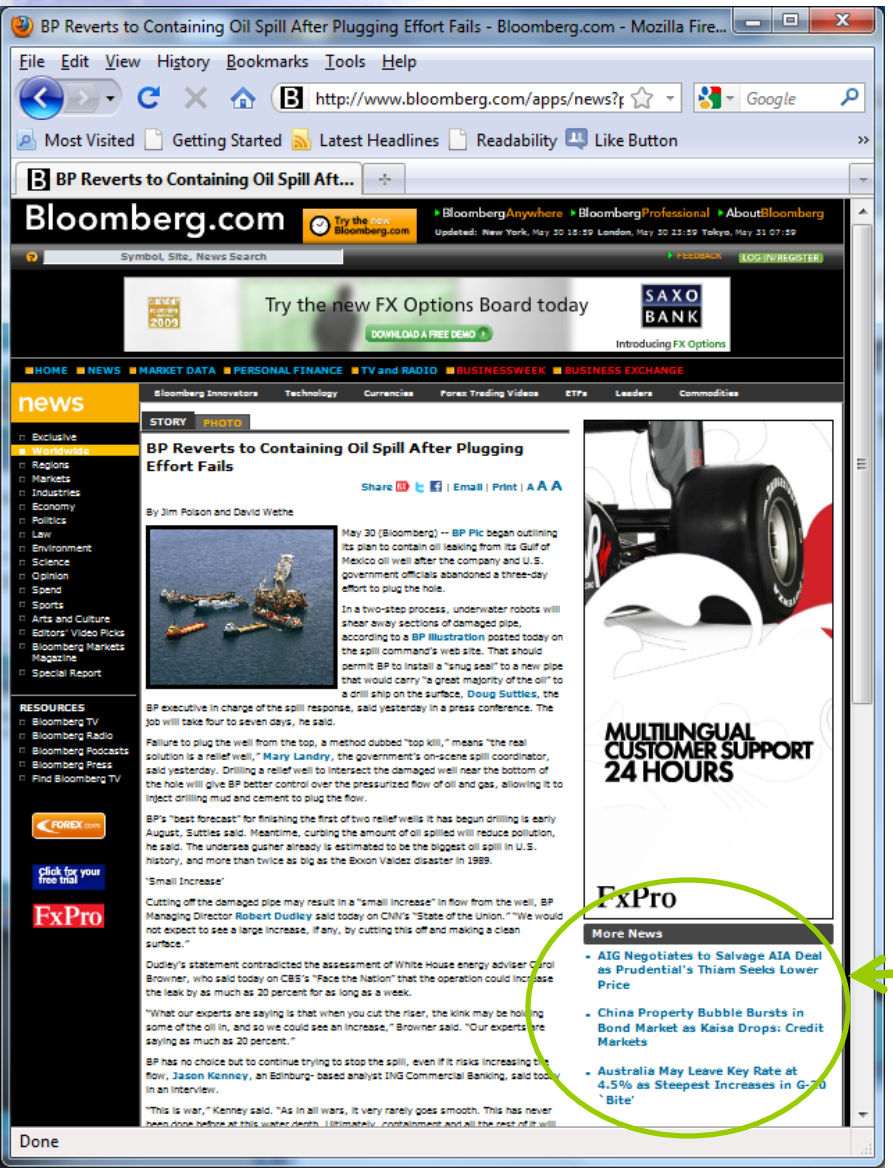
Income (≥10 visits)



System Overview



News Recommendation

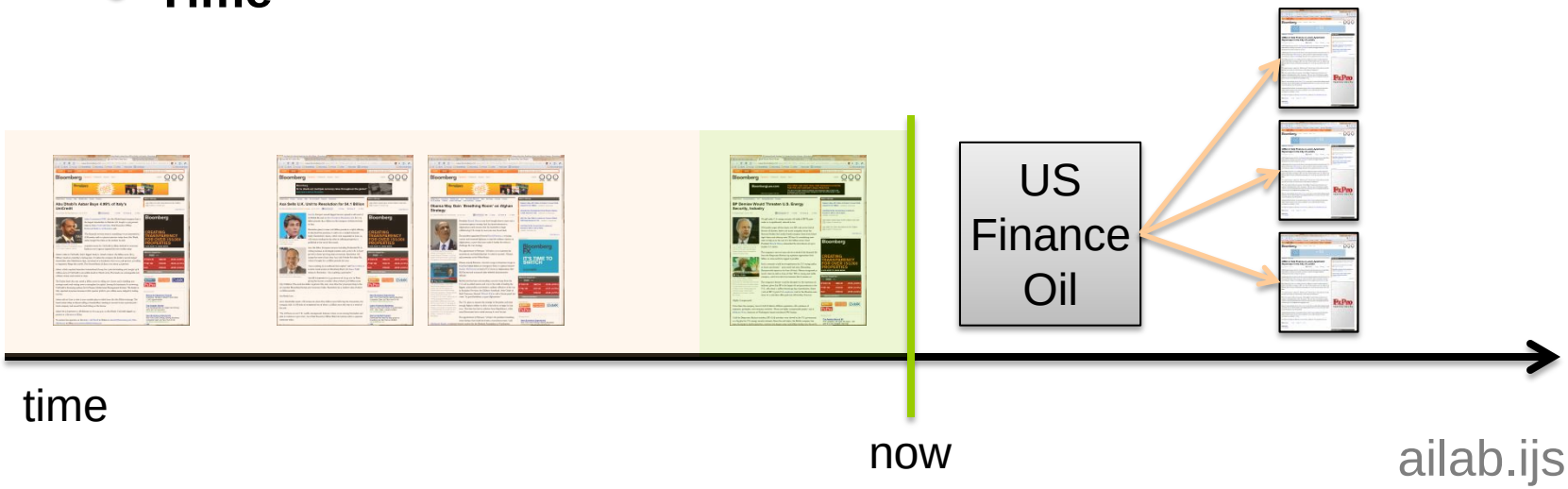


- Good recommendations can make a big difference when keeping a user on a web site
  - …the key is how rich context model a system is using to select information for a user

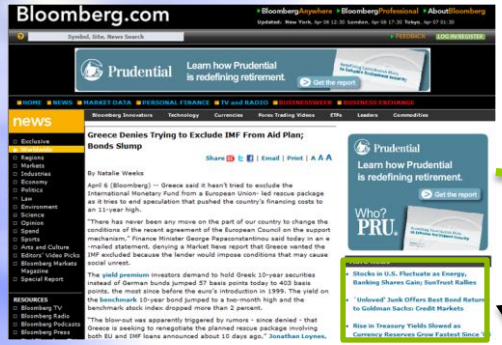
Contextual personalized recommendations generated in ~20ms

Recommendation

- Task:
  - At the moment of visit, predict the category of the next article
- Features:
  - **History** (user profile)
  - **Geo** (based on IP)
  - **Requested page** (where we serve recommendation)
  - **Referring URL**
  - **Time**



Real-time Architecture



Log visit

Logging

Store

Archive

Recommend

Update index

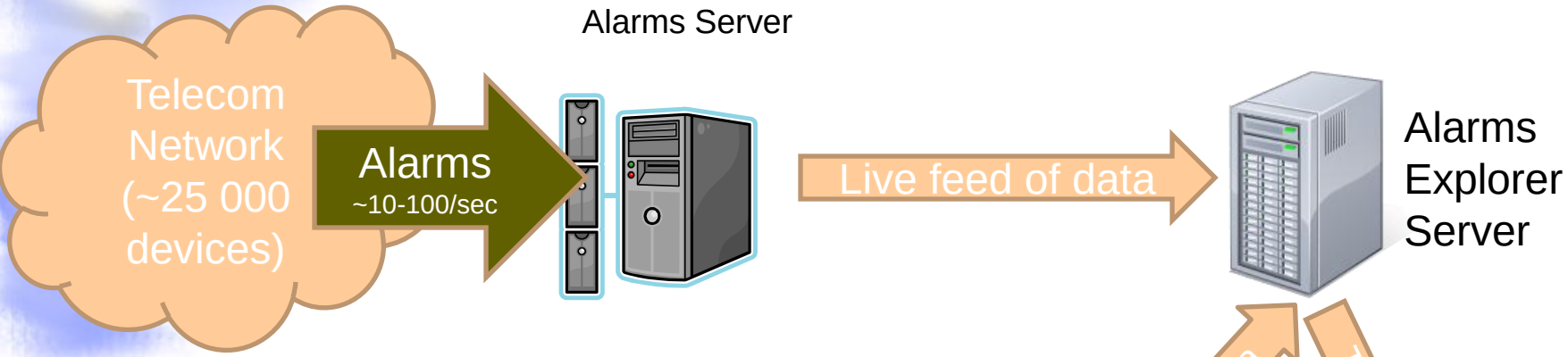
Recommendation Engine

Web

Amazon

TELECOMMUNICATION NETWORK MONITORING

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

CONCLUSIONS

Conclusions

- Presented streams of various modalities
  - Document streams
  - Dynamic social networks
  - Click streams
  - Event streams
- Future Challenges:
  - Commoditization still a challenge
    - Performing at high scale requires special handling of each modality
  - Dealing with unstructured data (e.g. documents, networks)
  - Semantic streams
    - Mapping low level events to higher level semantic concepts (e.g. sensor networks)

THANK YOU