



MEDNARODNA  
PODIPLOMSKA ŠOLA  
JOŽEFA STEFANA

JOŽEF STEFAN  
INTERNATIONAL  
POSTGRADUATE SCHOOL

## Text, web and multimedia mining

izr. prof. dr. Dunja Mladenić  
dr. Blaž Fortuna

Module Knowledge Technologies (ICT3)  
2011/2012

www.risi.si

## Overview

- Introduction
- Data representation
- Example tasks
- Algorithms
- References

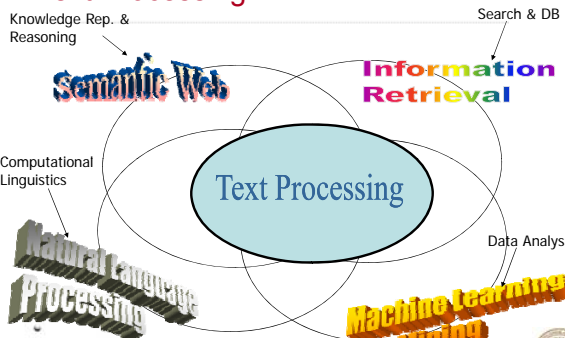
©Dunja Mladenic

## What are we talking about?

- "...finding **interesting** regularities in large **text, web or multimedia** data..." (Usama Fayad, adapted)
  - ...where **interesting** means: non-trivial, hidden, previously unknown and potentially useful
- finding semantic and abstract information from the raw data
  - surface form of text, bitmap of photos, graph structure
- finding regularities in web structure, web logs, web content
  - analysis and profiling of web customers based on web-server log files

©Dunja Mladenic

## Research areas contributing to Text Processing



Knowledge Rep. & Reasoning

Search & DB

Semantic Web

Information Retrieval

Computational Linguistics

Text Processing

Natural Language Processing

Machine Learning

Text Mining

Data Analysis

©Dunja Mladenic

## Dimensions in text analytics

- Three major dimensions of text analytics:
  - Representations
    - ...from character-level to first-order theories
  - Techniques
    - ...from manual work, over learning to reasoning
  - Tasks
    - ...from search, over (un-, semi-) supervised learning, to visualization, summarization, translation ...

©Dunja Mladenic

## Levels of text representations

- Character (character n-grams and sequences)
- Words (stop-words, stemming, lemmatization)
- Phrases (word n-grams, proximity features)
- Part-of-speech tags
- Taxonomies / thesauri
- Vector-space model
- Language models
- Full-parsing
- Cross-modality
- Collaborative tagging / Web2.0
- Templates / Frames
- Ontologies / First order theories

Lexical

Syntactic

Semantic

©Dunja Mladenic

## Levels of text representation

- Character (character n-grams and sequences)
- Words (stop-words, stemming, lemmatization)
- Phrases (word n-grams, proximity)
- Part-of-speech tags
- Taxonomies / thesauri
- Vector-space model
- Language models
- Full-parsing
- Cross-modality
- Collaborative tagging / Web2.0
- Templates / Frames
- Ontologies / First order theories

Language identification, Copy detection

Named entity extraction (names of people, places, organizations)

Text categorization, Clustering, Search, Summarization, ...

Spam filtering, Machine translation

Multilingual search, Associating text with images, ...

Unifying semantics of data sets

Reasoning, Semantic search

**Lexical**

**Synactic**

**Semantic**

## Levels of text representations

- **Character**
- Words
- Phrases
- Part-of-speech tags
- Taxonomies / thesauri
- Vector-space model
- Language models
- Full-parsing
- Cross-modality
- Collaborative tagging / Web2.0
- Templates / Frames
- Ontologies / First order theories

**Lexical**

**Synactic**

**Semantic**

## Character level

- Character level representation of a text consists from sequences of characters...
  - ...a document is represented by a frequency distribution of sequences
  - Usually we deal with contiguous strings...
  - ...each character sequence of length 1, 2, 3, ... represent a feature with its frequency

## Good and bad sides

- Representation has several important strengths:
  - ...it is very robust since avoids language morphology
    - (useful for e.g. language identification)
  - ...it captures simple patterns on character level
    - (useful for e.g. spam detection, copy detection)
  - ...because of redundancy in text data it could be used for many analytic tasks
    - (learning, clustering, search)
    - It is used as a basis for "string kernels" in combination with SVM for capturing complex character sequence patterns
- ...for deeper semantic tasks, the representation is too weak

## Levels of text representations

- Character
- **Words**
- Phrases
- Part-of-speech tags
- Taxonomies / thesauri
- Vector-space model
- Language models
- Full-parsing
- Cross-modality
- Collaborative tagging / Web2.0
- Templates / Frames
- Ontologies / First order theories

**Lexical**

**Synactic**

**Semantic**

## Word level

- The most common representation of text used for many techniques
  - ...there are many tokenization software packages which split text into the words
- Important to know:
  - Word is well defined unit in western languages – e.g. Chinese has different notion of semantic unit

## Words Properties

- Relations among word surface forms and their senses:
  - **Homonymy**: same form, but different meaning (e.g. bank: river bank, financial institution)
  - **Polysemy**: same form, related meaning (e.g. bank: blood bank, financial institution)
  - **Synonymy**: different form, same meaning (e.g. singer, vocalist)
  - **Hyponymy**: one word denotes a subclass of an another (e.g. breakfast, meal)
- Word frequencies in texts have **power distribution**:
  - ...small number of very frequent words
  - ...big number of low frequency words

## Stop-words

- Stop-words are words that from non-linguistic view do not carry information
  - ...they have mainly functional role
  - ...usually we remove them to help the methods to perform better
- Stop words are language dependent – examples:
  - **English**: A, ABOUT, ABOVE, ACROSS, AFTER, AGAIN, AGAINST, ALL, ALMOST, ALONE, ALONG, ALREADY, ...
  - **Dutch**: de, en, van, ik, te, dat, die, in, een, hij, het, niet, zijn, is, was, op, aan, met, als, voor, had, er, maar, om, hem, dan, zou, of, wat, mijn, men, dit, zo, ...
  - **Slovenian**: A, AH, AHA, ALI, AMPAK, BAJE, BODISI, BOJDA, BRŽKONE, BRŽČAS, BREZ, CELO, DA, DO, ...

## Stemming and lemmatization

- Different forms of the same word are usually problematic for text data analysis, because they have different spelling and similar meaning (e.g. learns, learned, learning,...)
- Stemming is a process of transforming a word into its stem
  - (universe, university, universities, university's, universal) → univers
- Lemmatization transforms word into its normalized form
  - universe → universe, (university, universities, university's) → university, universal → universal
- ...stemming provides an inexpensive mechanism to merge words with similar meaning

## Stemming

- For English is mostly used Porter stemmer at <http://www.tartarus.org/~martin/PorterStemmer/>
- Example cascade rules used in English Porter stemmer
 

– ATIONAL → ATE	relational → relate
– TIONAL → TION	conditional → condition
– ENCI → ENCE	valenci → valence
– ANCI → ANCE	hesitanci → hesitance
– IZER → IZE	digitizer → digitize
– ABLI → ABLE	conformabli → conformable
– ALLI → AL	radicalli → radical
– ENTLI → ENT	differentli → different
– ELI → E	vileli → vile
– OUSLI → OUS	analogousli → analogous

## Levels of text representations

- Character
- Words
- **Phrases**
  - Part-of-speech tags
  - Taxonomies / thesauri
- **Vector-space model**
- Language models
- Full-parsing
- Cross-modality
- Collaborative tagging / Web2.0
- Templates / Frames
- Ontologies / First order theories

Lexical

Syntactic

Semantic

## Phrase level

- Instead of having just single words we can deal with phrases
  - artificial intelligence, text mining, word for windows
- We use two types of phrases:
  - Phrases as frequent contiguous word sequences
  - Phrases as frequent non-contiguous word sequences
  - ...both types of phrases could be identified by simple dynamic programming algorithm
- The main effect of using phrases is to more precisely identify sense

## Google n-grams

- September 2006, Google released n-grams (sequences of up to n words)

Length of n-gram	Number of different n-grams
1	13,588,391
2	314,843,401
3	977,069,902
4	1,313,818,354
5	1,176,470,663
no. sentences	95,119,665,584
no. words	1,024,908,267,229

passive smoking increased the risk  
 cow eats grass  
 humans currently reside on earth  
 Iraq declared war  
 ship docked in the port  
 we use this a lot  
 for all the examples </S>  
 15th Century Book of Hours  
 170USD go thread ( 1  
 1395 0 BEA171 H 19

<http://googlerecherche.blogspot.com/2006/08/all-our-n-gram-are-belong-to-you.html#ixz5>

## Example: Google n-grams

- ceramics collectables collectibles 55
- ceramics collectables fine 130
- ceramics collected by 62
- ceramics collectible pottery 50
- ceramics collectibles cooking 45
- ceramics collection 144
- ceramics collection : 247
- ceramics collection </S> 120
- ceramics collection and 43
- ceramics collection at 52
- ceramics collection is 68
- ceramics collection of 76
- ceramics collection | 59
- ceramics collections , 66
- ceramics collections . 60
- ceramics combined with 46
- ceramics come from 69
- ceramics comes from 660
- ceramics community , 109
- ceramics community , 212
- ceramics community for 61
- ceramics companies . 53
- ceramics companies consultants 173
- ceramics company ! 4432
- ceramics company , 133
- serve as the incoming 92
- serve as the incubator 99
- serve as the independent 794
- serve as the index 223
- serve as the indication 72
- serve as the indicator 120
- serve as the indicators 45
- serve as the indispensable 111
- serve as the indispensable 40
- serve as the individual 234
- serve as the industrial 52
- serve as the industry 607
- serve as the info 42
- serve as the informal 102
- serve as the information 838
- serve as the informational 41
- serve as the infrastructure 500
- serve as the initial 5331
- serve as the initiating 125
- serve as the initiator 63
- serve as the initiator 81
- serve as the injector 56
- serve as the inlet 41
- serve as the inner 87
- serve as the input 1323

## Levels of text representations

- Character
- Words
- Phrases
- Part-of-speech tags**
- Taxonomies / thesauri
- Vector-space model

---

- Language models
- Full-parsing
- Cross-modality

---

- Collaborative tagging / Web2.0
- Templates / Frames
- Ontologies / First order theories





## Part-of-Speech level

- By introducing part-of-speech tags we introduce word-types enabling to differentiate words functions
  - For text-analysis part-of-speech information is used mainly for "information extraction" where we are interested in e.g. named entities which are "noun phrases"
  - Another possible use is reduction of the vocabulary (features)
    - ...it is known that nouns carry most of the information in text documents
- Part-of-Speech taggers are usually learned by HMM algorithm on manually tagged data

## Part-of-Speech Table

part of speech	function or "job"	example words	example sentences
<b>Verb</b>	action or state	(to) be, have, do, like, work, sing, can, must	EnglishClub.com <b>is</b> a web site. I <b>like</b> EnglishClub.com.
<b>Noun</b>	thing or person	pen, dog, work, music, town, London, teacher, John	This is my <b>dog</b> . He lives in my <b>town</b> . We live in <b>London</b> . <b>John</b>
<b>Adjective</b>	describes a noun	a/an, the, 69, some, good, big, red, well, interesting	My dog is <b>big</b> . I like <b>big</b> dogs.
<b>Adverb</b>	describes a verb, adjective or adverb	quickly, silently, well, badly, very, really	My dog eats <b>quickly</b> . When he is <b>very</b> hungry, he eats <b>really</b> quickly.
<b>Pronoun</b>	replaces a noun	I, you, he, she, some	Tara is Indian. <b>She</b> is beautiful.
<b>Preposition</b>	links a noun to another word	to, at, after, on, but	We went <b>to</b> school <b>on</b> Monday.
<b>Conjunction</b>	joins clauses or sentences or words	and, but, when	I like dogs <b>and</b> I like cats. I like cats <b>and</b> dogs. I like dogs <b>but</b> I don't like cats.
<b>Interjection</b>	short exclamation, sometimes inserted into a sentence	oh!, ouch!, hi, well	<b>Ouch!</b> That hurts! <b>Hi!</b> How are you? <b>Well</b> , I don't know.

[http://www.englishclub.com/grammar/parts-of-speech\\_1.htm](http://www.englishclub.com/grammar/parts-of-speech_1.htm)

## Part-of-Speech examples

verb	noun	verb
Stop!	John	works.

noun	verb	verb
John	is	working.

pronoun	verb	noun
She	loves	animals.

noun	verb	adjective	noun
Animals	like	kind	people.

noun	verb	noun	adverb
Tara	speaks	English	well.

noun	verb	adjective	noun
Tara	speaks	good	English.

pronoun	verb	preposition	adjective	noun	adverb
She	ran	to	the	station	quickly.

pron.	verb	adj.	noun	conjunction	pron.	verb	pron.
She	likes	big	snakes	but	I	hate	them.

Here is a sentence that contains every part of speech:

interjection	pron.	conj.	adj.	noun	verb	prep.	noun	adverb
Well,	she	and	young	John	walk	to	school	slowly.

[http://www.englishclub.com/grammar/parts-of-speech\\_2.htm](http://www.englishclub.com/grammar/parts-of-speech_2.htm)

## Levels of text representations

- Character
- Words
- Phrases
- Part-of-speech tags
- **Taxonomies / thesauri**
- **Vector-space model**
- Language models
- Full-parsing
- Cross-modality
- Collaborative tagging / Web2.0
- Templates / Frames
- Ontologies / First order theories

## Taxonomies/thesaurus level

- Thesaurus has a main function to connect different surface word forms with the same meaning into one sense (synonyms)
  - ...additionally we often use hypernym relation to relate general-to-specific word senses
  - ...by using synonyms and hypernym relation we compact the feature vectors
- The most commonly used general thesaurus is WordNet which exists in many languages (e.g. EuroWordNet)
  - <http://www.illc.uva.nl/EuroWordNet/>

## WordNet – database of lexical relations

- WordNet is the most well developed and widely used lexical database for English
  - ...it consist from 4 databases (nouns, verbs, adjectives, and adverbs)
- Each database consists from sense entries – each sense consists from a set of synonyms, e.g.:
  - musician, instrumentalist, player
  - person, individual, someone
  - life form, organism, being

Category	Unique Forms	Number of Senses
Noun	94474	116317
Verb	10319	22066
Adjective	20170	29881
Adverb	4546	5677

## WordNet relations

- Each WordNet entry is connected with other entries in the graph through relations
- Relations in the database of nouns:

Relation	Definition	Example
Hypernym	From lower to higher concepts	breakfast -> meal
Hyponym	From concepts to subordinates	meal -> lunch
Has-Member	From groups to their members	faculty -> professor
Member-Of	From members to their groups	copilot -> crew
Has-Part	From wholes to parts	table -> leg
Part-Of	From parts to wholes	course -> meal
Antonym	Opposites	leader -> follower

## Levels of text representations

- Character
- Words
- Phrases
- Part-of-speech tags
- Taxonomies / thesauri
- **Vector-space model**
- Language models
- Full-parsing
- Cross-modality
- Collaborative tagging / Web2.0
- Templates / Frames
- Ontologies / First order theories

## Vector-space model level

- The most common way to deal with documents is first to transform them into **sparse numeric vectors** and then deal with them with **linear algebra operations**
  - ...by this, we forget everything about the linguistic structure within the text
  - ...this is sometimes called “structural curse” because this way of forgetting about the structure doesn’t harm efficiency of solving many relevant problems
  - This representation is referred to also as “Bag-Of-Words” or “Vector-Space-Model”
  - Typical tasks on vector-space-model are classification, clustering, visualization etc.

## Representing documents as vectors

Having a set of documents, represent each as a feature vector:

1. divide text into units (eg., words), remove punctuation, (remove stop-words, stemming,...)
2. each unit becomes a feature having numeric weight as its value (eg., number of occurrences in the text - referred to as term frequency or TF)

Commonly used weight is TFIDF:

$$TFIDF(w) = tf(w) * \log\left(\frac{N}{df(w)}\right)$$

- $tf(w)$  – term frequency (no. of occurrences of word  $w$  in document)
- $df(w)$  – document frequency (no. of documents containing word  $w$ )
- $N$  – no. of all documents

## Example of document representation

Bob the builder is a children animated movie on a character Bob and his friends that include several vehicle characters. They face challenges and jointly solve them, such as, repair a roof or save Bob's cat from a tall tree...

Pixar has several short animated movies suitable for children. Locomotion is one of them showing train engine and a train wagon as two characters that face a challenge of crossing a half-broken bridge...

...

Simpson family provokes a smile on many adult and children faces showing everyday life of a family of four...

	bob	builder	children	animated	movie	character	friend	vehicle	...	...
3	1	1	1	1	1	2	1	1	...	...
0	0	1	1	1	1	1	0	0	...	...
...	...	...	...	...	...	...	...	...	...	...
0	0	1	0	0	0	0	0	0	...	...

## Similarity between document vectors

- Each document is represented as a vector of weights  $D = \langle x \rangle$
- Cosine similarity (dot product) is the most widely used similarity measure between two document vectors
  - ...calculates cosine of the angle between document vectors
  - ...efficient to calculate (sum of products of intersecting words)
  - ...similarity value between 0 (different) and 1 (the same)

$$Sim(D_1, D_2) = \frac{\sum_i x_{1i} \cdot x_{2i}}{\sqrt{\sum_j x_j^2} \sqrt{\sum_k x_k^2}}$$

## Levels of text representations

- Character
- Words
- Phrases
- Part-of-speech tags
- Taxonomies / thesauri
- Vector-space model
- **Language models**
- Full-parsing
- Cross-modality
- Collaborative tagging / Web2.0
- Templates / Frames
- Ontologies / First order theories

Lexical

Syntactic

Semantic

## Language model level

- Language modeling is about determining probability of a sequence of words
  - The task typically gets reduced to the estimating probabilities of a next word given two previous words (trigram model):

$$P(w_i | w_{i-2} w_{i-1}) \approx \frac{C(w_{i-2} w_{i-1} w_i)}{C(w_{i-2} w_{i-1})}$$

← Frequencies of word sequences

- It has many applications including speech recognition, OCR, handwriting recognition, machine translation and spelling correction

## Levels of text representations

- Character
- Words
- Phrases
- Part-of-speech tags
- Taxonomies / thesauri
- Vector-space model
- Language models
- **Full-parsing**
- Cross-modality
- Collaborative tagging / Web2.0
- Templates / Frames
- Ontologies / First order theories

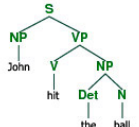
Lexical

Syntactic

Semantic

## Full-parsing level

- Parsing provides maximum structural information per sentence
- On the input we get a sentence, on the output we generate a parse tree
- For most of the methods dealing with the text data the information in parse trees is too complex



## Levels of text representations

- Character
- Words
- Phrases
- Part-of-speech tags
- Taxonomies / thesauri
- **Vector-space model**
- Language models
- Full-parsing
- **Cross-modality**
- Collaborative tagging / Web2.0
- Templates / Frames
- Ontologies / First order theories

Lexical

Synactic

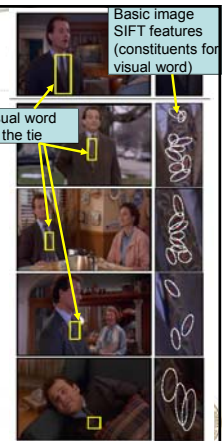
Semantic

## Cross-modality level

- It is very often the case that objects are represented with different data types:
  - Text documents
  - Multilingual texts documents
  - Images
  - Video
  - Social networks
  - Sensor networks
- ...the question is how to create mappings between different representation so that we can benefit using more information about the same objects

## Example: Aligning text with audio, images and video

- The word "tie" has several representations (<http://www.answers.com/tie&r=67>)
  - Textual
  - Multilingual text
    - (tie, kravata, krawatte, ...)
  - Audio
  - Image:
    - <http://images.google.com/images?hl=en&q=necktie>
  - Video (movie on the right)
- Out of each representation we can get set of features and the idea is to correlate them
  - KCCA (Kernel Correlation Analysis) method generates mappings between different representations into "modality neutral" data representation



## Levels of text representations

- Character
- Words
- Phrases
- Part-of-speech tags
- Taxonomies / thesauri
- **Vector-space model**
- Language models
- Full-parsing
- Cross-modality
- **Collaborative tagging / Web2.0**
- Templates / Frames
- Ontologies / First order theories

Lexical


Synactic

Semantic

## Collaborative tagging

- Collaborative tagging is a process of adding metadata to annotate content (e.g. documents, web sites, photos)
  - ...metadata is typically in the form of keywords
  - ...this is done in a collaborative way by many users from larger community collectively having good coverage of many topics
  - ...as a result we get annotated data where tags enable comparability of annotated data entries

### Example: flickr.com tagging



Tags entered by users annotating photos

### Example: del.icio.us tagging



Tags entered by users annotating Web sites

### Levels of text representations

- Character
- Words
- Phrases
- Part-of-speech tags
- Taxonomies / thesauri
- Vector-space model

---

- Language models
- Full-parsing
- Cross-modality

---

- Collaborative tagging / Web2.0
- **Templates / Frames**
- Ontologies / First order theories

**Lexical**

**Syntactic**

**Semantic**

### Template / frames level

- Templates are the mechanism for extracting the information from text
  - ...templates always focused on specific domain which includes consistent patterns on where specific information is positioned
  - Templates are one of the basic methods for information extraction

### Examples of templates of KnowItAll system

- Generic approach of extracting is described in
  - *Unsupervised named-entity extraction from the Web: An experimental study* [Oren Etzioni et al]
- KnowItAll system uses the following generic templates:
  - NP "and other" <class1>
  - NP "or other" <class1>
  - <class1> "especially" NPList
  - <class1> "including" NPList
  - <class1> "such as" NPList
  - "such" <class1> "as" NPList
  - NP "is a" <class1>
  - NP "is the" <class1>
- ...each template represents specific relationship between the words appearing in the variable slots
- From template patterns KnowItAll bootstraps new templates

### Levels of text representations

- Character
- Words
- Phrases
- Part-of-speech tags
- Taxonomies / thesauri
- Vector-space model

---

- Language models
- Full-parsing
- Cross-modality

---

- Collaborative tagging / Web2.0
- Templates / Frames
- **Ontologies / First order theories**

**Lexical**

**Syntactic**

**Semantic**

## Ontologies level

- Ontologies are the most general formalism for describing data objects
  - ...in the recent years ontologies got popular through Semantic Web and OWL standard
  - Ontologies can be of various complexity – from relatively simple ones (light weight described with simple relations) to heavy weight (described with first order theories).
  - Ontologies could be understood also as very generic data-models where we can store extracted information from text

## Example: text represented in the First Order Logic

### General Knowledge about Terrorism:

```
Terrorist groups are capable of directing assassinations:
(implies
  (isa ?GROUP TerroristGroup)
  (behaviorCapable ?GROUP AssassinatingSomeone directingAgent))
....
If a terrorist group considers an agent an enemy, that agent is vulnerable to an attack by that group:
(implies
  (and
    (isa ?GROUP TerroristGroup)
    (considersAsEnemy ?GROUP ?TARGET))
  (vulnerableTo ?GROUP ?TARGET TerroristAttack))
```

General Knowledge about Terrorism

Specific data, facts, and observations about terrorist groups and activities

MEDNARODNA PODIPLomsKA ŠOLA JOŽEFA STEFANA  
JOŽEF STEFAN INTERNATIONAL POSTGRADUATE SCHOOL

## Example tasks addressed in text mining

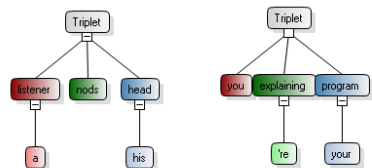
MEDNARODNA PODIPLomsKA ŠOLA JOŽEFA STEFANA  
JOŽEF STEFAN INTERNATIONAL POSTGRADUATE SCHOOL

## Extracting triplets from text

RUSU, Delia, FORTUNA, Blaž, GROBELNIK, Marko, MLADENIĆ, Dunja. Semantic graphs derived from triplets with application in document summarization. *Informatica (Ljublj.)*, 2009, vol. 33, no. 3, pp. 357-362.

## Task

- Extract (*subject, predicat, object*) triplets from text
- Example:
  - If a **listener** **nods** his **head** while **you're** **explaining** your **program**; wake him up.

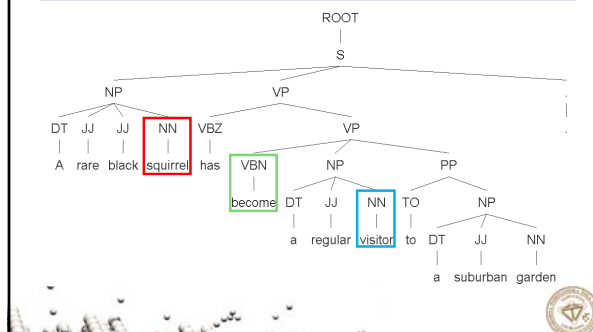


## Extraction of triplets using parsers

- Approach description:
  - Parse the sentence with a deep parser
  - Determine subject, object and predicate from the parse tree
- Advantage:
  - Many freely available parsers
- Disadvantage:
  - Solves much harder problem (deep parsing) in order to extract triplets

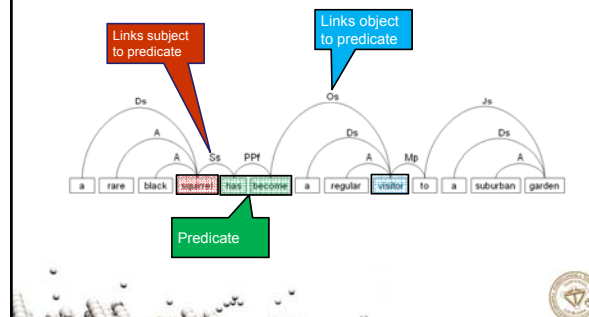
## Using OpenNLP

A rare black squirrel has become a regular visitor to a suburban garden.



## Using Linked Parser

A rare black squirrel has become a regular visitor to a suburban garden.



## Machine learning approach

- Triplet extraction can be defined as a binary classification problem
  - Set of tree words from a sentence can be positive (an actual triplet) or negative (not a triplet).
  - Classification algorithms, such as SVM, can be naturally applied to this task

## Document Summarization

## Document Summarization

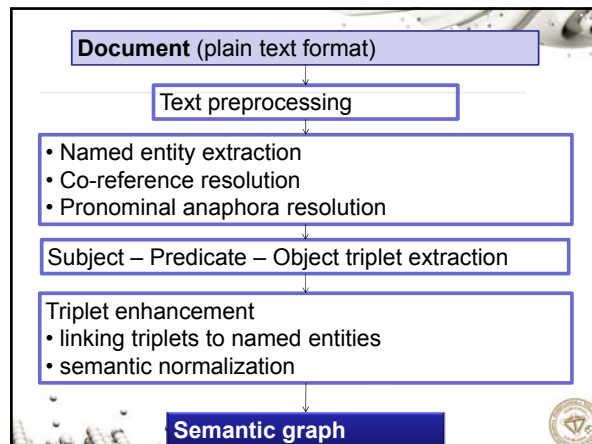
- **Task:** the task is to produce shorter, summary version of an original document
- Two main approaches to the problem:
  - **Selection based** – summary is selection of sentences from an original document
  - **Knowledge rich** – performing semantic analysis, representing the meaning and generating the text satisfying length restriction

## Selection based summarization

- Three main phases:
  - Analyzing the source text
  - Determining its important points (units)
  - Synthesizing an appropriate output
- Most methods adopt linear weighting model – each text unit (sentence) is assessed by the following formula:
  - $\text{Weight}(U) = \text{LocationInText}(U) + \text{CuePhrase}(U) + \text{Statistics}(U) + \text{AdditionalPresence}(U)$
  - ...lot of heuristics and tuning of parameters (also with Machine learning)
- ...output consists from topmost text units (sentences)

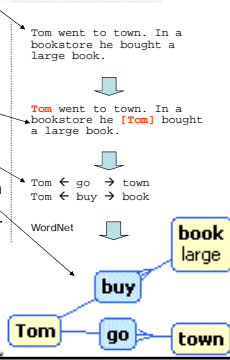
## Knowledge rich summarization

- To generate 'true' summary of a document we need to (at least partially) 'understand' the document text
  - ...the document is too small to count on statistics, we need to identify and use its linguistic and semantic structure
- On the next slides we show an approach from [Leskovec, Grobelnik, Milic-Frayling 2004] using 10 step procedure for extracting semantics from a document:
  - ...the approach was evaluated on "Document Understanding Conference" test set of documents and their summaries
  - ...the approach extracts semantic network from a document and tries to extract relevant part of the semantic network to represent summary
  - Results achieved 70% recall of and 25% precision on extracted Subject-Predicate-Object triples



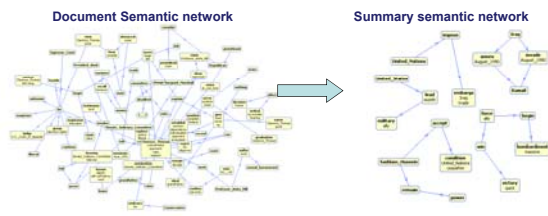
## Knowledge Rich Summarization Example

- Input document is split into sentences
- Each sentence is deep-parsed
- Name-entities are disambiguated:
  - Determining that 'Barac Obama' == 'Obama' == 'U.S. president'
- Performing Anaphora resolution:
  - Pronouns are connected with named-entities
- Extracting of **Subject-Predicate-Object** triples
- Constructing a **graph** from triples
- Each triple in the graph is described with features for learning
- Using machine learning train a model for classification of triples into the summary
- Generate a summary graph from selected triples
- From the summary graph generate textual summary document



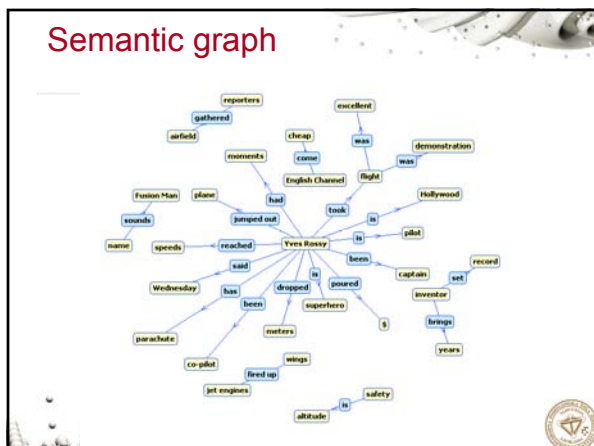
## Training of summarization model


- A model was trained deciding which **Subject-Predicate-Object** triple belongs into the target summary
- For training was used Support Vector Machine (SVM) on 400 statistic, linguistic and graph topological features



Leskovec, Grobelnik, Milic-Frayling, LinkKDD 2004 (Learning Sub-structures of Document Semantic Graphs for Document Summarization)  
 Rusu, Fortuna, Grobelnik, Mladenic, Informatica 2009 (Semantic Graphs Derived From Triples With Application in Document Summarization)

## Semantic graph





MEDNARODNA  
 PODIPLOMSKA ŠOLA  
 JOŽEFA STEFANA

JOŽEF STEFAN  
 INTERNATIONAL  
 POSTGRADUATE SCHOOL

## Question Answering

- DALI, Lorand, RUSU, Delia, FORTUNA, Blaž, MLADENIĆ, Dunja, GROBELNIK, Marko. *Question answering based on semantic graphs. WWW-2009 Workshop on Semantic Search 2009.*
- BRADEŠKO, Luka, DALI, Lorand, FORTUNA, Blaž, GROBELNIK, Marko, MLADENIĆ, Dunja, NOVALIJA, Inna, PAJNTAR, Boštjan. *Contextualized question answering, ITI-2010.*

### Question Answering

answer Art

where do tigers live  Ask

We found that

tigers	live	the following
Siberian tigers	surviving	world
tigers	live	Sumatra

Related documents

**world** CHINA: FEATURE - Tigers must earn their meat in China. With only about 300 Siberian tigers surviving in the world, and only 20 in the wild in China, that help must come soon, said Liu.

**Sumatra** INDONESIA: FEATURE - Chinese medicine threatens Sumatran tiger. Subijanto, a spokesman for the Forestry Ministry, said Indonesia was committed to protecting the tigers, which live within Sumatra's four designated conservation areas.

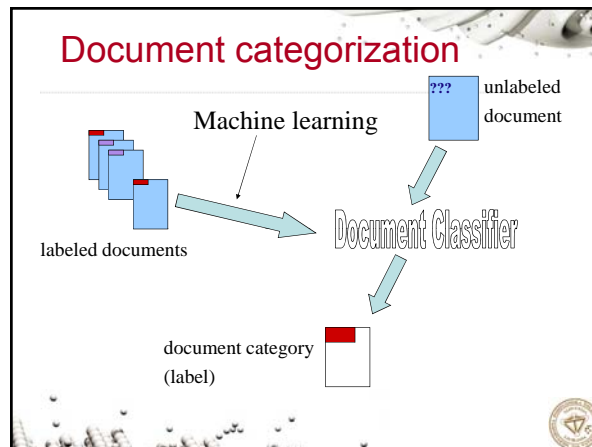
<http://answerart.net/>

MEDNARODNA PODIPLOMSKA ŠOLA JOŽEFA STEFANA JOŽEF STEFAN INTERNATIONAL POSTGRADUATE SCHOOL

## Supervised Learning

### Document Categorization Task

- **Given:** set of documents labeled with content categories
- **The goal:** to build a model which would automatically assign right content categories to new unlabeled documents.
- Content categories can be:
  - unstructured (e.g., Reuters) or
  - structured (e.g., Yahoo, DMOz, Medline)



### Measuring success - Model quality estimation

$Precision(M, targetC) = P(\overline{targetC} | targetC)$  ← The truth, and

$Recall(M, targetC) = P(targetC | \overline{targetC})$  ← ..the whole truth

$Accuracy(M) = \sum_i P(\overline{C}_i) \times Precision(M, C_i)$

$F_\beta(M, targetC) = \frac{(1 + \beta^2) Precision(M, targetC) \times Recall(M, targetC)}{\beta^2 Precision(M, targetC) + Recall(M, targetC)}$

- Classification accuracy
- Break-even point (precision=recall)
- F-measure (precision, recall = sensitivity)

©Dunja Mladenic

### Algorithms for learning document classifiers

- Popular algorithms for text categorization:
  - Support Vector Machines
  - Logistic Regression
  - Perceptron algorithm
  - Naive Bayesian classifier
  - Winnow algorithm
  - Nearest Neighbour
  - ....
- Unlike decision tree and rule learning algorithms, these are mainly non-symbolic learning algorithms

©Dunja Mladenic

### Example learning algorithm: Perceptron

**Input:**

- set of documents  $D$  in the form of (e.g. TFIDF) numeric vectors
- each document has label +1 (positive class) or -1 (negative class)

**Output:**

- linear model  $w_i$  (one weight per word from the vocabulary)

**Algorithm:**

- **Initialize** the model  $w_i$  by setting word weights to 0
- **Iterate** through documents  $N$  times
  - For document  $d$  from  $D$ 
    - // Using current model  $w_i$  classify the document  $d$
    - if  $\text{sum}(d_i * w_i) \geq 0$  then classify document as positive
    - else classify document as negative
    - if document classification is wrong then
      - // adjust weights of all words occurring in the document
      - $w_{w_i} = w_i + \text{sign}(\text{true-class}) * \text{Beta}$  (input parameter  $\text{Beta} > 0$ )
      - // where  $\text{sign}(\text{positive}) = 1$  and  $\text{sign}(\text{negative}) = -1$

### Categorization to flat categories

Example data set used in research:

- Documents are classified by editors into one or more categories
- Publicly available set of Reuter news mainly from 1987:
  - 120 categories giving the document content, such as: *earn, acquire, corn, rice, jobs, oilseeds, gold, coffee, housing, income,...*
- Larger dataset available for research from 2000 having 830,000 Reuters news documents

### Distribution of documents (Reuters-21578)

Top 20 categories of Reuter news in 1987-91

Unbalanced distribution

### Example of Perceptron model for Reuters category "Acquisition"

Feature	Positive Class Weight
STAKE	11.5
MERGER	9.5
TAKEOVER	9
ACQUIRE	9
ACQUIRED	8
COMPLETES	7.5
OWNERSHIP	7.5
SALE	7.5
OWNERSHIP	7.5
BUYOUT	7
ACQUISITION	6.5
UNDISCLOSED	6.5
BUYS	6.5
ASSETS	6
BID	6
BP	6
DIVISION	5.5

### SVM, Perceptron & Winnow text categorization performance on Reuters-21578 with different representations

Comparison of algorithms

Legend: SVM (blue), Perceptron (maroon), Winnow (yellow)

### Document categorization with only few labeled documents

- we have many documents but only some of them are labeled
- we may have a human available for a limited time to provide labels of documents

Approaches:

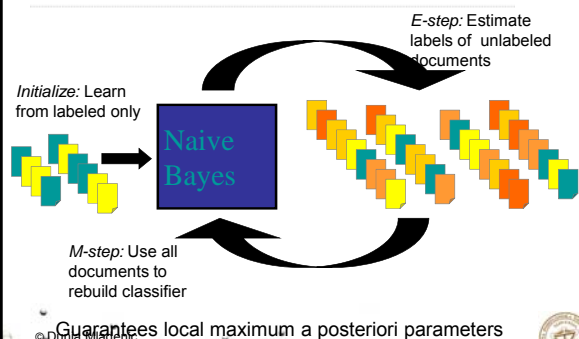
- Using unlabeled data
- Co-training
- Active learning

### Using unlabeled data [Nigam et al., 2000]

- Given: a small number of labeled examples and a large pool of unlabeled examples, no human available
  - e.g., classifying news article as interesting or not interesting
- Approach description (EM + Naive Bayes):
  - train a classifier with only labeled documents,
  - assign probabilistically-weighted class labels to unlabeled documents,
  - train a new classifier using all the documents
  - iterate until the classifier remains unchanged

©Dunja Mladenic

### Using Unlabeled Data with Expectation-Maximization (EM)



©Dunja Mladenic

### Co-training [Blum & Mitchell, 1998]

Theory behind co-training

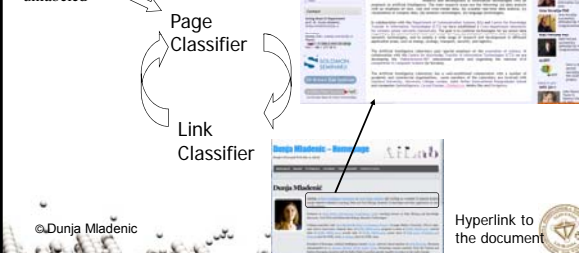
- Possible to learn from unlabeled examples
- Value of unlabeled data depends on
  - How (conditionally) independent are the two representations of the same data
    - The more the better
  - The number of redundant inputs (features)
    - Expected error decreases exponentially with this number
- Disagreement on unlabeled data predicts true error

Better performance on labelling unlabeled data compared to EM approach

©Dunja Mladenic

### Bootstrap Learning to Classify Web Pages

Given: set of documents where each document is described by two independent sets of features (e.g. document text + hyperlinks anchor text)  
few labeled and many unlabeled



©Dunja Mladenic

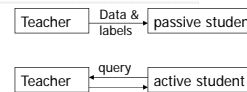


MEDNARODNA PODIPLOMSKA ŠOLA JOŽEFA STEFANA  
JOŽEF STEFAN INTERNATIONAL POSTGRADUATE SCHOOL

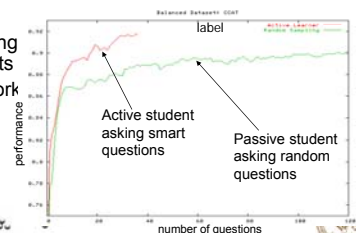
## Active Learning

### Active Learning

- We use this methods whenever hand-labeled data are rare or expensive to obtain
- Interactive method



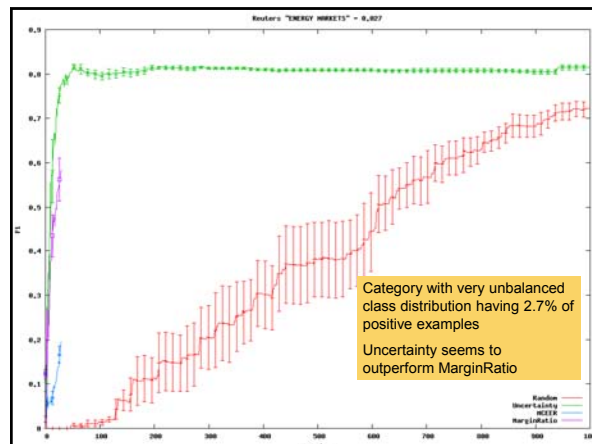
- Requests only labeling of "interesting" objects
- Much less human work needed for the same result compared to arbitrary labeling examples



©Dunja Mladenic

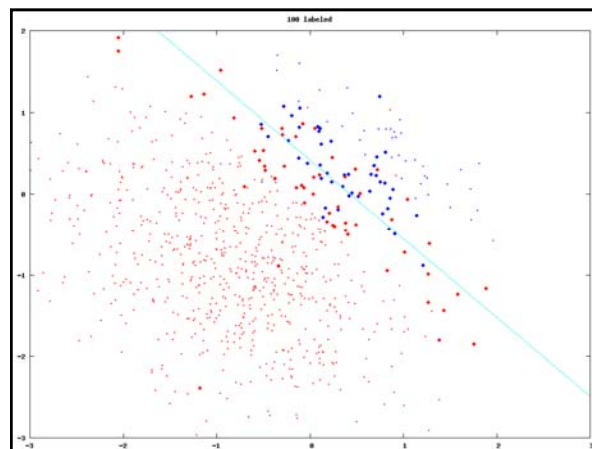
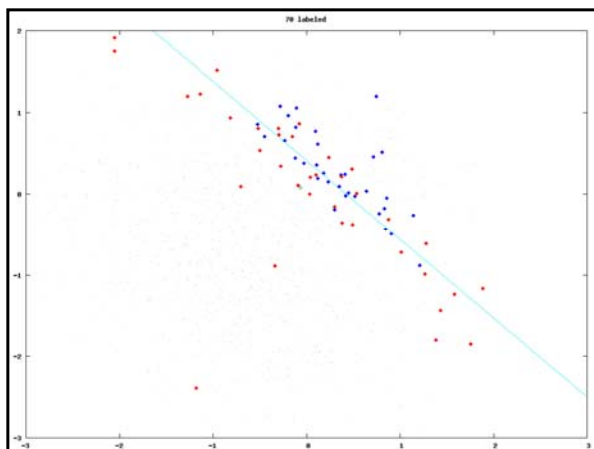
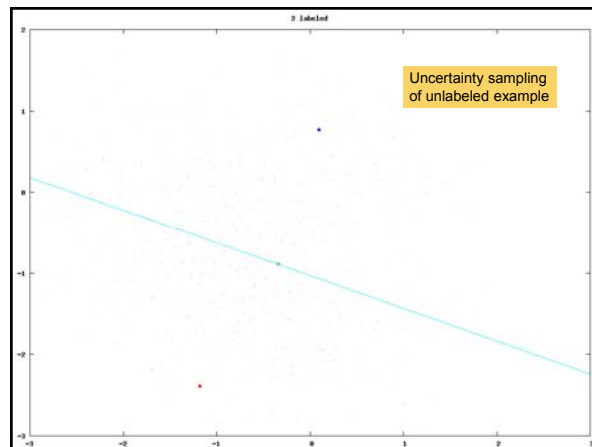
### Some approaches to Active Learning

- **Uncertainty sampling** (efficient)
    - select example closest to the decision hyperplane (or the one with classification probability closest to  $P=0.5$ ) [Tong & Koller 2000]
  - **Maximum margin ratio change**
    - select example with the largest predicted impact on the margin size if selected [Tong & Koller 2000]
  - **Monte Carlo Estimation of Error Reduction**
    - select example that reinforces our current beliefs [Roy & McCallum 2001]
  - **Random sampling** as baseline
- Experimental evaluation (using F1-measure) of the four listed approaches shown on three categories from Reuters-2000 dataset [Novak & Mladenic & Grobelnik, 2006]
- average over 10 random samples of 5000 training (out of 500k) and 10k testing (out of 300k) examples
  - two of the methods a rather time consuming, thus we run them for including the first 50 unlabeled examples
  - experiments show that active learning is especially useful for unbalanced data
- ©Dunja Mladenic



### Illustration of Active learning

- starting with one labeled example from each class (red and blue)
  - select one example for labeling (green circle)
  - request label and add re-generate the model using the extended labeled data
- Illustration of linear SVM model using
- arbitrary selection of unlabeled examples (random)
  - active learning selecting the most uncertain examples (closest to the decision hyperplane)
- ©Dunja Mladenic





MEDNARODNA  
PODIPLOMSKA ŠOLA  
JOŽEFA STEFANA

JOŽEF STEFAN  
INTERNATIONAL  
POSTGRADUATE SCHOOL

## Unsupervised Learning

www.risi.si

## Document clustering

- Given is a set of documents
- The goal is: to cluster the documents into several groups based on some similarity measure
  - documents inside the group should be similar while documents between the groups should be different

Similarity measure plays a crucial role in clustering, on documents we use cosine similarity:

$$\text{Cos}(d_1, d_2) = \frac{d_1 \cdot d_2}{\|d_1\| \|d_2\|} = \frac{\sum_i x_{1i} x_{2i}}{\sqrt{\sum_j x_j^2} \sqrt{\sum_k x_k^2}}$$

©Dunja Mladenic

## Clustering methods

- Hierarchical
  - agglomerative – at each step merge two or more groups
  - divisive – at each step break the selected group into two or more groups
- Non hierarchical
  - requires specification of the number of clusters
  - optimization of the initial clustering (e.g., maximize similarity of examples inside the same group)
- Geometrical
  - map multidimensional space into two- or three-dimensional (e.g., principal component analysis)
- Graph-theoretical

©Dunja Mladenic

## K-Means clustering algorithm

- **Given:**
  - set of examples (e.g., TFIDF vectors of documents),
  - distance measure (e.g., cosine)
  - **K** (number of groups)
- **For each of K** groups initialize its centroid with a random document
- **While** not converging
  - Each document is assigned to the nearest group (represented by its centroid)
  - For each group calculate new centroid (group mass point, average document in the group)

©Dunja Mladenic

## Web-log analysis

- Customer profiling is the most important application of Web-Mining:
  - ...the goal is to better understand our web customer behavior in order to optimize our e-services
  - The problem is how to get quality data about our users
  - ...even bigger problem is how to analyze the data...

©Dunja Mladenic

## Main source of the data:

### Log files

- Main source of the data about the activity of our web server are Log files
- Typical line of a Log file:
  - 2001-05-29 04:13:40 128.2.215.4 - W3SVC1 ASPIRE 194.249.231.167 80 GET /KddGarden/Grouper/Grouper.zi p - 206 64 1507568 551 1815813 HTTP/1.1 aspi re. i j s. si Mozi I I a/4.0+(compatibl e; +MSI E+5.5; +WI nd ows+NT+5.0) - http://aspi re. i j s. si /KddGarden/Grouper/
- E.g. Log files on WinNT/2000 reside at the \winnt\system32\logfiles\ system directory

©Dunja Mladenic

## Customer identification

- The most common way for identifying of the customers are:
  - **Cookies** – the information saved by a foreign web-server at the users local disk usually when first time using the web service
  - **Username and password** (explicit identification) – information input by the user at each e-service usage
- ...web customer identification could not be solved optimally (for all situations)

## Additional customer information

- What else do we know about the web customer/user?
  - The URL of the web page from which our user came to our web server written in the *Referrer* field in the Log file
  - The sequence of URLs or web services visited by our user (click-stream data) based on the *Referrer* field or *Session-Id*
  - How much time the user spent at the web page
  - The contents of the web page read by the user (text)
  - ...from additional sources we know the history of the users in the form of the past actions (purchases, visits, habits)
  - ...sometimes we have some demographical data etc.
- All the available information is hard to use in analysis

## Data analysis methods

- Log files include sequences of events (click-streams):
  - ...methods for analyzing *event sequences* are usually modified classical methods from the area of Data-Mining for analysis of very large databases
  - Basic methods are modified methods for induction of association rules, clustering, decision trees
- Other analytic methods are from the areas of Text-Mining, Statistics and Machine-Learning
- ...not enough time for details...

## What kind of problems do we solve?

- Personalization of web services:
  - Preparing offers (discounts, products, contents) customized for each particular user
- Understanding of what is going on at the web server:
  - Customer groups identification, behavioral patterns
  - ...the goal is to better organize web services
- Better "Banner Adds" selection to increase the probability to be clicked by the user
  - ...it is not hard to increase the probability for several 100%
- Building the psychological profiles based on the texts read by the user
  - ...to get more info about the user than he has about himself©

## Association rules in Web-logs

- Searching for rules that connect two or more events:
  - 60% of the users that visited **URL/company/product**, also visited **company/product/product1.html**
  - 30% of the users that visited **URL/company/special-offer/** also visited **company/product2.html**

## Profiling using time dimension

- Searching for rules that connect two or more events taking into account time dimension:
  - 30% of the users that visited **URL/company/product/product1.html** also searched in the last week words **W1** and **W2** on Yahoo
  - 60% of the users that ordered **product1** in the next 15 days also ordered **product2**

## Classification rules

- Identification of behavior for groups of users - additional information can be obtained from cookies, registration, etc.:
  - Users that frequently visit page **/company/products/product3.html** are from educational institutions
  - 50% of the users that visited **/company/products/product4.html** are in age group of 20-25 and live at the sea coast

## Real-Time Data-Analysis

- At some web servers there are too many hits to be saved and analyzed off-line:
  - ... we have a data stream – no time or space for off-line data analysis (e.g. search engines, shops, banks, news, ...)
  - ... we would like to understand what is going on to detect e.g. anomalies or changes in trends
- The solution is in using special type of methods for online event analysis:
  - Methods are able to analyze non-stationary data
  - At each moment results (models) are in human readable form (e.g. decision trees, rules, ...)
  - ...no need to save Log files


**MEDNARODNA  
PODIPLOMSKA ŠOLA  
JOŽEFA STEFANA**
**JOŽEF STEFAN  
INTERNATIONAL  
POSTGRADUATE SCHOOL**

## References

© Dunja Mladenec

## References to some of the Books



© Dunja Mladenec

## Requirements for this class

- Seminar as independent work following the provided instructions <http://blazfortuna.com/> under Teaching
- Report on the results of the seminar work to be sent via e-mail by 8.4.2012 to [Blaz.Fortuna@ijs.si](mailto:Blaz.Fortuna@ijs.si)
  - 5-10 page report per groups
- Presentation of the seminar results on 18.4.2012 15:00-19:00 at MPŠ
  - 5-10 slides presentation per group
- Oral exam