DEPARTMENT OF KNOWLEDGE TECHNOLOGIES

Jožef Stefan Institute

Text Mining for Knowledge Management

Dunja Mladenić, Blaž Fortuna

http://kt.ijs.si

Overview

- Example application using Text Mining to analyze collaboration and build competence map
- Levels of text representation
- Text Mining Algorithms
- References
- Ontology construction using OntoGen

Analysis of IST EU Projects database

SolEuNet

data mining & decision support

KNOWLEDGE TECHNOLOGIES

Jožef Stefan Institute

IST data description

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_PROJl_search.html)

IST 5FP has 2786 projects in which participate 7886 organizations

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

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

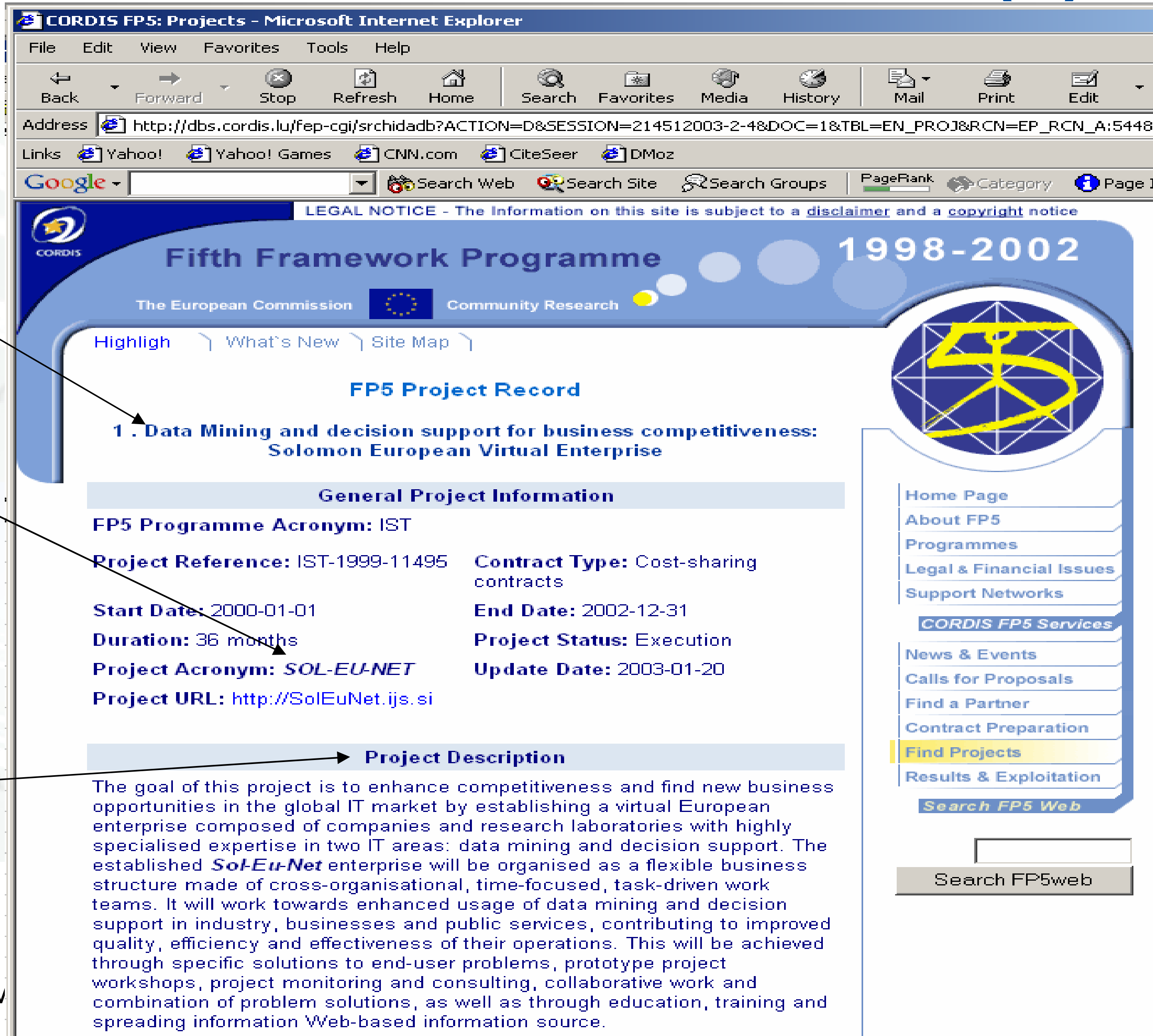
	A	B	C	D	E	F	G	H	I	J
1	Project Ref	Acronym	Domain / K	Unit	PO	Legal Name	Legal Country	Type of organisation	Participant role	
2	IST-1999-11495	SOL-EU-NET	KA2	C2	HANSEN RALF	ALARIX, D.O.O.	SLOVENIA	Private non research org.	CR	
3	IST-1999-11495	SOL-EU-NET	KA2	C2	HANSEN RALF	AUSTRIAN RESEARCH INS	AUSTRIA	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	

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

Project Title

Project Acronym

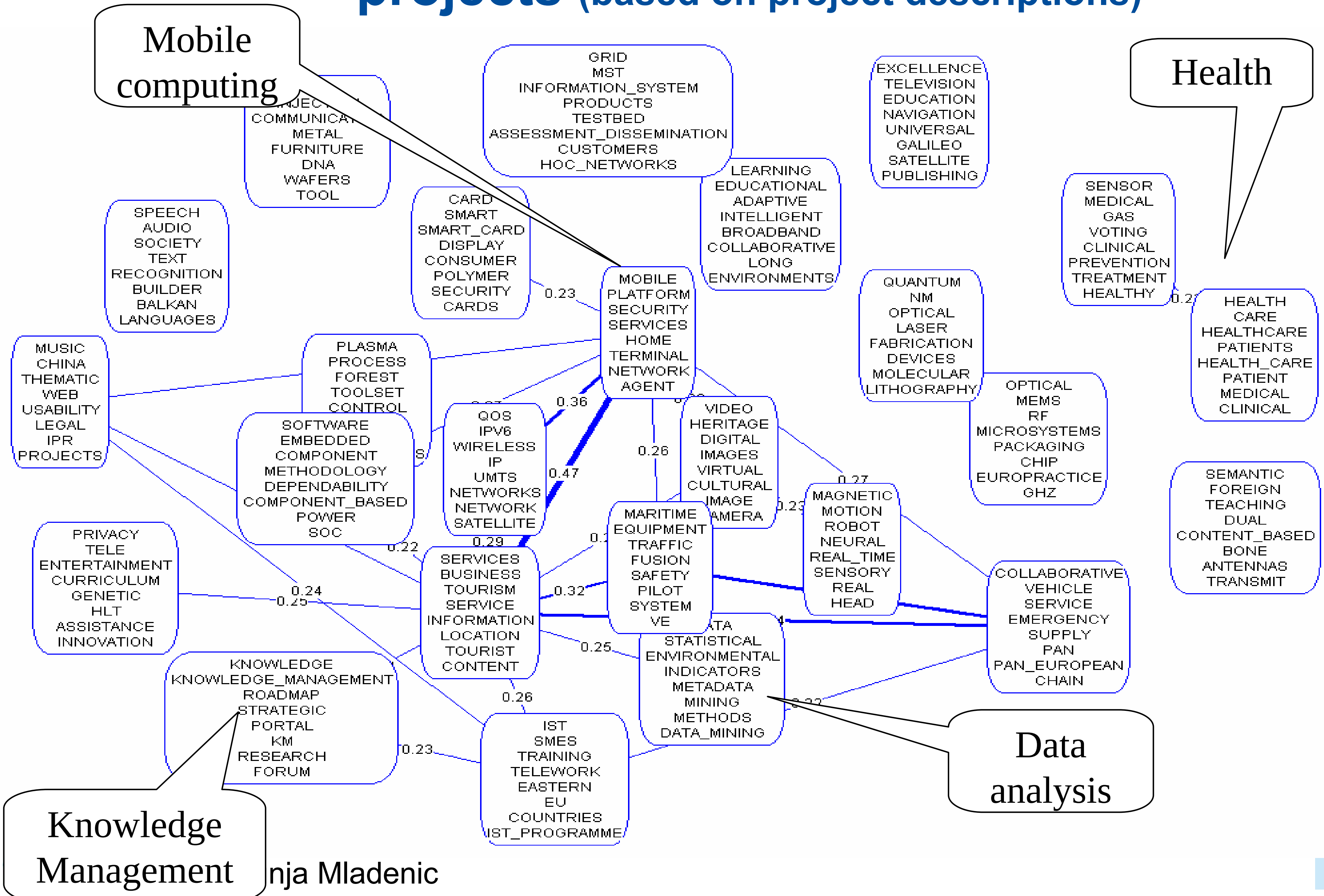
Project Description



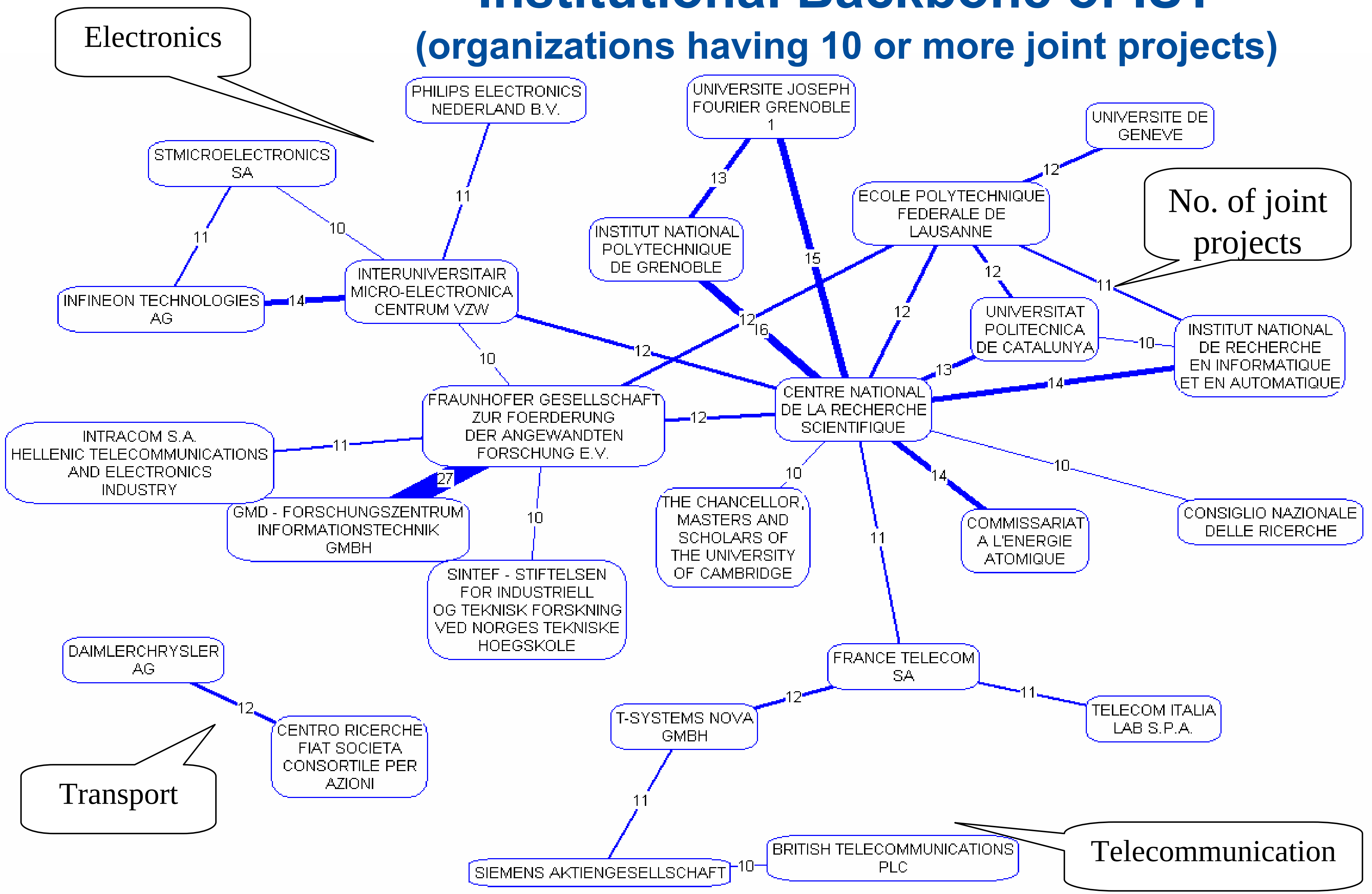
Analysis tasks

- Visualization of project topics
- Analysis of collaboration
- Community/clique identification
- Thematic consortia identification

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



Institutional Backbone of IST
(organizations having 10 or more joint projects)



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
  - …

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

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

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]

Project Intelligence Web site

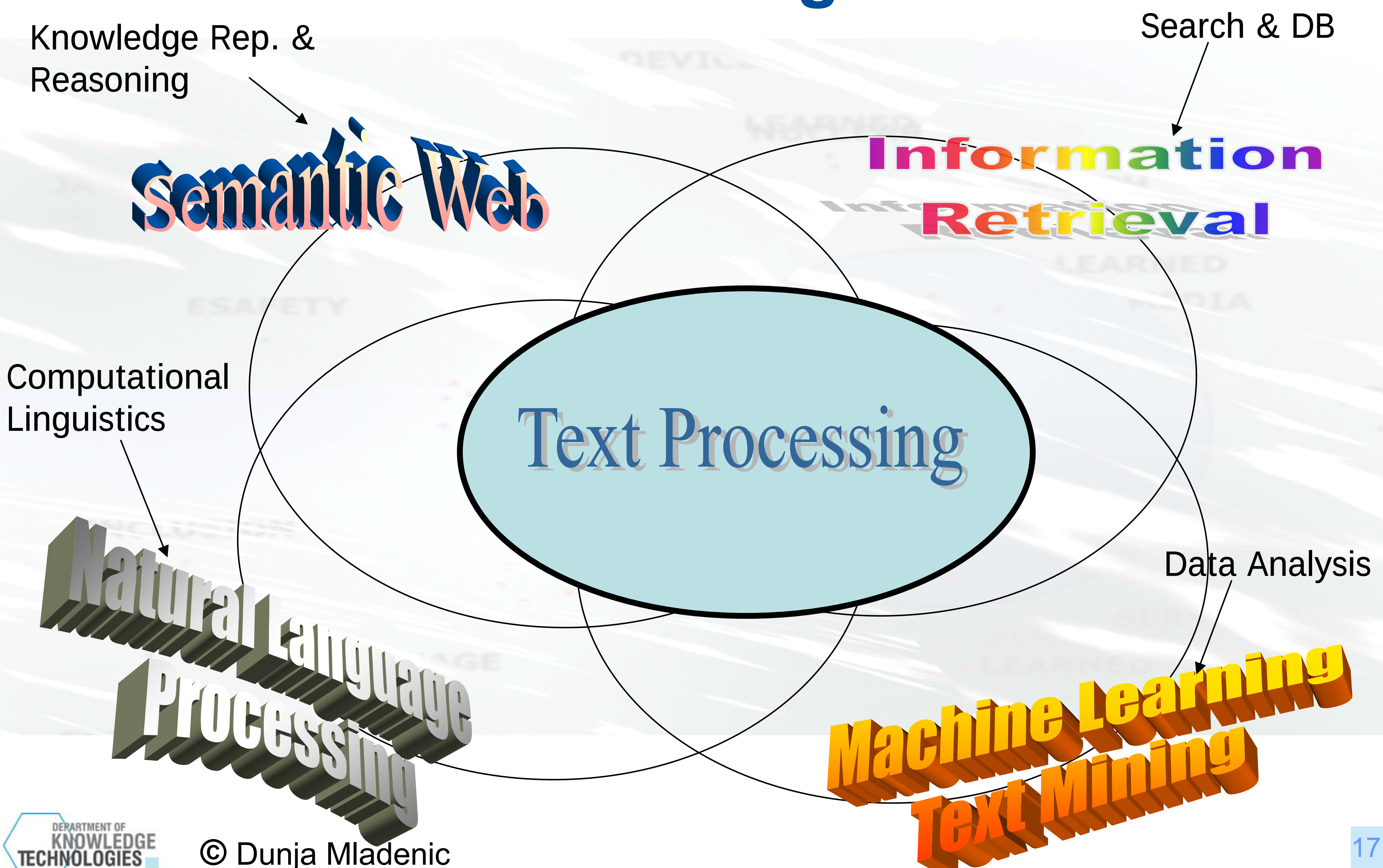
- All demos, reports and results available at the web at http://pi.ijs.si/

Text Mining Techniques

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…"

Which areas are active in Text Processing?



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

Lexical

- Vector-space model
- Language models
- Full-parsing
- Cross-modality

Syntactic

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

Semantic

Levels of text representations

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

Lexical

- Vector-space model
- Language models
- Full-parsing
- Cross-modality

Syntactic

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

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:
  - **Homonomy**: 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 usually problematic for text data analysis
  - because they have different spelling and similar meaning (e.g. learns, learned, learning,…)
  - usually treated as completely unrelated words
- Stemming is a process of transforming a word into its stem
  - cutting off a suffix (eg., smejala -> smej)
- Lemmatization is a process of transforming a word into its normalized form
  - replacing the word, most often replacing a suffix (eg., smejala -> smejati)

Stemming

- For English it is not a big problem - publicly available algorithms give good results
  - Most widely used is Porter stemmer at http://www.tartarus.org/~martin/PorterStemmer/
- In Slovenian language 10-20 different forms correspond to the same word:
  - (“to laugh” in Slovenian): smej, smejal, smejala, smejale, smejali, smejalo, smejati, smejejo, smejeta, smejete, smejeva, smeješ, smejemo, smejiš, smeje, smejoč, smejta, smejte, smejva

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
- Commonly used are two types of phrases:
  - Phrases as contiguous word sequences
  - Phrases as non-contiguous word sequences
  - …both types of phrases could be identified by a simple dynamic programming algorithm
- The main effect of using phrases is to more precisely identify sense

Google n-gram corpus

- In Sep 2006 Google announced availability of n-gram corpus:
  - http://googleresearch.blogspot.com/2006/08/all-our-n-gram-are-belong-to-you.html#links
  - Some statistics of the corpus:
    - File sizes: approx. 24 GB compressed (gzip'ed) text files
    - Number of tokens: 1,024,908,267,229
    - Number of sentences: 95,119,665,584
    - Number of unigrams: 13,588,391
    - Number of bigrams: 314,843,401
    - Number of trigrams: 977,069,902
    - Number of fourgrams: 1,313,818,354
    - Number of fivegrams: 1,176,470,663

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

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 other languages (e.g. EuroWordNet)
  - http://www.illc.uva.nl/EuroWordNet/

WordNet – a 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 consisting 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 a graph through relations.

Relations in the database of nouns:

Relation	Definition	Example
Hypernym	From concepts to subordinate	breakfast -> meal
Hyponym	From concepts to subtypes	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

Lexical

- **Vector-space model**
- Language models
- Full-parsing
- Cross-modality

Syntactic

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

Semantic

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 document as a vector

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

- divide text into units (eg., words), remove punctuation, (remove stop-words, stemming,…)
- 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 dokumentu)
- 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…

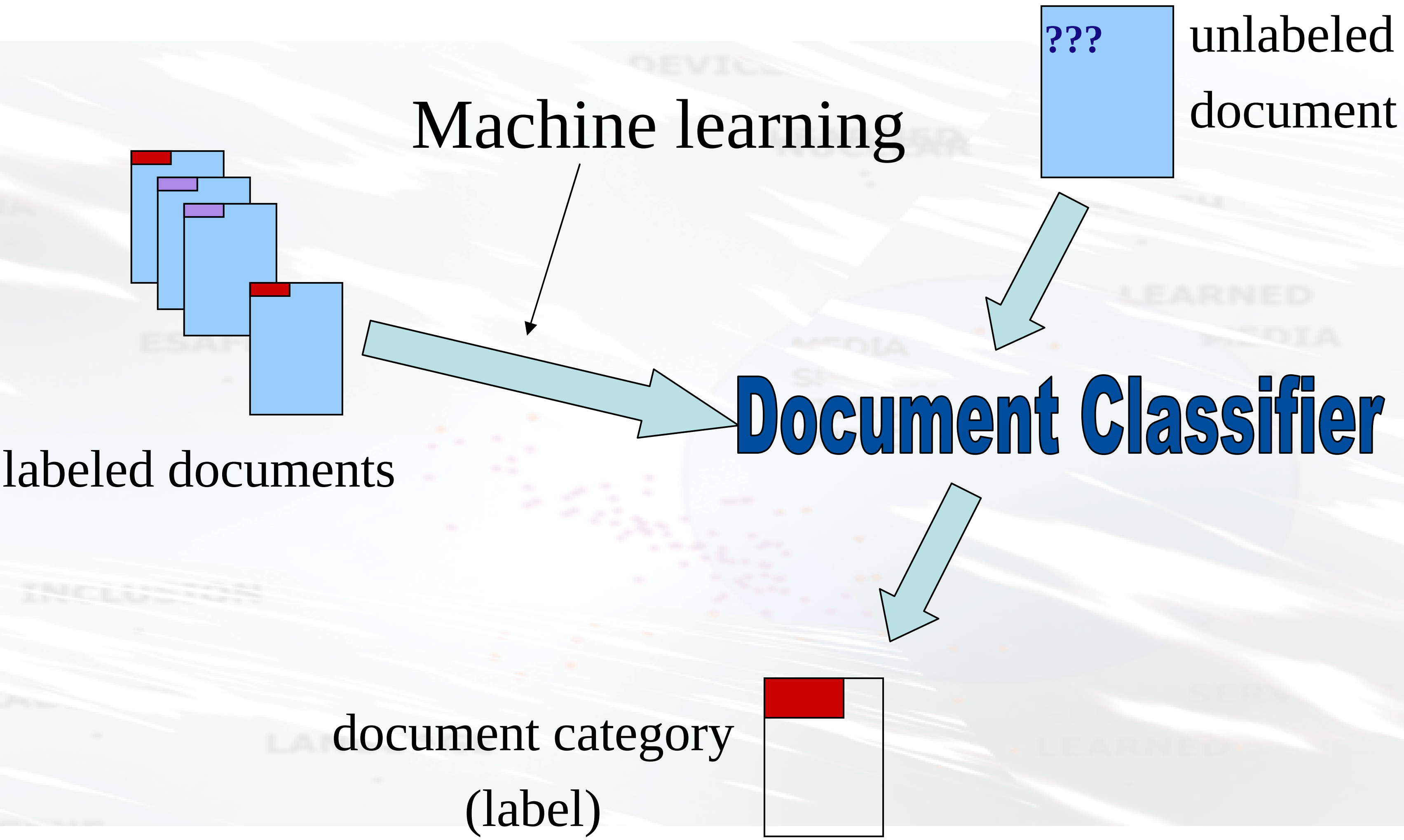
…

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	2	1	1	…	…
0	0	1	1	1	1	0	0	…	…
…	…	…	…	…	…	…	…	…	…
0	0	1	0	0	0	0	0	…	…

Document Categorization

Document categorization



Automatic Document Categorization

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

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

Measuring success - Model quality estimation

$$Precision(M,targetC) = P(targetC \mid \overline{targetC})$$

The truth, and

$$Recall(M,targetC) = P(\overline{targetC} \mid targetC)$$

..the whole truth

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

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

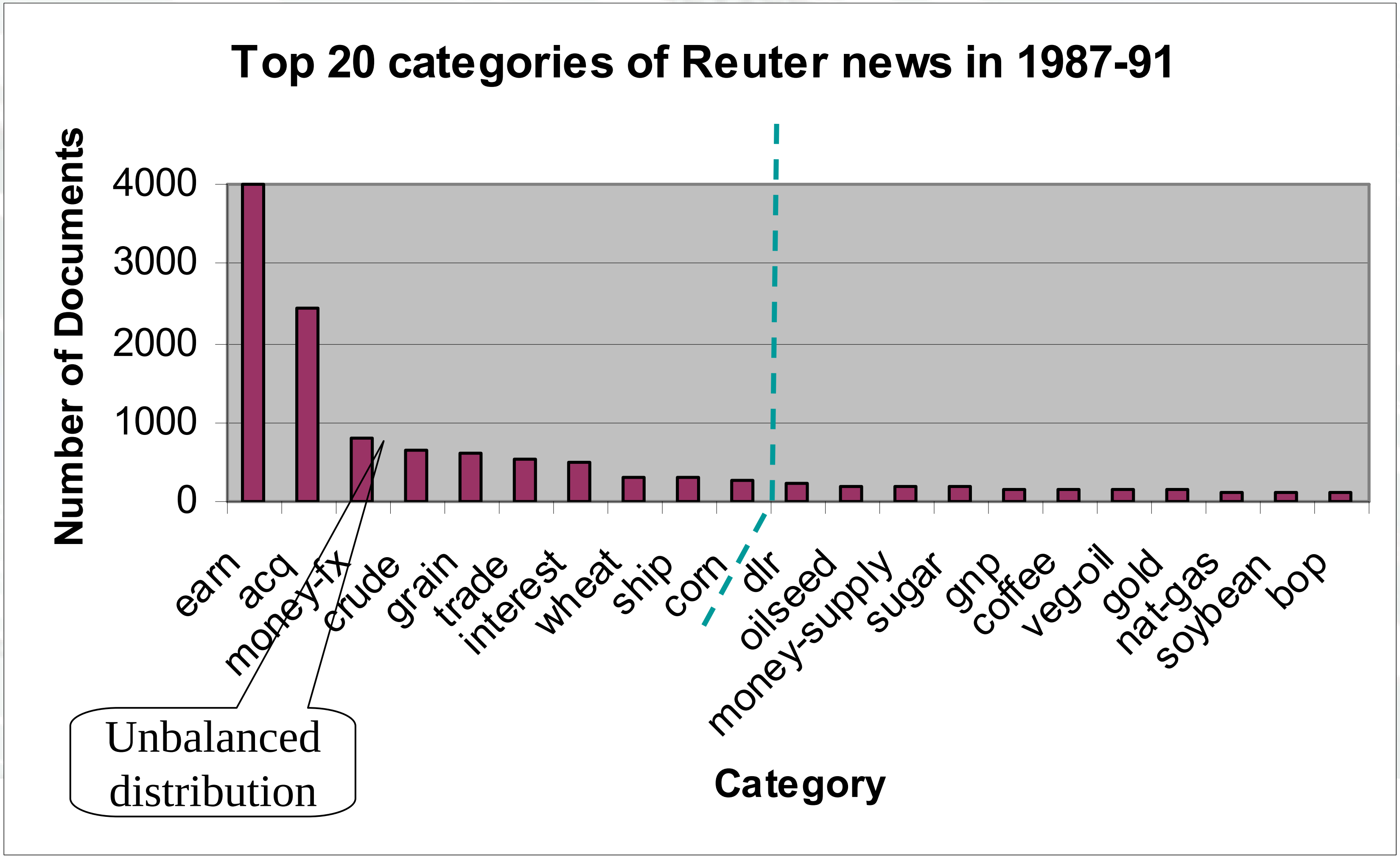
- Classification accuracy
- Break-even point (precision=recall)
- F-measure (precision, recall = sensitivity)

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

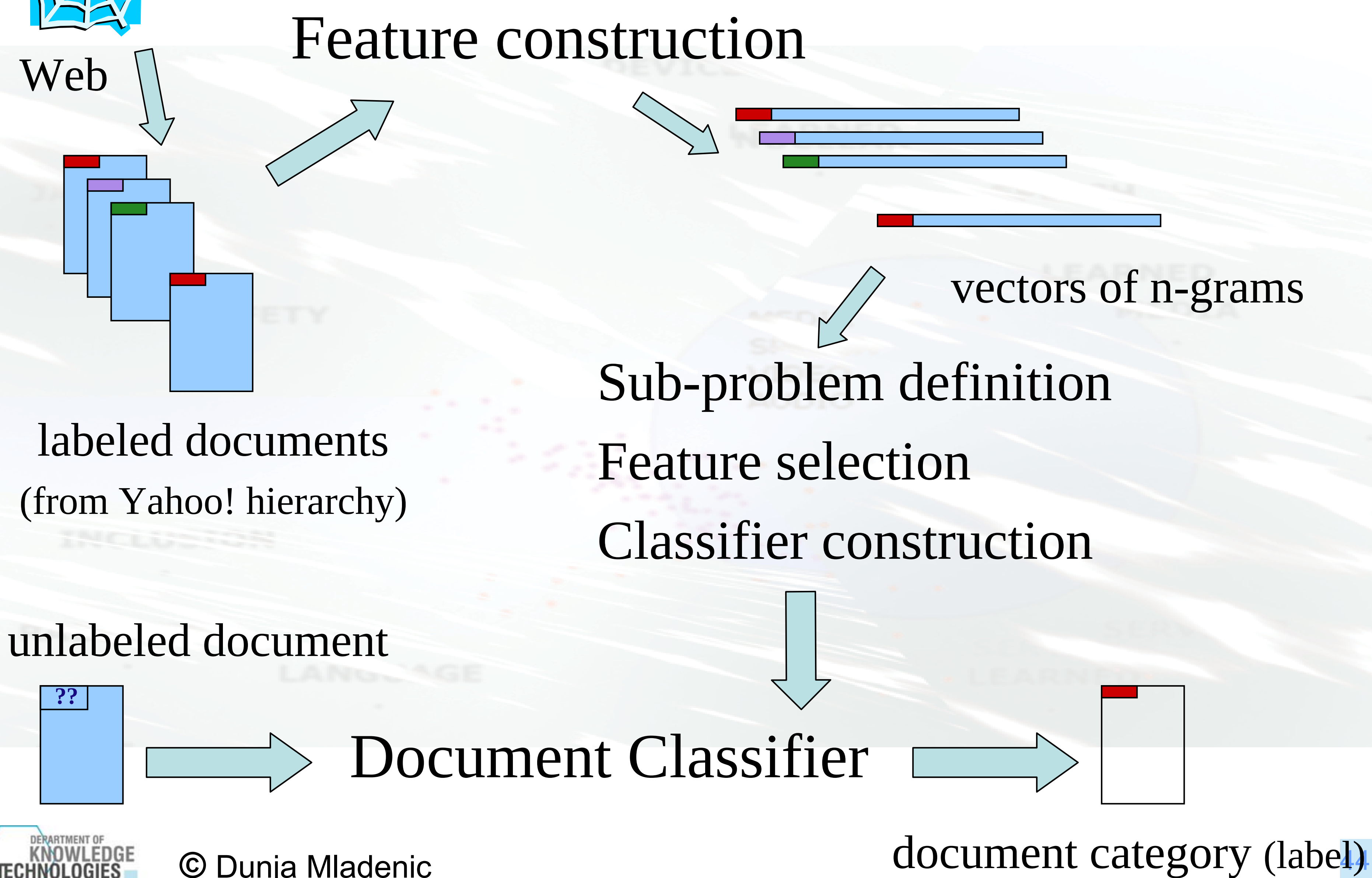
Distribution of documents
(Reuters-21578)



Categorization into hierarchy

- There are several hierarchies (taxonomies) of textual documents:
  - Yahoo, DMoz, Medline, …
- Different people use different approaches:
  - …series of hierarchically organized classifiers
  - …set of independent classifiers just for leaves
  - …set of independent classifiers for all nodes
- Example systems: Yahoo Planet [Mladenic & Grobelnik, 1998], WebClass [Ceci & Malerba, 2003]

Architecture of Yahoo Planet



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

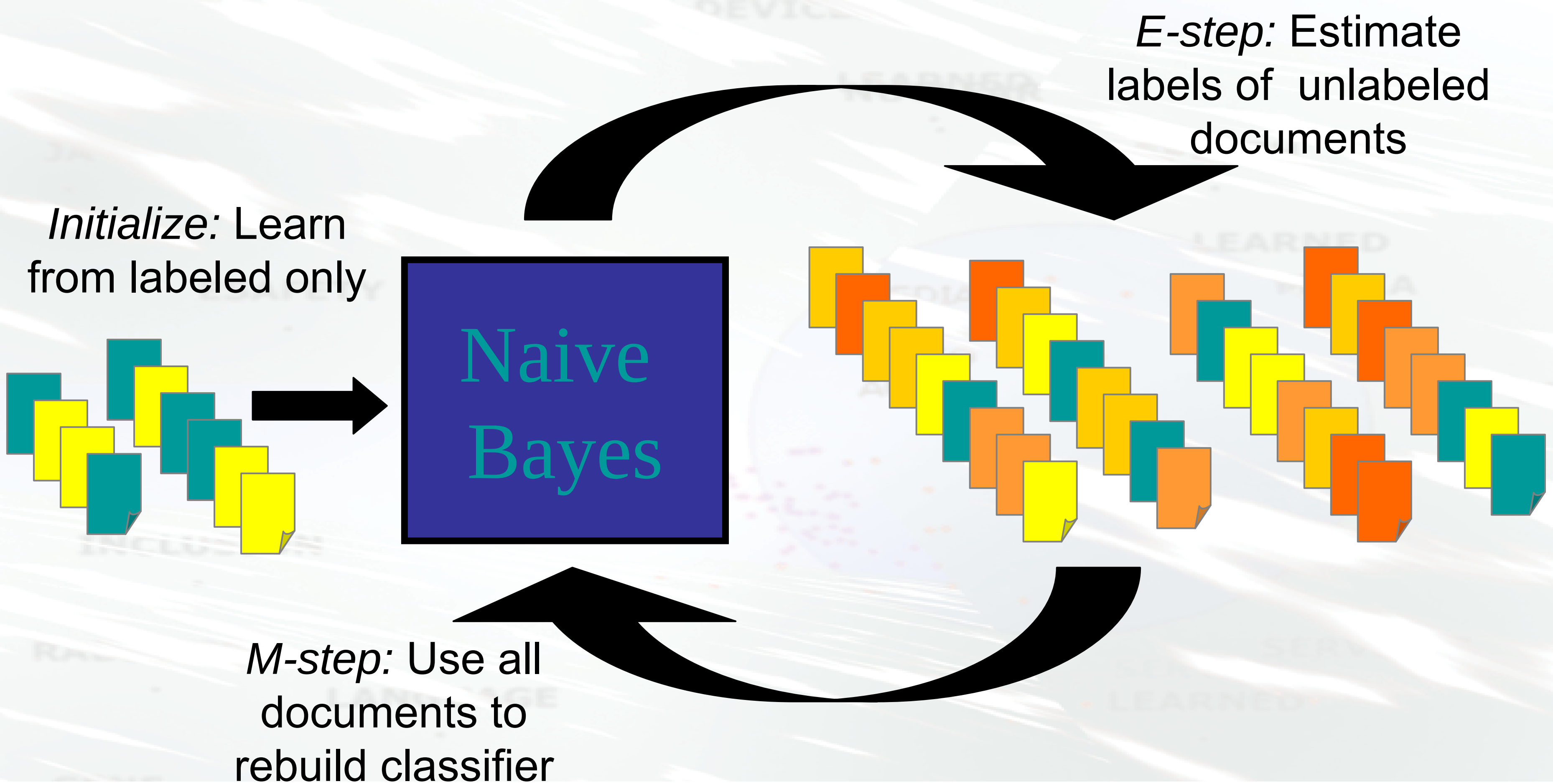
Approaches:

- 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

Using Unlabeled Data with Expectation-Maximization (EM)



Guarantees local maximum a posteriori parameters

Co-training [Blum & Mitchell, 1998]

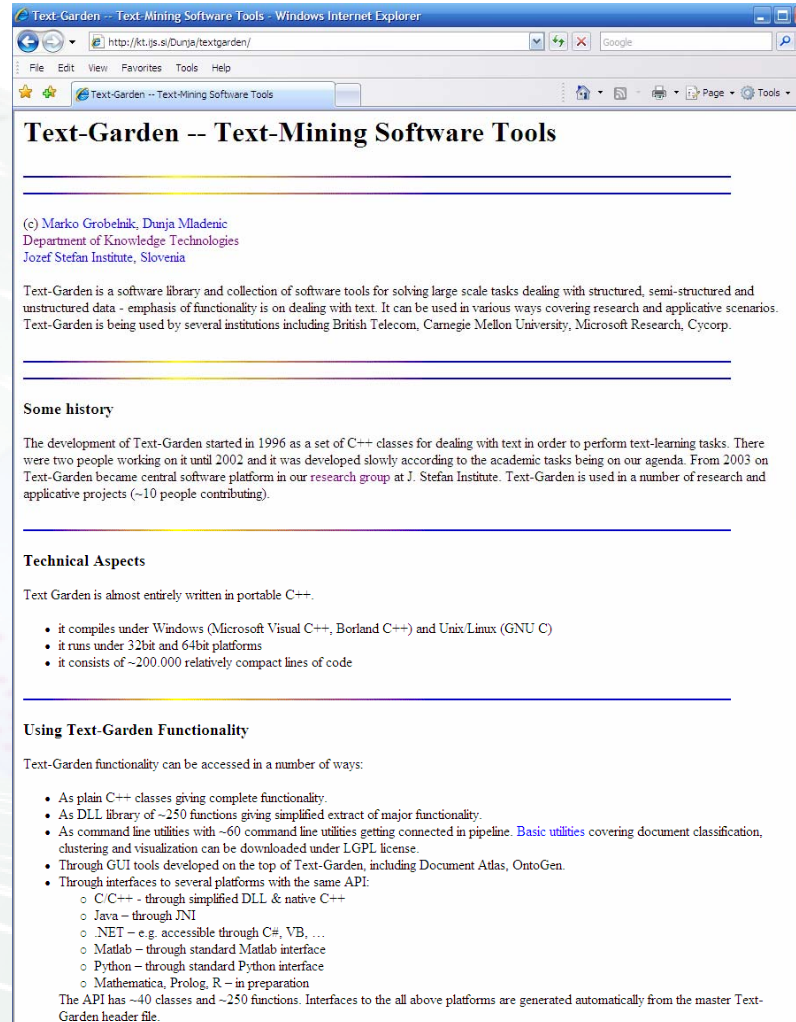
Theory behind co-training

- 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

Bootstrap Learning to Classify Web Pages

Document


Text-Garden -- Text-Mining Software Tools - Windows Internet Explorer

http://kt.ijs.si/Dunja/textgarden/

File Edit View Favorites Tools Help

Text-Garden -- Text-Mining Software Tools

(c) Marko Grobelnik, Dunja Mladenic
Department of Knowledge Technologies
Jozef Stefan Institute, Slovenia

Text-Garden is a software library and collection of software tools for solving large scale tasks dealing with structured, semi-structured and unstructured data - emphasis of functionality is on dealing with text. It can be used in various ways covering research and applicative scenarios. Text-Garden is being used by several institutions including British Telecom, Carnegie Mellon University, Microsoft Research, Cycorp.

Some history

The development of Text-Garden started in 1996 as a set of C++ classes for dealing with text in order to perform text-learning tasks. There were two people working on it until 2002 and it was developed slowly according to the academic tasks being on our agenda. From 2003 on Text-Garden became central software platform in our research group at J. Stefan Institute. Text-Garden is used in a number of research and applicative projects (~10 people contributing).

Technical Aspects

Text Garden is almost entirely written in portable C++.

- it compiles under Windows (Microsoft Visual C++, Borland C++) and Unix/Linux (GNU C)
- it runs under 32bit and 64bit platforms
- it consists of ~200.000 relatively compact lines of code

Using Text-Garden Functionality

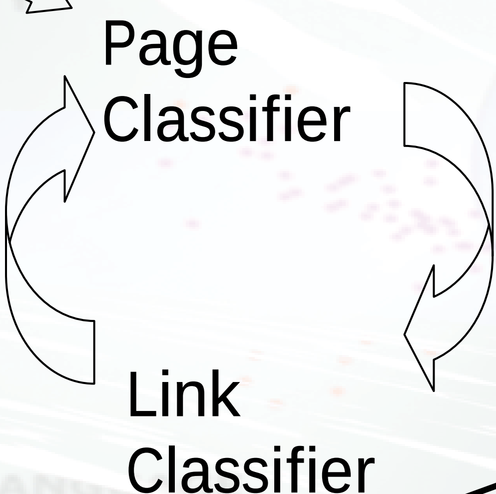
Text-Garden functionality can be accessed in a number of ways:

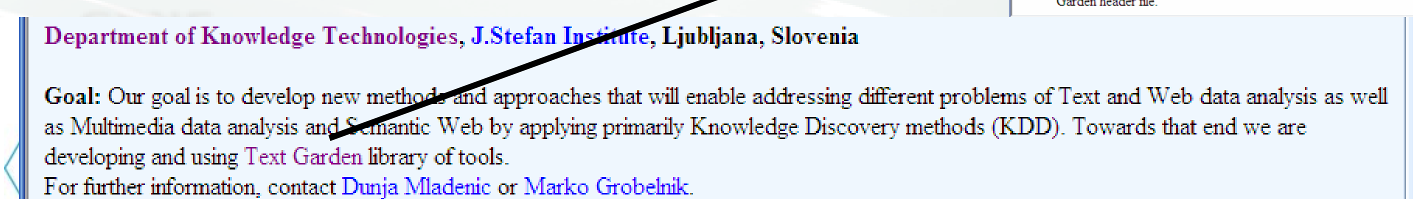
- As plain C++ classes giving complete functionality.
- As DLL library of ~250 functions giving simplified extract of major functionality.
- As command line utilities with ~60 command line utilities getting connected in pipeline. Basic utilities covering document classification, clustering and visualization can be downloaded under LGPL license.
- Through GUI tools developed on the top of Text-Garden, including Document Atlas, OntoGen.
- Through interfaces to several platforms with the same API:
  - C/C++ - through simplified DLL & native C++
  - Java – through JNI
  - .NET – e.g. accessible through C#, VB, …
  - Matlab – through standard Matlab interface
  - Python – through standard Python interface
  - Mathematica, Prolog, R – in preparation

The API has ~40 classes and ~250 functions. Interfaces to the all above platforms are generated automatically from the master Text-Garden header file.

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




Department of Knowledge Technologies, J.Stefan Institute, Ljubljana, Slovenia

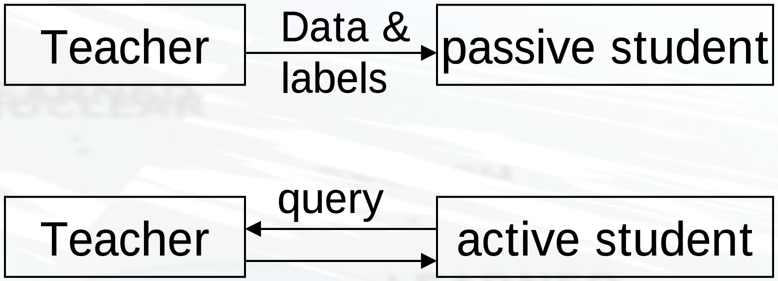
Goal: Our goal is to develop new methods and approaches that will enable addressing different problems of Text and Web data analysis as well as Multimedia data analysis and Semantic Web by applying primarily Knowledge Discovery methods (KDD). Towards that end we are developing and using Text Garden library of tools.
For further information, contact Dunja Mladenic or Marko Grobelnik.

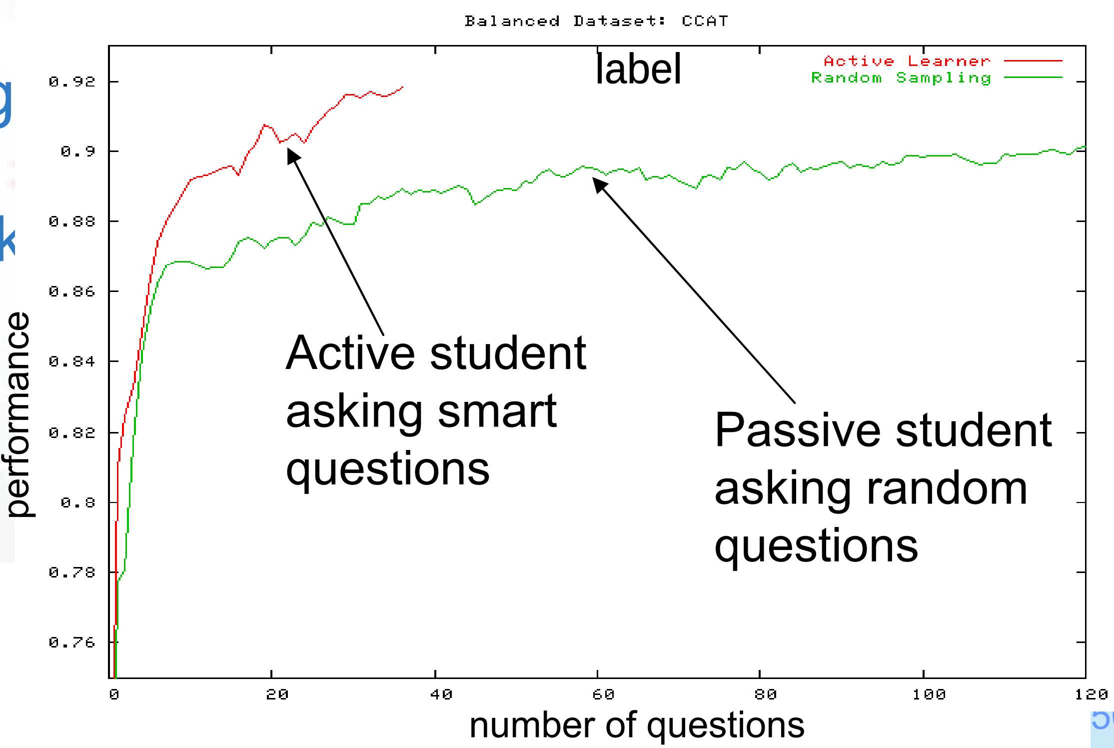
Hyperlink to the document

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

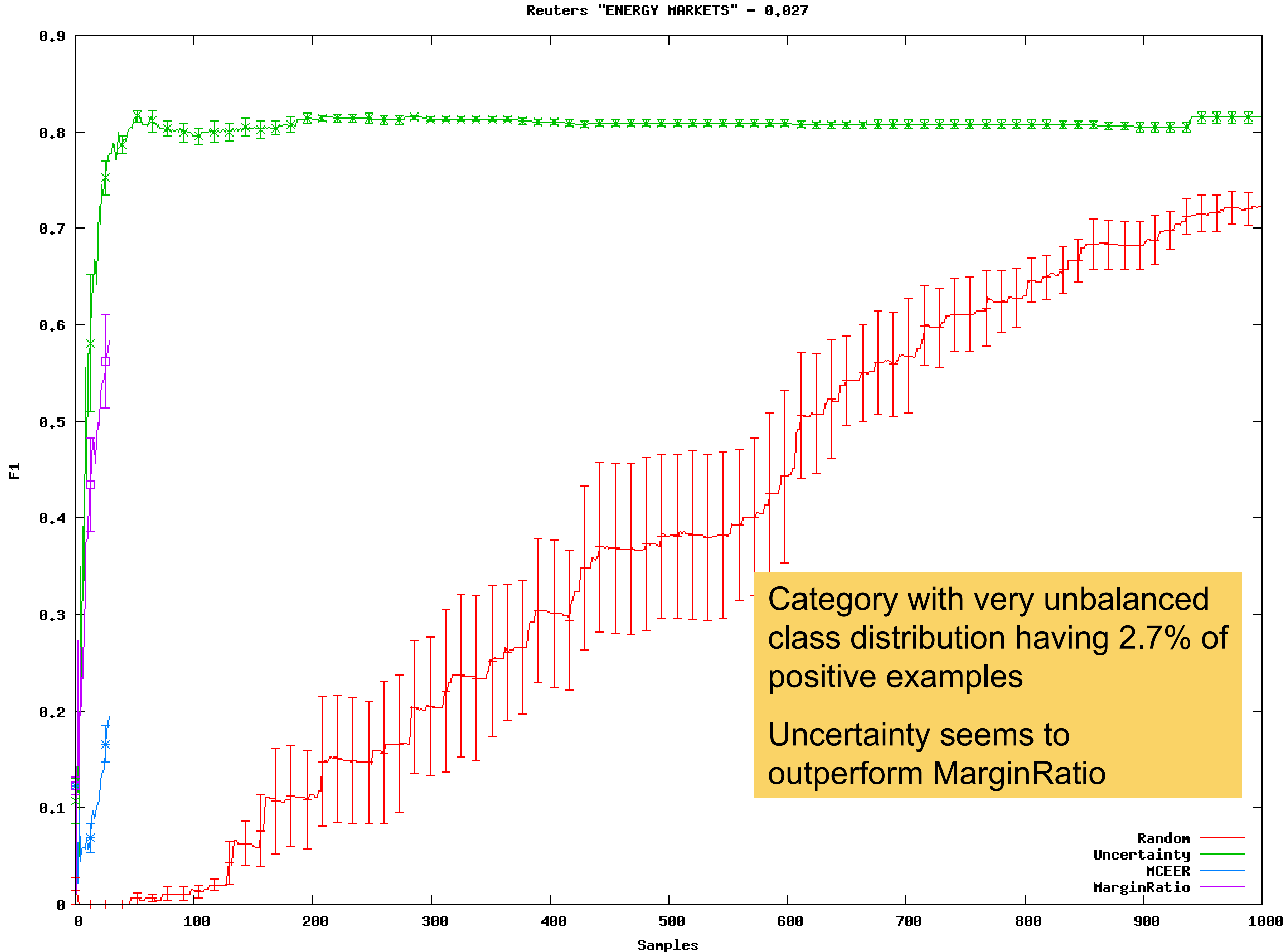




Approaches to Active Learning

- **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


Reuters "ENERGY MARKETS" - 0.027
F1
Samples
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
0
100
200
300
400
500
600
700
800
900
1000
Random
Uncertainty
MCEER
MarginRatio
Category with very unbalanced class distribution having 2.7% of positive examples
Uncertainty seems to outperform MarginRatio

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:

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

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

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)

Example of k-means clustering

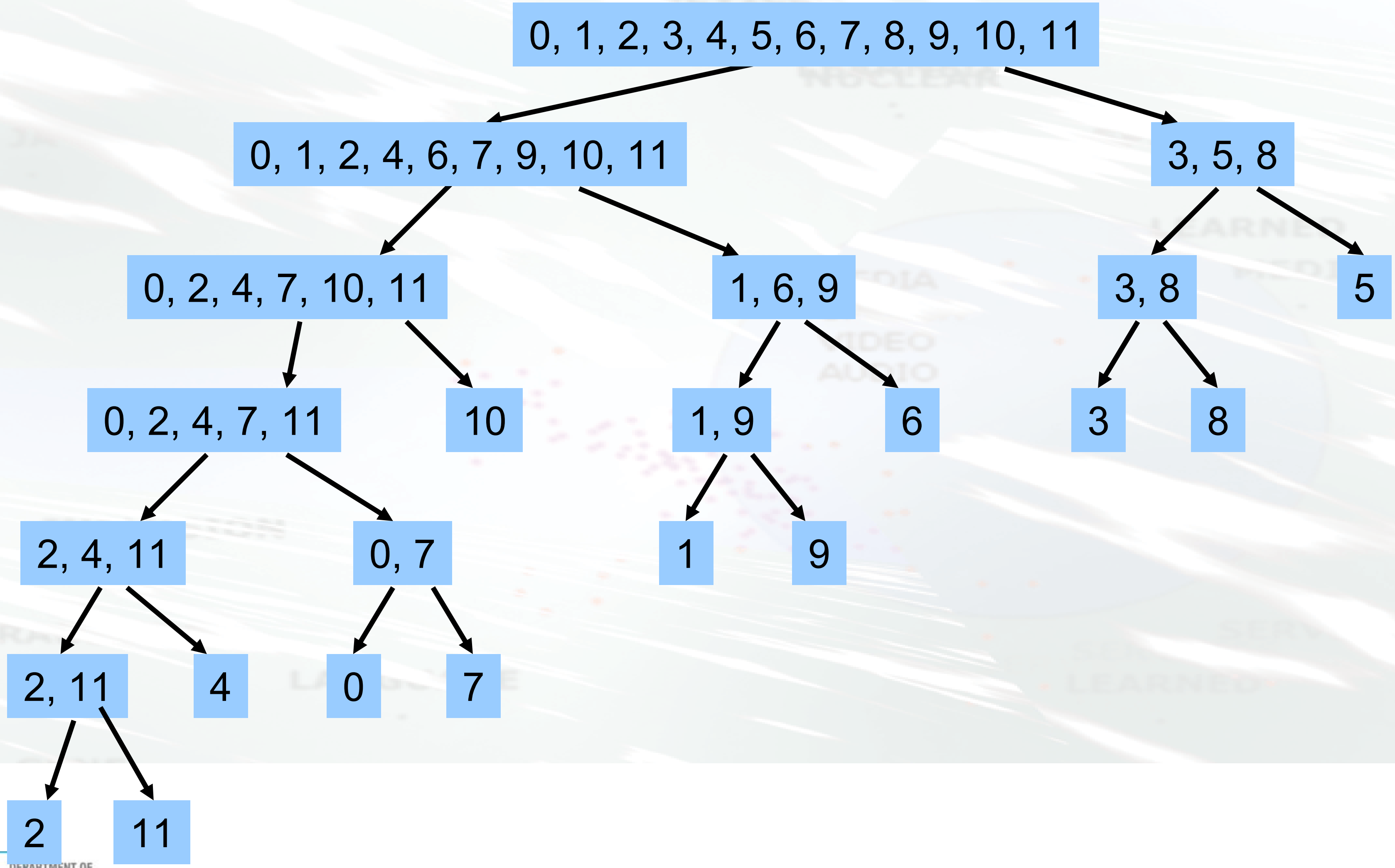
Examples:

- A: 1,0,1,0,1
- B: 1,0,0,0,1
- C: 1,0,1,0,0
- D: 0,0,0,1,0
- E: 0,1,0,1,0

K=2

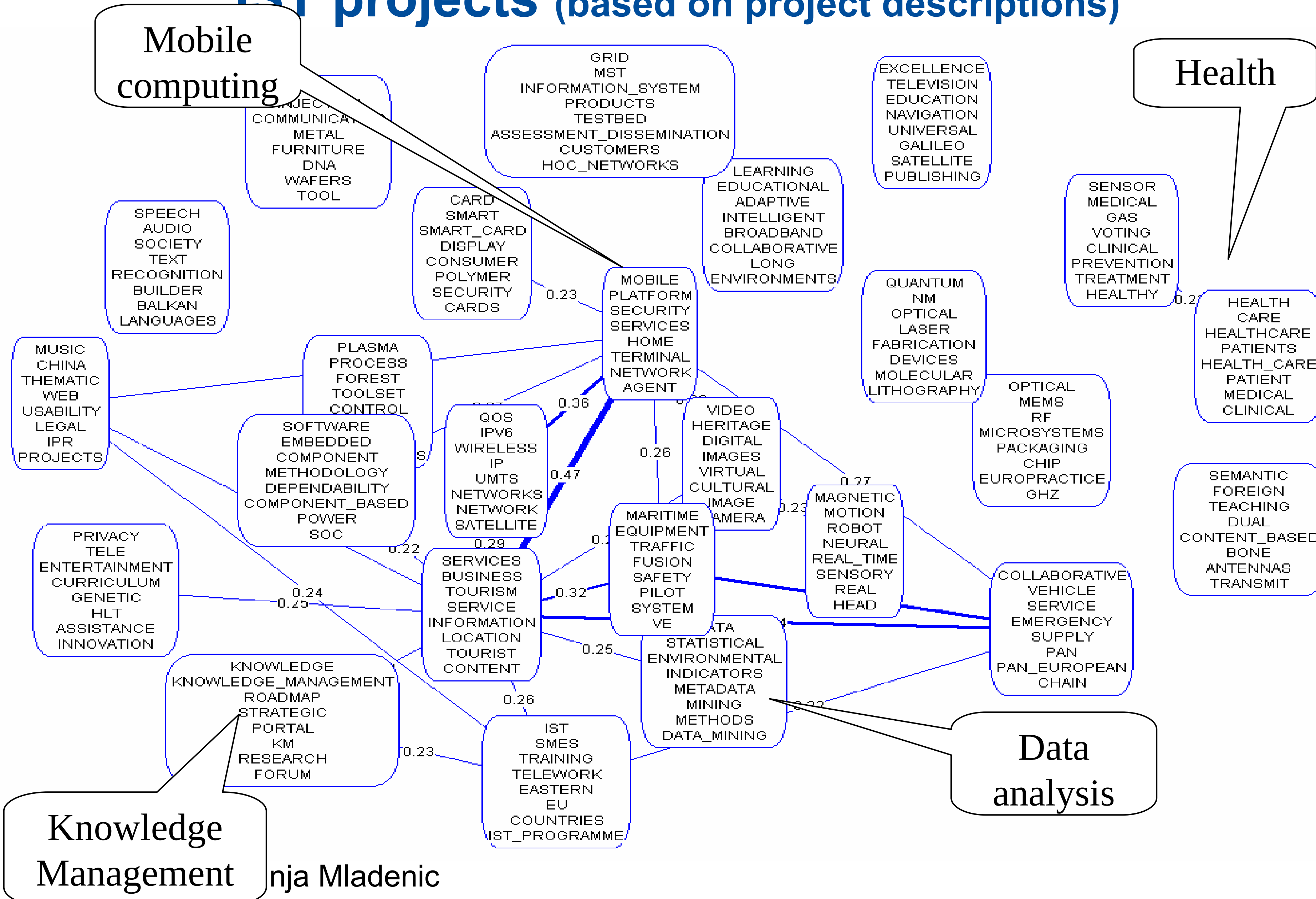
1. Randomly select two examples, e.g., A, D to be representatives of two clusters **I**: A, **II**: D
2. Calculate similarity of other examples to the them
   B,**I**= 0.82, B,**II**= 0, C,**I**= 0.82, C,**II**= 0, E,**I**= 0, E,**II**= 0.7
3. Assign examples to the most similar cluster
   I: (A,B,C)   **II**: (D,E)
4. Calculate the cluster centroid
   I: 1,0,0.67,0,0.67   **II**: 0,0.5,0,1,0
5. Calculate similarity of all the examples to the centroids A,**I**= 0.88, A,**II**= 0, B,**I**= 0.77, B,**II**= 0, C,**I**= 0.77, C,**II**= 0, D,**I**= 0, D,**II**= 0.82, E,**I**= 0, E,**II**= 0.87
6. Cluster the examples **I**: (A,B,C)   **II**: (D,E)
7. Stop as the clustering got stabilized

Example of hierarchical clustering (bisecting k-means)



DEPARTMENT OF KNOWLEDGE TECHNOLOGIES
Jožef Stefan Institute

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



References

<kt.ijs.si/Dunja/TextWebJSI>

More on our work

Text and Web Mining - IJS Group Page

Goal: Our goal is to develop new methods and approaches that will enable addressing different problems of Text and Web data analysis by applying primarily Machine Learning (ML) and Data Mining (DM) methods.
For further information, contact Dunja Mladenic or Marko Grobelnik.

Overview

The growing importance of electronic media for storing and exchanging text documents has led to a growing interest in tools and approaches for dealing with unstructured or semi-structured information included in the text documents. In addition to well-organized and maintained text databases, one of the important sources of textual information is the World Wide Web which is expected to continue to grow in the number of users and amount of information available. Connected to that is also a problem of Web access analysis, where different Web users show different behavior when browsing the same Web site (e.g. an e-commerce company Web site).

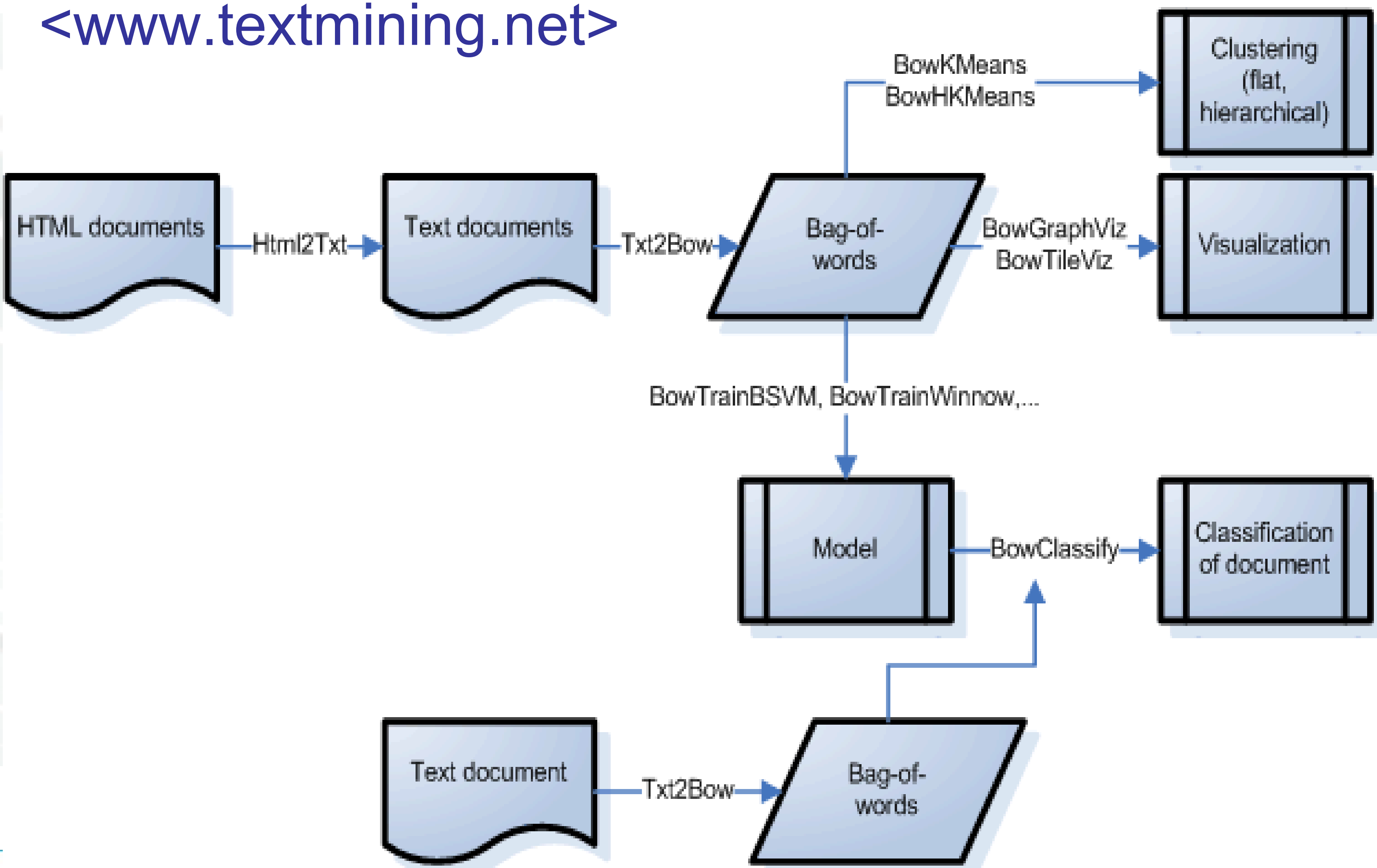
Methods developed for mining structured and unstructured data sets as well as text learning and natural language processing techniques are essential for analysis of textual data. While many approaches to text processing are based on statistics and thus only weakly dependent on the language the data is written in, those that involve deeper linguistic processing are typically aimed at English texts. Furthermore, an important step towards exploiting information from texts is automated information extraction from large document sets and building more or less domain specific knowledge bases.

Projects

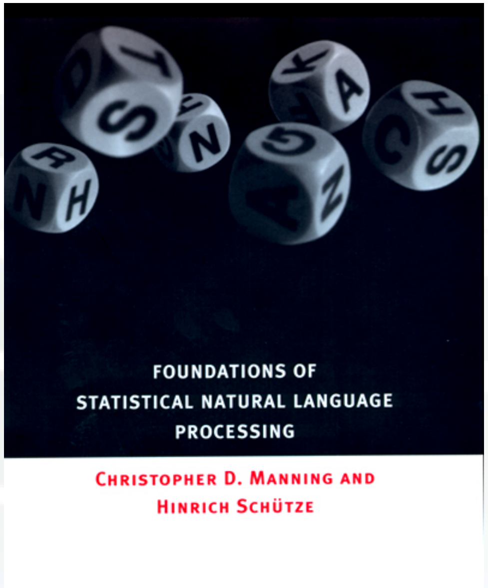
- Analysis of EU research projects and collaborations Project Intelligence
- European projects under support of EC
  - 6FP Integrated project SEKT: Semantically Enabled Knowledge Technologies (2004-2006) (IST-1-506826-IP)
  - 6FP Strategic targeted research project ALVIS: Superpeer Semantic Search Engine (2004-2006) (IST-1-002068-STP)
  - 6FP Network of Excellence PASCAL: Pattern Analysis, Statistical Modelling and Computational Learning (2003-2007) (IST-1-506778-NOE)
  - 6FP ERA project CEC-WYS: Central European Centre for Women and Youth in Science (2004-2006) (SAS6-CT-2004-003582)
  - 5FP RTD project SOL-EU-NET: Data Mining and Decision Support for Business Competitiveness: Solomon European Virtual Enterprise (2000-2003) (IST-1999-11495)
  - 5FP Network of Excellence KDNet: European Knowledge Discovery Network of Excellence (2002-2004) (IST-2001-33086)
  - 5FP Network of Excellence KMForum: European Knowledge Management Forum (2000-2003) (IST-2000-26393)
- Join projects with Microsoft Research, Cambridge, UK
  - Application of Advanced Natural Language Processing to Text Mining and Summarization (2002-2003)
  - Text Analysis using Natural Language Processing (2000-2001)
- Joint projects with CMU Text Learning Group, Pittsburgh, USA
  - Personal WebWatcher project
  - Yahoo Planet project
  - PhD thesis project: Machine Learning on non-homogeneous, distributed text data
  - Project on Analysis of Large Text Datasets
- National projects
  - Construction of archive for Slovenian Web publications, joint project with National and University Library of Slovenia (2002-2004)
  - Design and analysis of Slovenian digitalized electronic publications of national importance, joint project with National and University Library of Slovenia (2002-2004)

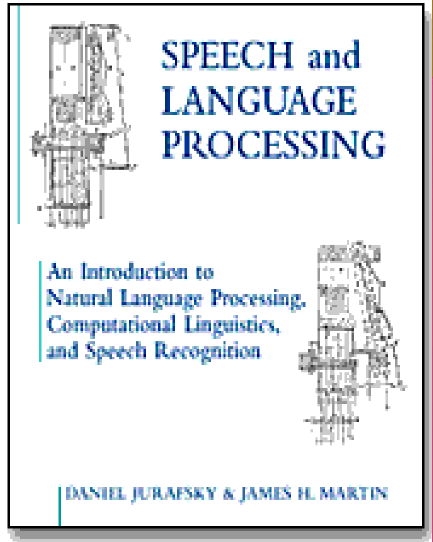
Text Garden (clustering, visualization, classification)

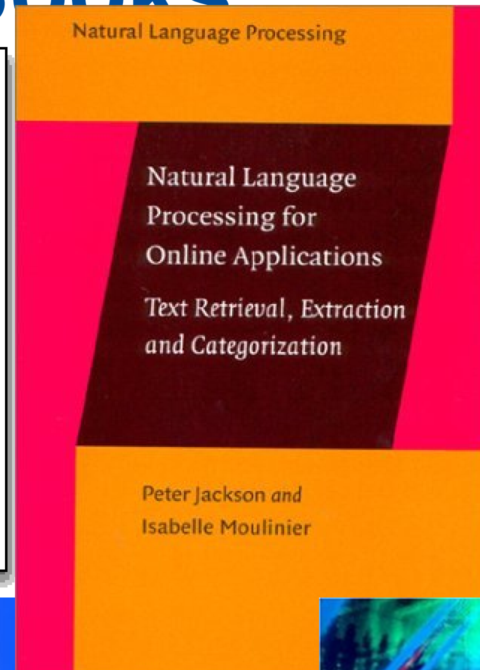
<www.textmining.net>

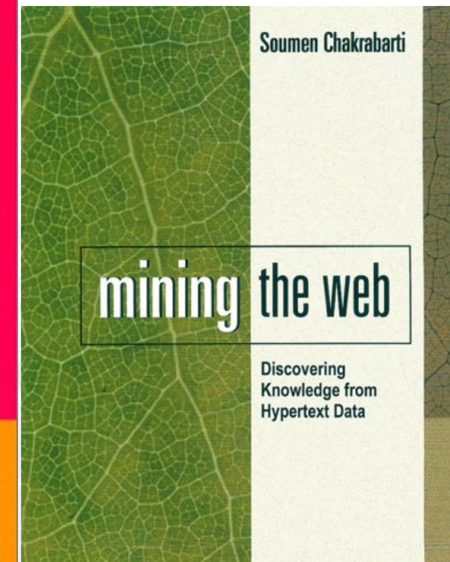


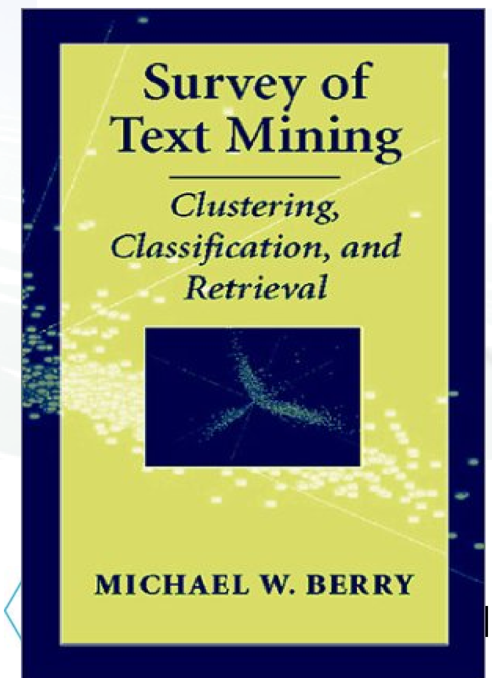
References to some of the Books

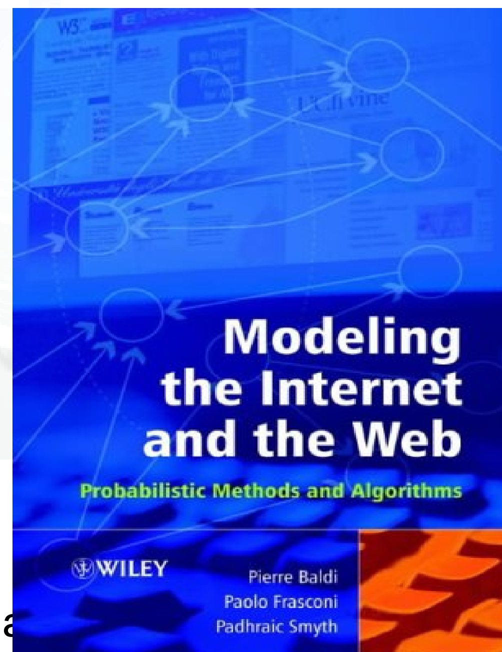


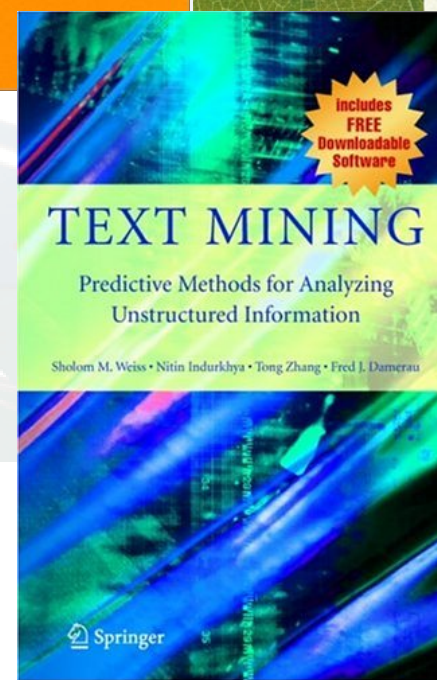










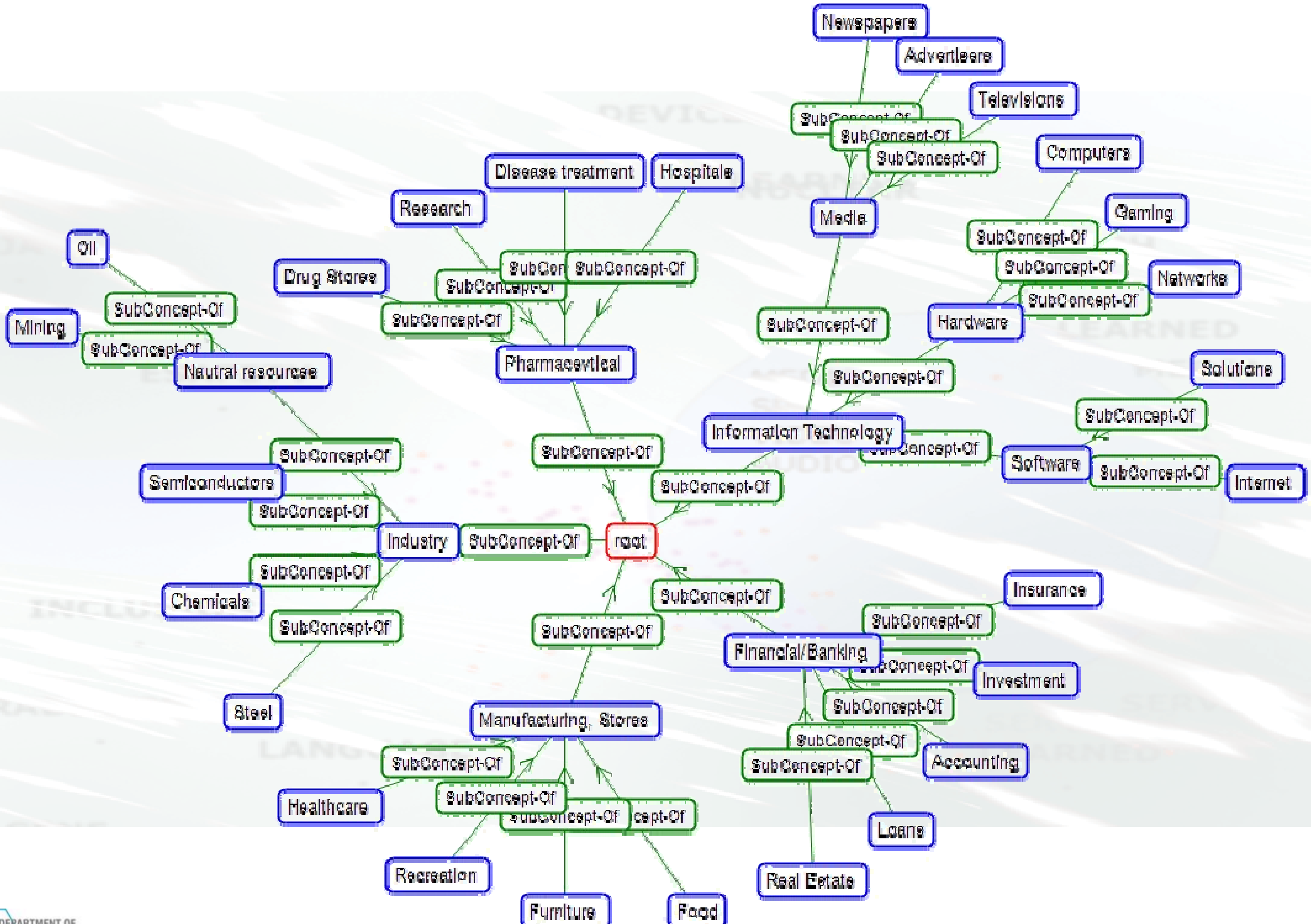


Ontology construction using OntoGen

Ontology

- Ontology is a data model that represents a set of concepts within a domain and the relationships between those concepts
- Ontology can be seen as a graph/network structure consisting from:
  - a set of concepts (vertices in a graph),
  - a set of instances assigned to a particular concepts (data records assigned to vertices in a graph)
  - a set of relationships connecting concepts (directed edges in a graph)

Example of a Topic Ontology



Ontology construction

One of the methodologies defined for ontology construction is a methodology for *semi-automatic ontology construction* analogous to the CRISP-DM methodology can be defined as consisting of the following interrelated phases:

1. *domain understanding* (what is the area we are dealing with?),
2. *data understanding* (what is the available data and its relation to semi-automatic ontology construction?),
3. *task definition* (based on the available data and its properties, define task(s) to be addressed),
4. *ontology learning* (semi-automated process addressing the task(s)
5. *ontology evaluation* (estimate quality of the solutions to the addressed task(s)),
6. *refinement with human in the loop* (perform any transformation needed to improve the ontology and return to any of the previous steps, as desired)

[Grobelnik, Mladenić 2006]

Ontology learning

- Define the ontology learning tasks in terms of mappings between ontology components, where some of the components are given and some are missing and we want to induce the missing ones.
- Some typical **scenarios in ontology learning** are the following:
  - Inducing concepts/clustering of instances (given instances)
  - Inducing relations (given concepts and the associated instances)
  - Ontology population (given an ontology and relevant, but not associated instances)
  - Ontology generation (given instances and any other background information)
  - Ontology updating/extending (given an ontology and background information, such as, new instances or the ontology usage patterns)

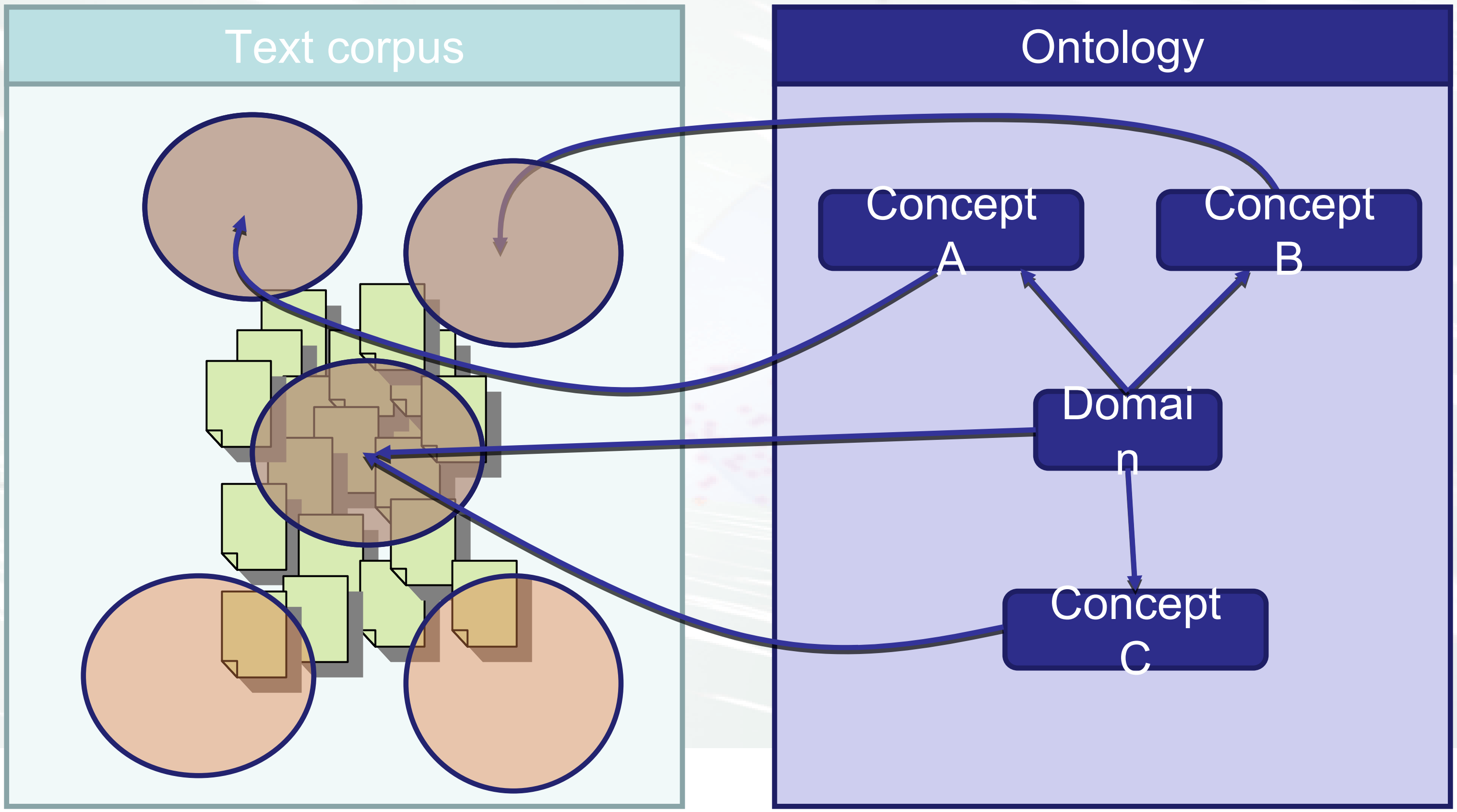
Ontology Learning with OntoGen
(developed on the top of Text Garden)

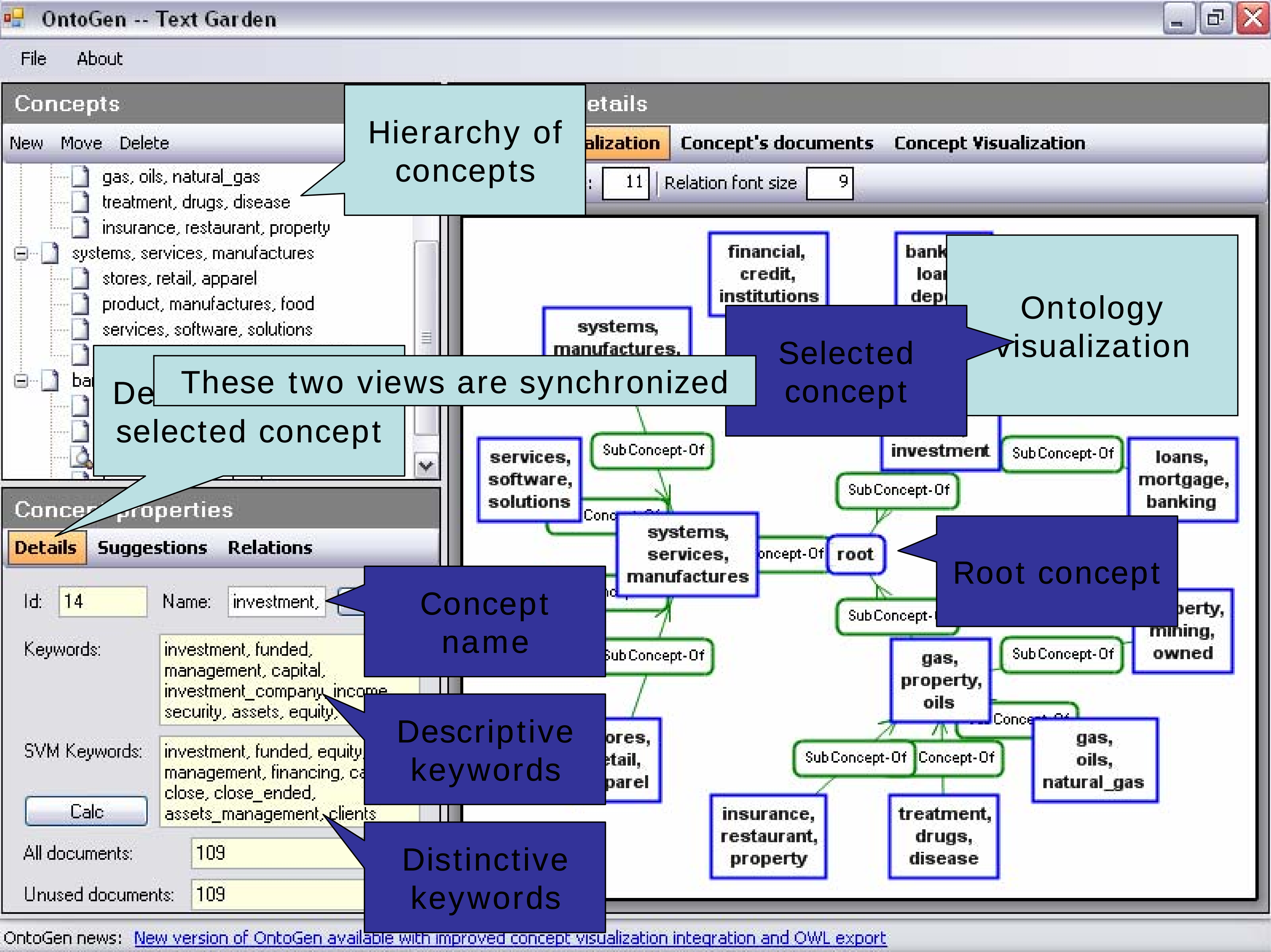
- Semi-Automatic
  - provide suggestions and insights into the domain
  - the user interacts with parameters of methods
  - final decisions taken by the user
- Data-Driven
  - most of the aid provided by the system is based on some underlying data
  - instances are described by features extracted from the data (eg., words-vectors)

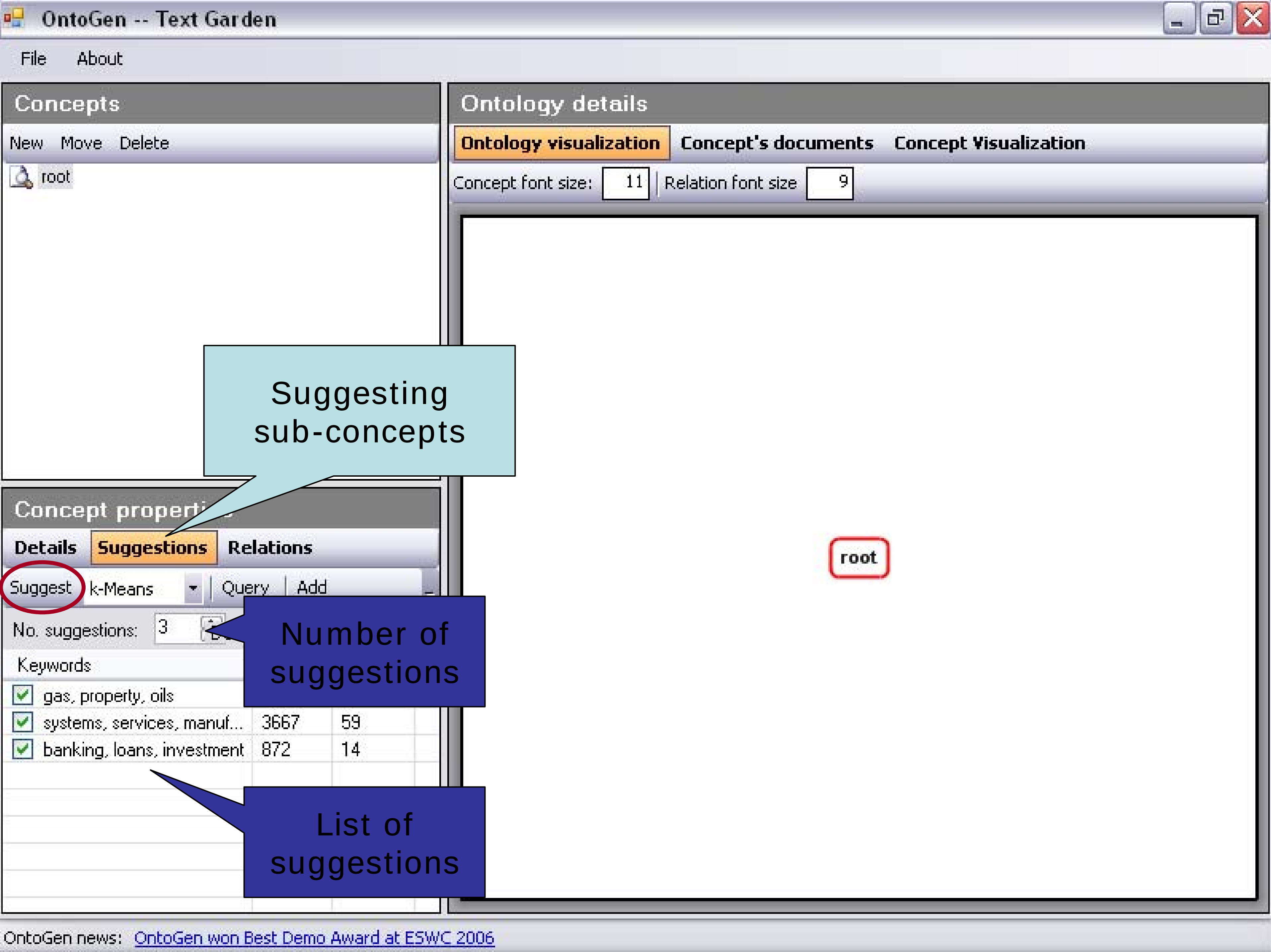
[Fortuna & Mladenić & Grobelnik, 2005]

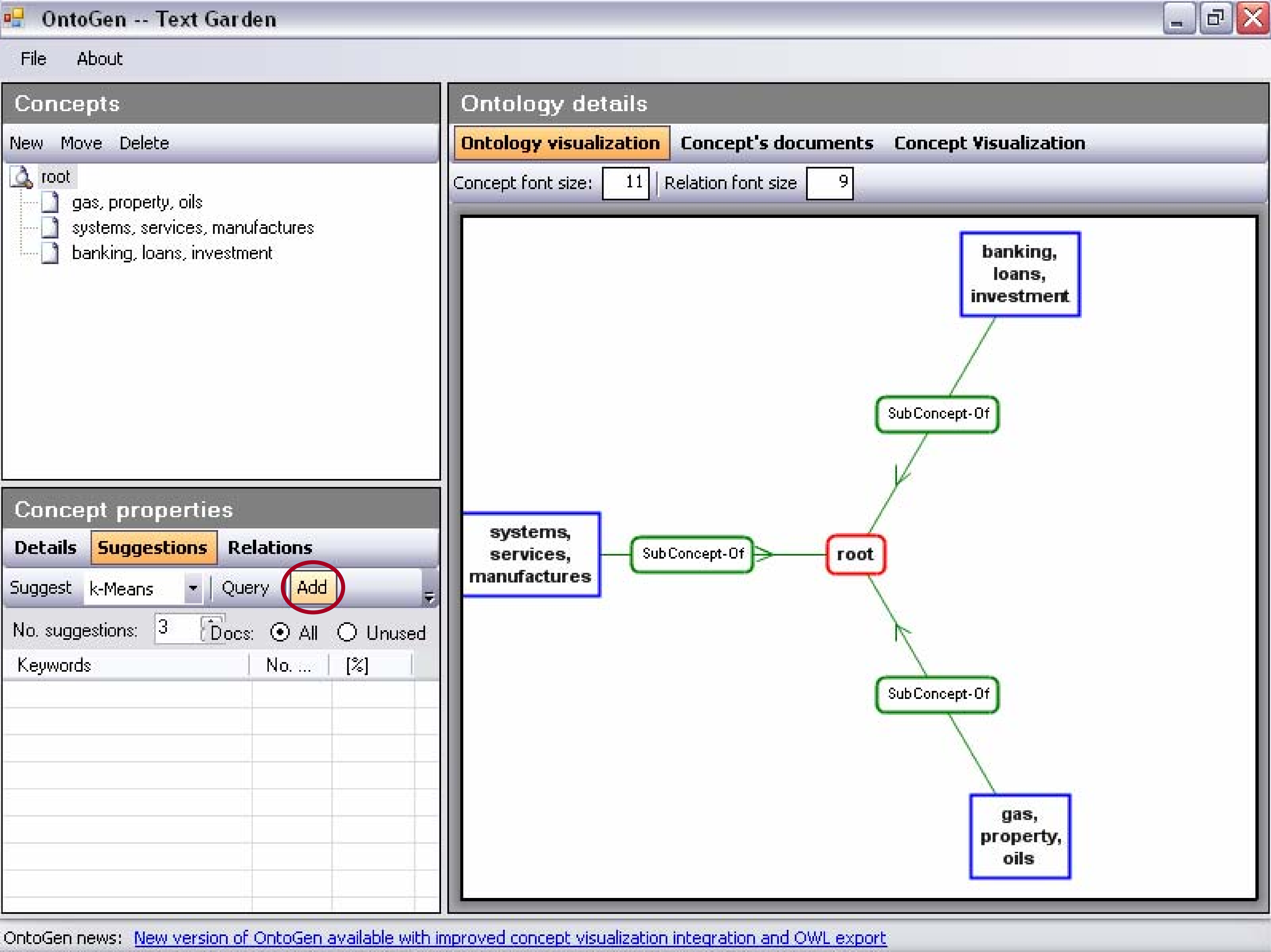
Installation package is publicly available in binaries at ontogen.ijs.si

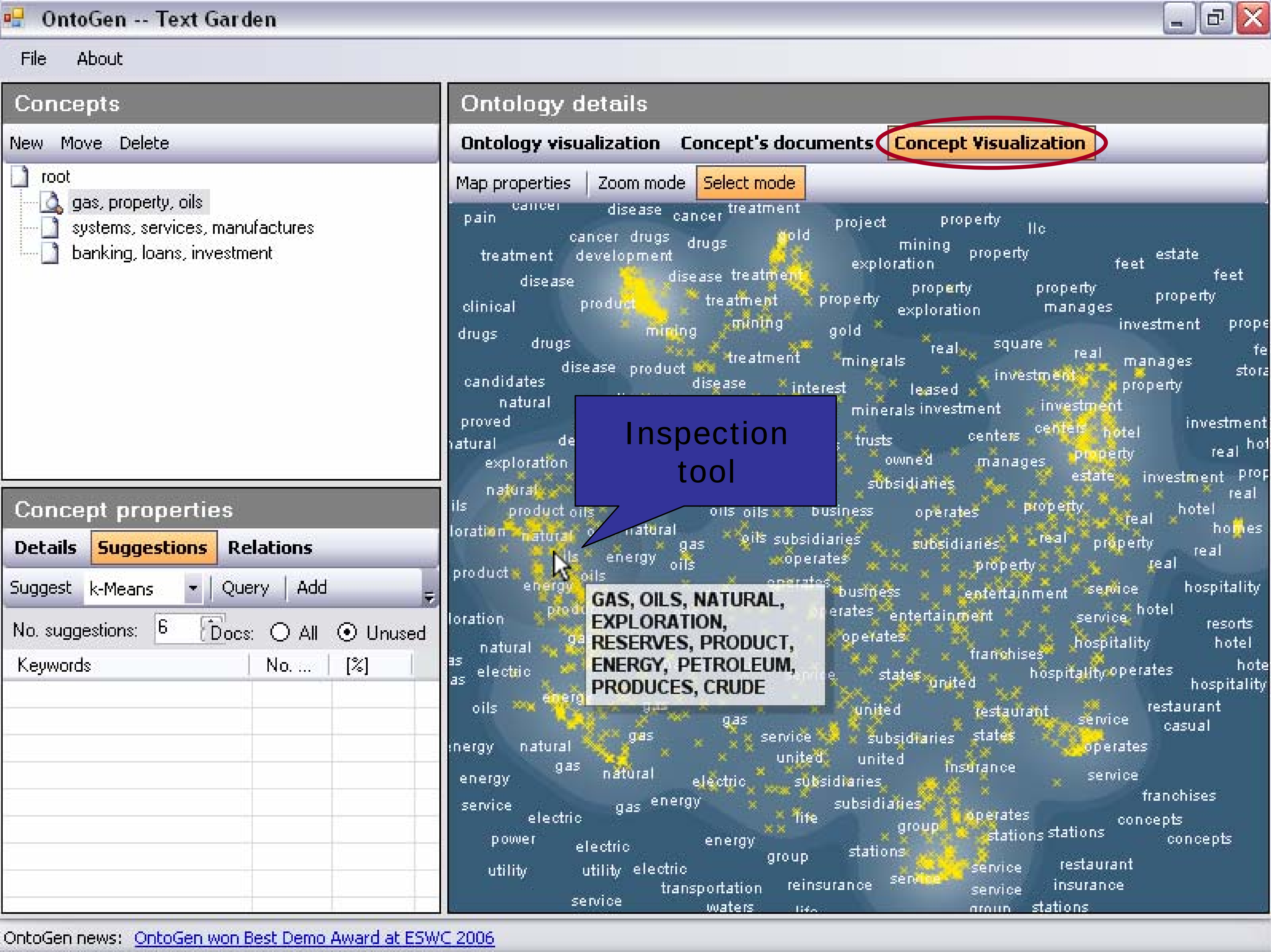
Basic idea behind OntoGen



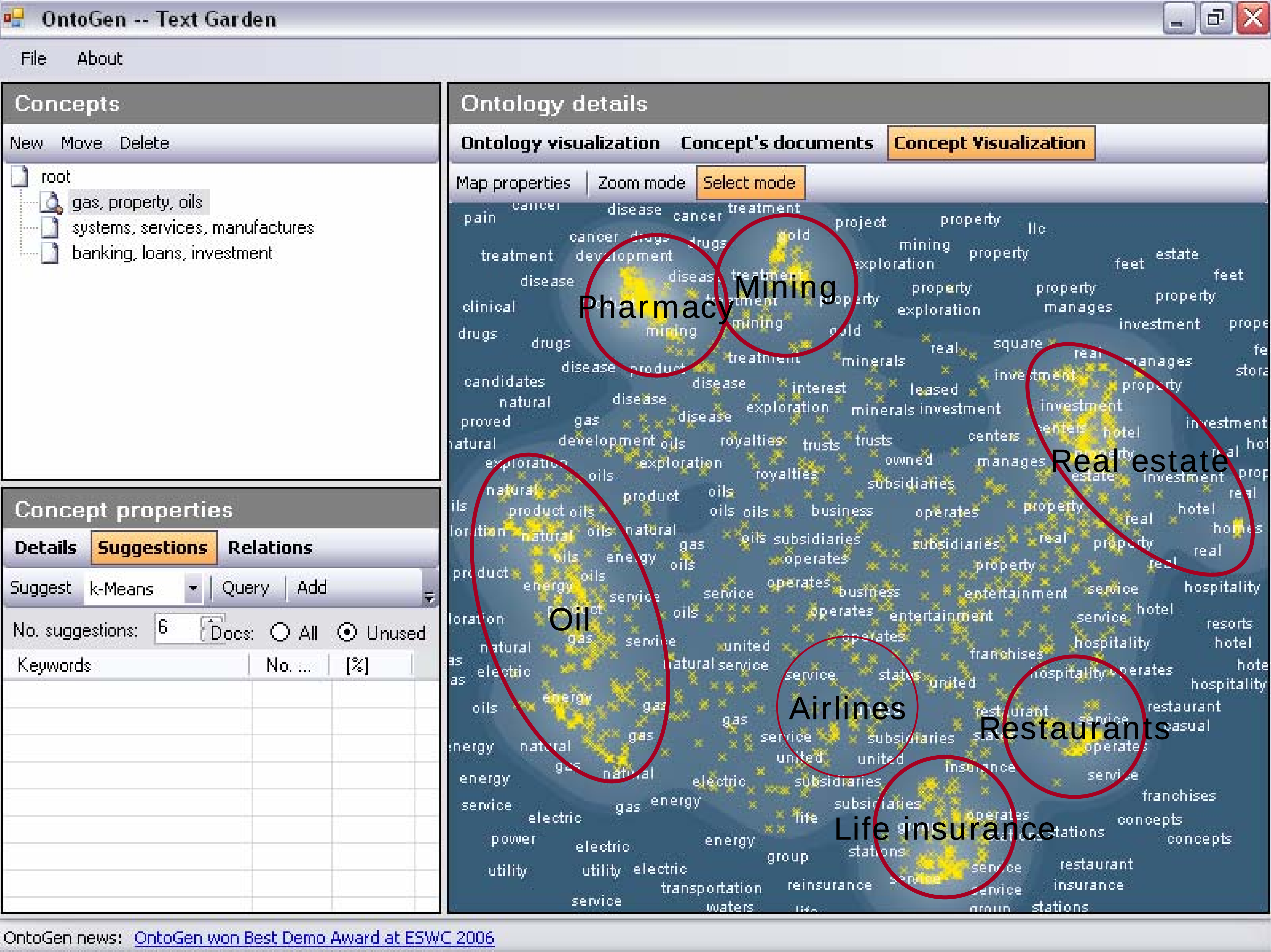

OntoGen -- Text Garden
File
About
Concepts
New
Move
Delete
gas, oils, natural_gas
treatment, drugs, disease
insurance, restaurant, property
systems, services, manufactures
stores, retail, apparel
product, manufactures, food
services, software, solutions
Hierarchy of concepts
Concept's documents
Concept Visualization
Relation font size
11
9
financial, credit, institutions
systems, manufactures,
Ontology visualization
Selected concept
These two views are synchronized
selected concept
investment
SubConcept-Of
services, software, solutions
loans, mortgage, banking
systems, services, manufactures
root
Root concept
Concept properties
Details
Suggestions
Relations
Id:
14
Name:
investment,
Concept name
Keywords:
investment, funded, management, capital, investment_company, income security, assets, equity,
Descriptive keywords
SVM Keywords:
investment, funded, equity, management, financing, close, close_ended, assets_management, clients
Distinctive keywords
Calc
All documents:
109
Unused documents:
109
gas, property, oils
gas, oils, natural_gas
insurance, restaurant, property
treatment, drugs, disease
OntoGen news:
New version of OntoGen available with improved concept visualization integration and OWL export

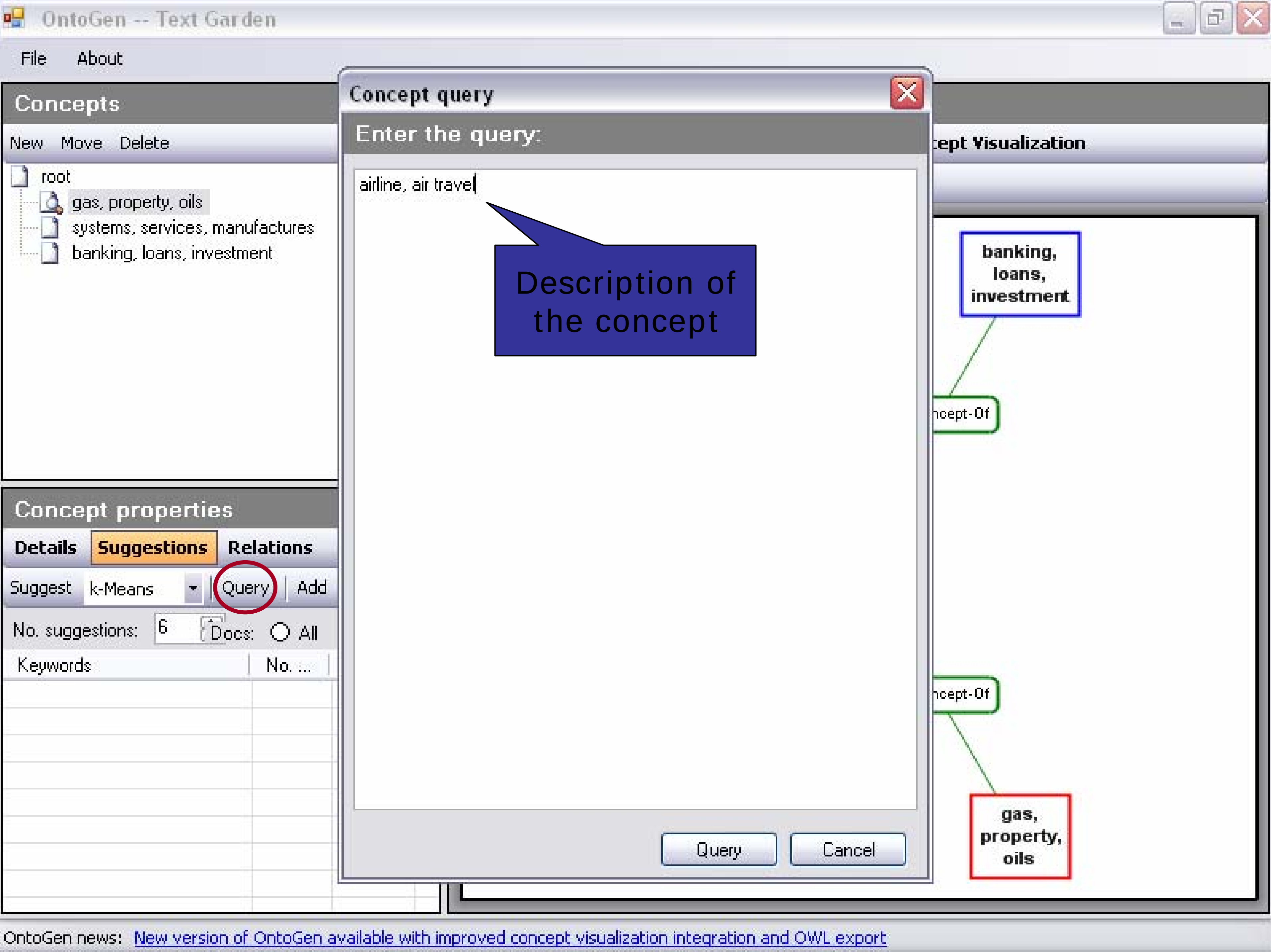

OntoGen -- Text Garden
File
About
Concepts
New
Move
Delete
root
Ontology details
Ontology visualization
Concept's documents
Concept Visualization
Concept font size:
11
Relation font size
9
root
Suggesting sub-concepts
Concept properties
Details
Suggestions
Relations
Suggest
k-Means
Query
Add
No. suggestions:
3
Number of suggestions
Keywords
gas, property, oils
systems, services, manuf... 3667 59
banking, loans, investment 872 14
List of suggestions
OntoGen news:
OntoGen won Best Demo Award at ESWC 2006


OntoGen -- Text Garden
File
About
Concepts
New
Move
Delete
root
gas, property, oils
systems, services, manufactures
banking, loans, investment
Concept properties
Details
Suggestions
Relations
Suggest
k-Means
Query
Add
No. suggestions:
3
Docs:
All
Unused
Keywords
No. ...
[%]
Ontology details
Ontology visualization
Concept's documents
Concept Visualization
Concept font size:
11
Relation font size
9
banking, loans, investment
SubConcept-Of
systems, services, manufactures
SubConcept-Of
root
SubConcept-Of
gas, property, oils
OntoGen news:
New version of OntoGen available with improved concept visualization integration and OWL export

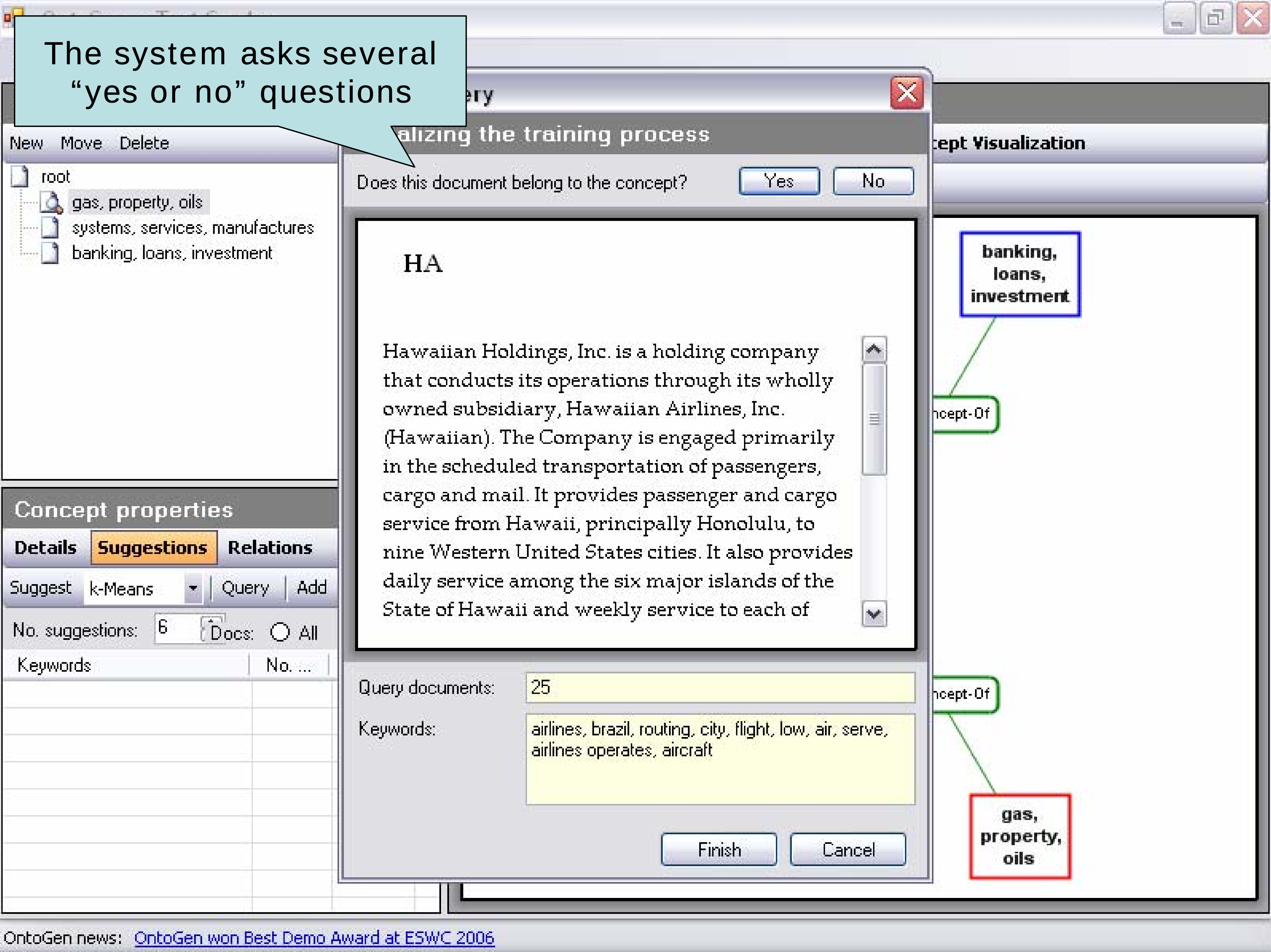


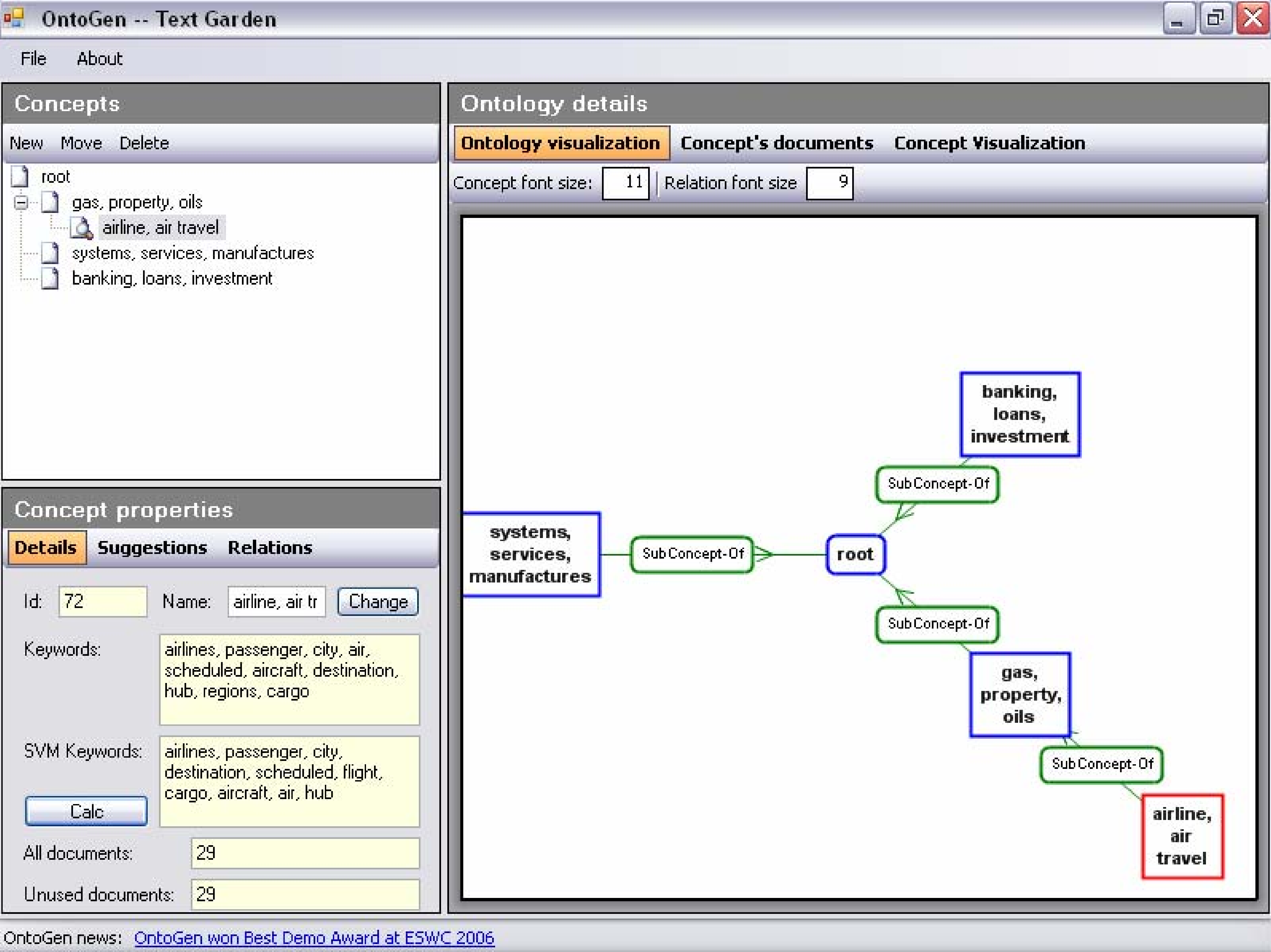
OntoGen -- Text Garden
File
About
Concepts
New
Move
Delete
root
gas, property, oils
systems, services, manufactures
banking, loans, investment
Concept properties
Details
Suggestions
Relations
Suggest
k-Means
Query
Add
No. suggestions:
6
Docs:
All
Unused
Keywords
No. ...
[%]
Ontology details
Ontology visualization
Concept's documents
Concept Visualization
Map properties
Zoom mode
Select mode
Inspection tool
GAS, OILS, NATURAL, EXPLORATION, RESERVES, PRODUCT, ENERGY, PETROLEUM, PRODUCES, CRUDE
OntoGen news:
OntoGen won Best Demo Award at ESWC 2006

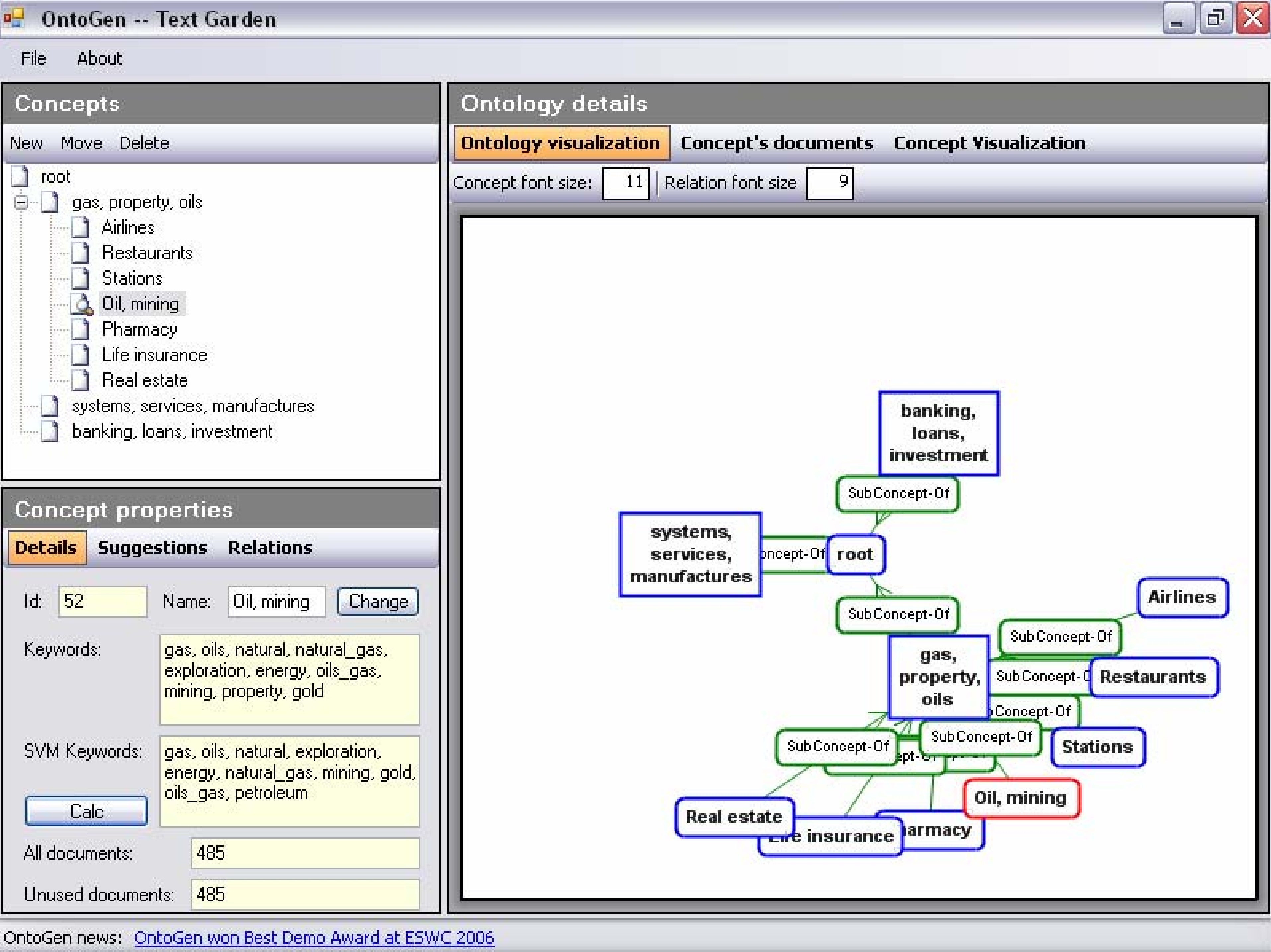

OntoGen -- Text Garden
File
About
Concepts
New
Move
Delete
root
gas, property, oils
systems, services, manufactures
banking, loans, investment
Concept properties
Details
Suggestions
Relations
Suggest
k-Means
Query
Add
No. suggestions:
6
Docs:
All
Unused
Keywords
No. ...
[%]
Ontology details
Ontology visualization
Concept's documents
Concept Visualization
Map properties
Zoom mode
Select mode
Pharmacy
Mining
Real estate
Oil
Airlines
Restaurants
Life insurance
OntoGen news:
OntoGen won Best Demo Award at ESWC 2006

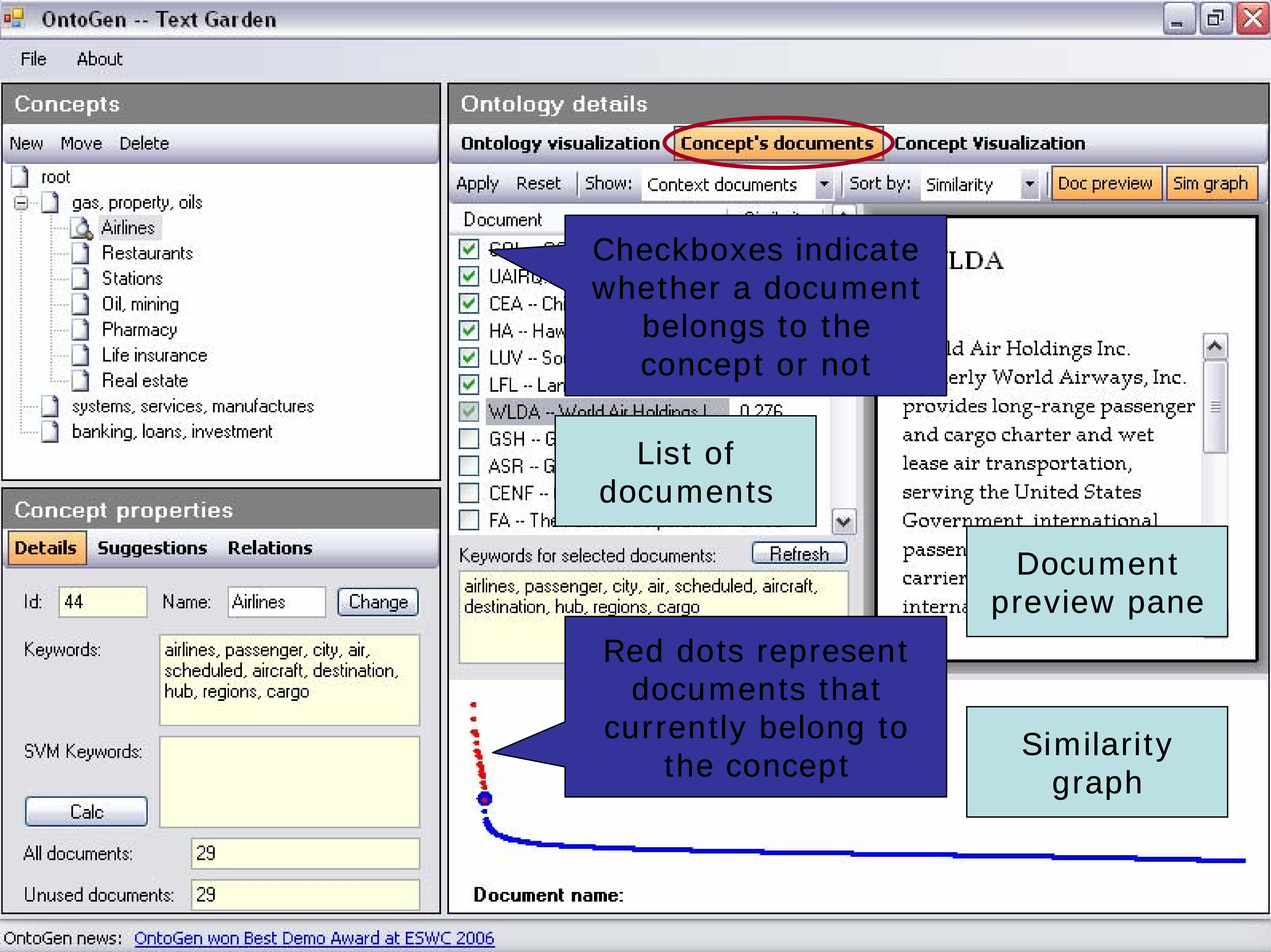

OntoGen -- Text Garden
File
About
Concepts
New
Move
Delete
root
gas, property, oils
systems, services, manufactures
banking, loans, investment
Concept properties
Details
Suggestions
Relations
Suggest
k-Means
Query
Add
No. suggestions:
6
Docs:
All
Keywords
No. ...
Concept query
Enter the query:
airline, air travel
Description of the concept
Query
Cancel
cept Visualization
banking, loans, investment
ncept-Of
ncept-Of
gas, property, oils
OntoGen news:
New version of OntoGen available with improved concept visualization integration and OWL export

The system asks several "yes or no" questions




OntoGen -- Text Garden
File
About
Concepts
New
Move
Delete
root
gas, property, oils
airline, air travel
systems, services, manufactures
banking, loans, investment
Concept properties
Details
Suggestions
Relations
Id:
72
Name:
airline, air tr
Change
Keywords:
airlines, passenger, city, air, scheduled, aircraft, destination, hub, regions, cargo
SVM Keywords:
airlines, passenger, city, destination, scheduled, flight, cargo, aircraft, air, hub
Calc
All documents:
29
Unused documents:
29
Ontology details
Ontology visualization
Concept's documents
Concept Visualization
Concept font size:
11
Relation font size
9
banking, loans, investment
SubConcept-Of
systems, services, manufactures
SubConcept-Of
root
SubConcept-Of
gas, property, oils
SubConcept-Of
airline, air travel
OntoGen news:
OntoGen won Best Demo Award at ESWC 2006


OntoGen -- Text Garden
File
About
Concepts
New
Move
Delete
root
gas, property, oils
Airlines
Restaurants
Stations
Oil, mining
Pharmacy
Life insurance
Real estate
systems, services, manufactures
banking, loans, investment
Concept properties
Details
Suggestions
Relations
Id:
52
Name:
Oil, mining
Change
Keywords:
gas, oils, natural, natural_gas, exploration, energy, oils_gas, mining, property, gold
SVM Keywords:
gas, oils, natural, exploration, energy, natural_gas, mining, gold, oils_gas, petroleum
Calc
All documents:
485
Unused documents:
485
Ontology details
Ontology visualization
Concept's documents
Concept Visualization
Concept font size:
11
Relation font size
9
banking, loans, investment
SubConcept-Of
systems, services, manufactures
root
Airlines
gas, property, oils
Restaurants
Stations
Oil, mining
Real estate
OntoGen news:
OntoGen won Best Demo Award at ESWC 2006


OntoGen -- Text Garden
File
About
Concepts
New
Move
Delete
root
gas, property, oils
Airlines
Restaurants
Stations
Oil, mining
Pharmacy
Life insurance
Real estate
systems, services, manufactures
banking, loans, investment
Concept properties
Details
Suggestions
Relations
Id:
44
Name:
Airlines
Change
Keywords:
airlines, passenger, city, air, scheduled, aircraft, destination, hub, regions, cargo
SVM Keywords:
Calc
All documents:
29
Unused documents:
29
Ontology details
Ontology visualization
Concept's documents
Concept Visualization
Apply
Reset
Show:
Context documents
Sort by:
Similarity
Doc preview
Sim graph
Document
Checkboxes indicate whether a document belongs to the concept or not
List of documents
Document preview pane
Keywords for selected documents:
Refresh
airlines, passenger, city, air, scheduled, aircraft, destination, hub, regions, cargo
Red dots represent documents that currently belong to the concept
Similarity graph
Document name:
OntoGen news:
OntoGen won Best Demo Award at ESWC 2006