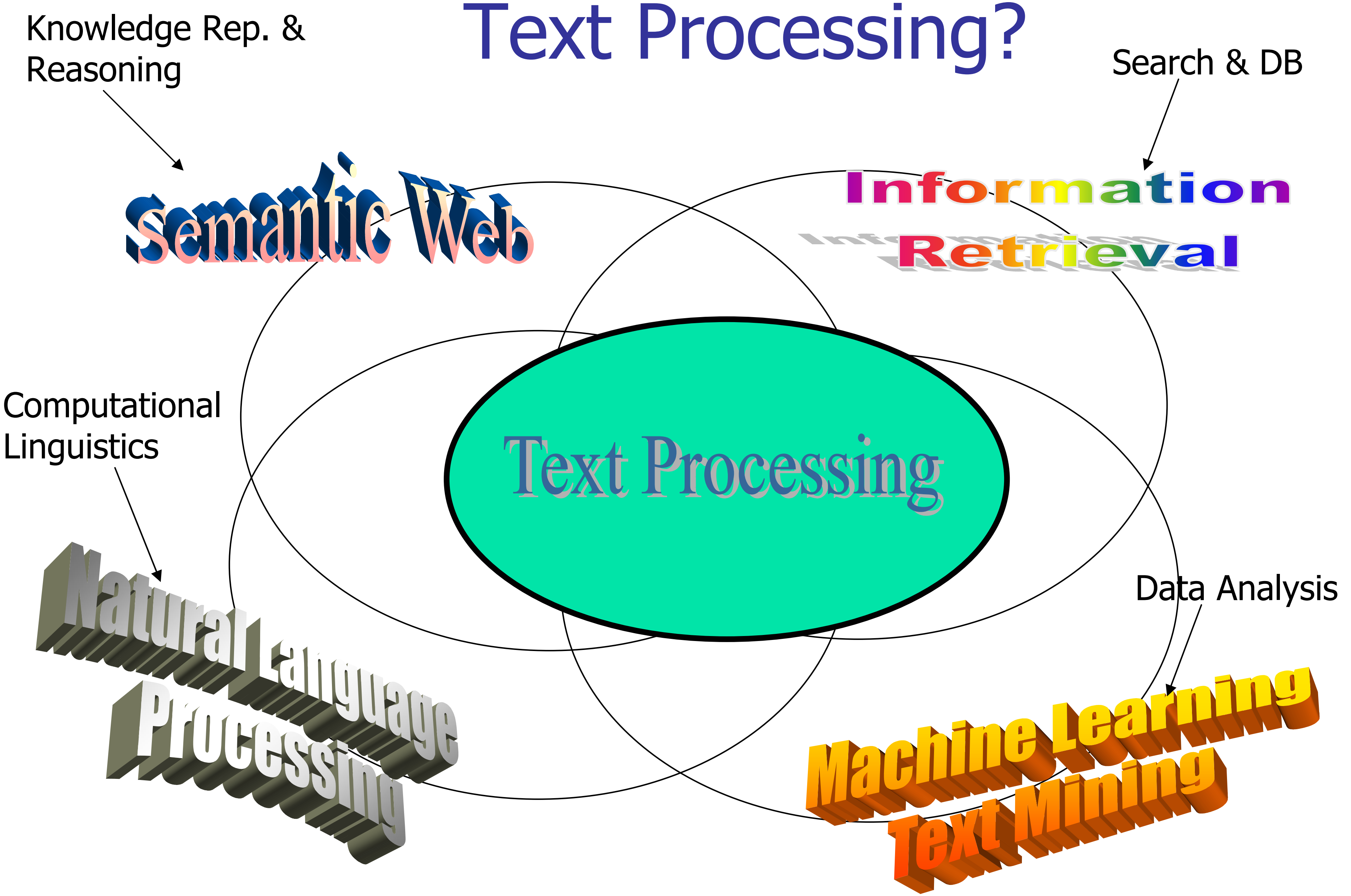
Text-Mining Tutorial

Marko Grobelnik, Dunja Mladenic
J. Stefan Institute, Slovenia

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?



Tutorial Contents

- Why Text is Easy and Why Tough?
- Levels of Text Processing
  - Word Level
  - Sentence Level
  - Document Level
  - Document-Collection Level
  - Linked-Document-Collection Level
  - Application Level
- References to Conferences, Workshops, Books, Products
- Final Remarks

Why Text is Tough? (M.Hearst 97)

- Abstract concepts are **difficult to represent**
- **“Countless” combinations** of subtle, abstract relationships among concepts
- **Many ways** to represent similar concepts
  - E.g. space ship, flying saucer, UFO
- Concepts are **difficult to visualize**
- **High dimensionality**
- **Tens or hundreds of thousands of features**

Why Text is Easy? (M.Hearst 97)

- **Highly redundant data**
  - …most of the methods count on this property
- **Just about any simple algorithm can get "good" results for simple tasks:**
  - Pull out "important" phrases
  - Find "meaningfully" related words
  - Create some sort of summary from documents

Levels of Text Processing 1/6

- **Word Level**
  - **Words Properties**
  - **Stop-Words**
  - **Stemming**
  - **Frequent N-Grams**
  - **Thesaurus (WordNet)**
- Sentence Level
- Document Level
- Document-Collection Level
- Linked-Document-Collection Level
- Application Level

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
- Natural language dependent – examples:
  - English: A, ABOUT, ABOVE, ACROSS, AFTER, AGAIN, AGAINST, ALL, ALMOST, ALONE, ALONG, ALREADY, ALSO, ...
  - Slovenian: A, AH, AHA, ALI, AMPAK, BAJE, BODISI, BOJDA, BRŽKONE, BRŽČAS, BREZ, CELO, DA, DO, ...
  - Croatian: A, AH, AHA, ALI, AKO, BEZ, DA, IPAK, NE, NEGO, ...

Original text

Information Systems Asia Web - provides research, IS-related commercial materials, interaction, and even research sponsorship by interested corporations with a focus on Asia Pacific region.

Survey of Information Retrieval - guide to IR, with an emphasis on web-based projects. Includes a glossary, and pointers to interesting papers.

After the stop-words removal

Information Systems Asia Web provides research IS-related commercial materials interaction research sponsorship interested corporations focus Asia Pacific region

Survey Information Retrieval guide IR emphasis web-based projects Includes glossary pointers interesting papers

Stemming (I)

- 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 (normalized form)

Stemming (II)

- 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/
- E.g. in Slovenian language 10-20 different forms correspond to the same word:
  - E.g. ("to laugh" in Slovenian): smej, smejal, smejala, smejale, smejali, smejalo, smejati, smejejo, smejeta, smejete, smejeva, smeješ, smejemo, smejiš, smeje, smejoč, smejta, smejte, smejva

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

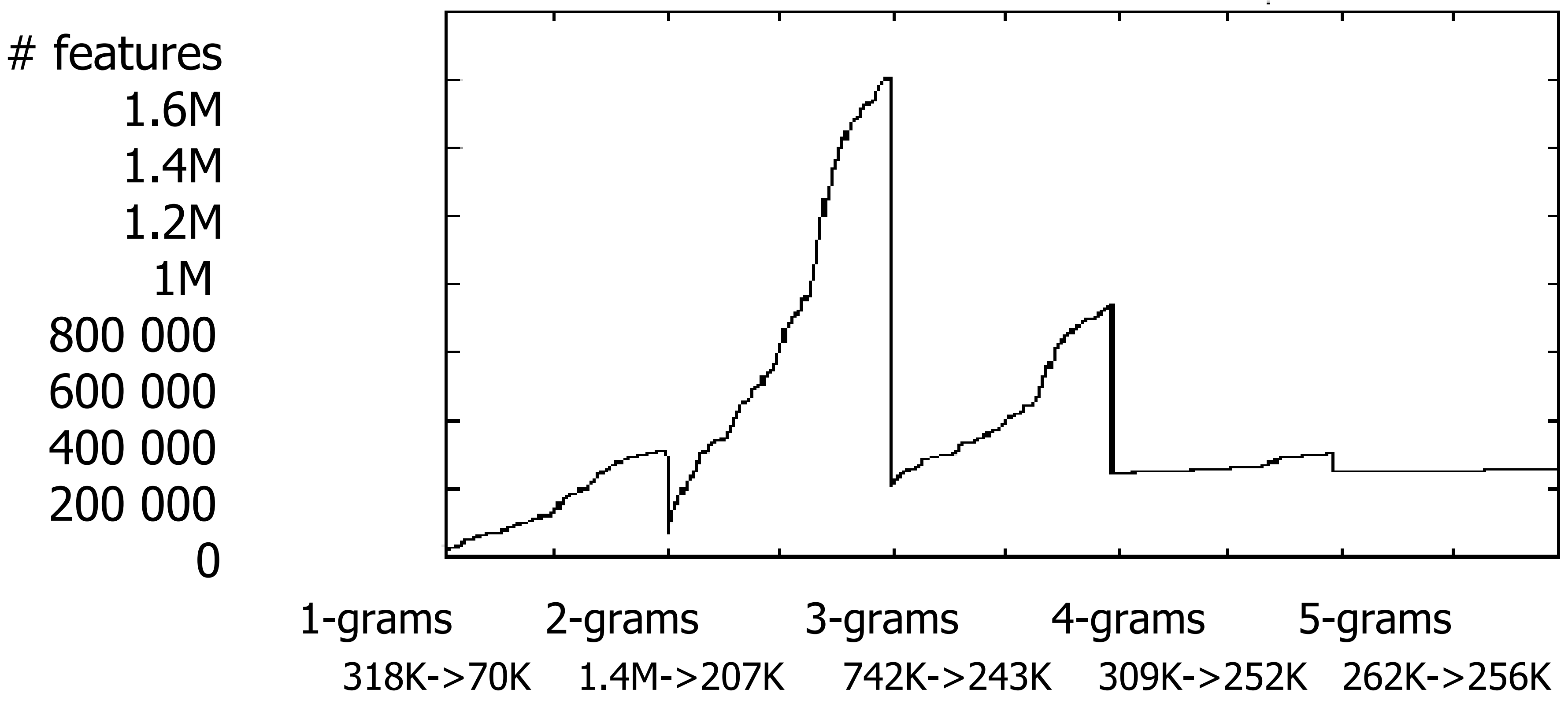
Rules automatically obtained for Slovenian language

- Machine Learning applied on Multext-East dictionary (http://nl.ijs.si/ME/)
- Two example rules:
  - Remove the ending "OM" if 3 last char is any of HOM, NOM, DOM, SOM, POM, BOM, FOM. For instance, ALAHOM, AMERICANOM, BENJAMINOM, BERLINOM, ALFREDOM, BEOGRADOM, DICKENSOM, JEZUSOM, JOSIPOM, OLIMPOM,... but not ALEKSANDROM (ROM -> ER)
  - Replace CEM by EC. For instance, ARABCEM, BAVARCEM, BOVCEM, EVROPEJCEM, GORENJCEM, ... but not FRANCEM (remove EM)

Phrases in the form of frequent N-Grams

- Simple way for generating phrases are frequent n-grams:
  - N-Gram is a sequence of n consecutive words (e.g. “machine learning” is 2-gram)
  - “Frequent n-grams” are the ones which appear in all observed documents MinFreq or more times
- N-grams are interesting because of the simple and efficient dynamic programming algorithm:
- Given:
  - Set of documents (each document is a sequence of words),
  - MinFreq (minimal n-gram frequency),
  - MaxNGramSize (maximal n-gram length)
- for Len = 1 to MaxNGramSize do
  - Generate candidate n-grams as sequences of words of size Len using frequent n-grams of length Len-1
  - Delete candidate n-grams with the frequency less then MinFreq

Generation of frequent n-grams for 50,000 documents from Yahoo



Original text on the Yahoo Web page:

1. Top:Reference:Libraries:Library and Information Science:Information Retrieval
2. UK Only
3. Idomeneus - IR \& DB repository - These pages mostly contain IR related resources such as test collections, stop lists, stemming algorithms, and links to other IR sites.
4. University of Glasgow - Information Retrieval Group - information on the resources and people in the Glasgow IR group.
5. Centre for Intelligent Information Retrieval (CIIR).
6. Information Systems Asia Web - provides research, IS-related commercial materials, interaction, and even research sponsorship by interested corporations with a focus on Asia Pacific region.
7. Seminar on Cataloging Digital Documents
8. Survey of Information Retrieval - guide to IR, with an emphasis on web-based projects. Includes a glossary, and pointers to interesting papers.
9. University of Dortmund - Information Retrieval Group

Document represented by n-grams:

1. "REFERENCE LIBRARIES LIBRARY INFORMATION SCIENCE (\#3 LIBRARY INFORMATION SCIENCE) INFORMATION RETRIEVAL (\#2 INFORMATION RETRIEVAL)"
2. "UK"
3. "IR PAGES IR RELATED RESOURCES COLLECTIONS LISTS LINKS IR SITES"
4. "UNIVERSITY GLASGOW INFORMATION RETRIEVAL (\#2 INFORMATION RETRIEVAL) GROUP INFORMATION RESOURCES (\#2 INFORMATION RESOURCES) PEOPLE GLASGOW IR GROUP"
5. "CENTRE INFORMATION RETRIEVAL (\#2 INFORMATION RETRIEVAL)"
6. "INFORMATION SYSTEMS ASIA WEB RESEARCH COMMERCIAL MATERIALS RESEARCH ASIA PACIFIC REGION"
7. "CATALOGING DIGITAL DOCUMENTS"
8. "INFORMATION RETRIEVAL (\#2 INFORMATION RETRIEVAL) GUIDE IR EMPHASIS INCLUDES GLOSSARY INTERESTING"
9. "UNIVERSITY INFORMATION RETRIEVAL (\#2 INFORMATION RETRIEVAL) GROUP"

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 Processing 2/6

- Word Level
- **Sentence Level**
- Document Level
- Document-Collection Level
- Linked-Document-Collection Level
- Application Level

Levels of Text Processing 3/6

- Word Level
- Sentence Level
- **Document Level**
  - **Summarization**
  - **Single Document Visualization**
  - **Text Segmentation**
- Document-Collection Level
- Linked-Document-Collection Level
- Application Level

Summarization

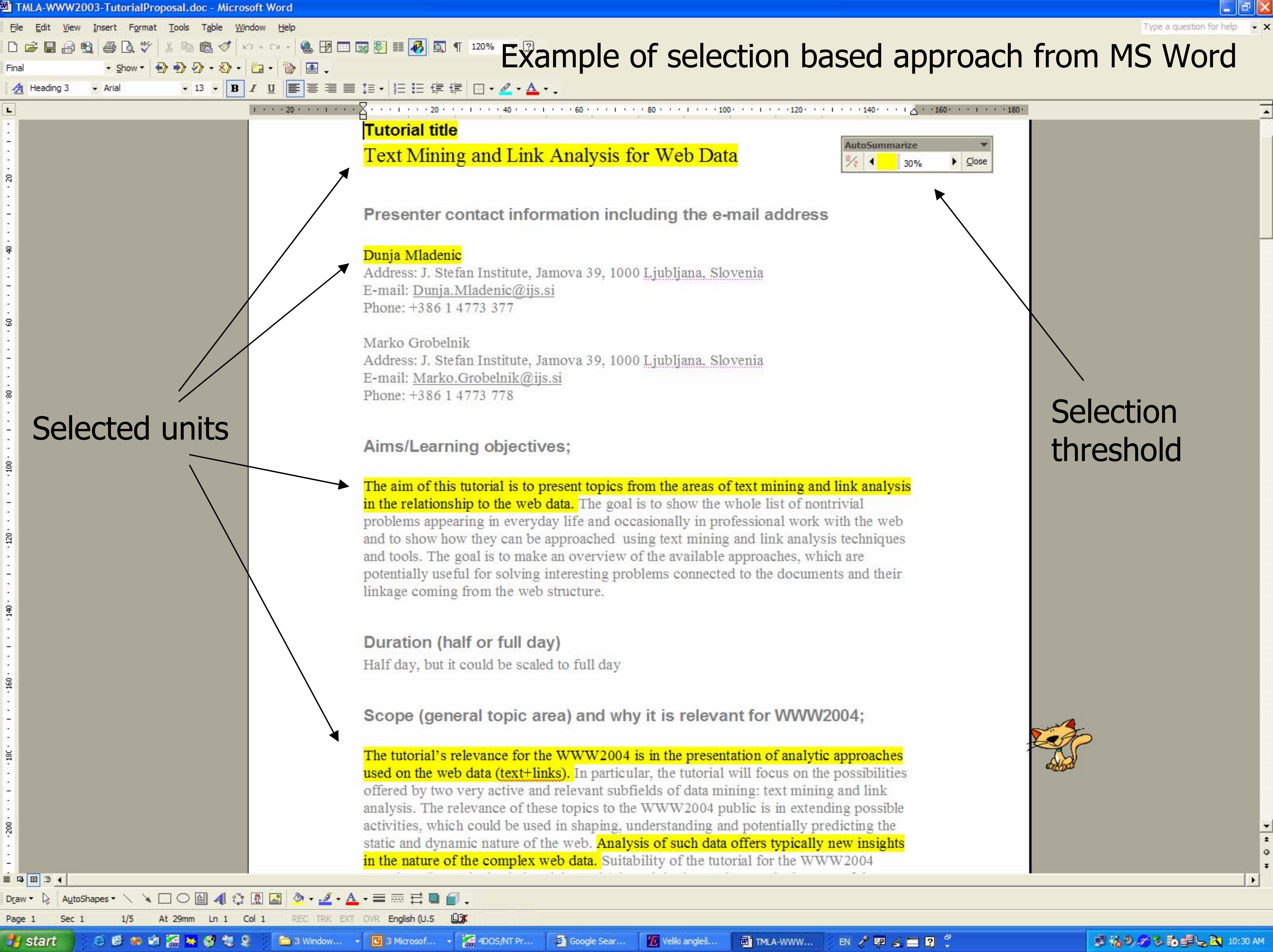
Summarization

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

Selection based summarization

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

Example of selection based approach from MS Word



Tutorial title

Text Mining and Link Analysis for Web Data

Presenter contact information including the e-mail address

Dunja Mladenic
Address: J. Stefan Institute, Jamova 39, 1000 Ljubljana, Slovenia
E-mail: Dunja.Mladenic@ijs.si
Phone: +386 1 4773 377

Marko Grobelnik
Address: J. Stefan Institute, Jamova 39, 1000 Ljubljana, Slovenia
E-mail: Marko.Grobelnik@ijs.si
Phone: +386 1 4773 778

Aims/Learning objectives;

The aim of this tutorial is to present topics from the areas of text mining and link analysis in the relationship to the web data. The goal is to show the whole list of nontrivial problems appearing in everyday life and occasionally in professional work with the web and to show how they can be approached using text mining and link analysis techniques and tools. The goal is to make an overview of the available approaches, which are potentially useful for solving interesting problems connected to the documents and their linkage coming from the web structure.

Duration (half or full day)

Half day, but it could be scaled to full day

Scope (general topic area) and why it is relevant for WWW2004;

The tutorial's relevance for the WWW2004 is in the presentation of analytic approaches used on the web data (text+links). In particular, the tutorial will focus on the possibilities offered by two very active and relevant subfields of data mining: text mining and link analysis. The relevance of these topics to the WWW2004 public is in extending possible activities, which could be used in shaping, understanding and potentially predicting the static and dynamic nature of the web. Analysis of such data offers typically new insights in the nature of the complex web data. Suitability of the tutorial for the WWW2004

Visualization of a single document

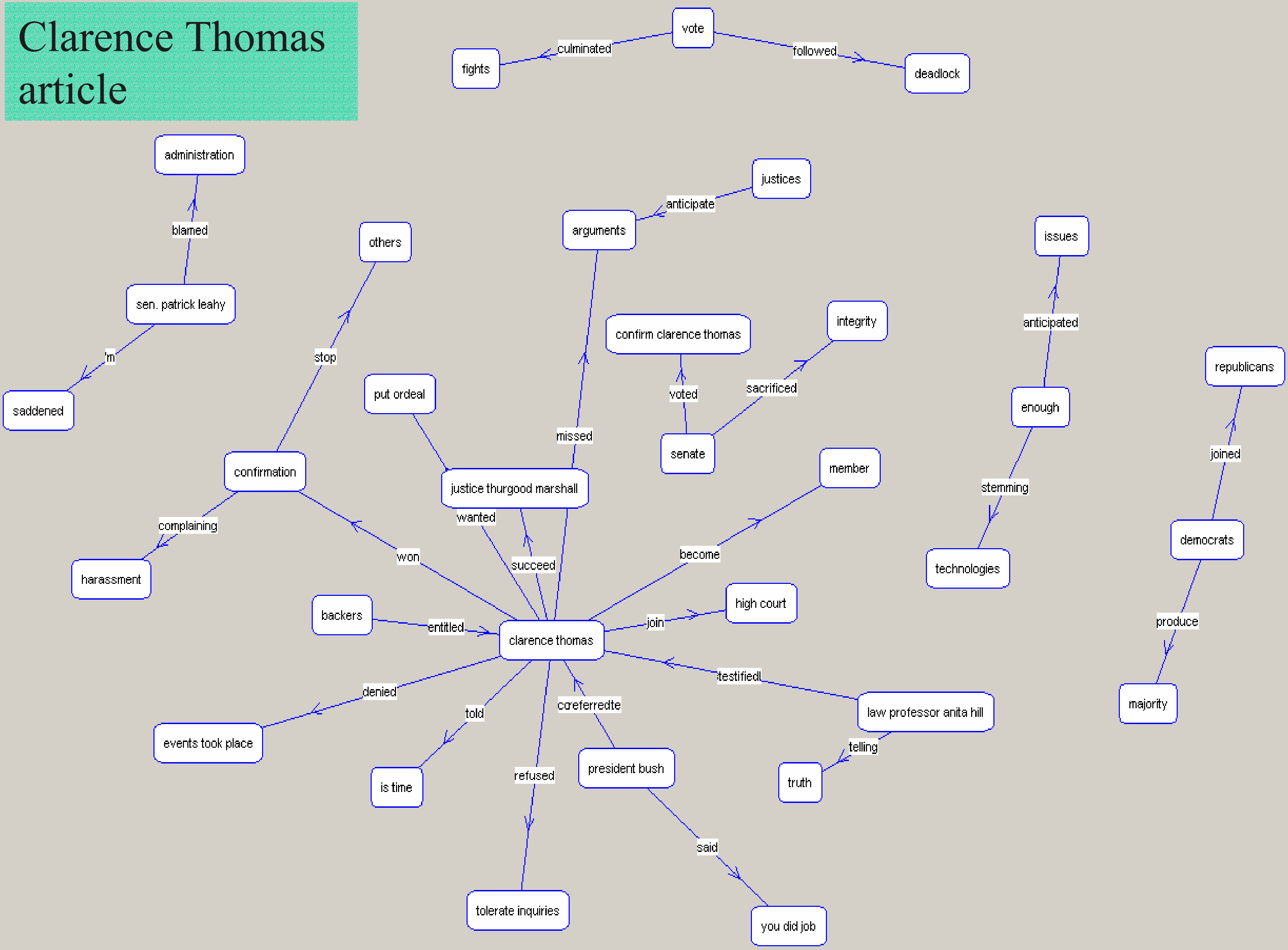
Why visualization of a single document is hard?

- Visualizing of big text corpora is easier task because of the big amount of information:
  - ...statistics already starts working
  - ...most known approaches are statistics based
- Visualization of a single (possibly short) document is much harder task because:
  - ...we can not count of statistical properties of the text (lack of data)
  - ...we must rely on syntactical and logical structure of the document

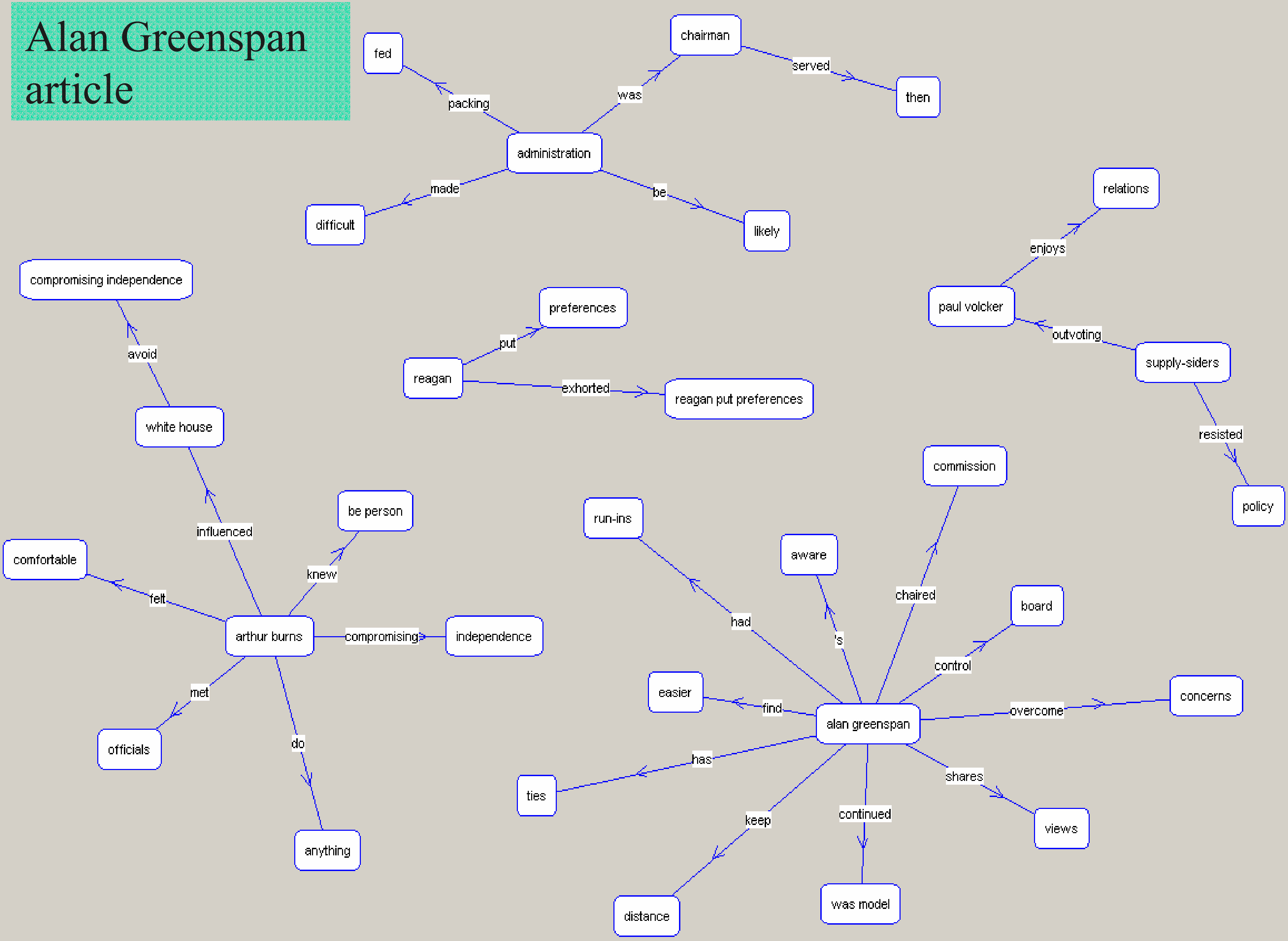
Simple approach

1. The text is split into the sentences.
2. Each sentence is deep-parsed into its logical form
   - we are using Microsoft's NLPWin parser
3. Anaphora resolution is performed on all sentences
   - ...all 'he', 'she', 'they', 'him', 'his', 'her', etc. references to the objects are replaced by its proper name
4. From all the sentences we extract [Subject-Predicate-Object triples] (SPO)
5. SPOs form links in the graph
6. ...finally, we draw a graph.

Clarence Thomas article



Alan Greenspan article



Text Segmentation

Text Segmentation

- **Problem**: divide text that has no given structure into segments with similar content
- Example applications:
  - topic tracking in news (spoken news)
  - identification of topics in large, unstructured text databases

Algorithm for text segmentation

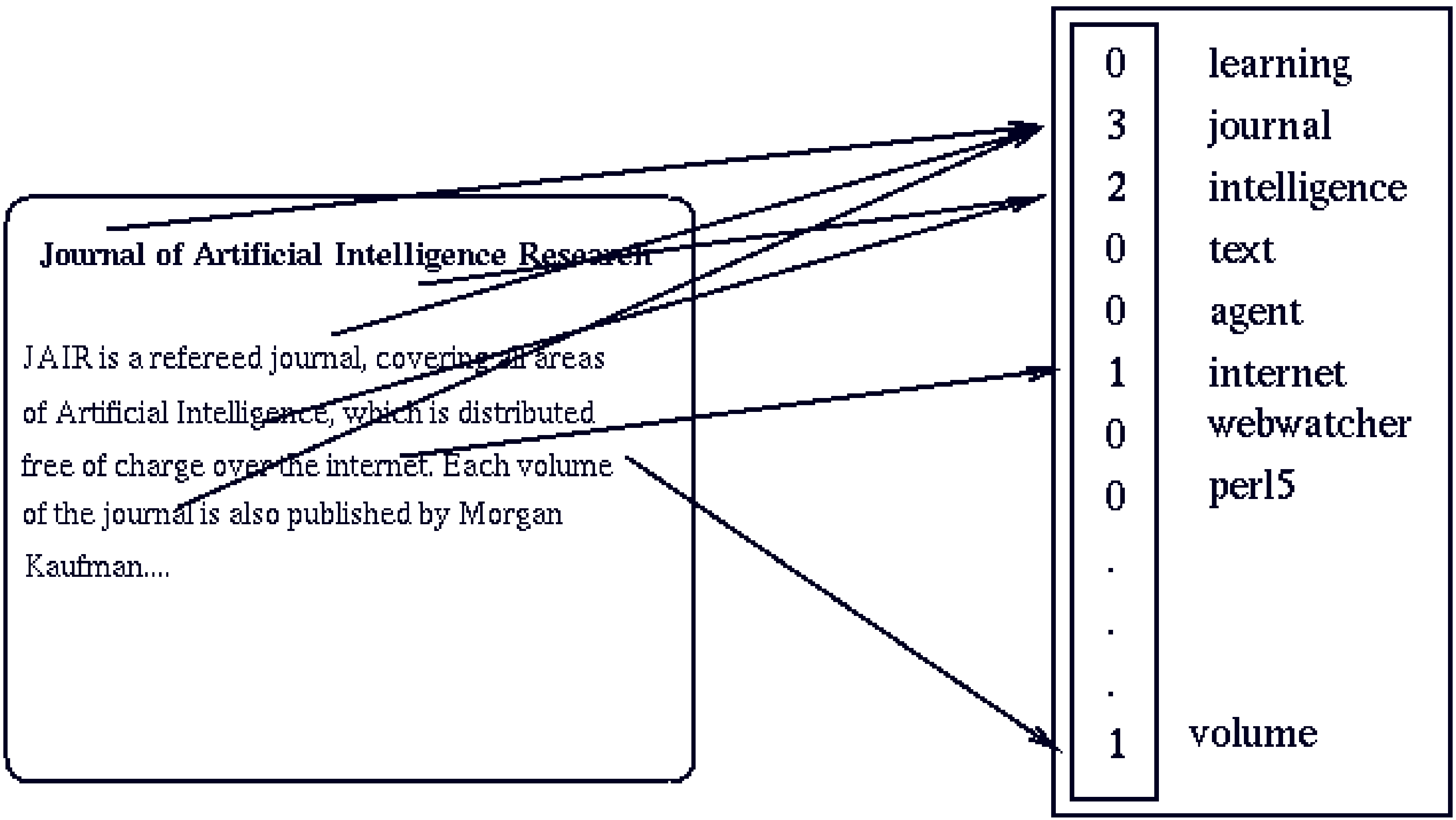
- **Algorithm:**
  - Divide text into sentences
  - Represent each sentence with words and phrases it contains
  - Calculate similarity between the pairs of sentences
  - Find a segmentation (sequence of delimiters), so that the similarity between the sentences inside the same segment is maximized and minimized between the segments
- …the approach can be defined either as optimization problem or as sliding window

Levels of Text Processing 4/6

- Word Level
- Sentence Level
- Document Level
- **Document-Collection Level**
  - **Representation**
  - **Feature Selection**
  - **Document Similarity**
  - **Representation Change (LSI)**
  - **Categorization (flat, hierarchical)**
  - **Clustering (flat, hierarchical)**
  - **Visualization**
  - **Information Extraction**
- Linked-Document-Collection Level
- Application Level

Representation

Bag-of-words document representation



Word weighting

- In bag-of-words representation each word is represented as a separate variable having numeric weight.
- The most popular weighting schema is normalized word frequency TFIDF:

$$tfidf(w) = tf \, . \log(\frac{N}{df(w)})$$

- Tf(w) – term frequency (number of word occurrences in a document)
- Df(w) – document frequency (number of documents containing the word)
- N – number of all documents
- Tfidf(w) – relative importance of the word in the document

The word is more important if it appears several times in a target document

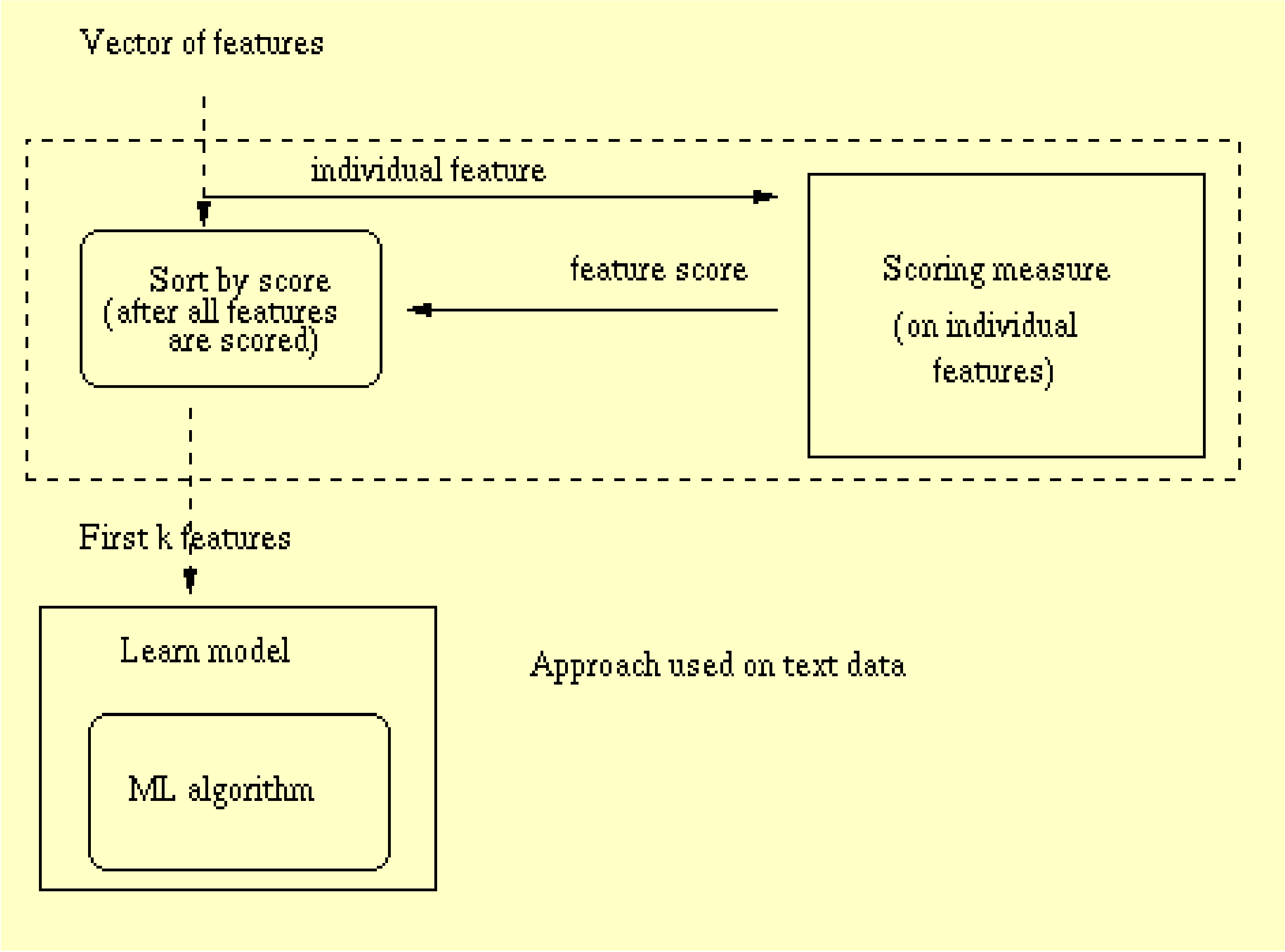
The word is more important if it appears in less documents

Example document and its vector representation

- TRUMP MAKES BID FOR CONTROL OF RESORTS Casino owner and real estate Donald Trump has offered to acquire all Class B common shares of Resorts International Inc, a spokesman for Trump said. The estate of late Resorts chairman James M. Crosby owns 340,783 of the 752,297 Class B shares. Resorts also has about 6,432,000 Class A common shares outstanding. Each Class B share has 100 times the voting power of a Class A share, giving the Class B stock about 93 pct of Resorts' voting power.
- [RESORTS:0.624] [CLASS:0.487] [TRUMP:0.367] [VOTING:0.171] [ESTATE:0.166] [POWER:0.134] [CROSBY:0.134] [CASINO:0.119] [DEVELOPER:0.118] [SHARES:0.117] [OWNER:0.102] [DONALD:0.097] [COMMON:0.093] [GIVING:0.081] [OWNS:0.080] [MAKES:0.078] [TIMES:0.075] [SHARE:0.072] [JAMES:0.070] [REAL:0.068] [CONTROL:0.065] [ACQUIRE:0.064] [OFFERED:0.063] [BID:0.063] [LATE:0.062] [OUTSTANDING:0.056] [SPOKESMAN:0.049] [CHAIRMAN:0.049] [INTERNATIONAL:0.041] [STOCK:0.035] [YORK:0.035] [PCT:0.022] [MARCH:0.011]

Feature Selection

Feature subset selection



Feature subset selection

- Select only the best features (different ways to define "the best"-different feature scoring measures)
  - the most frequent
  - the most informative relative to the all class values
  - the most informative relative to the positive class value,…

Scoring individual feature

- InformationGain: $$\sum_{F=W,\overline{W}} P(F) \sum_{C=pos,neg} P(C|F)\log\frac{P(C|F)}{P(C)}$$
- CrossEntropyTxt: $$P(W) \sum_{C=pos,neg} P(C|W)\log\frac{P(C|W)}{P(C)}$$
- MutualInfoTxt: $$\sum_{C=pos,neg} P(C)\log\frac{P(W|C)}{P(W)}$$

- WeightOfEvidTxt: $$\sum_{C=pos,neg} P(C)P(W)\left|\log\frac{P(C|W)(1-P(C)}{P(C)(1-P(C|W)}\right|$$
- OddsRatio: $$\log\frac{P(W|pos)\times(1-P(W|neg))}{(1-P(W|pos))\times P(W|neg)}$$
- Frequency: $$Freq(W)$$

Example of the best features

Odds Ratio

feature score [P(F|pos), P(F|neg)]

IR 5.28 [0.075, 0.0004]

INFORMATION RETRIEVAL 5.13...

RETRIEVAL 4.77 [0.075, 0.0007]

GLASGOW 4.72 [0.03, 0.0003]

ASIA 4.32 [0.03, 0.0004]

PACIFIC 4.02 [0.015, 0.0003]

INTERESTING 4.02[0.015, 0.0003]

EMPHASIS 4.02 [0.015, 0.0003]

GROUP 3.64 [0.045, 0.0012]

MASSACHUSETTS 3.46 [0.015, ...]

COMMERCIAL 3.46 [0.015,0.0005]

REGION 3.1 [0.015, 0.0007]

Information Gain

feature score [P(F|pos), P(F|neg)]

LIBRARY 0.46 [0.015, 0.091]

PUBLIC 0.23 [0, 0.034]

PUBLIC LIBRARY 0.21 [0, 0.029]

UNIVERSITY 0.21 [0.045, 0.028]

LIBRARIES 0.197 [0.015, 0.026]

INFORMATION 0.17 [0.119, 0.021]

REFERENCES 0.117 [0.015, 0.012]

RESOURCES 0.11 [0.029, 0.0102]

COUNTY 0.096 [0, 0.0089]

INTERNET 0.091 [0, 0.00826]

LINKS 0.091 [0.015, 0.00819]

SERVICES 0.089 [0, 0.0079]

Document Similarity

Cosine similarity between document vectors

- Each document is represented as a vector of weights D = <x>
- Similarity between vectors is estimated by the similarity between their vector representations (cosine of the angle between vectors):

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

Representation Change: Latent Semantic Indexing

Latent Semantic Indexing

- LSI is a statistical technique that attempts to estimate the hidden content structure within documents:
  - …it uses linear algebra technique Singular-Value-Decomposition (SVD)
  - …it discovers statistically most significant co-occurences of terms

LSI Example

Original document-term mantrix

	d1	d2	d3	d4	d5	d6
cosmonaut	1	0	1	0	0	0
astronaut	0	1	0	0	0	0
moon	1	1	0	0	0	0
car	1	0	0	1	1	0
truck	0	0	0	1	0	1

Rescaled document matrix,
Reduced into two dimensions

	d1	d2	d3	d4	d5	d6
Dim1	-1.62	-0.60	-0.04	-0.97	-0.71	-0.26
Dim2	-0.46	-0.84	-0.30	1.00	0.35	0.65

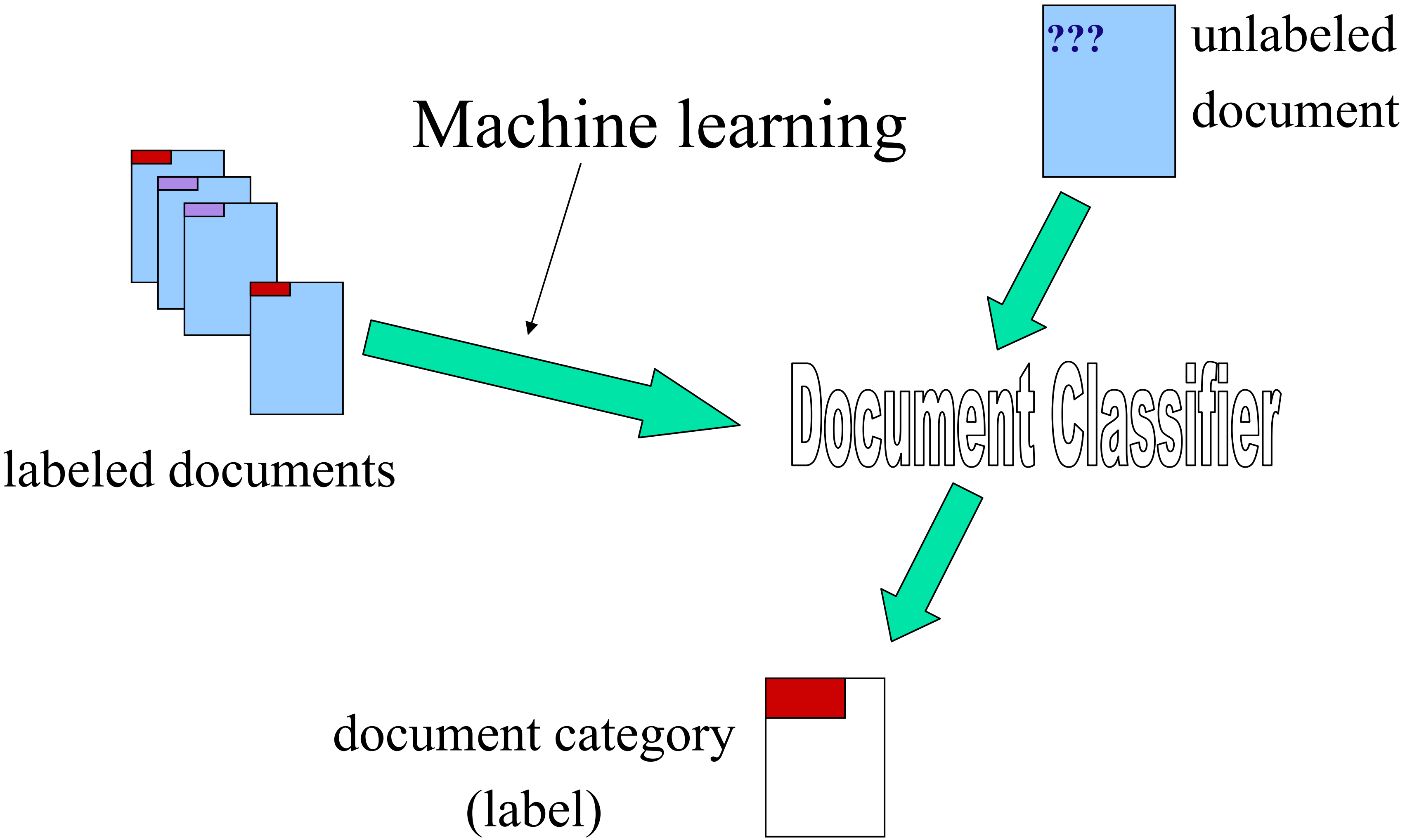
High correlation although d2 and d3 don't share any word

Correlation matrix

	d1	d2	d3	d4	d5	d6
d1	1.00					
d2	0.8	1.00				
d3	0.4	0.9	1.00			
d4	0.5	-0.2	-0.6	1.00		
d5	0.7	0.2	-0.3	0.9	1.00	
d6	0.1	-0.5	-0.9	0.9	0.7	1.00

Text Categorization

Document categorization



Automatic Document Categorization Task

- Given is a set of documents labeled with content categories.
- The goal is: 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)

Algorithms for learning document classifiers

- Popular algorithms for text categorization:
  - Support Vector Machines
  - Logistic Regression
  - Perceptron algorithm
  - Naive Bayesian classifier
  - Winnow algorithm
  - Nearest Neighbour
  - ....

Perceptron algorithm

Input: set of pre-classified documents

Output: model, one weight for each word from the vocabulary

Algorithm:

- initialize the model by setting word weights to 0
- iterate through documents N times
  - classify the document X represented as bag-of-words

    $\sum_{i=1}^{V} x_i w_i \geq 0$ predict positive class

    else predict negative class
  - if document classification is wrong then adjust weights of all words occurring in the document

$w_{t+1} = w_t + sign(trueClass)\beta; \beta > 0$

sign(positive) = 1

sign(negative) =-1

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

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

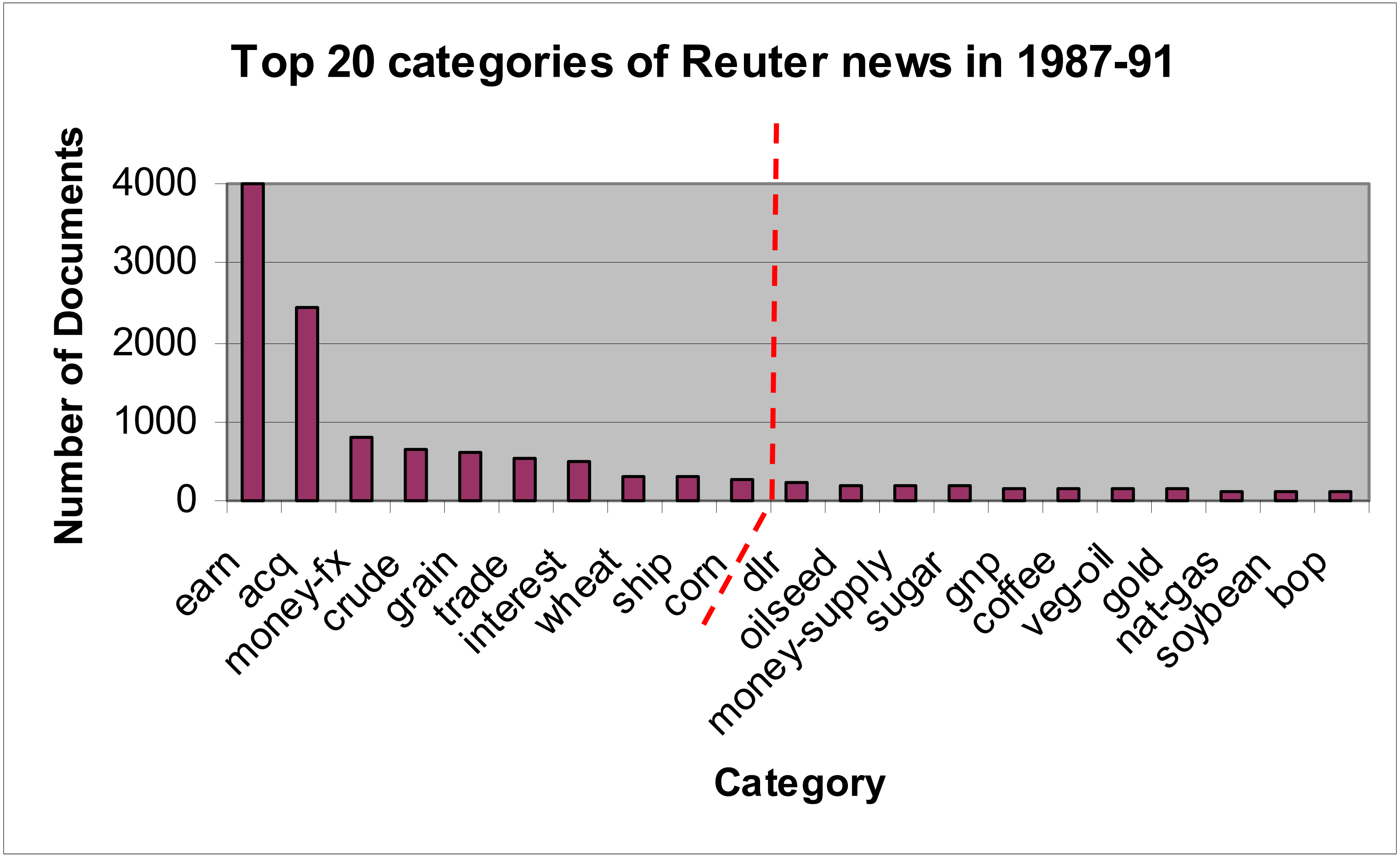
$$F_\beta(M, targetC) = \frac{(1+\beta^2) Precision(M, targetC) \times Recall(M, targetC)}{\beta^2 Precision(M, targetC) + Recall(M, targetC)}$$

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

Reuters dataset – Categorization to flat categories

- Documents 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,...*
- ...from 2000 is available new dataset of 830,000 Reuters documents available fo research

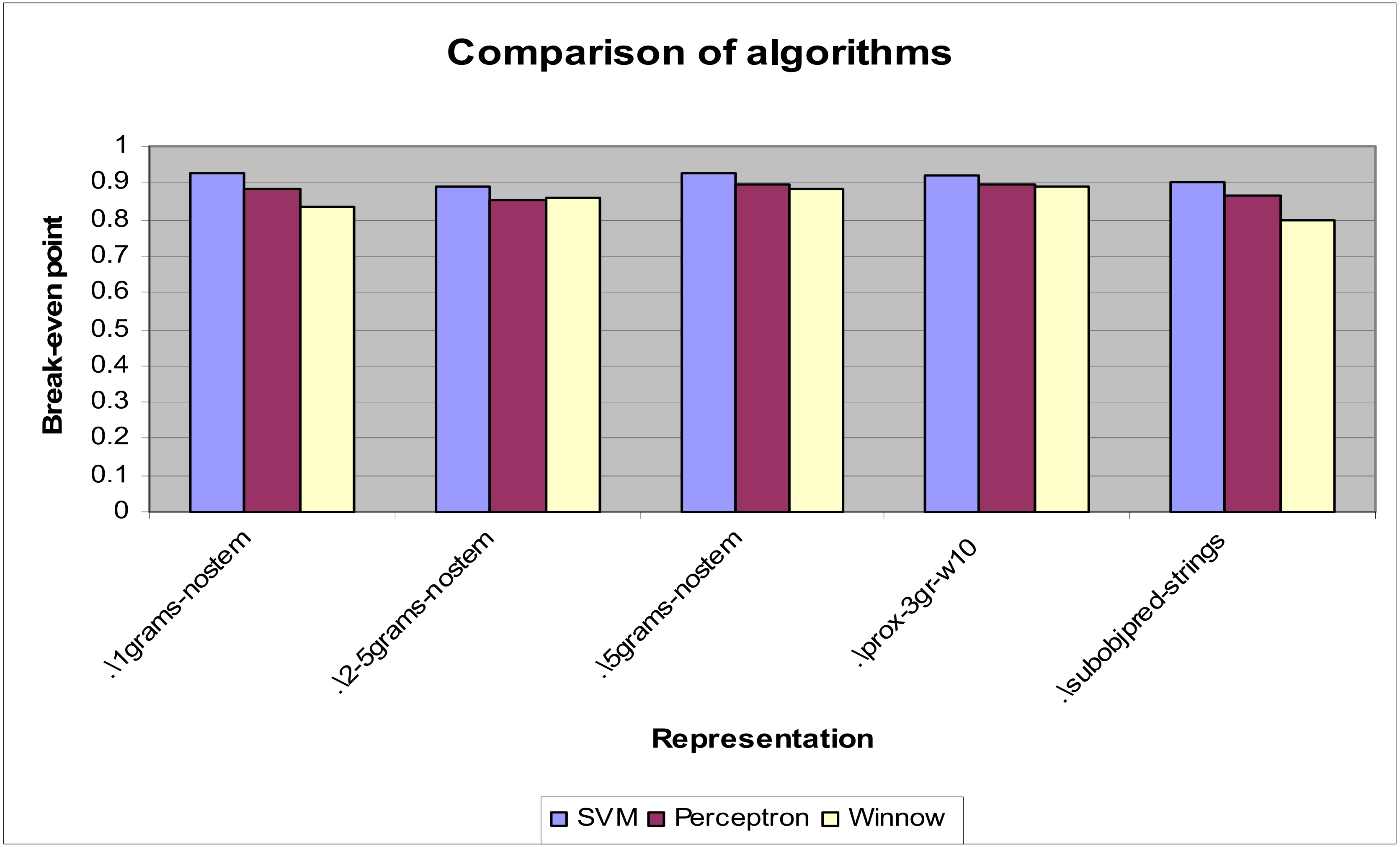
Distribution of documents (Reuters-21578)



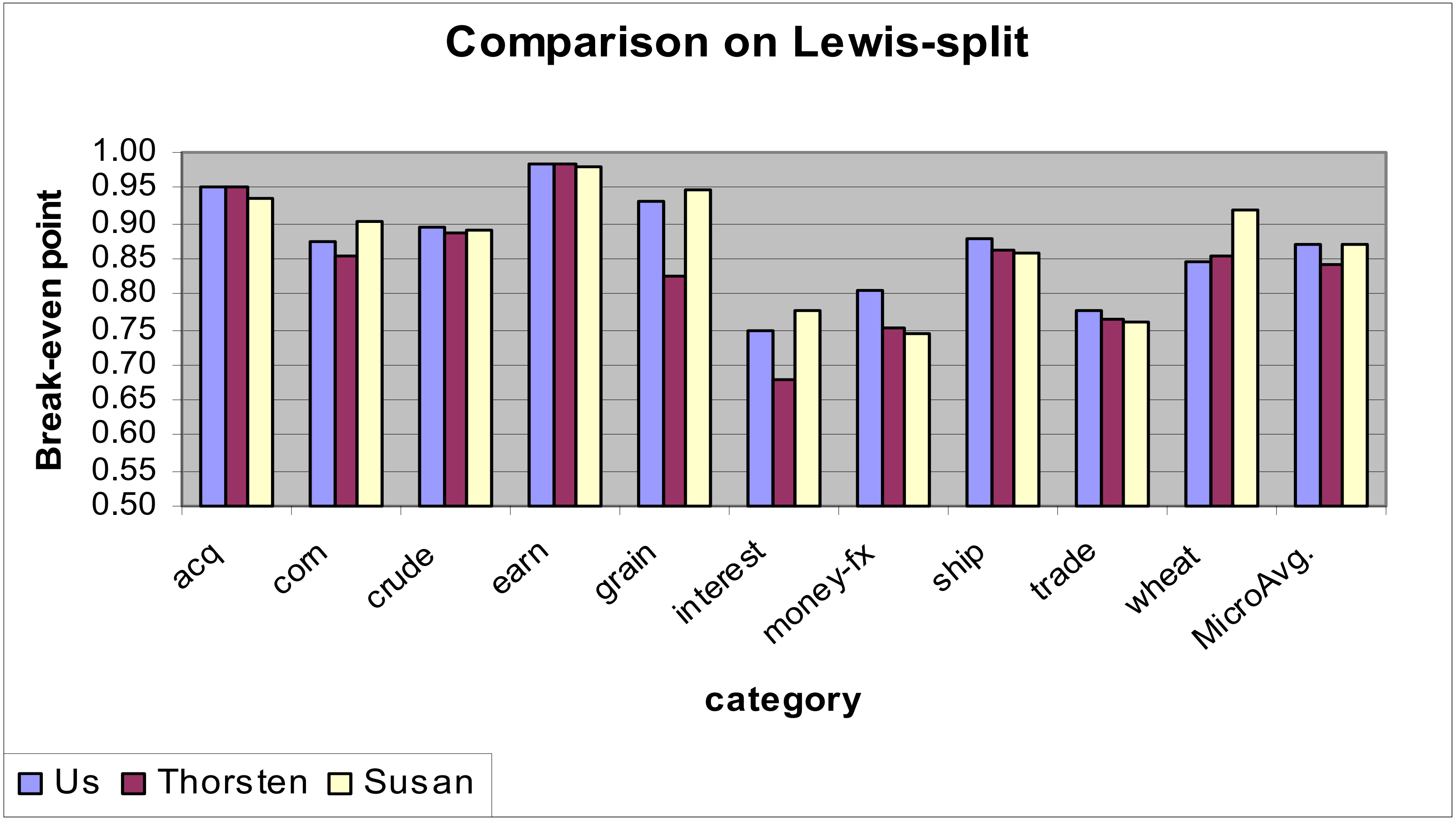
Example of Perceptron model for Reuters category “Acquisition”

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 on using SVM on stemmed 1-grams with related results

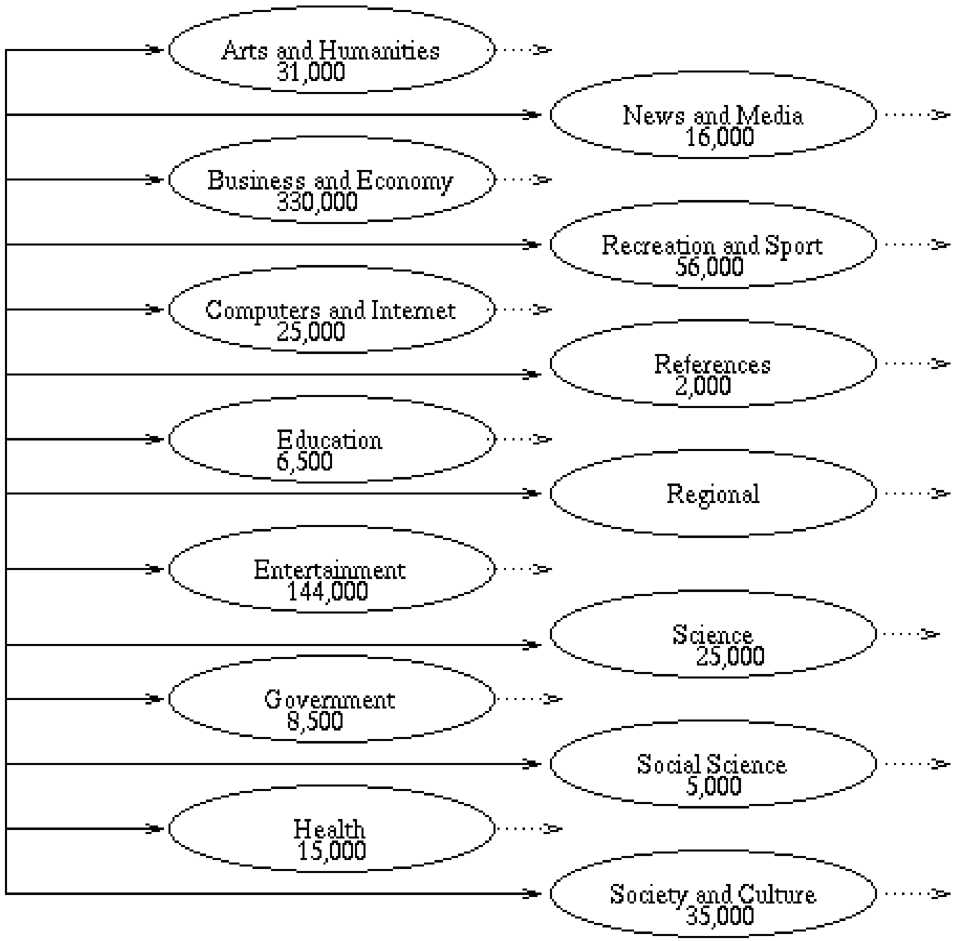


Text Categorization into hierarchy of categories

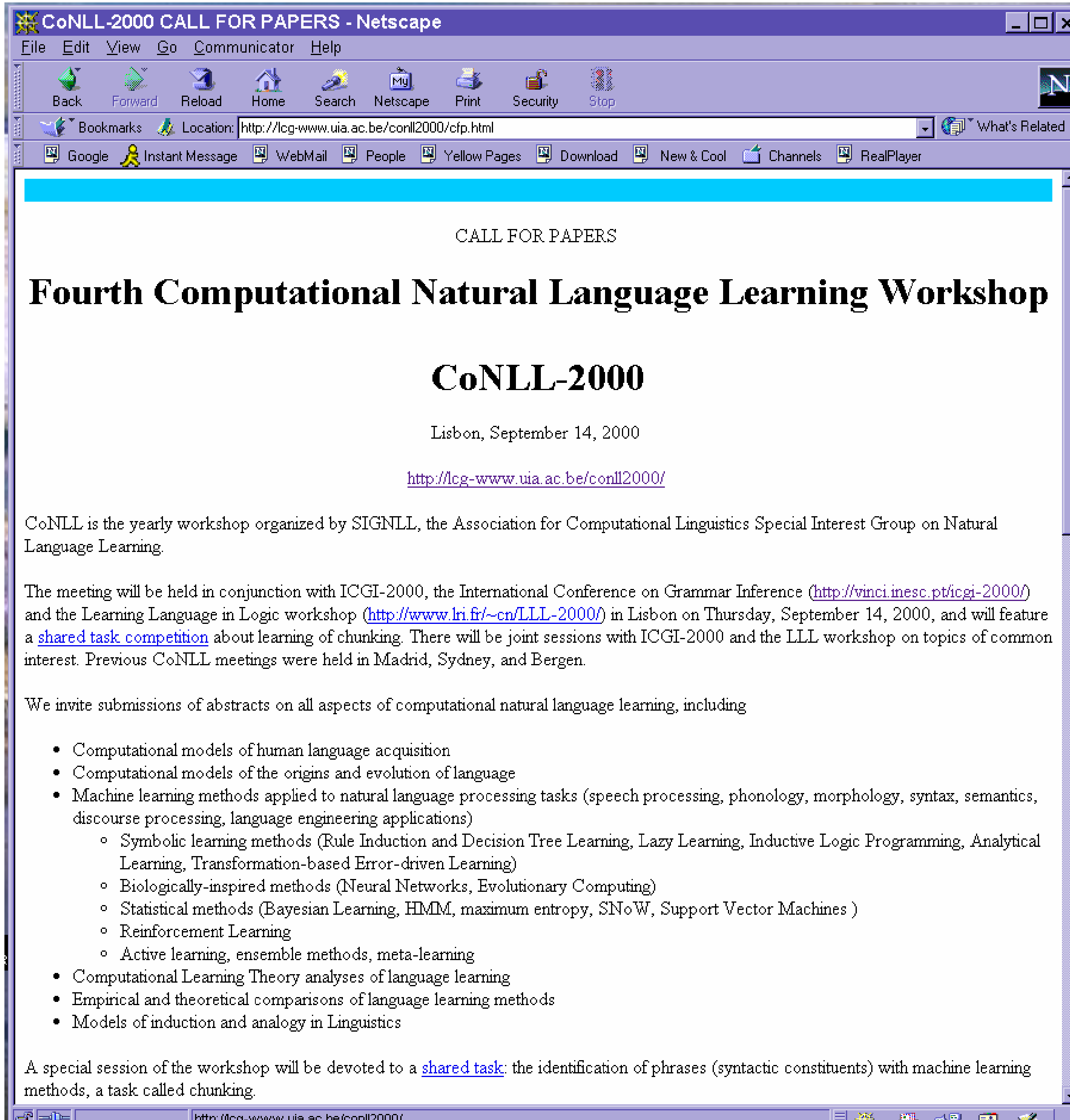
- There are several hierarchies (taxonomies) of textual documents:
  - Yahoo, DMoz, Medline, …
- Different people use different approaches:
  - …series of hierarchically organized classifiers
  - …set of independent classifiers just for leaves
  - …set of independent classifiers for all nodes

Yahoo! hierarchy (taxonomy)

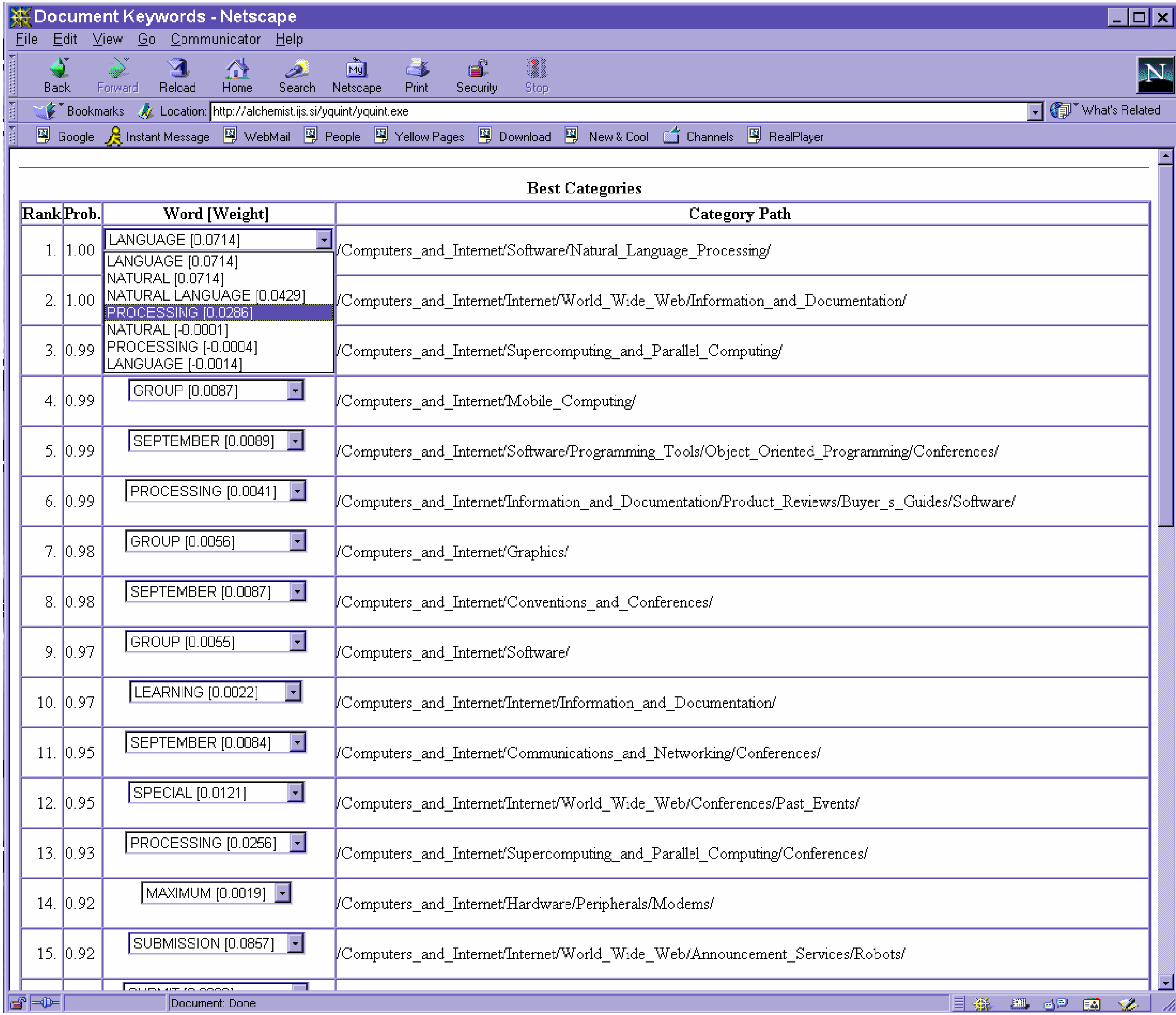
- human constructed hierarchy of Web-documents
- exists in several languages (we use English)
- easy to access and regularly updated
- captures most of the Web topics
- English version includes over 2M pages categorized into 50,000 categories
- contains about 250Mb of HTML files



Document to categorize:
CFP for CoNLL-2000



Some predicted categories

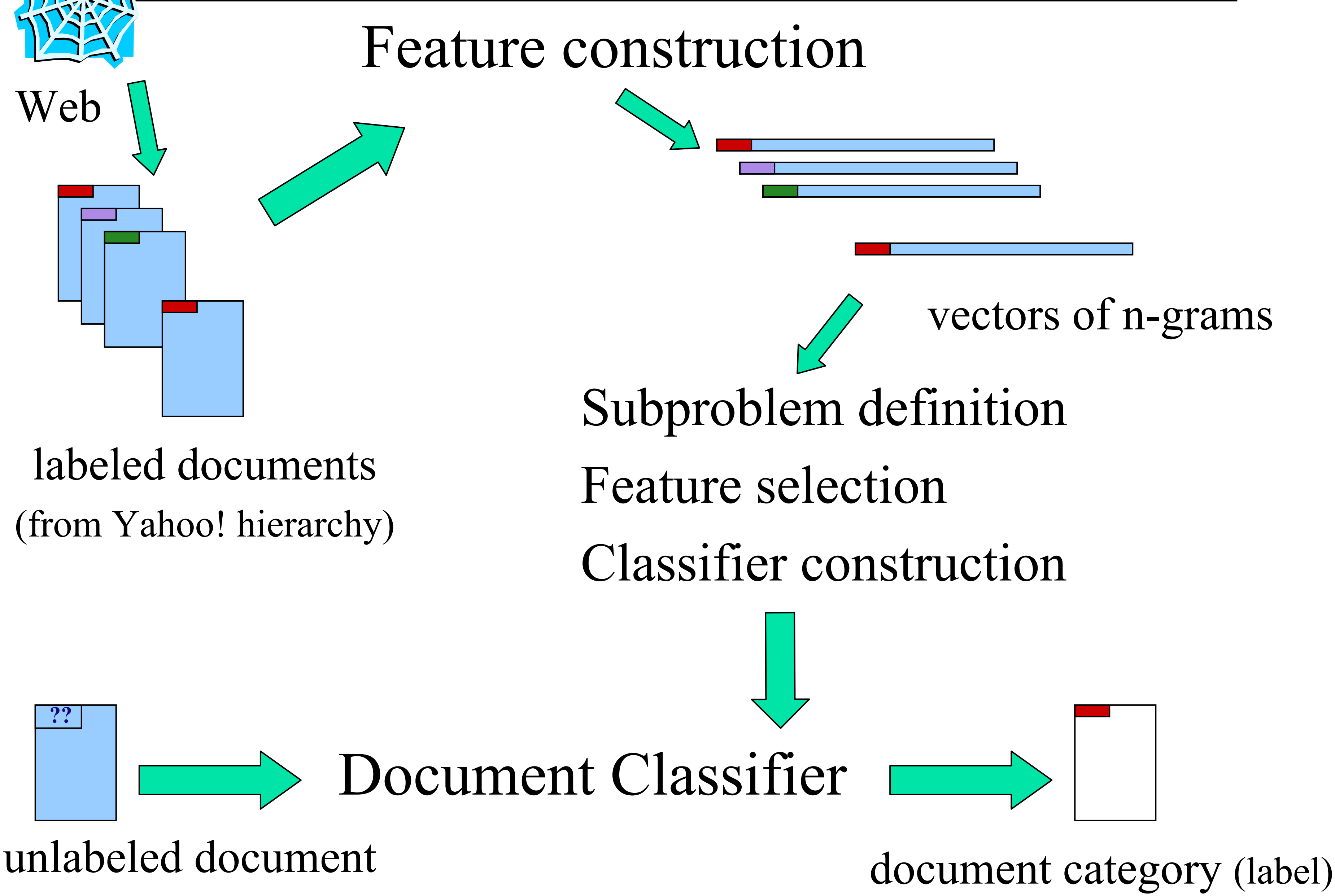


Best Categories

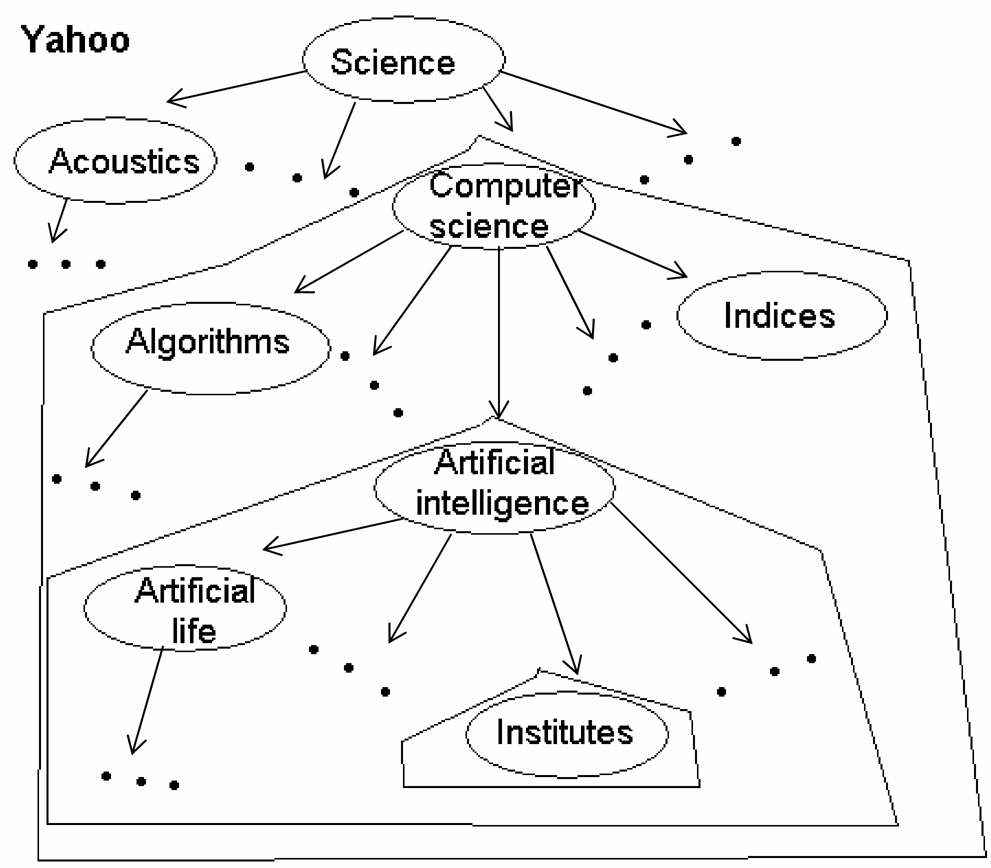
Rank	Prob.	Word [Weight]	Category Path
1.	1.00	LANGUAGE [0.0714]	/Computers_and_Internet/Software/Natural_Language_Processing/
2.	1.00		/Computers_and_Internet/Internet/World_Wide_Web/Information_and_Documentation/
3.	0.99		/Computers_and_Internet/Supercomputing_and_Parallel_Computing/
4.	0.99	GROUP [0.0087]	/Computers_and_Internet/Mobile_Computing/
5.	0.99	SEPTEMBER [0.0089]	/Computers_and_Internet/Software/Programming_Tools/Object_Oriented_Programming/Conferences/
6.	0.99	PROCESSING [0.0041]	/Computers_and_Internet/Information_and_Documentation/Product_Reviews/Buyer_s_Guides/Software/
7.	0.98	GROUP [0.0056]	/Computers_and_Internet/Graphics/
8.	0.98	SEPTEMBER [0.0087]	/Computers_and_Internet/Conventions_and_Conferences/
9.	0.97	GROUP [0.0055]	/Computers_and_Internet/Software/
10.	0.97	LEARNING [0.0022]	/Computers_and_Internet/Internet/Information_and_Documentation/
11.	0.95	SEPTEMBER [0.0084]	/Computers_and_Internet/Communications_and_Networking/Conferences/
12.	0.95	SPECIAL [0.0121]	/Computers_and_Internet/Internet/World_Wide_Web/Conferences/Past_Events/
13.	0.93	PROCESSING [0.0256]	/Computers_and_Internet/Supercomputing_and_Parallel_Computing/Conferences/
14.	0.92	MAXIMUM [0.0019]	/Computers_and_Internet/Hardware/Peripherals/Modems/
15.	0.92	SUBMISSION [0.0857]	/Computers_and_Internet/Internet/World_Wide_Web/Announcement_Services/Robots/

LANGUAGE [0.0714]
NATURAL [0.0714]
NATURAL LANGUAGE [0.0429]
PROCESSING [0.0286]
NATURAL [-0.0001]
PROCESSING [-0.0004]
LANGUAGE [-0.0014]

System architecture



Content categories



- For each content category generate a separate classifier that predicts probability for a new document to belong to its category

Considering promising categories only (classification by Naive Bayes)

$$P(C \mid Doc) = \frac{P(C) \prod_{W \in Doc} P(W \mid C)^{Freq(W, Doc)}}{\sum_i P(C_i) \prod_{W_l \in Doc} P(W_l \mid C_i)^{Freq(W_l, Doc)}}$$

- Document is represented as a set of word sequences W
- Each classifier has two distributions: P(W|pos), P(W|neg)
- Promising category:
  - calculated P(pos|Doc) is high meaning that the classifier has P(W|pos)>0 for at least some W from the document (otherwise, the prior probability is returned, P(neg) is about 0.90)

Summary of experimental results

Domain	probability	rank	precision	recall
Entertain.	0.96	16	0.44	0.80
Arts	0.99	10	0.40	0.83
Computers	0.98	12	0.40	0.84
Education	0.99	9	0.57	0.65
Reference	0.99	3	0.51	0.81

Document Clustering

Document Clustering

- Clustering is a process of finding natural groups in data in a unsupervised way (no class labels preassigned to documents)
- Most popular clustering methods are:
  - K-Means clustering
  - Agglomerative hierarchical clustering
  - EM (Gaussian Mixture)
  - …

K-Means clustering

- Given:
  - set of documents (e.g. TFIDF vectors),
  - 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)

Visualization

Why text visualization?

- ...to have a top level view of the topics in the corpora
- ...to see relationships between the topics in the corpora
- ...to understand better what's going on in the corpora
- ...to show highly structured nature of textual contents in a simplified way
- ...to show main dimensions of highly dimensional space of textual documents
- ...because it's fun!

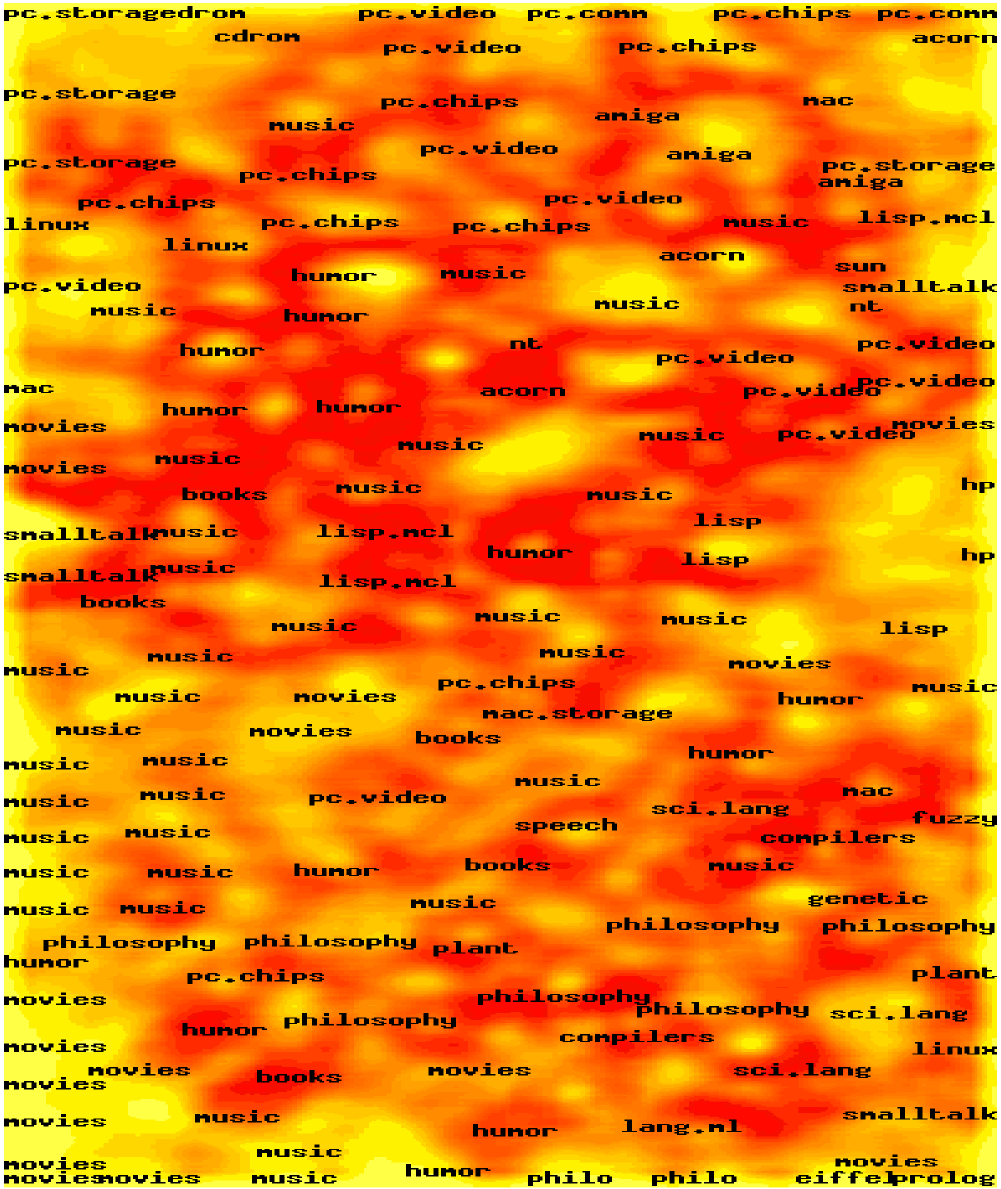
Examples of Text Visualization

- Text visualizations
  - WebSOM
  - ThemeScape
  - Graph-Based Visualization
  - Tiling-Based Visualization
  - …
- … collection of approaches at http://nd.loopback.org/hyperd/zb/

WebSOM

- Self-Organizing Maps for Internet Exploration
  - An ordered map of the information space is provided: similar documents lie near each other on the map
  - …algorithm that automatically organizes the documents onto a two-dimensional grid so that related documents appear close to each other
  - … based on Kohonen's Self-Organizing Maps
  - Demo at http://websom.hut.fi/websom/

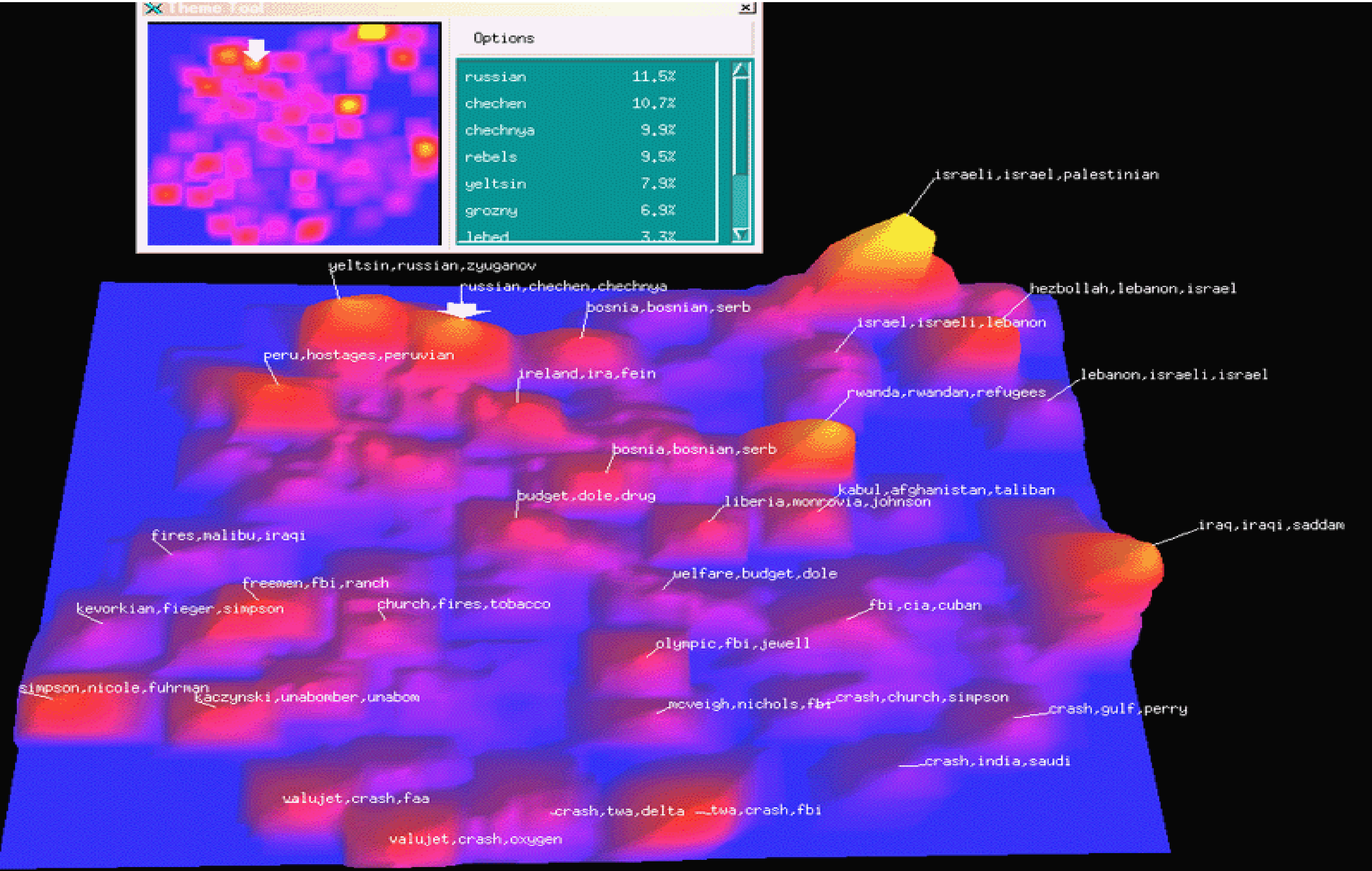
WebSOM visualization



ThemeScape

- Graphically displays images based on word similarities and themes in text
- Themes within the document spaces appear on the computer screen as a relief map of natural terrain
  - The mountains in indicate where themes are dominant - valleys indicate weak themes
  - Themes close in content will be close visually based on the many relationships within the text spaces.
- … similar techniques for visualizing stocks (http://www.webmap.com./trademapdemo.html)

ThemeScape Document visualization

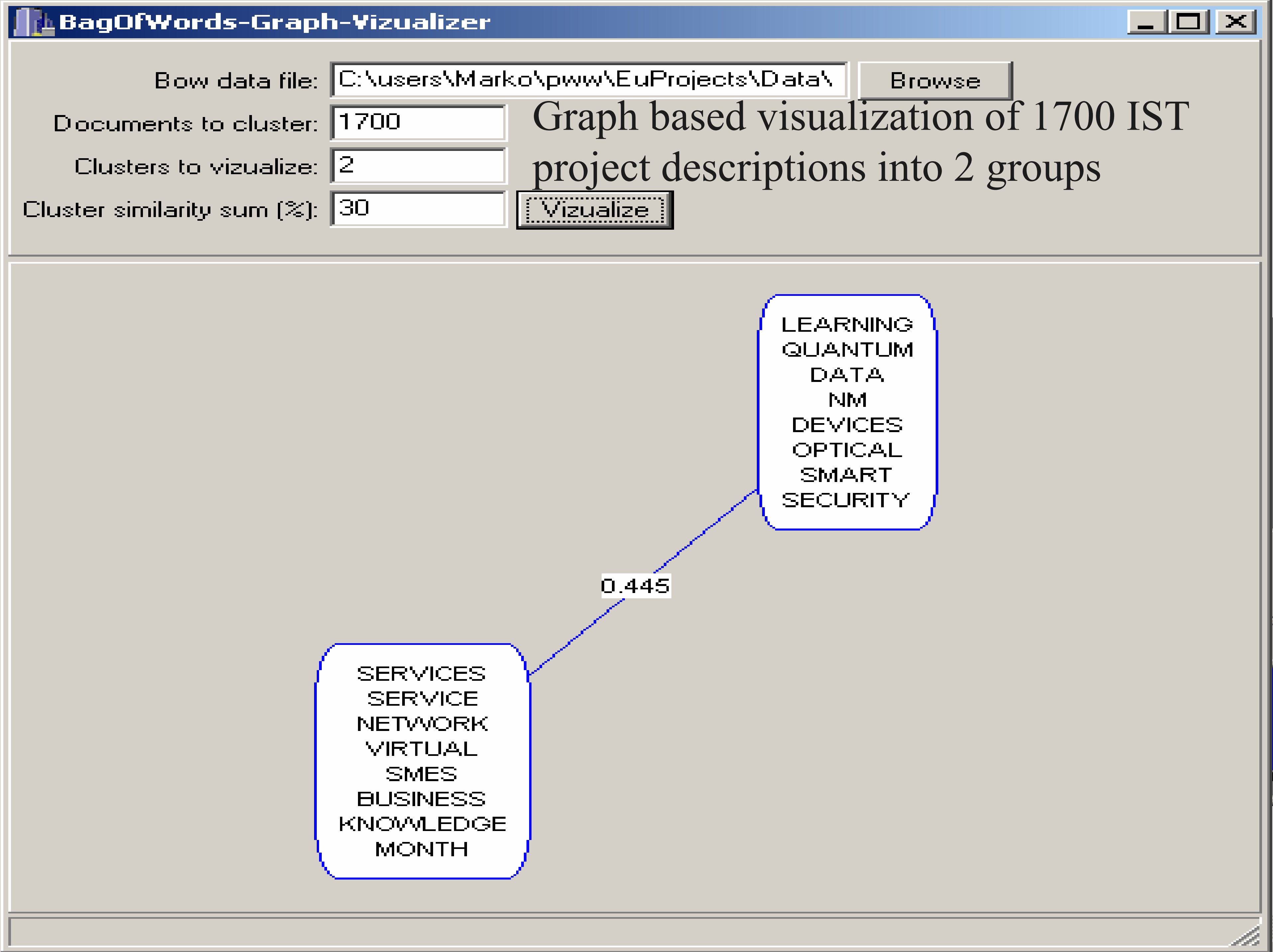


Graph based visualization

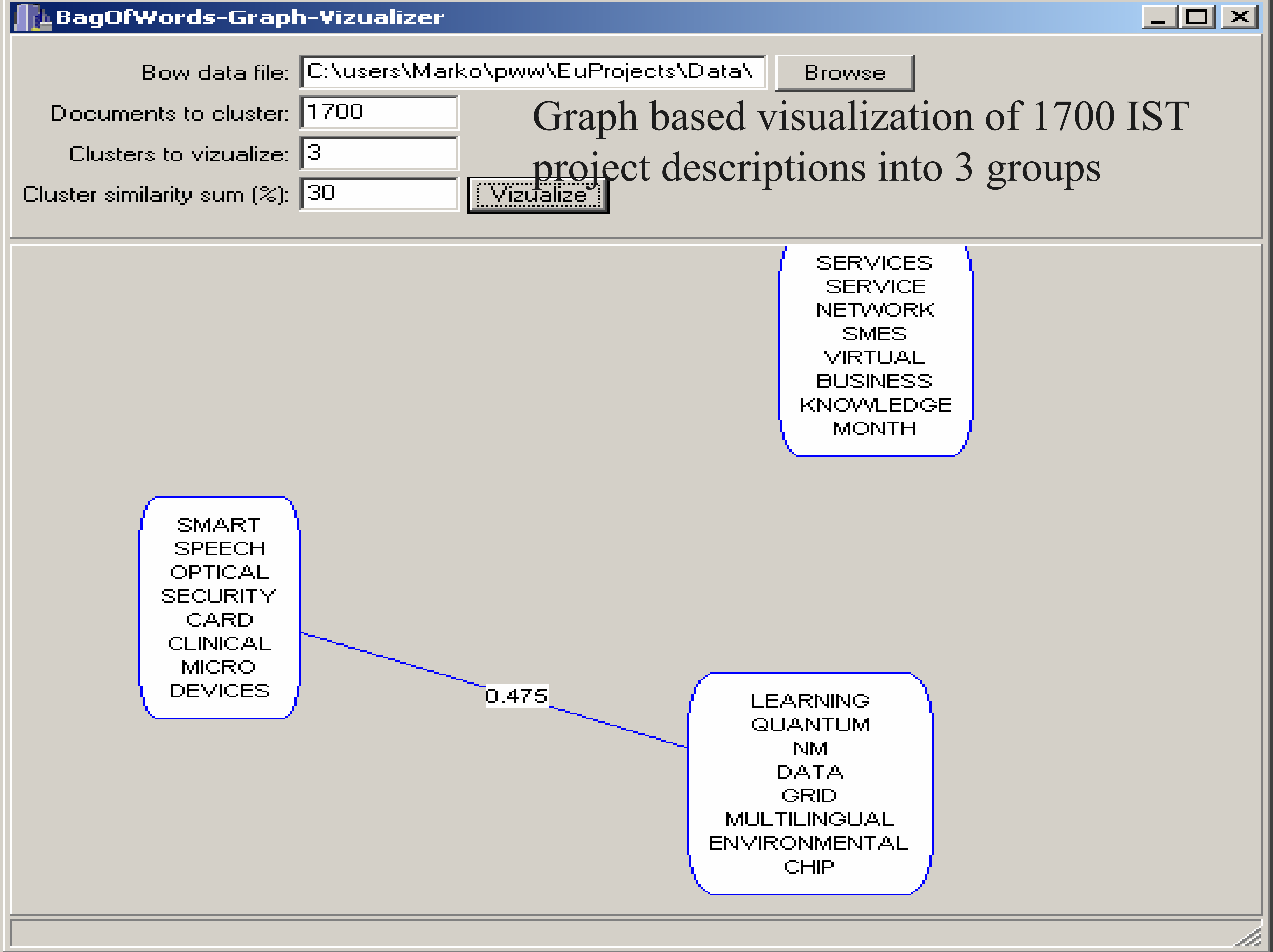
- The sketch of the algorithm:
  1. Documents are transformed into the bag-of-words sparse-vectors representation
     - Words in the vectors are weighted using TFIDF
  2. K-Means clustering algorithm splits the documents into K groups
     - Each group consists from similar documents
     - Documents are compared using cosine similarity
  3. K groups form a graph:
     - Groups are nodes in graph; similar groups are linked
     - Each group is represented by characteristic keywords
  4. Using simulated annealing draw a graph

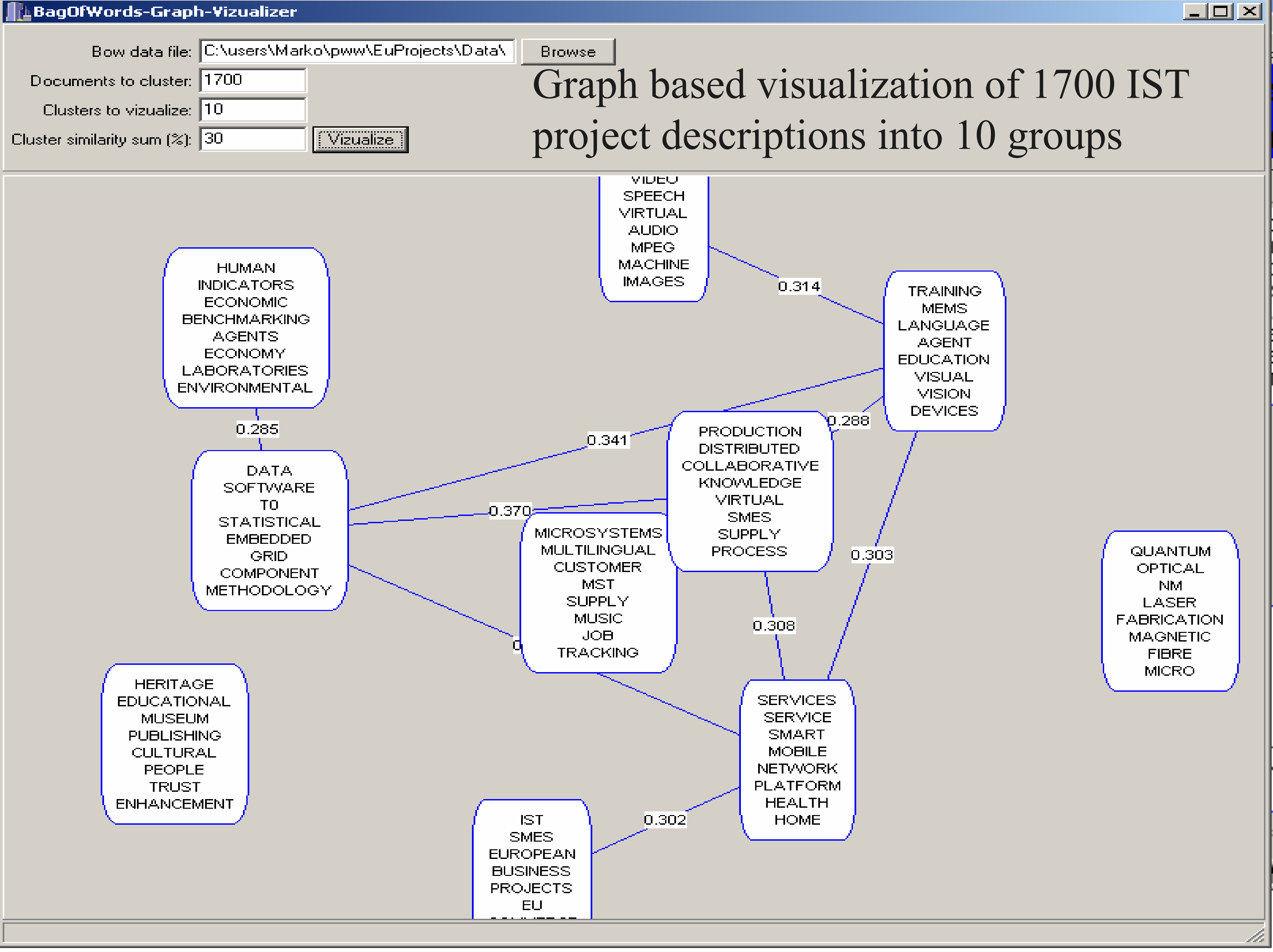
Example of visualizing Eu IST projects corpora

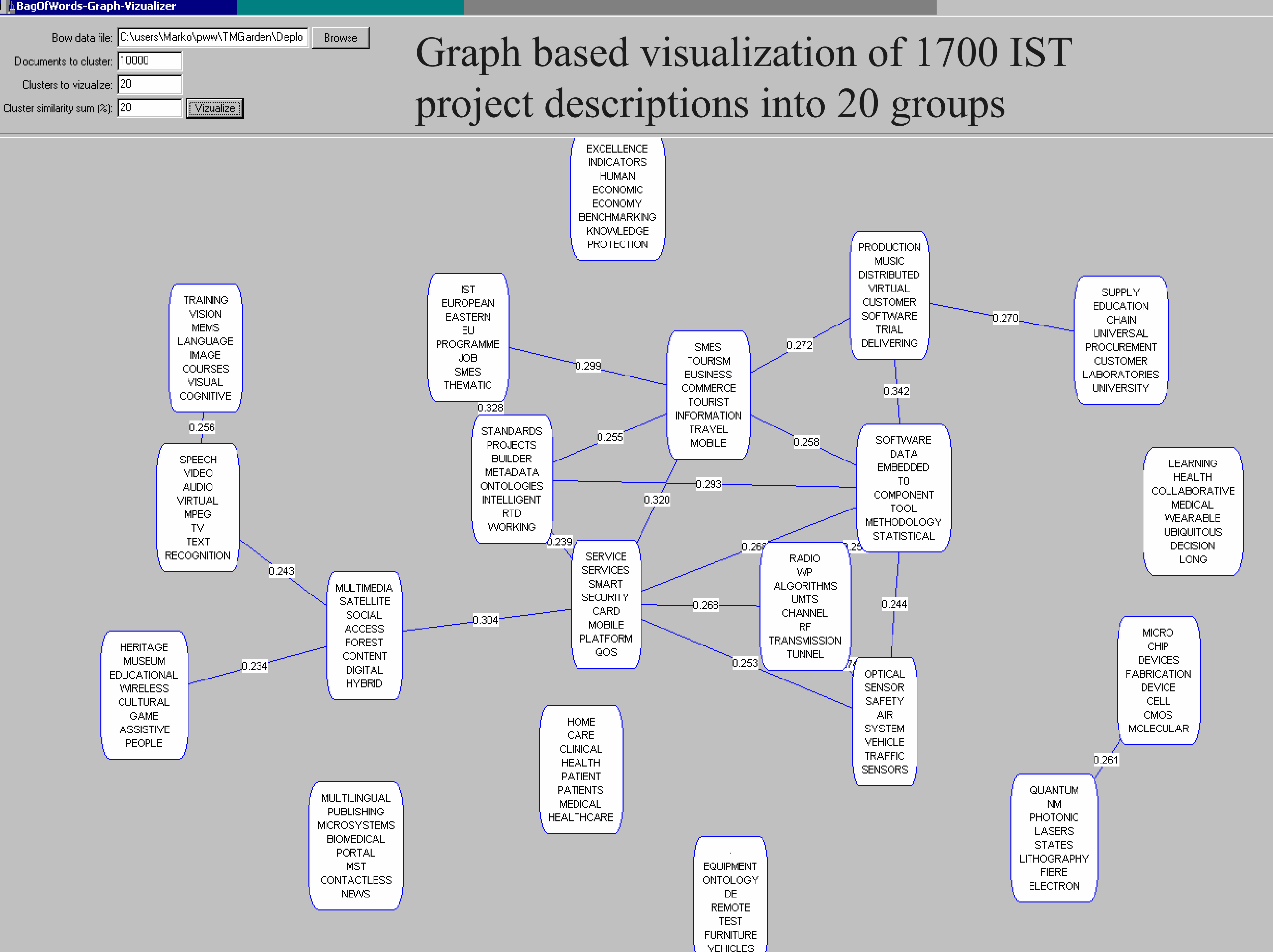
- Corpus of 1700 Eu IST projects descriptions
  - Downloaded from the web http://www.cordis.lu/
  - Each document is few hundred words long describing one project financed by EC
  - ...the idea is to understand the structure and relations between the areas EC is funding through the projects
- ...the following slides show different visualizations with the graph based approach


BagOfWords-Graph-Vizualizer
Bow data file:
C:\users\Marko\pww\EuProjects\Data\
Browse
Documents to cluster:
1700
Clusters to vizualize:
2
Cluster similarity sum (%):
30
Vizualize
Graph based visualization of 1700 IST project descriptions into 2 groups
LEARNING
QUANTUM
DATA
NM
DEVICES
OPTICAL
SMART
SECURITY
0.445
SERVICES
SERVICE
NETWORK
VIRTUAL
SMES
BUSINESS
KNOWLEDGE
MONTH

Graph based visualization of 1700 IST project descriptions into 3 groups




BagOfWords-Graph-Vizualizer
Bow data file: C:\users\Marko\pww\EuProjects\Data\
Browse
Documents to cluster: 1700
Clusters to vizualize: 10
Cluster similarity sum (%): 30
Vizualize
Graph based visualization of 1700 IST project descriptions into 10 groups
VIDEO SPEECH VIRTUAL AUDIO MPEG MACHINE IMAGES
HUMAN INDICATORS ECONOMIC BENCHMARKING AGENTS ECONOMY LABORATORIES ENVIRONMENTAL
TRAINING MEMS LANGUAGE AGENT EDUCATION VISUAL VISION DEVICES
0.314
0.285
0.341
0.288
DATA SOFTWARE TO STATISTICAL EMBEDDED GRID COMPONENT METHODOLOGY
PRODUCTION DISTRIBUTED COLLABORATIVE KNOWLEDGE VIRTUAL SMES SUPPLY PROCESS
0.370
MICROSYSTEMS MULTILINGUAL CUSTOMER MST SUPPLY MUSIC JOB TRACKING
0.303
QUANTUM OPTICAL NM LASER FABRICATION MAGNETIC FIBRE MICRO
0.308
HERITAGE EDUCATIONAL MUSEUM PUBLISHING CULTURAL PEOPLE TRUST ENHANCEMENT
SERVICES SERVICE SMART MOBILE NETWORK PLATFORM HEALTH HOME
0.302
IST SMES EUROPEAN BUSINESS PROJECTS EU


BagOfWords-Graph-Vizualizer
Bow data file: C:\users\Marko\pww\TMGarden\Deplo
Browse
Documents to cluster: 10000
Clusters to vizualize: 20
Cluster similarity sum (%): 20
Vizualize
Graph based visualization of 1700 IST project descriptions into 20 groups
EXCELLENCE INDICATORS HUMAN ECONOMIC ECONOMY BENCHMARKING KNOWLEDGE PROTECTION
PRODUCTION MUSIC DISTRIBUTED VIRTUAL CUSTOMER SOFTWARE TRIAL DELIVERING
SUPPLY EDUCATION CHAIN UNIVERSAL PROCUREMENT CUSTOMER LABORATORIES UNIVERSITY
TRAINING VISION MEMS LANGUAGE IMAGE COURSES VISUAL COGNITIVE
IST EUROPEAN EASTERN EU PROGRAMME JOB SMES THEMATIC
SMES TOURISM BUSINESS COMMERCE TOURIST INFORMATION TRAVEL MOBILE
STANDARDS PROJECTS BUILDER METADATA ONTOLOGIES INTELLIGENT RTD WORKING
SOFTWARE DATA EMBEDDED TO COMPONENT TOOL METHODOLOGY STATISTICAL
LEARNING HEALTH COLLABORATIVE MEDICAL WEARABLE UBIQUITOUS DECISION LONG
SPEECH VIDEO AUDIO VIRTUAL MPEG TV TEXT RECOGNITION
SERVICE SERVICES SMART SECURITY CARD MOBILE PLATFORM QOS
RADIO WP ALGORITHMS UMTS CHANNEL RF TRANSMISSION TUNNEL
MULTIMEDIA SATELLITE SOCIAL ACCESS FOREST CONTENT DIGITAL HYBRID
HERITAGE MUSEUM EDUCATIONAL WIRELESS CULTURAL GAME ASSISTIVE PEOPLE
OPTICAL SENSOR SAFETY AIR SYSTEM VEHICLE TRAFFIC SENSORS
MICRO CHIP DEVICES FABRICATION DEVICE CELL CMOS MOLECULAR
HOME CARE CLINICAL HEALTH PATIENT PATIENTS MEDICAL HEALTHCARE
QUANTUM NM PHOTONIC LASERS STATES LITHOGRAPHY FIBRE ELECTRON
MULTILINGUAL PUBLISHING MICROSYSTEMS BIOMEDICAL PORTAL MST CONTACTLESS NEWS
EQUIPMENT ONTOLOGY DE REMOTE TEST FURNITURE VEHICLES
0.270
0.272
0.342
0.299
0.328
0.256
0.255
0.258
0.293
0.320
0.239
0.243
0.304
0.268
0.244
0.234
0.253
0.261

How do we extract keywords?

- Characteristic keywords for a group of documents are the most highly weighted words in the centroid of the cluster
  - ...centroid of the cluster could be understood as an “average document” for specific group of documents
  - ...we are using the effect provided by the TFIDF weighting schema for weighting the importance of the words
  - ...efficient solution

TFIDF words weighting in vector representation

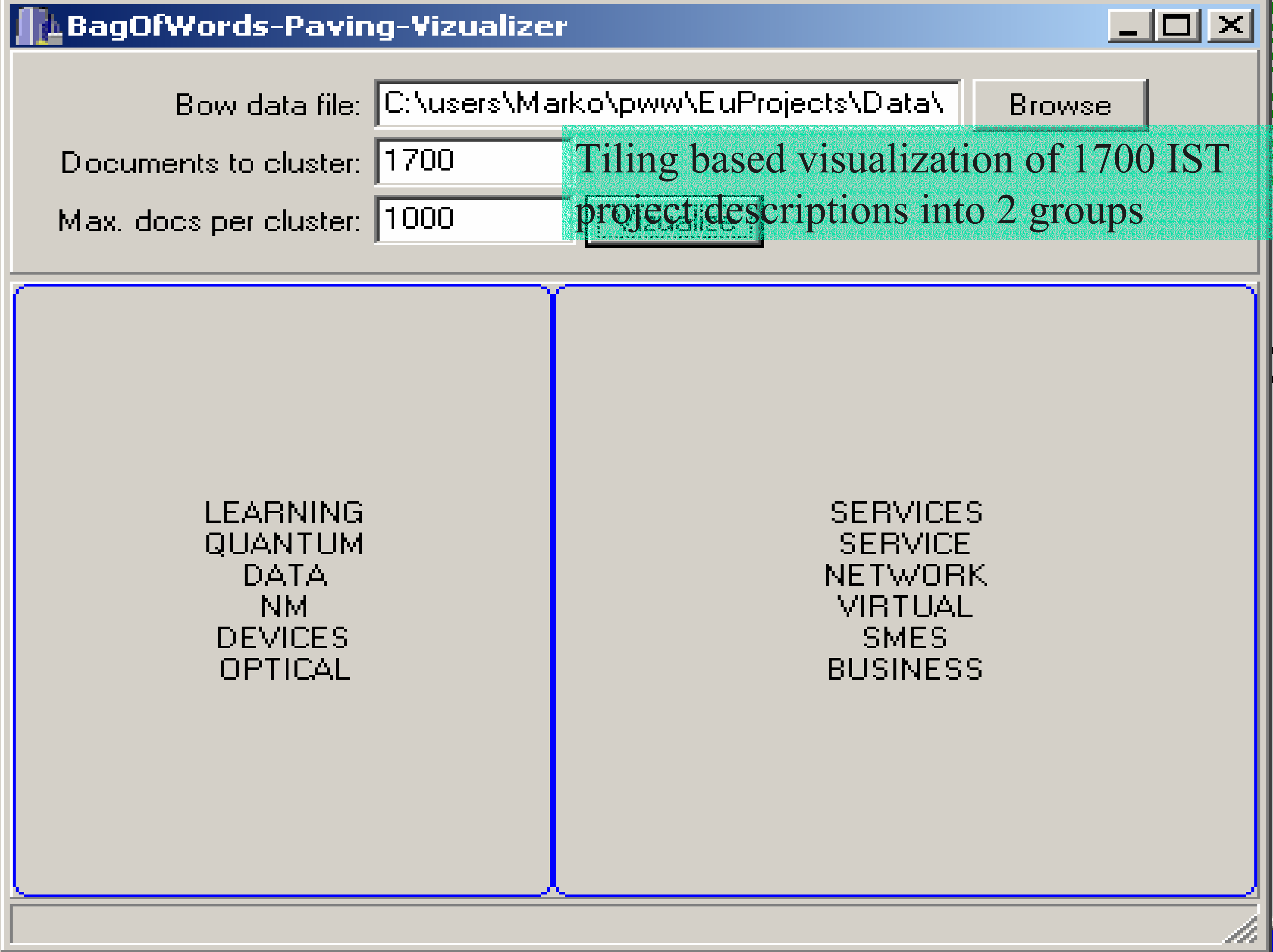
- In Information Retrieval, the most popular weighting schema is normalized word frequency TFIDF:

$$tfidf(w) = tf \,.\, \log\left(\frac{N}{df(w)}\right)$$

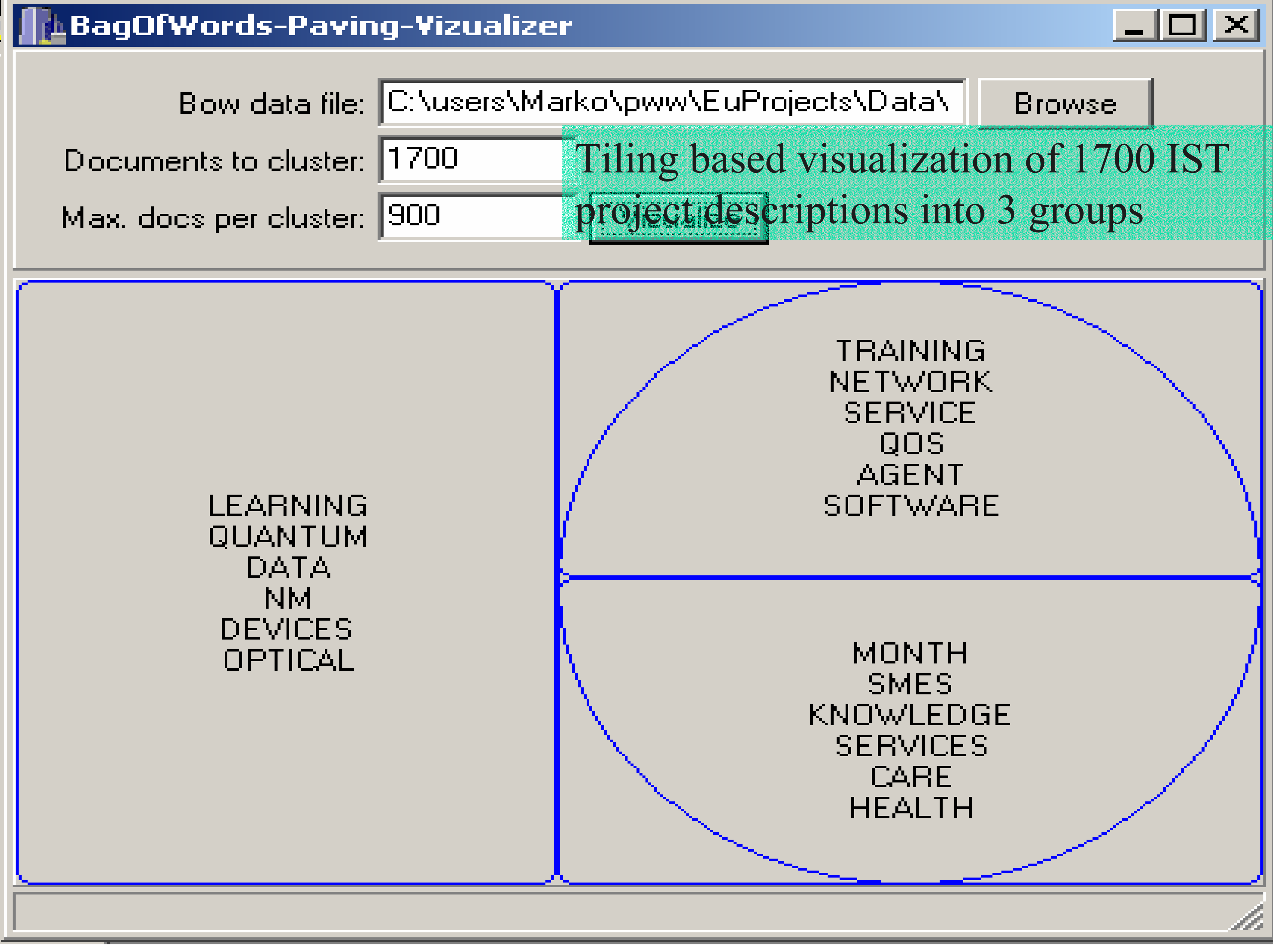
- Tf(w) – term frequency (number of word occurrences in a document)
- Df(w) – document frequency (number of documents containing the word)
- N – number of all documents
- Tfidf(w) – relative importance of the word in the document

Tiling based visualization

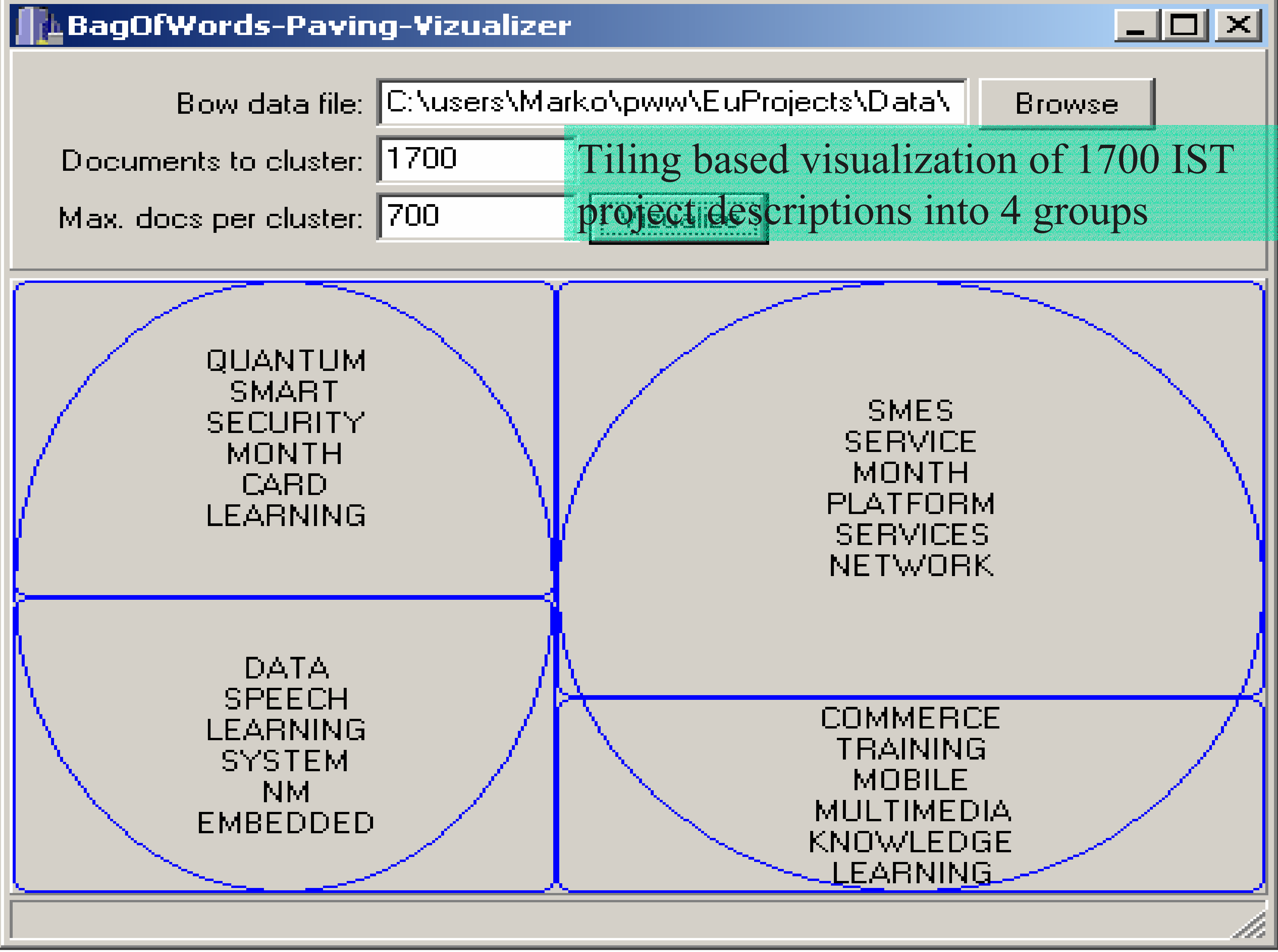
- The sketch of the algorithm:
  1. Documents are transformed into the bag-of-words sparse-vectors representation
     - Words in the vectors are weighted using TFIDF
  2. Hierarchical top-down two-wise K-Means clustering algorithm builds a hierarchy of clusters
     - The hierarchy is an artificial equivalent of hierarchical subject index (Yahoo like)
  3. The leaf nodes of the hierarchy (bottom level) are used to visualize the documents
     - Each leaf is represented by characteristic keywords
     - Each hierarchical binary split splits recursively the rectangular area into two sub-areas



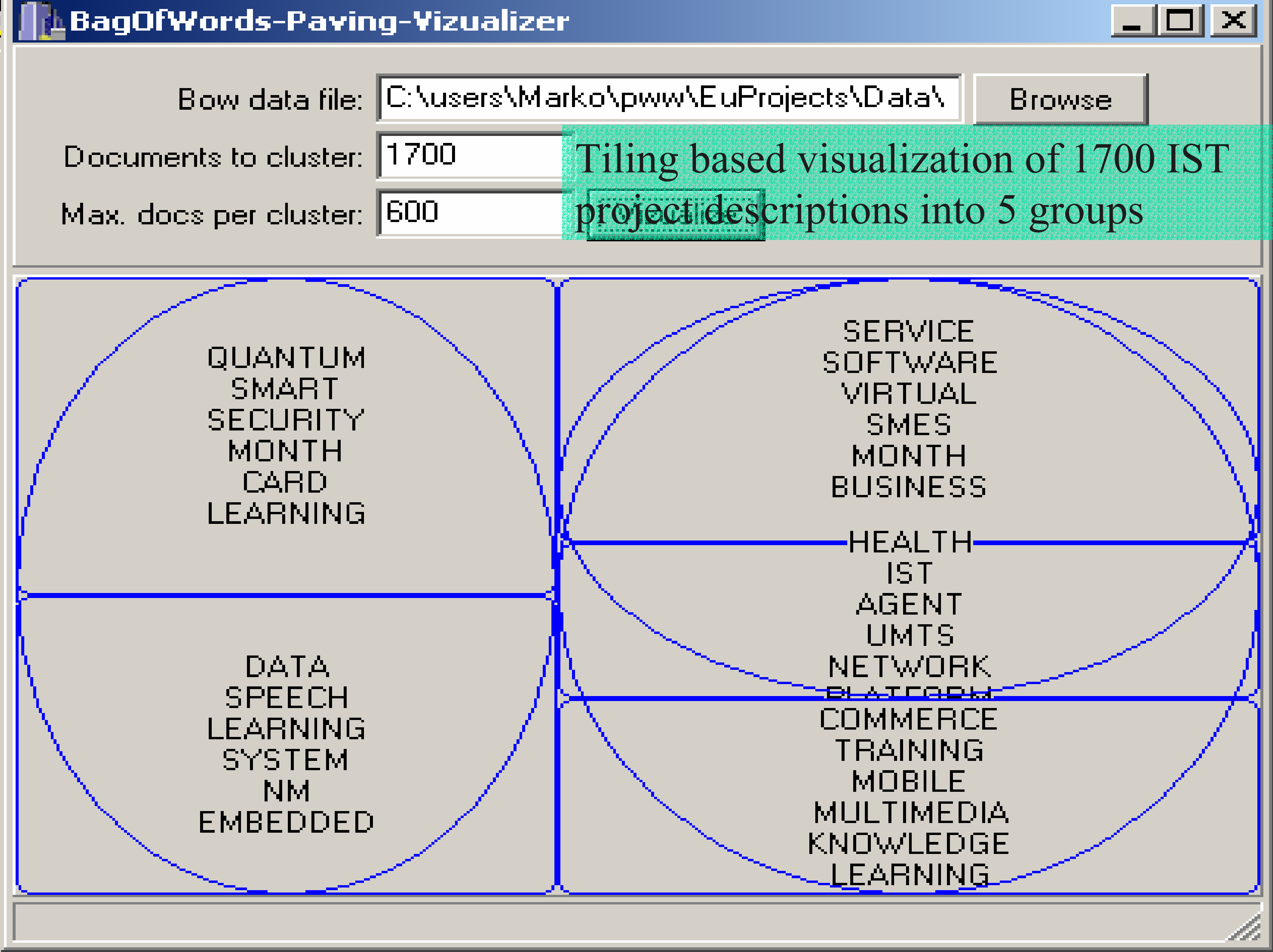
Tiling based visualization of 1700 IST project descriptions into 2 groups



Tiling based visualization of 1700 IST project descriptions into 3 groups

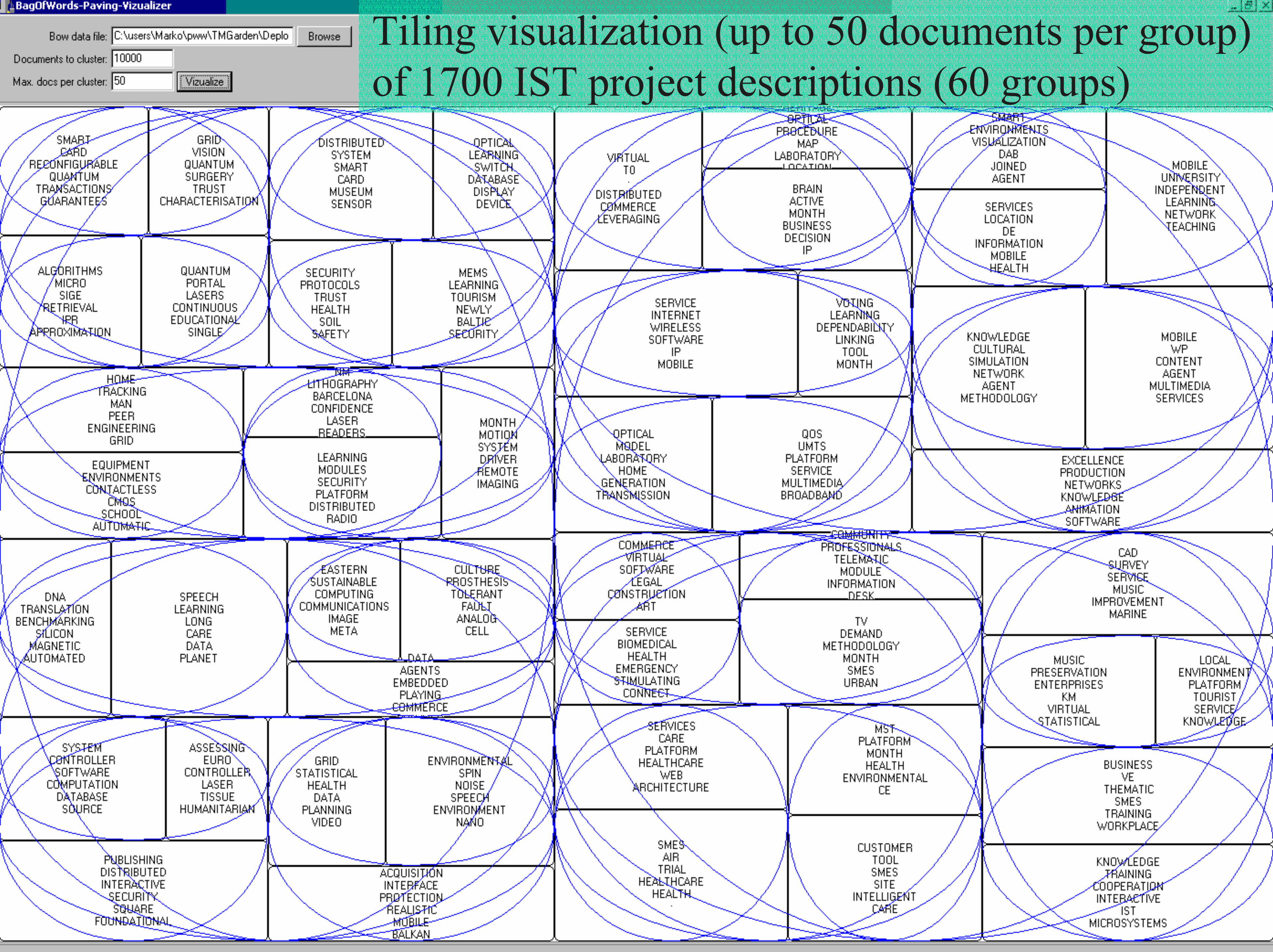


Tiling based visualization of 1700 IST project descriptions into 4 groups



Tiling based visualization of 1700 IST project descriptions into 5 groups

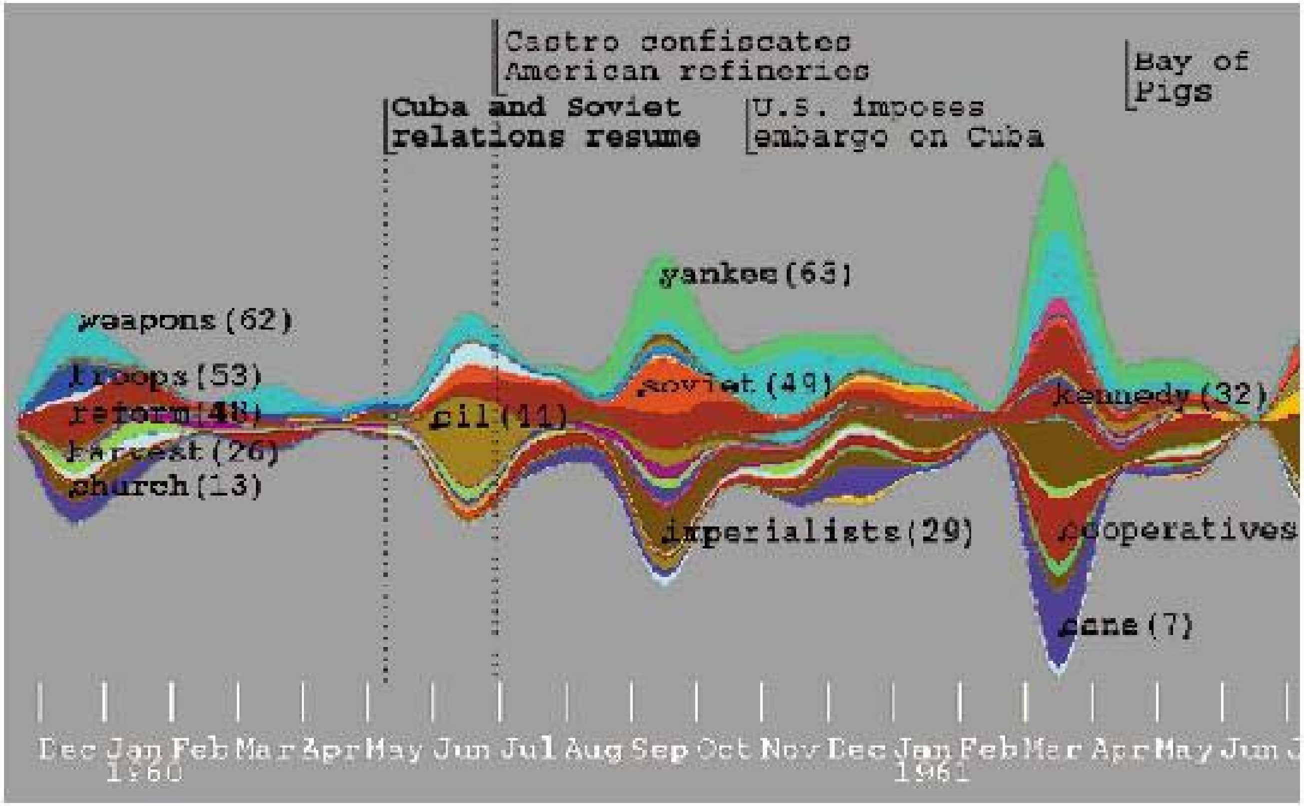
Tiling visualization (up to 50 documents per group) of 1700 IST project descriptions (60 groups)



ThemeRiver

- System that visualizes thematic variations over time across a collection of documents
  - The “river” flows through time, changing width to visualize changes in the thematic strength of documents temporally collocated
  - Themes or topics are represented as colored “currents” flowing within the river that narrow or widen to indicate decreases or increases in the strength of a topic in associated documents at a specific point in time.
  - Described in paper at http://www.pnl.gov/infoviz/themeriver99.pdf

ThemeRiver topic stream

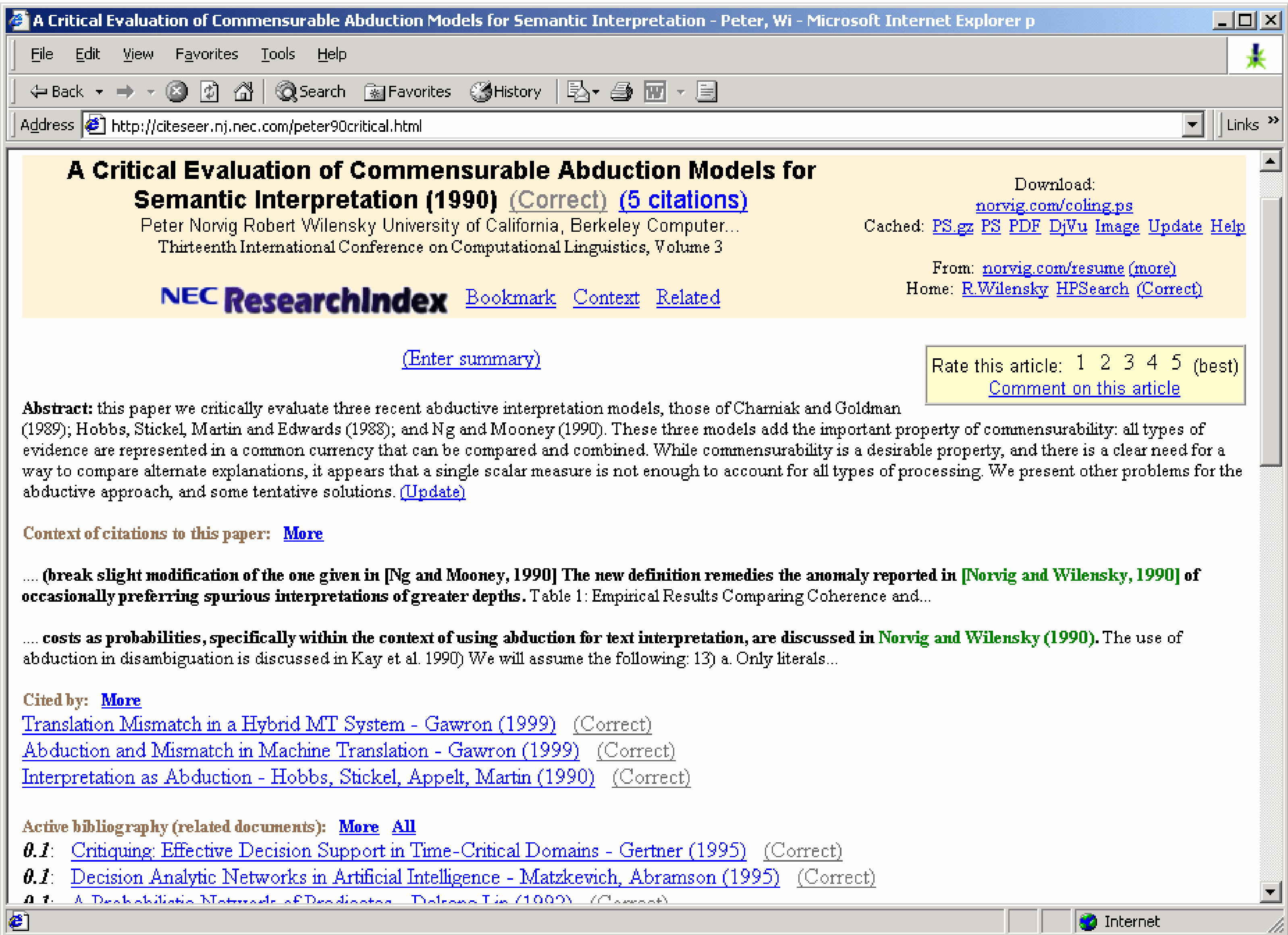


Information Extraction

(slides borrowed from
William Cohen’s Tutorial on IE)

Extracting Job Openings from the Web

IE from Research Papers



What is “Information Extraction”

As a task: Filling slots in a database from sub-segments of text.

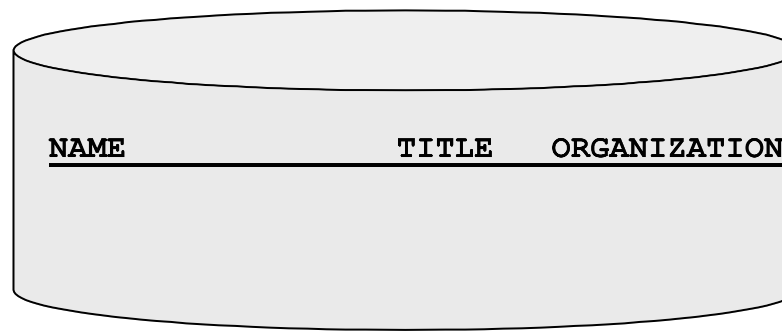
October 14, 2002, 4:00 a.m. PT

For years, Microsoft Corporation CEO Bill Gates railed against the economic philosophy of open-source software with Orwellian fervor, denouncing its communal licensing as a "cancer" that stifled technological innovation.

Today, Microsoft claims to "love" the open-source concept, by which software code is made public to encourage improvement and development by outside programmers. Gates himself says Microsoft will gladly disclose its crown jewels--the coveted code behind the Windows operating system--to select customers.

"We can be open source. We love the concept of shared source," said Bill Veghte, a Microsoft VP. "That's a super-important shift for us in terms of code access.“

Richard Stallman, founder of the Free Software Foundation, countered saying…



What is “Information Extraction”

As a task: **Filling slots in a database from sub-segments of text.**

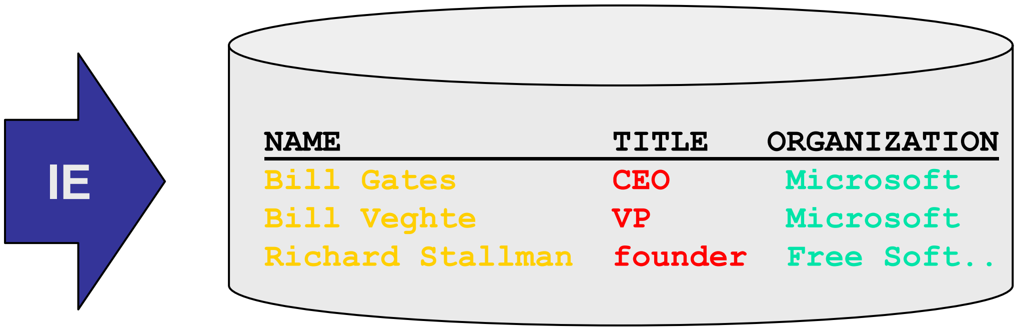
October 14, 2002, 4:00 a.m. PT

For years, Microsoft Corporation CEO Bill Gates railed against the economic philosophy of open-source software with Orwellian fervor, denouncing its communal licensing as a "cancer" that stifled technological innovation.

Today, Microsoft claims to "love" the open-source concept, by which software code is made public to encourage improvement and development by outside programmers. Gates himself says Microsoft will gladly disclose its crown jewels--the coveted code behind the Windows operating system--to select customers.

"We can be open source. We love the concept of shared source," said Bill Veghte, a Microsoft VP. "That's a super-important shift for us in terms of code access.“

Richard Stallman, founder of the Free Software Foundation, countered saying…



What is "Information Extraction"

As a family of techniques:

Information Extraction = segmentation + classification + clustering + association

October 14, 2002, 4:00 a.m. PT

For years, Microsoft Corporation CEO Bill Gates railed against the economic philosophy of open-source software with Orwellian fervor, denouncing its communal licensing as a "cancer" that stifled technological innovation.

Today, Microsoft claims to "love" the open-source concept, by which software code is made public to encourage improvement and development by outside programmers. Gates himself says Microsoft will gladly disclose its crown jewels--the coveted code behind the Windows operating system--to select customers.

"We can be open source. We love the concept of shared source," said Bill Veghte, a Microsoft VP. "That's a super-important shift for us in terms of code access.“

Richard Stallman, founder of the Free Software Foundation, countered saying…

Microsoft Corporation
CEO
Bill Gates
Microsoft
Gates
Microsoft
Bill Veghte
Microsoft
VP
Richard Stallman
founder
Free Software Foundation

aka "named entity extraction"

What is "Information Extraction"

As a family of techniques:

Information Extraction =
segmentation + classification + association + clustering

October 14, 2002, 4:00 a.m. PT

For years, Microsoft Corporation CEO Bill Gates railed against the economic philosophy of open-source software with Orwellian fervor, denouncing its communal licensing as a "cancer" that stifled technological innovation.

Today, Microsoft claims to "love" the open-source concept, by which software code is made public to encourage improvement and development by outside programmers. Gates himself says Microsoft will gladly disclose its crown jewels--the coveted code behind the Windows operating system--to select customers.

"We can be open source. We love the concept of shared source," said Bill Veghte, a Microsoft VP. "That's a super-important shift for us in terms of code access.“

Richard Stallman, founder of the Free Software Foundation, countered saying…

Microsoft Corporation
CEO
Bill Gates
Microsoft
Gates
Microsoft
Bill Veghte
Microsoft
VP
Richard Stallman
founder
Free Software Foundation

What is "Information Extraction"

As a family of techniques:

Information Extraction = segmentation + classification + association + clustering

October 14, 2002, 4:00 a.m. PT

For years, Microsoft Corporation CEO Bill Gates railed against the economic philosophy of open-source software with Orwellian fervor, denouncing its communal licensing as a "cancer" that stifled technological innovation.

Today, Microsoft claims to "love" the open-source concept, by which software code is made public to encourage improvement and development by outside programmers. Gates himself says Microsoft will gladly disclose its crown jewels--the coveted code behind the Windows operating system--to select customers.

"We can be open source. We love the concept of shared source," said Bill Veghte, a Microsoft VP. "That's a super-important shift for us in terms of code access."

Richard Stallman, founder of the Free Software Foundation, countered saying…

Microsoft Corporation CEO Bill Gates
Microsoft Gates
Microsoft
Bill Veghte Microsoft VP
Richard Stallman founder Free Software Foundation

What is "Information Extraction"

As a family of techniques:

Information Extraction =
segmentation + classification + association + clustering

October 14, 2002, 4:00 a.m. PT

For years, Microsoft Corporation CEO Bill Gates railed against the economic philosophy of open-source software with Orwellian fervor, denouncing its communal licensing as a "cancer" that stifled technological innovation.

Today, Microsoft claims to "love" the open-source concept, by which software code is made public to encourage improvement and development by outside programmers. Gates himself says Microsoft will gladly disclose its crown jewels--the coveted code behind the Windows operating system--to select customers.

"We can be open source. We love the concept of shared source," said Bill Veghte, a Microsoft VP. "That's a super-important shift for us in terms of code access."

Richard Stallman, founder of the Free Software Foundation, countered saying…

* Microsoft Corporation
CEO
Bill Gates

* Microsoft
Gates

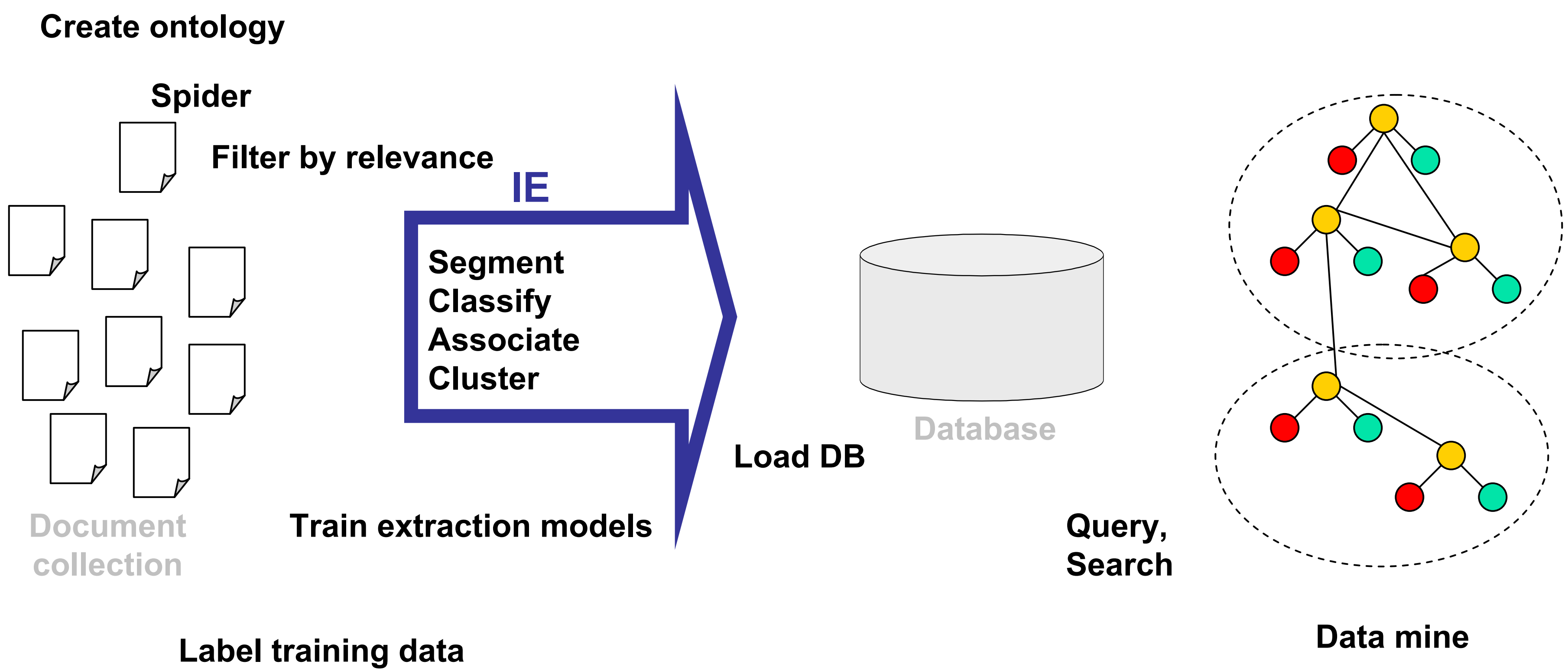
* Microsoft

Bill Veghte
* Microsoft
VP

Richard Stallman
founder
Free Software Foundation

NAME	TITLE	ORGANIZATION
Bill Gates	CEO	Microsoft
Bill Veghte	VP	Microsoft
Richard Stallman	founder	Free Soft..

IE in Context



Typical approaches to IE

- Hand-built rules/models for extraction
- Machine learning used on manually labeled data:
  - Classification problem on sliding window
    - ...examples are taken from sliding window
    - ...models classify short segments of text such as title, name, institution, ...
    - ...limitation of sliding window because it does not take into account sequential nature of text
  - Training stochastic finite state machines (e.g. HMM)
    - ...probabilistic reconstruction of parsing sequence

Levels of Text Processing 5/6

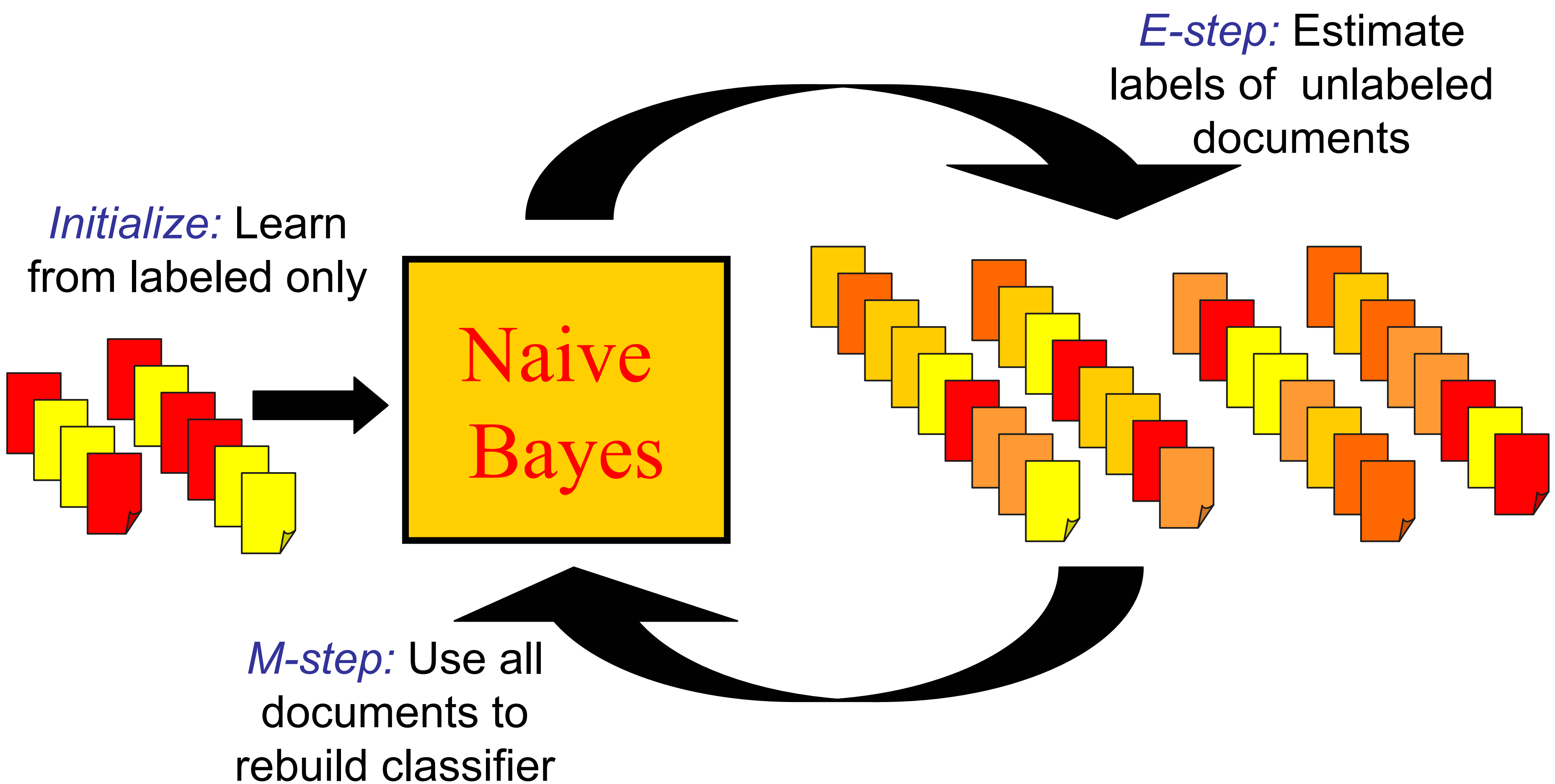
- Word Level
- Sentence Level
- Document Level
- Document-Collection Level
- **Linked-Document-Collection Level**
  - **Labelling unlabeled data**
  - **Co-training**
- Application Level

Labelling unlabeled data

Using unlabeled data
(Nigam et al., ML Journal 2000)

- small number of labeled documents and a large pool of unlabeled documents, eg., classify an article in one of the 20 News groups, classify Web page as student, faculty, course, project,...
- 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

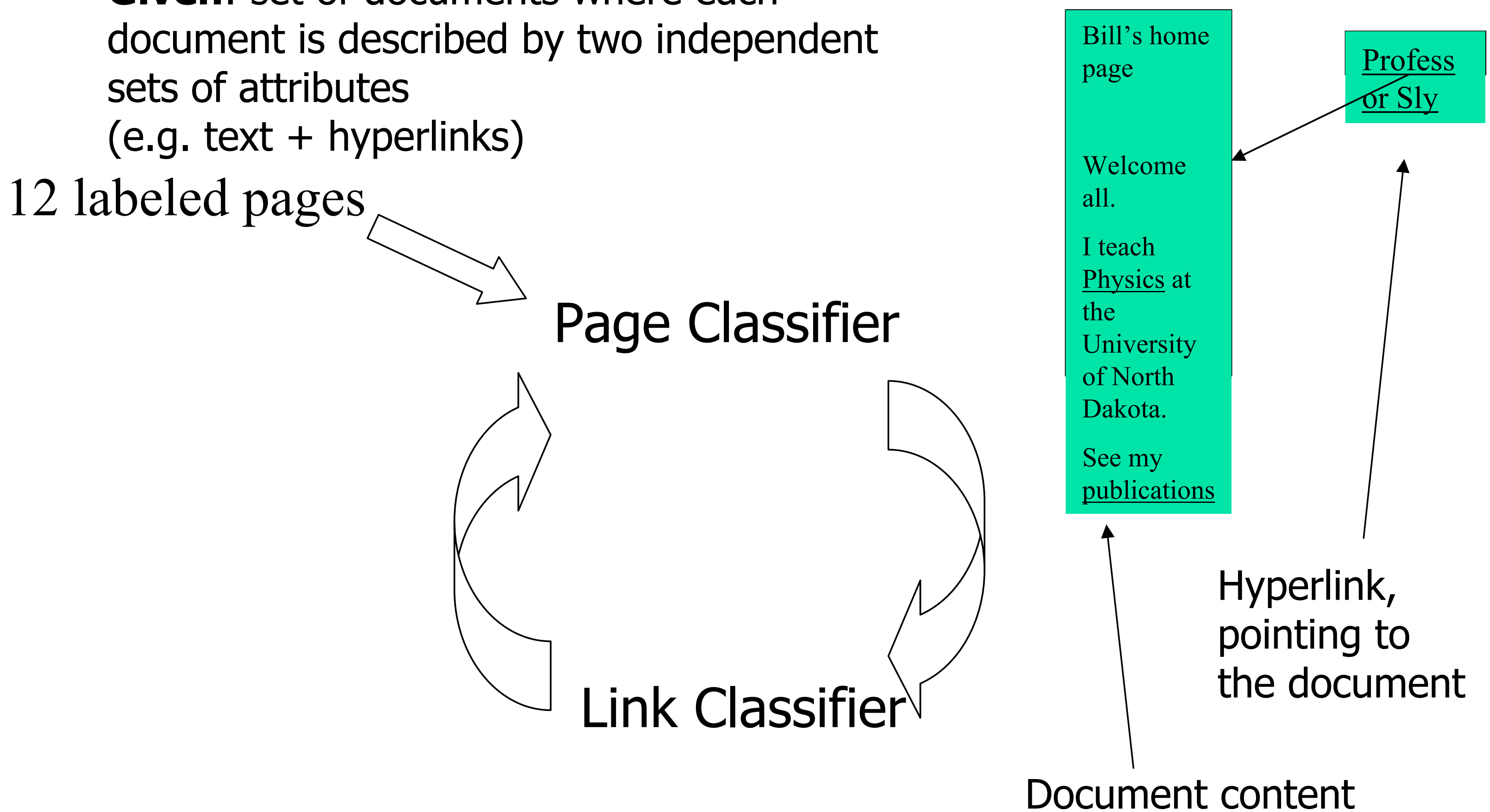
Co-training

Co-training

- Better performance on labelling unlabeled data compared to EM approach

Bootstrap Learning to Classify Web Pages (co-training)

Given: set of documents where each document is described by two independent sets of attributes
(e.g. text + hyperlinks)



Levels of Text Processing 6/6

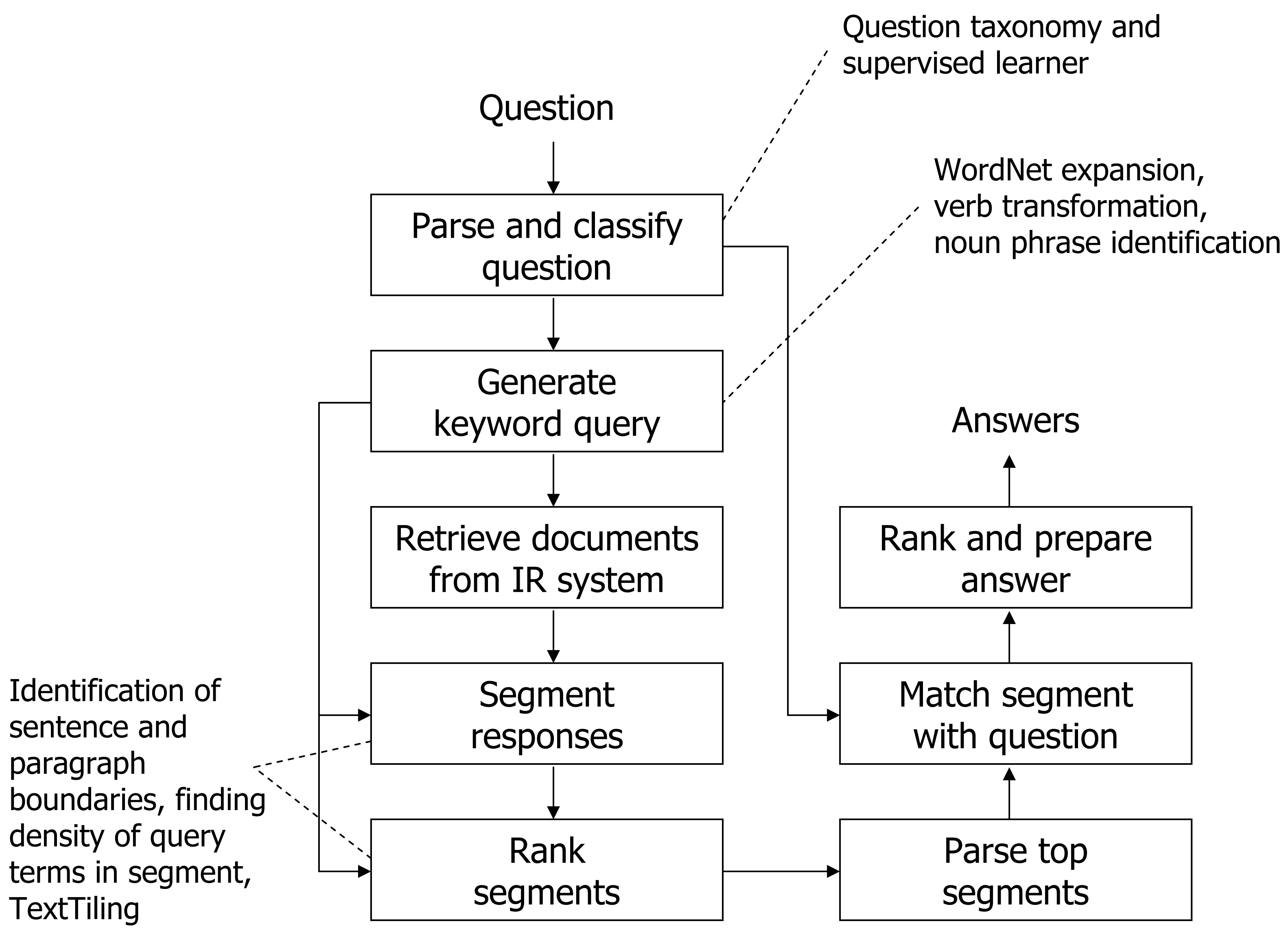
- Word Level
- Sentence Level
- Document Level
- Document-Collection Level
- Linked-Document-Collection Level
- **Application Level**
  - **Question-Answering**
  - **Mixing Data Sources (KDD Cup 2003)**

Question-Answering

Question Answering

- QA Systems are returning short and accurate replies to the well-formed natural language questions such as:
  - *What is the hight of Mount Everest?*
  - *After which animal is the Canary Island named?*
  - *How many liters are there in to a gallon?*
- QA Systems can be classified into following levels of sophistication:
  - Slot-filling – easy questions, IE technology
  - Limited-Domain – handcrafted dictionaries & ontologies
  - Open-domain – IR, IE, NL parsing, inferencing

Question Answering Architecture



Question Answering Example

- Example question and answer:
  - Q:What is the color of grass?
  - A: Green.
- …the answer may come from the document saying: "*grass is green*" without mentioning "*color*" with the help of WordNet having hypernym hierarchy:
  - green, chromatic color, color, visual property, property

Mixing Data Sources (KDD Cup 2003)

borrowed from

Janez Brank & Jure Leskovec

The Dataset on KDD Cup 2003

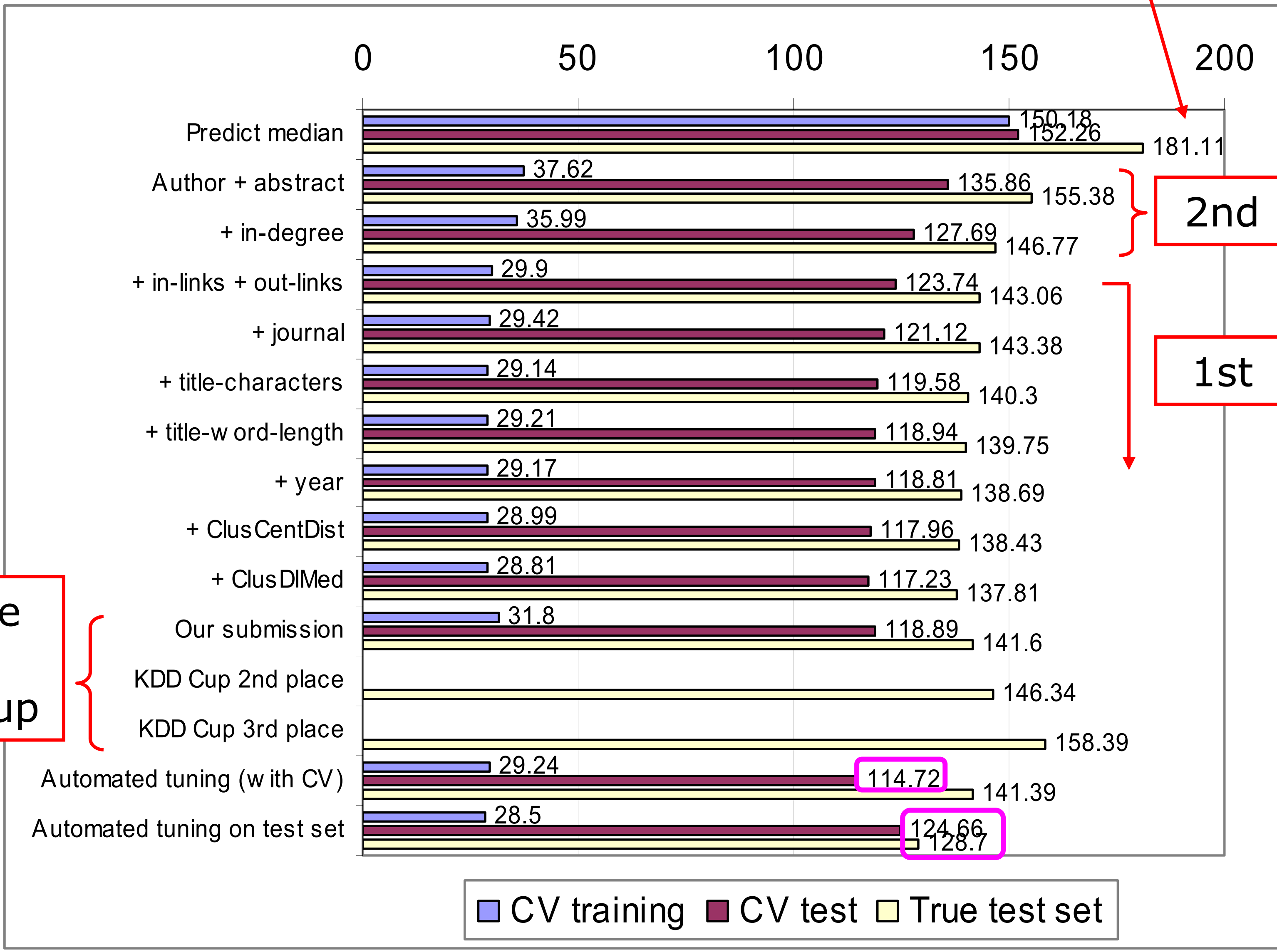
- Approx. 29000 papers from the "high energy physics – theory" area of arxiv.org
- For each paper:
  - Full text (TeX file, often very messy)
    Avg. 60 KB per paper. Total: 1.7 GB.
  - Metadata in a nice, structured file (authors, title, abstract, journal, subject classes)
- The citation graph
- **Task**: How many times have certain papers been downloaded in the first 60 days since publication in the arXiv?

Solution

- Textual documents have traditionally been treated as "bags of words"
  - The number of occurrences of each word matters, but the order of the words is ignored
  - Efficiently represented by sparse vectors
- We extend this to include other items besides words ("bag of X")
  - Most of our work was spent trying various features and adjusting their weight (more on that later)
- Use support vector regression to train a linear model, which is then used to predict the download counts on test papers
- Submitted solution was based on the model trained on the following representation:
  - AA + 0.005 in-degree + 0.5 in-links +
    0.7 out-links + 0.3 journal +
    0.004 title-chars. + 0.6 (year – 2000) +
    0.15 ClusDlAvg

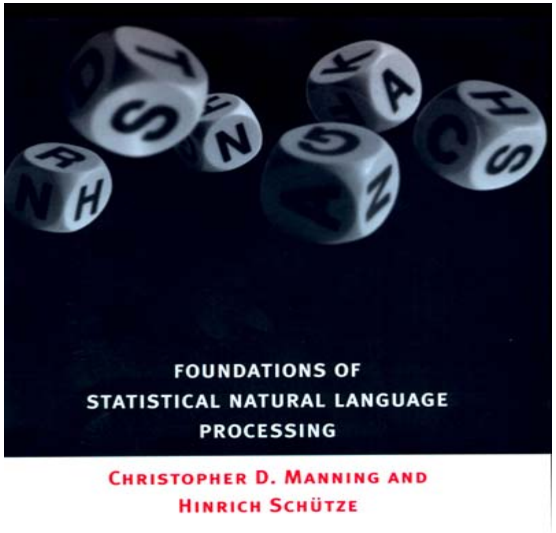
A Look Back...

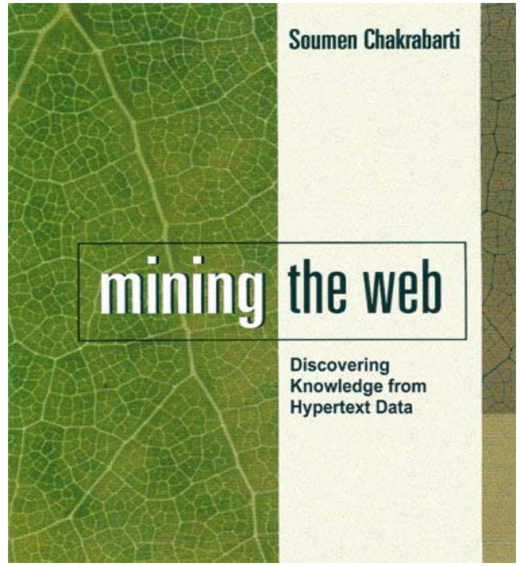
If we'd submitted this, we'd have been 8th or 9th

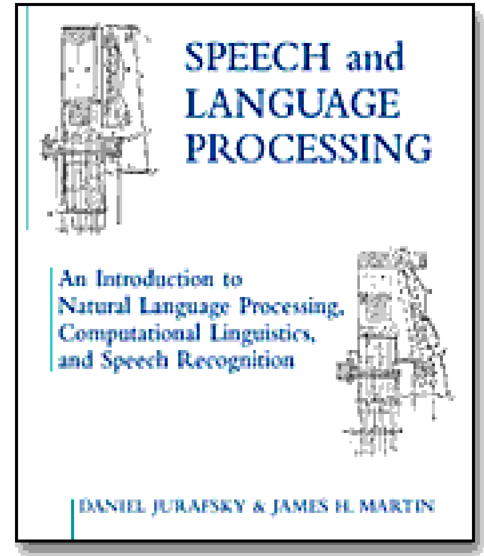


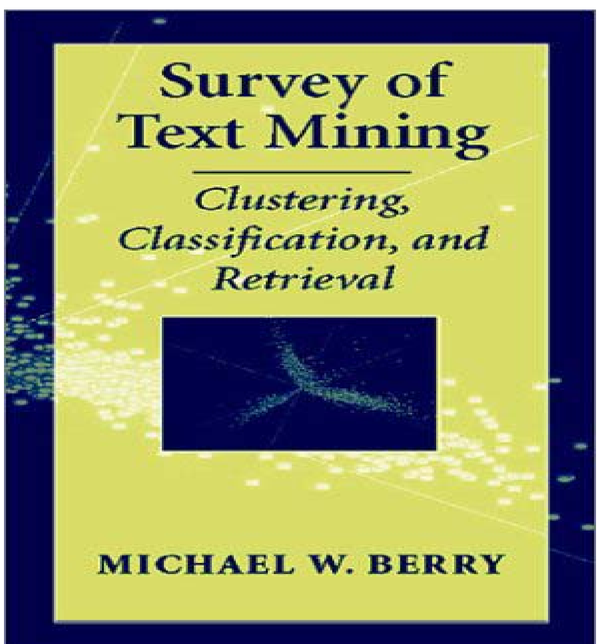
Best three entries on KDD Cup

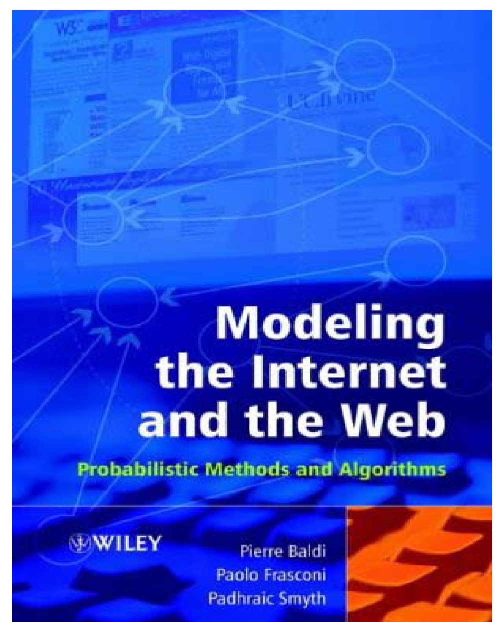
References to some of the Books











References to Conferences

- Information Retrieval: SIGIR, ECIR
- Machine Learning/Data Mining: ICML, ECML/PKDD, KDD, ICDM, SCDM
- Computational Linguistics: ACL, EACL, NAACL
- Semantic Web: ISWC, ESSW

References to some of the TM workshops (available online)

- ICML-1999 Workshop on Machine Learning in Text Data Analysis (TextML-1999) (http://www-ai.ijs.si/DunjaMladenic/ICML99/TLWsh99.html) at International Conference on Machine Learning, Bled 1999
- KDD-2000 Workshop on Text Mining (TextKDD-2000) (http://www.cs.cmu.edu/~dunja/WshKDD2000.html) at ACM Conference on Knowledge Discovery on Databases, Boston 2000
- ICDM-2001 Workshop on Text Mining (TextKDD-2001) (http://www-ai.ijs.si/DunjaMladenic/TextDM01/), at IEEE International Conference on Data Mining, San Jose 2001
- ICML-2002 Workshop on Text Learning (TextML-2002) (http://www-ai.ijs.si/DunjaMladenic/TextML02/) at International Conference on Machine Learning, Sydney 2002
- IJCAI-2003 Workshop on Text-Mining and Link-Analysis (Link-2003) (http://www.cs.cmu.edu/~dunja/TextLink2003/), at International Joint Conference on Artificial Intelligence, Acapulco 2003
- KDD-2003 Workshop on Workshop on Link Analysis for Detecting Complex Behavior (LinkKDD2003) (http://www.cs.cmu.edu/~dunja/LinkKDD2003/) at ACM Conference on Knowledge Discovery on Databases, Washington DC 2003

Some of the Products

- Authonomy
- ClearForest
- Megaputer
- SAS/Enterprise-Miner
- SPSS - Clementine
- Oracle - ConText
- IBM - Intelligent Miner for Text

Final Remarks

- In the future we can expect stronger integration and **bigger overlap** between TM, IR, NLP and SW…
- …the technology and it's solutions will try to **capture deeper semantics** within the text,
- …**integration of various** data sources (including text) is becoming increasingly important.