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
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_PROJ1_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

<table>
<thead>
<tr>
<th></th>
<th>A</th>
<th>B</th>
<th>C</th>
<th>D</th>
<th>E</th>
<th>F</th>
<th>G</th>
<th>H</th>
<th>I</th>
<th>J</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Project Ref</td>
<td>Acronym</td>
<td>Domain / Unit</td>
<td>PO</td>
<td>Legal Name</td>
<td>Legal Country</td>
<td>Type of organisation</td>
<td>Participant role</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>IST-1999-11495</td>
<td>SCL-EU-NET</td>
<td>KA2</td>
<td>C2</td>
<td>HANSEN RALF</td>
<td>ALARIX, D.O.O.</td>
<td>SLOVENIA</td>
<td>Private non research org</td>
<td>CR</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>IST-1999-11495</td>
<td>SCL-EU-NET</td>
<td>KA2</td>
<td>C2</td>
<td>HANSEN RALF</td>
<td>AUSTRIAN RESEARCH INS AUSTRIA</td>
<td>Research centres</td>
<td>CR</td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>IST-1999-11495</td>
<td>SCL-EU-NET</td>
<td>KA2</td>
<td>C2</td>
<td>HANSEN RALF</td>
<td>CZECH TECHNICAL UNIVERSITY CZECH REPUBLIC</td>
<td>Higher education</td>
<td>CR</td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>IST-1999-11495</td>
<td>SCL-EU-NET</td>
<td>KA2</td>
<td>C2</td>
<td>HANSEN RALF</td>
<td>DIALOGIS SOFTWARE &amp; S GERMANY</td>
<td>Private non research org</td>
<td>CR</td>
<td></td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>IST-1998-11495</td>
<td>SCL-EU-NET</td>
<td>KA2</td>
<td>C2</td>
<td>HANSEN RALF</td>
<td>FACHHOCHSCHULE BONN GERMANY</td>
<td>Higher education</td>
<td>CR</td>
<td></td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>IST-1999-11495</td>
<td>SCL-EU-NET</td>
<td>KA2</td>
<td>C2</td>
<td>HANSEN RALF</td>
<td>FRAUNHOFER GESELLSCHE GERMANY</td>
<td>Research centres</td>
<td>CR</td>
<td></td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>IST-1999-11495</td>
<td>SCL-EU-NET</td>
<td>KA2</td>
<td>C2</td>
<td>HANSEN RALF</td>
<td>GMD - FORSCHUNGSZENTRUM GERMANY</td>
<td>Research centres</td>
<td>CR</td>
<td></td>
<td></td>
</tr>
<tr>
<td>9</td>
<td>IST-1999-11495</td>
<td>SCL-EU-NET</td>
<td>KA2</td>
<td>C2</td>
<td>HANSEN RALF</td>
<td>INSTITUT JOZEF STEFAN SLOVENIA</td>
<td>Research centres</td>
<td>CR</td>
<td></td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>IST-1999-11495</td>
<td>SCL-EU-NET</td>
<td>KA2</td>
<td>C2</td>
<td>HANSEN RALF</td>
<td>KATHOLIEKE UNIVERSITEIT BELGIUM</td>
<td>Higher education</td>
<td>CR</td>
<td></td>
<td></td>
</tr>
<tr>
<td>11</td>
<td>IST-1999-11495</td>
<td>SCL-EU-NET</td>
<td>KA2</td>
<td>C2</td>
<td>HANSEN RALF</td>
<td>STUDIO PHI D.O.O., COMM SLOVENIA</td>
<td>Private non research org</td>
<td>CR</td>
<td></td>
<td></td>
</tr>
<tr>
<td>12</td>
<td>IST-1999-11495</td>
<td>SCL-EU-NET</td>
<td>KA2</td>
<td>C2</td>
<td>HANSEN RALF</td>
<td>TEMIDA D.O.O., COMPANY SLOVENIA</td>
<td>Private non research org</td>
<td>CR</td>
<td></td>
<td></td>
</tr>
<tr>
<td>13</td>
<td>IST-1999-11495</td>
<td>SCL-EU-NET</td>
<td>KA2</td>
<td>C2</td>
<td>HANSEN RALF</td>
<td>THE CHANCELLOR, MASTERS UNITED KINGDOM</td>
<td>Higher education</td>
<td>CR</td>
<td></td>
<td></td>
</tr>
<tr>
<td>14</td>
<td>IST-1999-11495</td>
<td>SCL-EU-NET</td>
<td>KA2</td>
<td>C2</td>
<td>HANSEN RALF</td>
<td>UNIVERSIDADE DO PORTO PORTUGAL</td>
<td>Private non research org</td>
<td>CR</td>
<td></td>
<td></td>
</tr>
<tr>
<td>15</td>
<td>IST-1999-11495</td>
<td>SCL-EU-NET</td>
<td>KA2</td>
<td>C2</td>
<td>HANSEN RALF</td>
<td>UNIVERSITY OF BRISTOL UNITED KINGDOM</td>
<td>Higher education</td>
<td>CR</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

© Dunja Mladenec
Example of data for Sol-Eu-Net (2)

Project Title

Project Acronym

Project Description

The goal of this project is to enhance competitiveness and find new business opportunities in the global IT market by establishing a virtual European enterprise composed of companies and research laboratories with highly specialised expertise in two IT areas: data mining and decision support. The established Sol-Eu-Net enterprise will be organised as a flexible business structure made of cross-organisational, time-focused, task-driven work teams. It will work towards enhanced usage of data mining and decision support in industry, businesses and public services, contributing to improved quality, efficiency and effectiveness of their operations. This will be achieved through specific solutions to end-user problems, prototype project workshops, project monitoring and consulting, collaborative work and combination of problem solutions, as well as through education, training and spreading information Web-based information source.
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 - [SPINI, 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, SPINI, 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 - [SPINI, 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) KUNGGLIGA 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?

- Knowledge Rep. & Reasoning
- Computational Linguistics
- Semantic Web
- Natural Language Processing
- Information Retrieval
- Search & DB
- Data Analysis
- Machine Learning
- Text Mining

© Dunja Mladenic
Levels of text representations

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

© Dunja Mladenic
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

<table>
<thead>
<tr>
<th>Category</th>
<th>Unique Forms</th>
<th>Number of Senses</th>
</tr>
</thead>
<tbody>
<tr>
<td>Noun</td>
<td>94474</td>
<td>116317</td>
</tr>
<tr>
<td>Verb</td>
<td>10319</td>
<td>22066</td>
</tr>
<tr>
<td>Adjective</td>
<td>20170</td>
<td>29881</td>
</tr>
<tr>
<td>Adverb</td>
<td>4546</td>
<td>5677</td>
</tr>
</tbody>
</table>
WordNet relations

Each WordNet entry is connected with other entries in a graph through relations.

Relations in the database of nouns:

<table>
<thead>
<tr>
<th>Relation</th>
<th>Definition</th>
<th>Example</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hypernym</td>
<td>From concepts to subordinate</td>
<td>breakfast -&gt; meal</td>
</tr>
<tr>
<td>Hyponym</td>
<td>From concepts to subtypes</td>
<td>meal -&gt; lunch</td>
</tr>
<tr>
<td>Has-Member</td>
<td>From groups to their members</td>
<td>faculty -&gt; professor</td>
</tr>
<tr>
<td>Member-Of</td>
<td>From members to their groups</td>
<td>copilot -&gt; crew</td>
</tr>
<tr>
<td>Has-Part</td>
<td>From wholes to parts</td>
<td>table -&gt; leg</td>
</tr>
<tr>
<td>Part-Of</td>
<td>From parts to wholes</td>
<td>course -&gt; meal</td>
</tr>
<tr>
<td>Antonym</td>
<td>Opposites</td>
<td>leader -&gt; follower</td>
</tr>
</tbody>
</table>
Levels of text representations

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

---

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

---

- Collaborative tagging / Web2.0
- Templates / Frames
- Ontologies / First order theories
Vector-space model level

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

<table>
<thead>
<tr>
<th>bob</th>
<th>builder</th>
<th>children</th>
<th>animated</th>
<th>movie</th>
<th>character</th>
<th>friend</th>
<th>vehicle</th>
<th>...</th>
<th>...</th>
</tr>
</thead>
<tbody>
<tr>
<td>3</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>...</td>
<td>...</td>
</tr>
</tbody>
</table>

© Dunja Mladenic
Document Categorization
Document categorization

Machine learning

labeled documents

Document Classifier

unlabeled document

document category (label)
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

\[
\begin{align*}
\text{Precision}(M, \text{targetC}) &= P(\text{targetC} | \text{targetC}) \\
\text{Recall}(M, \text{targetC}) &= P(\text{targetC} | \text{targetC}) \\
\text{Accuracy}(M) &= \sum_{i} P(C_i) \times \text{Precision}(M, C_i) \\
F_\beta(M, \text{targetC}) &= \frac{(1 + \beta^2) \text{Precision}(M, \text{targetC}) \times \text{Recall}(M, \text{targetC})}{\beta^2 \text{Precision}(M, \text{targetC}) + \text{Recall}(M, \text{targetC})}
\end{align*}
\]

- 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)

Top 20 categories of Reuter news in 1987-91

Unbalanced distribution
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

Feature construction

Sub-problem definition
Feature selection
Classifier construction

Vectors of n-grams

Labeled documents
(from Yahoo! hierarchy)

Unlabeled document

Document Classifier

Document category (label)

© Dunja Mladenic
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)

Initialize: Learn from labeled only

Naive Bayes

M-step: Use all documents to rebuild classifier

E-step: Estimate labels of unlabeled documents

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

**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

Page Classifier

Link Classifier

**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 Danja Mladenic or Marko Grobelnik.
Active Learning

- We use this method 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.

![Diagram of Active Learning process]

© Dunja Mladenic
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 are rather time consuming, thus we run them for including the first 50 unlabeled examples
  – experiments show that active learning is especially useful for unbalanced data

© Dunja Mladenic
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

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

Knowledge Management

Data analysis

Health

Mobile computing
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
- 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)
- Join 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

- Foundations of Statistical Natural Language Processing
  - Christopher D. Manning and Hinrich Schütze

- Speech and Language Processing
  - Daniel Jurafsky & James H. Martin

- Natural Language Processing for Online Applications
  - Peter Jackson and Isabelle Moulinier

- Mining the Web
  - Soumen Chakrabarti

- Survey of Text Mining
  - Clustering, Classification, and Retrieval
  - Michael W. Berry

- Modeling the Internet and the Web
  - Probabilistic Methods and Algorithms
  - Pierre Beldi, Paolo Frasconi, Padhraic Smyth

- Text Mining
  - Predictive Methods for Analyzing Unstructured Information
  - Shubham M. Vaze, Nitesh Indurkhya, Tong Zhang, Fred L. Damper
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 concept (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 (e.g., words-vectors)
  [Fortuna & Mladenović & Grobelnik, 2005]

Installation package is publicly available in binaries at [ontogen.ijs.si](http://ontogen.ijs.si)
Basic idea behind OntoGen
Hierarchy of concepts

Ontology visualization

These two views are synchronized

Selected concept

Root concept

Concept name

Descriptive keywords

Distinctive keywords

Id: 14 Name: investment

Keywords: investment, funded, management, capital, investment_company, income, security, assets, equity

SVM Keywords: investment, funded, equity, management, financing, close, close-ended, assets_management, clients

All documents: 109
Unused documents: 109
Suggesting sub-concepts

Number of suggestions

List of suggestions
Inspection tool
Description of the concept
The system asks several “yes or no” questions.
OntoGen -- Text Garden

Concepts
- root
  - gas, property, oils
  - Airlines
  - Restaurants
  - Stations
  - Oil, mining
  - Pharmacy
  - Life insurance
  - Real estate
  - systems, services, manufactures
  - banking, loans, investment

Ontology details
- Ontology visualization
- Concept's documents
- Concept Visualization
  - Concept font size: 11
  - Relation font size: 9

Concept properties
- Details
  - Id: 52
  - Name: Oil, mining
  - 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
  - All documents: 485
  - Unused documents: 485

OntoGen news: OntoGen won Best Demo Award at ESWC 2006
Checkboxes indicate whether a document belongs to the concept or not.

List of documents

Red dots represent documents that currently belong to the concept.

Document preview pane

Similarity graph