SuperComputing Applications and Innovation

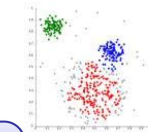
Text Mining

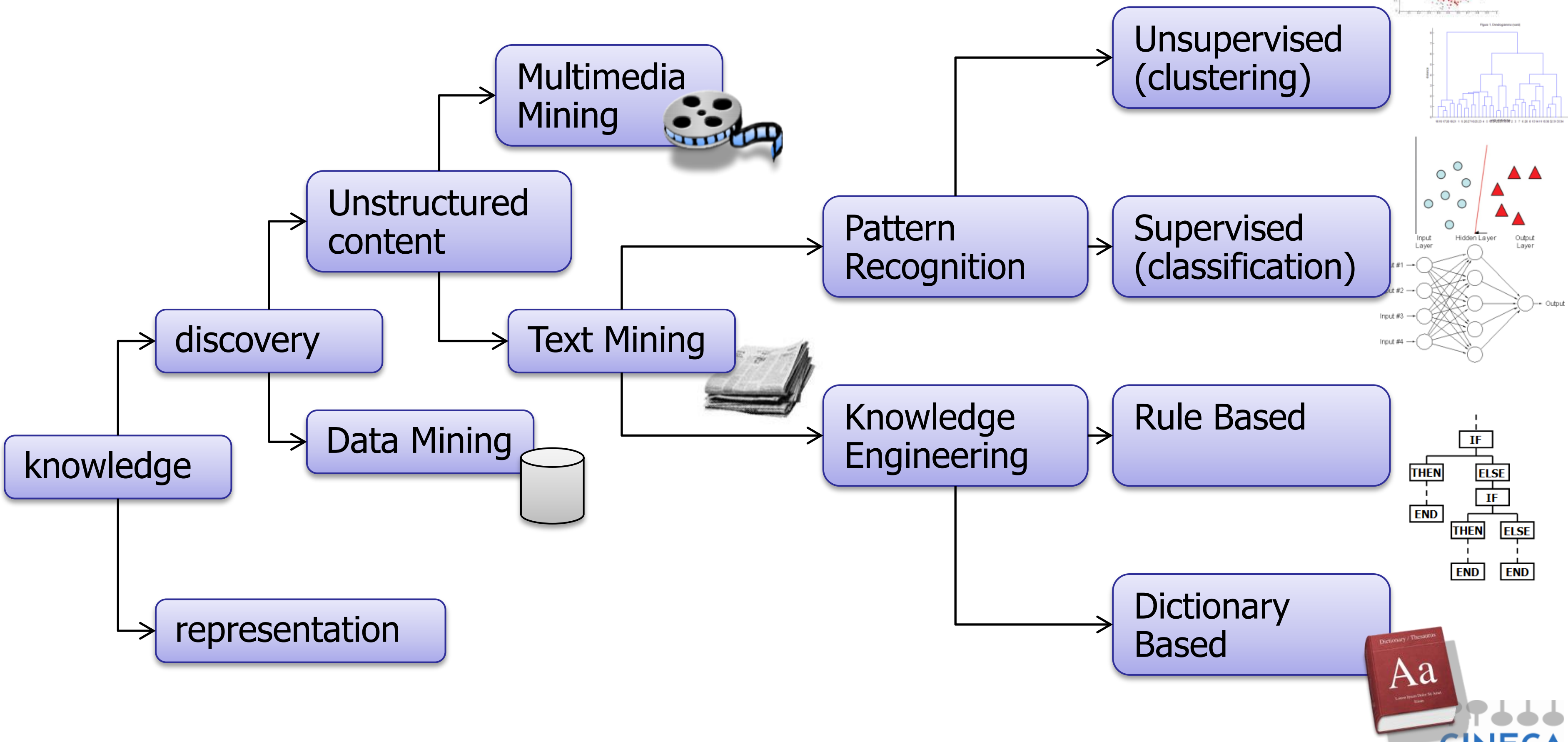
School on Scientific Data Analytics and Visualization

Roberta Turra, *Cineca*

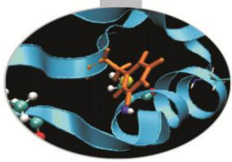
9 June 2015

Overview





knowledge
discovery
representation
Unstructured content
Data Mining
Multimedia Mining
Text Mining
Pattern Recognition
Knowledge Engineering
Unsupervised (clustering)
Supervised (classification)
Rule Based
Dictionary Based
IF
THEN
ELSE
END
IF
THEN
ELSE
END
END
Aa



Information Retrieval

Selects documents that match a string (query) and ranks them according to term weights

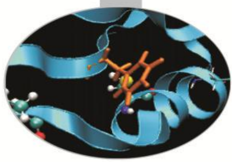
Information Extraction

Extracts entities and relations appropriate to a specific task e.g. terrorist attacks (who attacked whom, where and when), corporate acquisitions and mergers, …

Text Mining

Word patterns indicate similarities among documents and establish relations without any predefinition of query terms or entities to be looked for

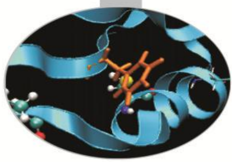
SuperComputing Applications and Innovation



Text Mining

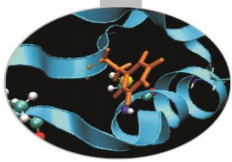
Application of Data Mining techniques to unstructured texts (news, web pages, e-mails, ...) with the aim of :

- finding the main **thematic groups** (Clustering)
- **classifying documents** in predefined categories (Machine Learning)
- discovering **hidden associations** (links among topics, or among authors, trends, ...)
- **extracting specific information** like genes, companies names, ... (Information Extraction)
- automatic metadata generation / **semantic annotation** (tagging)
- indexing information for **semantic search engine**
- extracting concepts and relations for **ontologies creation** (ontology learning), **ontologies population** (ontology feeding / content mapping)



The process

- collecting
- indexing
- mining
- evaluation

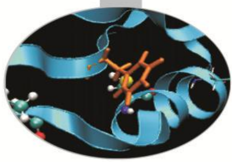


The Process

Phase1: collecting

- document selection
  - Document collection from multiple sources
    - retreiving from DBs (query)
    - downloading (through API)
    - web crawling / web scraping

- pre – processing
  - parsing
  - integration
  - transformation to a common format



The Process

Phase2: indexing

- document preparation (**indexing**)
  - tokenization
  - Part Of Speech tagging
  - selection of terms (nouns, verbs, adjectives, …)
  - stemming / lemmatization
  - chunking (n-grams, nominal phrases)
  - weighting (binary, frequencies, tfidf, …)
  - stop-words filtering
  - dimensionality reduction
  - meta-information tagging

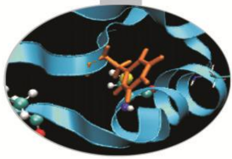
The Process

Phase2: indexing

The result of the indexing phase is a document vector (a sequence of terms and tags).

All document vectors are then converted to a common format: the analysis matrix.

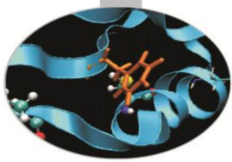
	team	coach	play	ball	score
Document 1	3	0	5	0	2
Document 2	0	7	0	2	1
Document 3	0	1	0	0	1



The Process

Phase3: mining

- document clustering / classification (**mining**)
  - computation of distances (similarity index, euclidian distance, cosine similarity, …)
  - selection of the appropriate clustering / learning algorithm (hierarchical/partitive, K-means, Self Organizing Maps, …)
  - parameter tuning
  - training



The Process

Phase4: evaluation

- result interpretation (clustering)
  - balanced / unbalanced clusters
  - internal and external similarity
  - meaning (interpretation)
  - links between clusters
  - meta – information among clusters
- result evaluation (classification)
  - define a score threshold
  - apply the classification model on the test set
  - compare the classifier results with document labels
  - compute efficacy measures (accuracy, precision, recall, F-score, …)
  - application to unlabeled corpus

Use case: SAE news

Society of Automotive Engineers

Technical news published in 1999-2000 were selected (business news were not relevant for our customer): 3262 documents

- Mark up the different parts of a document

TITLE
TEXT

Textual part

SUBJECTS
SOURCE
DATE
STATE
LANGUAGE

Meta-information

- Linguistic analysis (keyword extraction and stemming)
- Information Extraction (meta-information integration)

Original document

>>> 35:TOYOTA: Avalon Receives Top Score in Frontal Offset Crash Tests

Toyota Motor Corp.'s Avalon received the top score -- a "good" rating earning a "best pick" -- in the 40 mile per hour frontal offset crash tests on new or updated vehicles. The tests were conducted by the Insurance Institute for Highway Safety, a nonprofit group funded by automobile insurers. Nissan Motor Co. Ltd.'s Maxima midsize sedan and Infiniti I30 luxury sedan, the Nissan Sentra small car and Mazda Motor Corp.'s Mazda MPV minivan all scored "average" marks. Isuzu Motors Ltd..'s Rodeo sport utility, also sold by Honda Motor Co. Ltd. as the Honda Passport, earned a "poor" rating due to high crash forces recorded on the crash dummy's head, indicating an increased likelihood of injury. In the crash tests, the vehicles were driven into a deformable barrier at 40 mph, with the driver's side of the vehicle taking the impact. The tests measured the potential for injury to the head, neck, chest and foot areas, and the risk of intrusion into the passenger compartment.

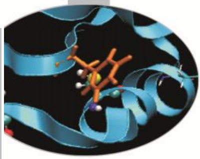
SUBJECTS: Japan; Safety; Passenger Vehicles;
SOURCE: Reuters, June 21, 2000;Japan;English

TI	**TOYOTA: Avalon Receives Top Score in Frontal Offset Crash Tests**
TXT	Toyota Motor Corp.'s Avalon received the top score -- a "good" rating earning a "best pick" -- in the 40 mile per hour frontal offset crash tests on new or updated vehicles. The tests were conducted by the Insurance Institute for Highway Safety, a nonprofit group funded by automobile insurers. Nissan Motor Co. Ltd.'s Maxima midsize sedan and Infiniti I30 luxury sedan, the Nissan Sentra small car and Mazda Motor Corp.'s Mazda MPV minivan all scored "average" marks. Isuzu Motors Ltd..'s Rodeo sport utility, also sold by Honda Motor Co. Ltd. as the Honda Passport, earned a "poor" rating due to high crash forces recorded on the crash dummy's head, indicating an increased likelihood of injury. In the crash tests, the vehicles were driven into a deformable barrier at 40 mph, with the driver's side of the vehicle taking the impact. The tests measured the potential for injury to the head, neck, chest and foot areas, and the risk of intrusion into the passenger compartment.
SUBJECTS	Japan; Safety; Passenger Vehicles;
SOURCE	Reuters
STATE	Japan
LANGUAGE	English
DATE	6/21/2000

ORG

TI	**TOYOTA: Avalon Receives Top Score in Frontal Offset Crash Tests**
TXT	Toyota Motor Corp.'s Avalon received the top score -- a "good" rating earning a "best pick" -- in the 40 mile per hour frontal offset crash tests on new or updated vehicles. The tests were conducted by the Insurance Institute for Highway Safety, a nonprofit group funded by automobile insurers. Nissan Motor Co. Ltd.'s Maxima midsize sedan and Infiniti I30 luxury sedan, the Nissan Sentra small car and Mazda Motor Corp.'s Mazda MPV minivan all scored "average" marks. Isuzu Motors Ltd..'s Rodeo sport utility, also sold by Honda Motor Co. Ltd. as the Honda Passport, earned a "poor" rating due to high crash forces recorded on the crash dummy's head, indicating an increased likelihood of injury. In the crash tests, the vehicles were driven into a deformable barrier at 40 mph, with the driver's side of the vehicle taking the impact. The tests measured the potential for injury to the head, neck, chest and foot areas, and the risk of intrusion into the passenger compartment.
SUBJECTS	Japan; Safety; Passenger Vehicles;
SOURCE	Reuters
STATE	Japan
LANGUAGE	English
DATE	6/21/2000

SuperComputing Applications and Innovation



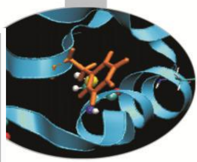
ORG

NAME

TERM

TI	**TOYOTA: Avalon Receives Top Score in Frontal Offset Crash Tests**
TXT	Toyota Motor Corp.'s Avalon received the top score -- a "good" rating earning a "best pick" -- in the 40 mile per hour frontal offset crash tests on new or updated vehicles. The tests were conducted by the Insurance Institute for Highway Safety, a nonprofit group funded by automobile insurers. Nissan Motor Co. Ltd.'s Maxima midsize sedan and Infiniti I30 luxury sedan, the Nissan Sentra small car and Mazda Motor Corp.'s Mazda MPV minivan all scored "average" marks. Isuzu Motors Ltd..'s Rodeo sport utility, also sold by Honda Motor Co. Ltd. as the Honda Passport, earned a "poor" rating due to high crash forces recorded on the crash dummy's head, indicating an increased likelihood of injury. In the crash tests, the vehicles were driven into a deformable barrier at 40 mph, with the driver's side of the vehicle taking the impact. The tests measured the potential for injury to the head, neck, chest and foot areas, and the risk of intrusion into the passenger compartment.
SUBJECTS	Japan; Safety; Passenger Vehicles;
SOURCE	Reuters
STATE	Japan
LANGUAGE	English
DATE	6/21/2000

SCAI SuperComputing Applications and Innovation



TERM

nouns

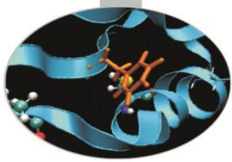
adjectives

verbs

TI	**TOYOTA: Avalon Receives Top Score in Frontal Offset Crash Tests**
TXT	Toyota Motor Corp.'s Avalon received the top score -- a "good" rating earning a "best pick" -- in the 40 mile per hour frontal offset crash tests on new or updated vehicles. The tests were conducted by the Insurance Institute for Highway Safety, a nonprofit group funded by automobile insurers. Nissan Motor Co. Ltd.'s Maxima midsize sedan and Infiniti I30 luxury sedan, the Nissan Sentra small car and Mazda Motor Corp.'s Mazda MPV minivan all scored "average" marks. Isuzu Motors Ltd.'s Rodeo sport utility, also sold by Honda Motor Co. Ltd. as the Honda Passport, earned a "poor" rating due to high crash forces recorded on the crash dummy's head, indicating an increased likelihood of injury. In the crash tests, the vehicles were driven into a deformable barrier at 40 mph, with the driver's side of the vehicle taking the impact. The tests measured the potential for injury to the head, neck, chest and foot areas, and the risk of intrusion into the passenger compartment.
SUBJECTS	Japan; Safety; Passenger Vehicles;
SOURCE	Reuters
STATE	Japan
LANGUAGE	English
DATE	6/21/2000

CINECA

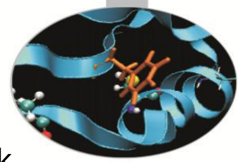
TI	**TOYOTA: Avalon Receives Top Score in Frontal Offset Crash Tests**
TXT	Toyota Motor Corp.'s Avalon received the top **score** -- a "good" **rating** earning a "best pick" -- in the 40 **mile** per **hour** frontal **offset crash tests** on new or updated **vehicles**. The **tests** were conducted by the Insurance Institute for Highway Safety, a nonprofit **group** funded by **automobile insurers**. Nissan Motor Co. Ltd.'s Maxima **midsize sedan** and Infiniti I30 **luxury sedan**, the Nissan Sentra small **car** and Mazda Motor Corp.'s Mazda MPV **minivan** all scored "**average**" **marks**. Isuzu Motors Ltd..'s Rodeo **sport utility**, also sold by Honda Motor Co. Ltd. as the Honda Passport, earned a "poor" **rating due** to high **crash forces** recorded on the **crash dummy**'s **head**, indicating an increased **likelihood** of **injury**. In the **crash tests**, the **vehicles** were driven into a deformable **barrier** at 40 **mph**, with the **driver**'s **side** of the **vehicle** taking the **impact**. The **tests** measured the **potential** for **injury** to the **head**, **neck**, **chest** and **foot areas**, and the **risk** of **intrusion** into the **passenger compartment**.
SUBJECTS	Japan; Safety; Passenger Vehicles;
SOURCE	Reuters
STATE	Japan
LANGUAGE	English
DATE	6/21/2000



tn.5.26.35 SOURCE Reuters
tn.5.26.35 DATE 6/21/2000
tn.5.26.35 MONTHYEAR 2000_06
tn.5.26.35 SUBJECTS Japan
tn.5.26.35 SUBJECTS Passenger_Vehicles
tn.5.26.35 SUBJECTS Safety
tn.5.26.35 STATE Japan
tn.5.26.35 LANGUAGE English
tn.5.26.35 ORG2 TOYOTA
tn.5.26.35 NN area
tn.5.26.35 NN automobile
tn.5.26.35 NN average
tn.5.26.35 NN barrier
tn.5.26.35 NN car
tn.5.26.35 NN chest
tn.5.26.35 NN compartment
tn.5.26.35 NN crash
tn.5.26.35 NN driver
tn.5.26.35 NN dummy
tn.5.26.35 NN foot
tn.5.26.35 NN force
tn.5.26.35 NN group
tn.5.26.35 NN head

tn.5.26.35 NN hour
tn.5.26.35 NN impact
tn.5.26.35 NN injury
tn.5.26.35 NN insurer
tn.5.26.35 NN intrusion
tn.5.26.35 NN likelihood
tn.5.26.35 NN luxury
tn.5.26.35 NN mark
tn.5.26.35 NN mile
tn.5.26.35 NN neck
tn.5.26.35 NN offset
tn.5.26.35 NN passenger
tn.5.26.35 NN potential
tn.5.26.35 NN rating
tn.5.26.35 NN risk
tn.5.26.35 NN safety
tn.5.26.35 NN score
tn.5.26.35 NN sedan
tn.5.26.35 NN side
tn.5.26.35 NN sport
tn.5.26.35 NN test
tn.5.26.35 NN utility
tn.5.26.35 NN vehicle

tn.5.26.35 UTERM crash_test
tn.5.26.35 UTERM top_score
tn.5.26.35 ORG honda_motor_co
tn.5.26.35 ORG insurance_institute for …
tn.5.26.35 ORG isuzu_motors
tn.5.26.35 ORG mazda_motor
tn.5.26.35 ORG nissan_motor
tn.5.26.35 ORG toyota_motor
tn.5.26.35 UNAME avalon
tn.5.26.35 UNAME honda_passport
tn.5.26.35 UNAME infiniti_i30
tn.5.26.35 UNAME maxima
tn.5.26.35 UNAME mazda_mpv
tn.5.26.35 UNAME rodeo



	W_1	W_2	...									W_m
Doc i	1	1	1	1	0	1	1	0	1	0	1	0
Doc j	1	0	0	1	1	1	0	1	0	0	0	1

$$N_{11}=\sum_{k=1}^{m} x_{ik}\, x_{jk}$$

$$N_{10}=\sum_{k=1}^{m} x_{ik}\,(1-x_{jk})$$

$$N_{01}=\sum_{k=1}^{m} (1-x_{ik})\, x_{jk}$$

$$N_{00}=\sum_{k=1}^{m} (1-x_{ik})\,(1-x_{ik})$$

- Similarity index

$$\mathbf{s(i,j)} = \frac{\boldsymbol{a\, N_{11}}}{\boldsymbol{b\, N_{11} + c\,(N_{10} + N_{01})}} \Rightarrow$$

- **Condorcet** ***a*=*b*=1 *c*=1/2**
- **Dice** ***a*=*b*=1 *c*=1/4**

- Similarity threshold

if s(i,j) > α $\Rightarrow$ **Doc_i e Doc_j are similar** α in [0,1]

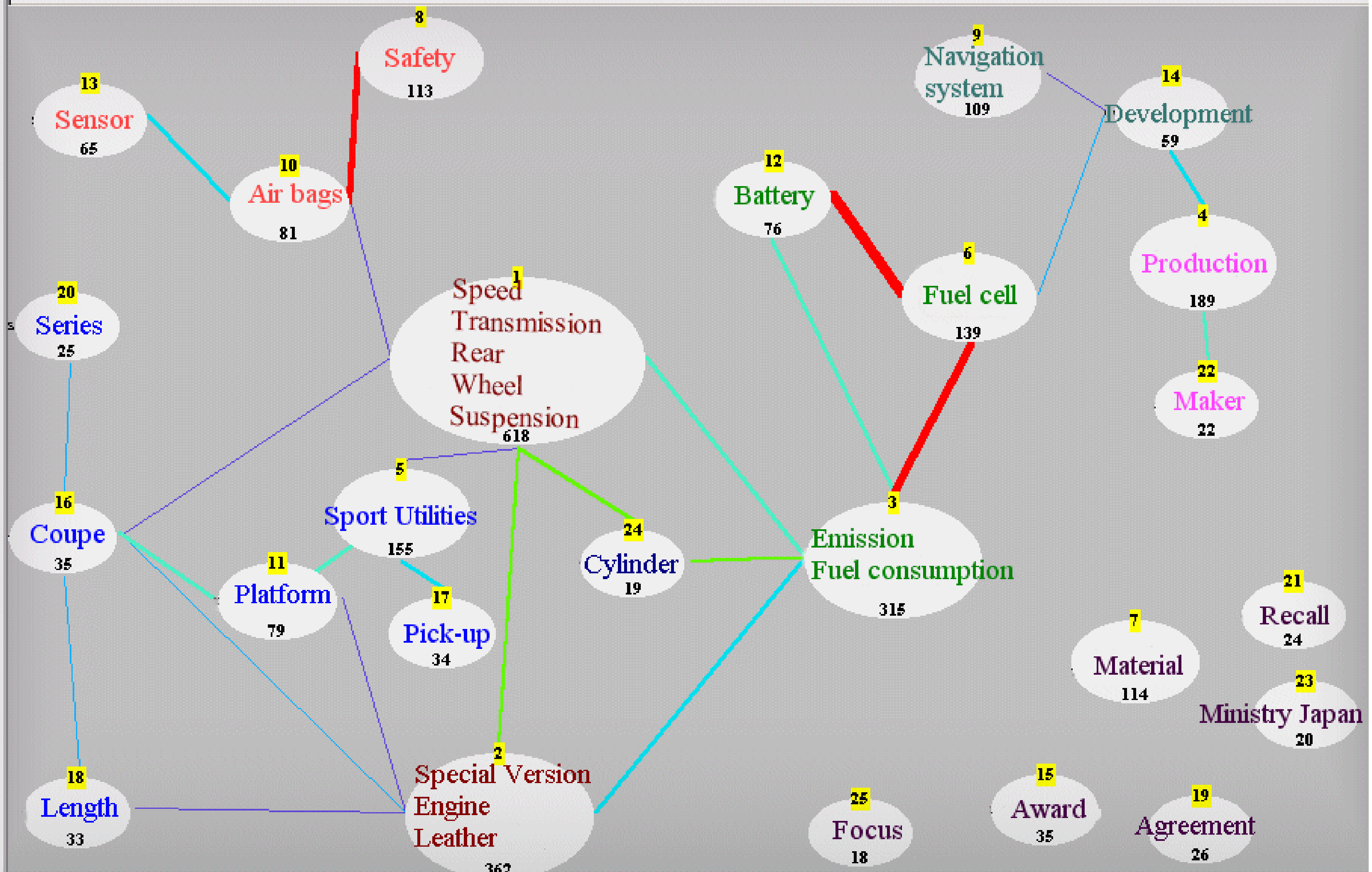
- default: $\alpha = \mathbf{0.5}$

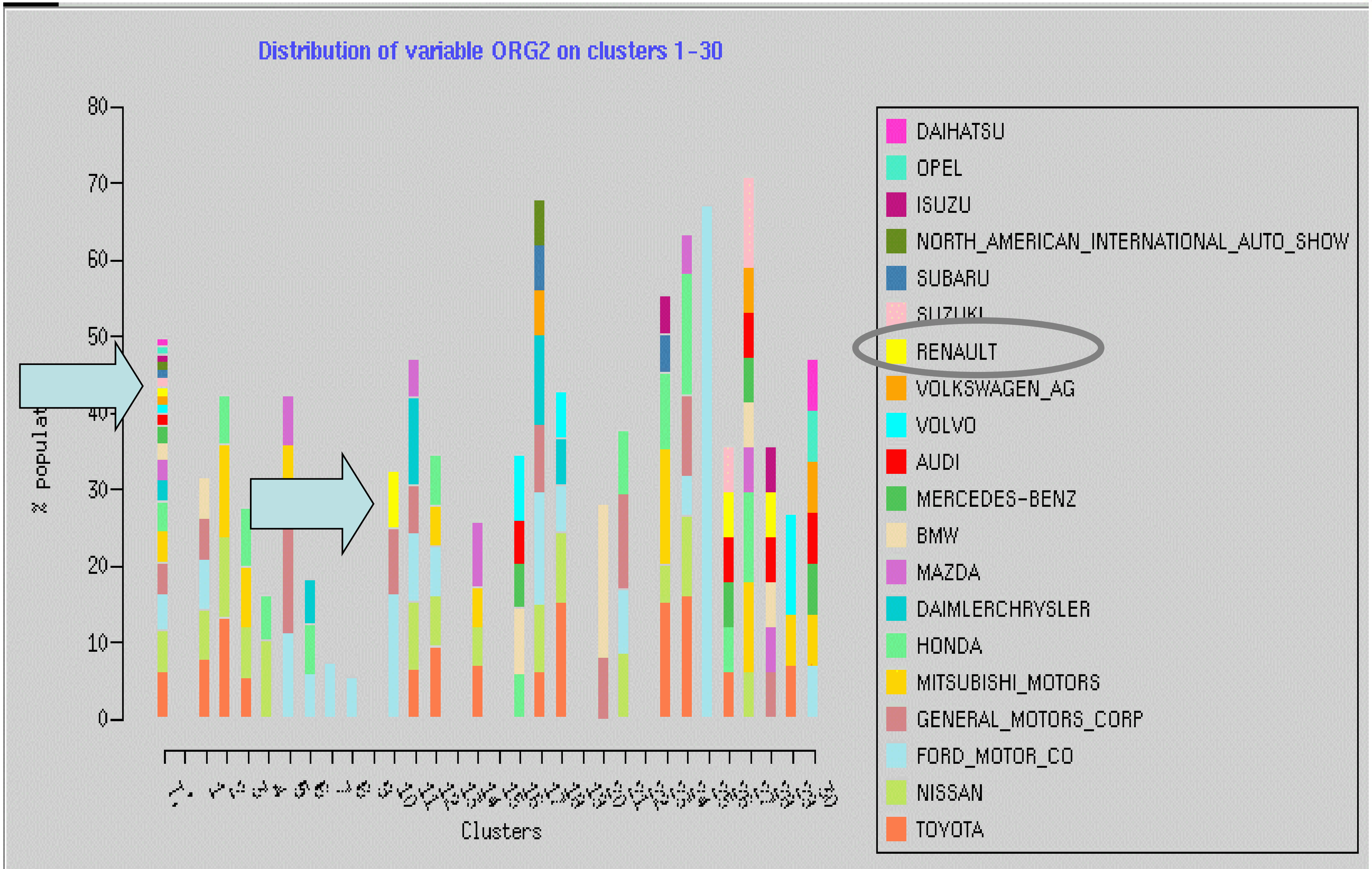
- Weighting system

$$N_{11}=\sum_{k=1}^{m} x_{ik}\, x_{jk}\, \boldsymbol{w}_k \quad (N_{10}=..\ N_{01}=..) \Rightarrow$$

- $\boldsymbol{w_k = 1 / x_{.k}}$
- $\boldsymbol{w_k = \log(N / x_{.k})}$

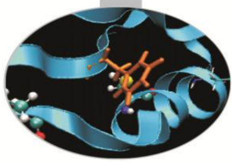
25 first clusters




Distribution of variable ORG2 on clusters 1-30
80
70
60
50
40
30
20
10
0
% populat
Clusters
DAIHATSU
OPEL
ISUZU
NORTH_AMERICAN_INTERNATIONAL_AUTO_SHOW
SUBARU
SUZUKI
RENAULT
VOLKSWAGEN_AG
VOLVO
AUDI
MERCEDES-BENZ
BMW
MAZDA
DAIMLERCHRYSLER
HONDA
MITSUBISHI_MOTORS
GENERAL_MOTORS_CORP
FORD_MOTOR_CO
NISSAN
TOYOTA

10-air bags

Use case: Medline abstracts



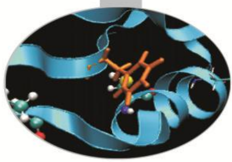
Gene "Dictionary":

gene name	**alias**
CDKN1B	P27KIP1
IFI27	P27
P27	P27

Information extraction

PMID	10532419
DA	19991124
DCOM	19991124
LR	20001218
IS	1433-7398
VI	16
IP	1
DP	1999
TI	**Anomalous p27kip1 expression in a subset of malignant gliomas.**
	Find similar documents
PG	17-21
AB	p27Kip (p27) expression was immunohistochemically investigated in 28 astrocytic tumors, and compared with the cell proliferation index (MIB-1 staining index). Normal rat brains and surgical specimens from human nonneoplastic brain lesions were used as controls. In the rat brains, the astrocytes were exclusively p27-positive. The reactive astrocytes in various disease processes sometimes lacked p27 expression. The distribution of p27-positive cells was uniform in low-grade astrocytomas and heterogeneous in high-grade tumors. Double staining of p27 and MIB-1 showed a reciprocal pattern in most cases. The frequency of p27 expression was inversely correlated with MIB-1 staining index and tumor grade. However, several malignant gliomas showed high p27 expression in spite of high MIB-1 staining indices. In such cases, MIB-1-positive cells were occasionally p27-positive. In this paper we discuss the etiology of the anomalous p27 expression in a subset of malignant gliomas.
ORG	Department of Neurosurgery, Shiga University of Medical Science, Ohtsu, Japan. snakasu@belle.shiga-med.ac.jp
AU	Nakasu S Nakajima M Handa J
LG	eng
PT	Journal Article
CY	JAPAN
TA	Brain Tumor Pathol
JC	CYI
JID	9716507
RN	0 (Cyclin-Dependent Kinases) 0 (Enzyme Inhibitors) 0 (Microtubule-Associated Proteins)

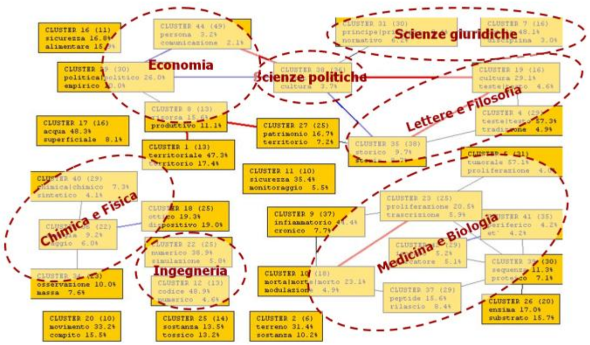
20000219 gene CDKN1B
20000219 gene IFI27
20000219 gene P27



Text mining: thematic clustering

MOLE

Mining On-Line Expert



- MOLE allows to :
- organize documents into thematic groups, providing an overview of the content
- identify new topics, relationships between subject areas, time trends, …

- Examples:

MedMole, PatMole, GiuriMole

- http://giurimole.cineca.it/
- http://genmole.cineca.it/

Text mining: automatic classification

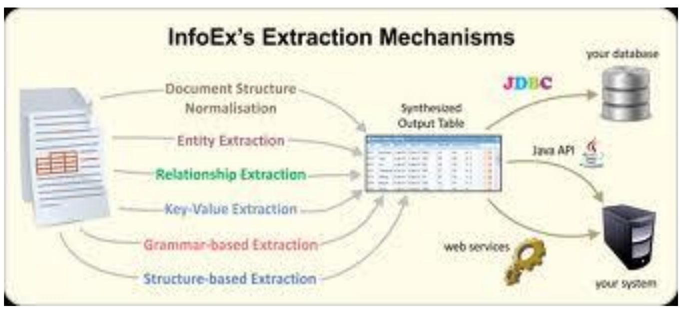
AreaMapping
IPCmapping
GiuriClass

- Technology that automatically learns accurate classification rules with respect to user-defined categories. These rules are exportable to classify automatically and efficiently large quantities of documents.

- Examples:
- http://eric-class.test.cineca.it/
- Classification of research projects in areas and disciplines (AreaMapping) and in classes and subclasses of the International Patent Classification (IPCmapping)

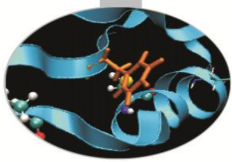
- http://classificatore.test.cineca.it/
- Court decisions classification into the main topics of civil matter (GiuriClass)

Text mining: information extraction



- Extraction of specific information from unstructured texts (Named Entities, names of genes, ... any other information required by the application).
- Based on automatic techniques (machine learning) or manual coding of knowledge (knowledge engineering).

- Examples:
- Astrea (GiuriMole)
- Medline (MedMole)
- Finance Police reports (VerbaliGdF)

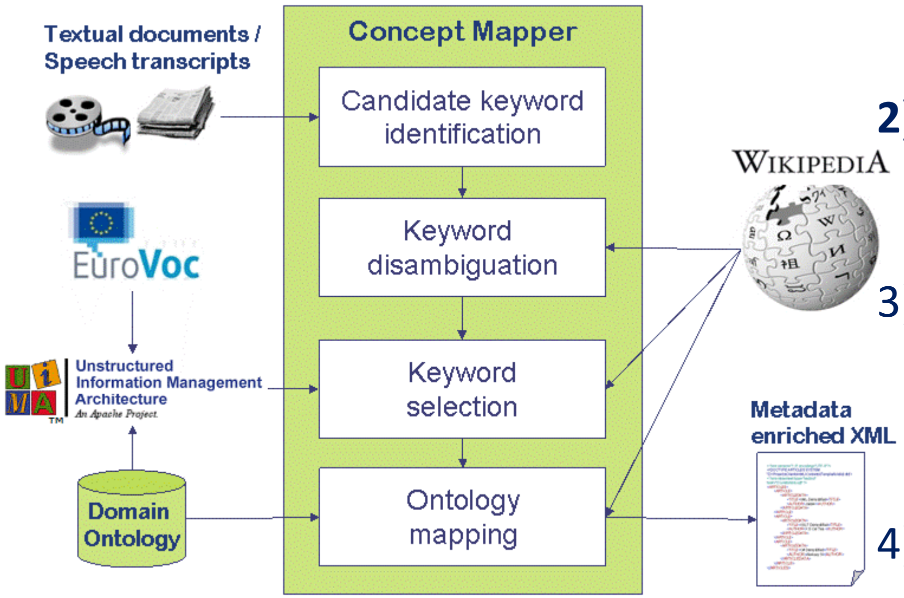


Text mining: semantic tagging

- http://conceptmapper.cineca.it/

- Concept Mapper allows to extract the most relevant concepts from a document. It can be used to automatically annotate texts and to generate semantic metadata or as an intermediate step to index documents, group, classify or map them to a domain ontology.

Concept Mapper – the components



1) **Identification** of nominal phrases (e.g. "turbines", "carbon monoxide", "President of the United States"), through linguistic analysis;
2) **Validation** through identification of the corresponding concept/s in Wikipedia (anchor search);
3) Word sense **disambiguation** to the correct underlying concept (when more than one page is retrieved) based on concept commonness and context relatedness;
4) Concept **selection** on the basis of relevance measures (internal and external relatedness, frequency, …);
5) Concept **mapping** to an ontology, on the basis of the group relatedness, which exploits the ontology structure and the relations among ontology concepts to define a new context.

Semantic relatedness

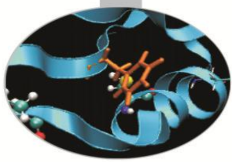
The semantic relatedness between two concepts is based on the number of outgoing and incoming links shared by the respective Wikipedia pages.

$$relatedness(a,b) = \frac{\log(\max(|A|,|B|)) - \log(|A \cap B|)}{\log(|W|) - \log(\min(|A|,|B|))}$$

where A, B and W are the set of links incoming and outgoing from the pages a, b and from all pages of Wikipedia

The semantic relatedness of a concept to a context the weighted average of its relatedness to each semantic concept that makes up the context.

The context can be internal or external to the document (item relatedness or domain relatedness).



Concept Mapper – what it does

- **Identification of the most relevant concepts** in textual documents (or speech transcripts), where relevance can be with respect to the document context only, or with respect to a reference domain context. This feature enables validation of the keywords extracted from the text (assessing meaningfulness) and avoids selecting the irrelevant ones.
- **Automatic annotation** of texts with concepts and links to the corresponding Wikipedia page.
- **Semantic metadata generation** that enrich the content with new information semantically related to it, such as Wikipedia categories, redirects and anchors, translation to other languages, …
- **Mapping document content to an ontology** by automatically identifying the correct correspondences between the identified relevant concepts (Wikipedia-like annotation of the document) and their formalization in the ontology (ontology annotation).

Tools and linguistic resources for Text Mining

(among many others)

Apache Tika

Tree Tagger

Wikipedia

Stanford University NLP Natural Language Processing

R

Unstructured Information Management Architecture
An Apache Project

Cluto

Rainbow