Text Mining
School on Scientific Data Analytics and Visualization

Roberta Turra, Cineca

9 June 2015
Overview

Unstructured content

Multimedia Mining

Unsupervised (clustering)

Supervised (classification)

Pattern Recognition

Knowledge Engineering

Rule Based

Dictionary Based
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
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
Phase 1: collecting

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

- pre-processing
  - parsing
  - integration
  - transformation to a common format
The Process
Phase 2: 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
Phase 2: 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.

<table>
<thead>
<tr>
<th></th>
<th>team</th>
<th>coach</th>
<th>pla</th>
<th>ball</th>
<th>score</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Document 1</strong></td>
<td>3</td>
<td>0</td>
<td>5</td>
<td>0</td>
<td>2</td>
</tr>
<tr>
<td><strong>Document 2</strong></td>
<td>0</td>
<td>7</td>
<td>0</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td><strong>Document 3</strong></td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
</tbody>
</table>
The Process
Phase 3: 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
Phase 4: 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

Textual part

Meta-information

Linguistic analysis (keyword extraction and stemming)

Information Extraction (meta-information integration)
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.
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
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.
**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
**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                              |
| STATE    | Japan                                |
| LANGUAGE | English                              |
| DATE     | 6/21/2000                            |
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
STATE Japan
LANGUAGE English
DATE 6/21/2000
- **Similarity index**
  
  \[
  s(i,j) = \frac{a N_{11}}{b N_{11} + c (N_{10} + N_{01})}
  \]

- **Similarity threshold**
  
  if \( s(i,j) > \alpha \)  
  Doc\(i\) e Doc\(j\) are similar

  \(\alpha\) in \([0,1]\)

- **Weighting system**
  
  \[
  N_{11} = \sum_{k=1}^{m} x_{ik} x_{jk} w_k \quad (N_{10}=.., N_{01}=..)
  \]

  \(w_k = 1 / x_k\)

  \(w_k = \log( N / x_k)\)

- **Condorcet**  
  \(a=b=1, c=1/2\)

- **Dice**  
  \(a=b=1, c=1/4\)
10-air bags
**Use case: Medline abstracts**

<table>
<thead>
<tr>
<th>PMID</th>
<th>DA</th>
<th>DCOM</th>
<th>LR</th>
<th>IS</th>
<th>VI</th>
<th>IP</th>
<th>DP</th>
<th>TI</th>
</tr>
</thead>
<tbody>
<tr>
<td>10532419</td>
<td>19991124</td>
<td>19991124</td>
<td>20001218</td>
<td>1433-7398</td>
<td>16</td>
<td>1</td>
<td>1999</td>
<td>Anomalous \textit{p27kip1} expression in a subset of malignant gliomas.</td>
</tr>
</tbody>
</table>

**Gene “Dictionary”:**

<table>
<thead>
<tr>
<th>gene name</th>
<th>alias</th>
</tr>
</thead>
<tbody>
<tr>
<td>CDKN1B</td>
<td>P27KIP1</td>
</tr>
<tr>
<td>IFI27</td>
<td>P27</td>
</tr>
<tr>
<td>P27</td>
<td>P27</td>
</tr>
</tbody>
</table>

\textit{p27Kip1} 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 \textit{p27}-positive. The reactive astrocytes in various disease processes sometimes lacked \textit{p27} expression. The distribution of \textit{p27}-positive cells was uniform in low-grade astrocytomas and heterogeneous in high-grade tumors. Double staining of \textit{p27} and MIB-1 showed a reciprocal pattern in most cases. The frequency of \textit{p27} expression was inversely correlated with MIB-1 staining index and tumor grade. However, several malignant gliomas showed high \textit{p27} expression in spite of high MIB-1 staining indices. In such cases, MIB-1-positive cells were occasionally \textit{p27}-positive. In this paper, we discuss the etiology of the anomalous \textit{p27} expression in a subset of malignant gliomas.

**ORC**

Department of Neurosurgery, Shiga University of Medical Science, Ohtsu, Japan. snakasu@belle.shiga-med.ac.jp

**AU**

Nakasu S
Nakajima M
Hamba T

**LG**

eng

**PT**

Journal Article

**CY**

JAPAN

**TA**

Brain Tumor Pathol

**JC**

CVI

**JID**

9716307

**RN**

0 (Cyclin-Dependent Kinases)
0 (Enzyme Inhibitors)
0 (Microtubule-Associated Proteins)
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://giurimole.cineca.it/)
- [http://genmole.cineca.it/](http://genmole.cineca.it/)
Text mining: automatic classification

- 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 is 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)