Introducci' on the T'ecnicas Miner'ıa Text

Maria-Amparo Vila

vila@decsai.ugr.es

Investigaci'on group Databases and Intelligent Systems Informaci'on https://idbis.ugr.es/

Department of Sciences Computaci' on artificial Intelligence University of Granada

Master in Ingenier'ıa

Outline other slideshow

The problem Miner'ıa Text

The concept of "Miner'ıa Text" (TM) is something that is in discusi'on todav'ıa. It podr'ıa de fi ned as:

"Extracci'on process knowledge or patterns, previously unknown, non-trivial and interesting (potentially ' utiles) and understandable by users from of unstructured text documents. ".

Text Mining is a data mining extensi'on where the discovery was made from unstructured databases.

The problem Miner'ıa Text

should not be confused with Text Mining Recuperaci'on of Informaci'on from textual databases.

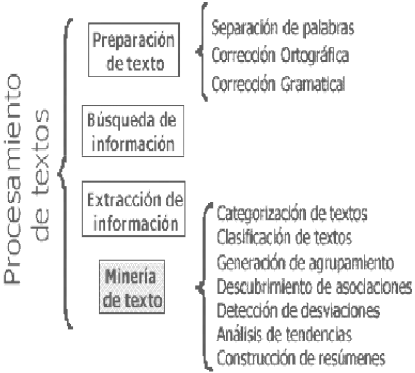
The recuperaci'on of informaci'on looking "documents" according to some requirements.

TM seek:

- unknown knowledge
- Understandable by users
- nontrivial
- Interesting

The problem Miner'ıa Text

It is considered that the TM is' last step in processing texts:



The problem Miner'ıa Text

Podr'ıa think is that you can directly apply the DM cl'asicas t'ecnicas the textual informaci'on. Nothing is further from reality . The DM works with databases with known schema. Each text document is an ordered colecci'on of words and signs of separaci'on meaning associated with situaci'on in the text which is determined by restrictions sint'actico and sem'antico type. There are semi-structured texts such as documents written in XML. In TM, the data include:

- inherently unstructured
  - impl'ıcita structure
  - Much greater wealth than in structured cases
- ambiguous
- multi-lingual

The problem Miner'ıa Text

? This lack of structure is the biggest problem of the TM and involves the need to preprocess the texts, passing them to a

intermediately

- Bags (bags) of t'erminos
- matrix structures (datasets)
- sem'anticas conceptual graphs or networks.
- Type structures "ontolog'ıa"

Preparaci'on phase data inherent in any process of extracci' on knowledge, it is crucial.

- The obtenci'on of the "parison" in turn involves use of extracci' t'ecnicas on knowledge,
- In many cases, it is not clear then, where the process ends preparaci'on data and begins to miner'ıa.

The problem Miner'ıa Text

A simple scheme of this ser'ıa process:

The problem Miner'ıa Text

? Some authors simply define the TM as an interdisciplinary field that includes elements of:

- Recover of information
- Extracci'on of informaci'on by LING computational uistica
- Grouping (clustering)
- Categorizaci'on,
- Other t'ecnicas DM

The problem Miner'ıa Text

? All authors agree that est'an process TM includes the following phases:

one. I preprocessing

two. Miner'ıa (proper)

3. Visualizaci'on In

funci' on the complexity of the preprocessing we find the possibility that this step occurs or not a complex intermediate form.

Text preprocessing Miner'ıa

preprocessing Sint'actico

b'asica Idea

It is free text processing so that the output can be treated in an automated manner.

We must move from unstructured data structure data. At first T'erminos bag, possibly annotated

Stages in the processing sint'actico

one. Tokenizaci'on

two. Recognition / Eliminaci'on signs of puntuaci'on

3. Recognition / Eliminaci'on of "vac'ıas words" (stopwords)

Four. M'recognition of words ultiple. (N-grams)

5. Sint'acticos recognition types (POS)

6. Lematizaci'on (Steaming)

Text preprocessing Miner'ıa

preprocessing Sint'actico

Tokenizaci'on

- This is the text from "tokens" strings that represent words.
- Several "tokens" represent the same t'ermino or word.
- In the end we will have bags of words with its occurrence in each t'ermino
- Common problems:

  Prepare the document to extract text (PDF, HTML, XML). Different areas of the document may have to be treated differently.

  Words "composite" requires preprocessing sem'antico we will see later

Text preprocessing Miner'ıa

preprocessing Sint'actico

Recognition / eliminaci'on signs of puntuaci'on

Problems:

- It is detected when a sign is truly puntuaci'on

  ()?!? "Are delimiters and can be tokens.

  - ',; Y . They may form part of or be t'ermino delimiters. Habr'a likely to make an in-depth an'alisis if there are many signs.

- It is very dependent on language
- At the end of the process they are removed from the set of t'erminos document.

Text preprocessing Miner'ıa

preprocessing Sint'actico

Recognition / Eliminaci'on of "vac'ıas words" (stopwords)

- They are words that do not provide informaci'on from a point of view not LING u'ısitico
- They are essentially functional role
- Seg' are removed from the text a A given list
- It is very dependent on language

http://www.navigla.es/posicionamiento-seo/palabrasstopwords-seo-espanol/

You can find the list of words vac'ıas espa~ nol using Google's engine B' earch

M'recognition of words ultiple.

- There are groups of words that have meaning for themselves: Data base, etc. Operating System these groups should be treated as t'erminos. They can be regarded as a particular case of n-grams (n-grams)

- His autom'atica detecci'on is based on the idea that a ser'am'as n-gram n often together any words.

Text preprocessing Miner'ıa

preprocessing Sint'actico

M'recognition of words ultiple. To detect ngram one possible algorithm:

one. M'aximo set a value of n (2,3, ..) of words to be taken together. Setting a threshold measurement aparici'on n-gram: Minmes

two. For k = 2..N

2.1 k-sequences of consecutive words are discussed in the text T_k.

2.2 those are selected T_k such that:

$$AM(T_k) = k \left(\log_{10} freq(T_k)\right) \frac{freq(T_k)}{\sum_{Word_i \, two \, T_k} freq(word_i)} \quad minmes$$

2.3 If there alg' a $T_{k\,one}$, $k >$ two I est'e included in T_k remove $T(k$ one) the set of n-grams selected

Text preprocessing Miner'ıa

preprocessing Sint'actico

Recognition / Eliminaci'on of sint'acticos types (Part of Speech, POS)

- For certain applications it is necessary to recognize the grammatical functions of t'erminos name, own name, verbs in different times etc.
- There are algorithms that allow you to tag the t'erminos in this regard.

  LING algorithms based on rules u'ısticas. M'as are old.
  Autom'atico learning algorithms. They are based on recognizing a categor'ıa for t'ermino and assign the next word his M'as categor'ıa likely seg'
  one language. PE in English ANN (adjective name name) is a likely sequence in espa~ nol NNA is likely M'as

- Obviously they depend on the language

Text preprocessing Miner'ıa

preprocessing Sint'actico

Recognition / Eliminaci'on of sint'acticos types (Part of Speech, POS)

Some categories of POS in ingl'es

Tag	Description
CC	Coordinating conjunction
CD	Cardinal number
DT	Determiner
EX	Existential there
FW	Foreign word
IN	Preposition or subordinating conjunction
JJ	Adjective
JJR	Adjective, comparative
JJS	Adjective, superlative
LS	List item marker
MD	Modal
NN	Noun, singular or mass
NNS	Noun, plural
POS	Possessive ending
UH	Interjection
VB	Verb, base form
VBD	Verb, past tense
VBG	Verb, gerund or present participle
VBN	Verb, past participle
VBP	Verb, non-3rd person singular present
WDT	Wh-determiner

Text preprocessing Miner'ıa

preprocessing Sint'actico

Stemming (Steaming)

- The different ways the same word usually problem'aticas for an'alisis word, as they have different ortograf'ıa and similar meaning (for example, learn, learn, learning)

- we mean by *Lematizaci'on (Steaming or Lematization)* the process transform a word in his "standarizada root"
- For Ingl'es is not a big problem, there are algorithms available p' ublicamente that give good results. M'as is known Porter algorithm

- for espa~ nol no algorithm generally recognized as the Porter but there are several available algorithms.

Text preprocessing Miner'ıa

preprocessing Sint'actico

Stemming (Steaming)

- Is not always Useful carry out the lematizaci'on.

  Overall for cases in which they work frequently t'erminos relizarlo interesting because it summarizes several t'erminos a s'olo and increases frequency of this.

  But it can cause problems if a subsequent sem'antico preprocessing is done

- In any case the lematizaci'on habr'a to carry out labeling and eliminaci'on despu'es if the sint'acticas forms (POS) unwanted because lematizaci'on changes the function of the terms.

Text preprocessing Miner'ıa

preprocessing Sint'actico

Stemming (Steaming)

Lematizaci'on rules in ingl'es

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

Text preprocessing Miner'ıa

preprocessing Sem'antico

b'asica Idea

Once clean and labeled the t'erminos relations sem'antico type are used to reduce them again.

Possible relationships between t'erminos

- *Synonymy :* differently, the same meaning (class lecci'on)
- *homonymy :* same meaning different form (bank financial institution, sit site)
- *Polysemy :* Similarly, different signi fi ed related (Bank Blood Bank)
- *hyponymy* A word is a subclass of another (dog, animal)

Text preprocessing Miner'ıa

preprocessing Sem'antico

Desambiguaci'on

Process by which is assigned to multiple t'erminos s'olo one that has the mOsmol meaning that all of them

- There are tools that help you work with sin'onimos and hip'onimos, reaching assign each set of terms to a class semantics (desambiguaci'on).
- The famous M'as is Wordnet; but there is no algorithm that solves the problem completely and less in several languages.

The desambiguaci'on is an open problem and has to be made "by hand" for every problem, if needed

Reducci' on variables

b'asicas ideas

problems

- Colecci'on given a document each est'a¡ characterized by t'erminos.

- T'erminos assembly is called a colecci'on dictionary

  and may have several thousand items.

- The REPRESENTATION of each document is based on the dictionary

- T'ecnicas of reducci'on the aim of t'erminos *reduce n'*

  umber of t'erminos Dictionary

Reducci' on variables

b'asicas ideas

direct T'ecnicas

They are used directly in the preprocessing. Reduce Dictionary

one. The eliminaci'on of vac'ıas words

two. M'recognizing words ultiple

3. The eliminaci'on types of sint'acticos no signi fi cant

Four. the lematizaci'on

5. the desambiguaci'on

Reducci' on variables

T'ecnicas based on measurements

b'asico process

- For each t'ermino k and each document i a "measure of importance" is set w_{ik} the t'ermino k in the document i.
- To measure the importance of a t'ermino in a document colecci'on set $w_k = Agg_n$

  $_{i=\text{one}}$ w_{ik} where Agg is a AGGREGATION measure may be the m'aximo, m'ınimo, average, sum etc.

- the seg' t'erminos are ordered a w_k and selected the m first or a percentage of the total (ie 75% better) etc.

Reducci' on variables

T'ecnicas based on measurements

Measures based on frequencies

- Be n_{ik} = n' umber of times the t'ermino appears k in the document i, Y n_i n' umber of associated t'erminos i
- The frequency of each t'ermino k in i It is given by $F_{ik} = \frac{n_{ik}}{n_i}$
- If we work with frequencies take measures $w_{ik} = F_{ik}. w_k$ can be: $w_k = t_k = \text{P}_{ni=\text{one}} F_{ik}$ Total or frequency $w_k = \text{P}_{ni=\text{one}} F_{ik} / n$ average frequency.

Reducci' on variables

T'ecnicas based on measurements

Discriminaci'on measures

issue

Are frequent M'as t'erminos those who best represent a document colecci'on ?.

T'erminos no better representar'an certain documents appearing in some and not in others, discriminate some of the documents?

As tf * idf Yes d_k It is the n' Number of documents in which it appears

k, the larger less discriminates *k.* It is defined then:

$$idf_k = log_{two}(N / A_k) + \text{one Y } w_{ik} = tf_{\leftarrow} idf_{ik} = F_{ik}\, idf_k$$

Reducci' on variables

T'ecnicas based on measurements

Discriminaci'on measures

Noise measurement For each t'ermino it is defined:

$$n_k = \sum_i (F_{ik}/t_k) Log_{two}(F_{ik}/t_k) Y s_k = log_{two} t_k n_k$$

$$\text{so } w_{ik} = F_{ik} s_k$$

Measures based on similarity of documents For colecci'on

document it defines the average similarity between them

, for each t'ermino k is defined $_k$ as the *average similarity suppressing t'ermino k documents.*

$_k = _k$ discriminator measures the power k.

$$w_{ik} = F_{ik\,k}$$

Reducci' on variables

T'ecnicas based on principal component selecci'on

Initial model: *The vector model documents / t'erminos*

Every document is represented as a vector: $d = (w_{one}, .. w_M)$ where each w_i represents the weight that the t'ermino i in document d.

issue

Although it has made a *cleaning process* "empty words", eliminaci'on of sin'onimos etc .., and *process frequency reducci'on* we have hundreds of variables, ie M can be very large. Much weights can be zero in a particular document, that is, the vectors are very large and "empty"

To do a process of reducci'on dimensions

The process reducci'on dimensional transforms a vector ($w_{one}, .. w_M$) where d is very large in another (v one, .. v_m) where m may be fixed.

Reducci' on variables

T'ecnicas based on principal component selecci'on

The problem

Given a variable n dimensional¯ $x = (x_{\text{one}}, .. x_n)$ Y m values the same x_{ij}, i two { one, , m}, j two { one, .., n}, find standard linear transformations (SLC) that "summarize" the best possible data, capturing most variance thereof.

intuitive idea

If the items are considered as a point cloud R_n, all can be enclosed in an ellipsoid, the average center, whose matrix is the covariance matrix.

The axes of the ellipsoid are a rectangular coordinate system, if we make a change to this coordinate system, the point spread along axis

Reducci' on variables

T'ecnicas based on principal component selecci'on

The model matem'atico

Be ¯ μ the average ¯ x Y ^ its covariance matrix, is find a linear transformaci'on ¯ $y = 0(¯ x ¯ \mu)$ such that new coordinate axes are the axes of the ellipsoid. It is proved to be a matrix such that:

$$0 \,\hat{}\, = \leftarrow$$

where,

0BBBBBBB @ 1CCCCCCCA

one

$$\leftarrow = diag(¯) = \ddots \quad \ddots \quad \ddots \quad n$$

Reducci' on variables

T'ecnicas based on principal component selecci'on

Vector values ¯ veri fi ed one , . . $_n$ and are the "Eigenvalue" of the covariance matrix and associated with each $_j$ there is a "eigenvector" ¯ $_j$ that is the jth column of the matrix and verify c'andose:

$$8\ j\ two\ \{\text{one}, \ , n\}\ and_j = \bar{\ }_j{}^0(\bar{x} \quad \mu)$$

Y_j It is called jth principal component.

$$8\ k, j\ two\ \{\text{one}, \ , n\}\ Cov\ (and_j, Y_k) = 0 \quad Var\ (and_j) = _j$$

$$Var\ (and_{\text{one}}) \qquad . . \ Var\ (and_n)$$

Reducci' on variables

T'ecnicas based on principal component selecci'on

Proporci'on explained variance

? The proporci'on of variance explained by *k* factors is

(1 + .. + *k)* / 1 + .. + *n)* and it allows us to reduce the dimensionality of space. That is express fen'omeno less variables.

How many components make ?:

- At least 90% of variance explained
- All eigenvalues that are greater than the average of the same.

- If the matrix is used instead Correlation of the covariance matrix eigenvalues greater than 1.

  The proporci'on of variaci'on explained variable j by k component is given by → (x j, Y k) = r ik, with these values can identify components and identify one semantics for them.

Reducci' on variables

T'ecnicas based on principal component selecci'on

New REPRESENTATION of documents yt'erminos

The expresi'on:

$$\bar{y} = 0 (\bar{\ } \quad x \bar{\ } \quad \mu)$$

We can get:

$$\bar{x} = \leftarrow 0$$

So that, if ⇤ matrix represents the new "factors", then

REPRESENTATION represents each document in the

new space factor and the REPRESENTATION each t'ermino

in funci' on of the factors. The same ocurres if the first k elements are taken ⇤, the first k

columns and the first k

columns .

An'alisis latent semantics

In our case the model identi fi ca c'omo:

- The items are documents
- T'erminos weights in each document are the variables
- New combinations of factors are interrelated terms

- Representations of documents funci'on new factors can be used in many different ways

Reducci' on variables

T'ecnicas based on principal component selecci'on

An'alisis latent semantics

- the fi c gr'a representations according factors can give an idea of a first cluster of documents
- T'erminos representations in funci'on factors can give an idea of the semantics of them. To the extent that some authors speak of new "concepts" identi fi ed by the factors.

Previous experiences show that:

- They can be reduced to 200 factors, several thousand spaces t'erminos M'as retaining 90% of the variance
- The reducci'on dimensions substantially not affect processes posteriones TM as clustering

Regrouping and TM: some approaches

cl'asico model

- Input data: *vector or vector model reduced*
- Distance used: *The measure cosine*

  Since each document is a vector is calculated cosine'angulo forming. Yes $t_1 = (w_{\text{eleven}} \ldots w_{\text{one } d})$ Y $t_2 = (w_{\text{twenty-one}} \ldots w_{\text{two } d})$ are two vectors, then:

$$cos\,(t_{\text{one}}, t_2) = (t_{\text{one}} \qquad t_2) / |t_1||t_2|$$

where represents the scalar product and |. | the modulus, ie:

$$cos\,(t_{\text{one}}, t_2) = \frac{P_{dj=\text{one}}\, w_{\text{one } j}\, w_{\text{two } j}}{\sqrt{P_{dj=\text{one}}\, w_{\text{two}}^{\text{one } j}}\ \sqrt{P_{dj=\text{one}}\, w_{\text{two}}^{\text{two } j}}} \qquad \text{(one)}$$

- If a factor model used can be used without problems eucl'ıdea distance

Regrouping and TM: some approaches

? It is very common to use the 's method k-means with cosine distance. some variants that allow improved m'etodo are used:

- Selecci' on centroids by
  - An initial cluster jer'arquico a set peque~ do not give documents.
  - An'alisis in previous and subsequent components REPRESENTATION
  - Other t'ecnicas fractionation set of documents
- Utilizaci'on improvement of the k-means:
  - 'S method k-means continuous continuously centroids are calculated not at the end of each step

Regrouping and TM: some approaches

- Utilizaci'on improvement of the k-means:
  - 'S method k-means in two phases, where a refinement of the process is performed at each stage of the algorithm to avoid falling into "local'optimo" of the function of coherence.
  - 'S method k-means by biparticí'on.

? There are approaches that recommend the use of M'as t'ecnicas particional elaborate grouping such as m'etodos Medoids (CLARANS) om'etodos an'alsis based on density, advanced jer'arquicos om'etodos (BIRCH).

Regrouping and TM: discovery events

issue

We are surrounded by texts "have events", ie "news". The discovery of events in the context of news is the identi fi caci'on of "stories" that correspond to new events or previously identi fi ed. The detecci'on of events can be done in two ways:

- Retrospectively
- on line

Regrouping and TM: discovery events

New events occur continuously and news relating to them are usually in the form of "blowout" or "explosions" with the following caracter'ısticas:

one. You are grouped in time,

two. The "explosion" associated with new events are separated for a certain amount of time,

3. Every new event is usually acompa~ swim a change in the type of t'erminos and vocabulary used to describe the associated news,

Four. The events usually take a time window of between one and duraci'on four weeks.

These four observations seem to suggest that any t'ecnica grouping should give good results in detecci'on new events, both in the retrospective case as in the case "on line".

Regrouping and TM: discovery events

Model

- Each news is an "object" characterized by a set of attributes (t'erminos)

- Each event is a "prototype" of a set of similar news funci'
  on their t'erminos.

In the case of retrospective detecci'on "Explosions" are looking at a hist'orico file by:

- ⇧ Sudden changes in distribuci'on of t'erminos over time

- ⇧ L'exicas similarities and proximity in time of the news texts.

Regrouping and TM: discovery events

In the detecci'on "on line" the idea is to process the news as

they arrive. The algorithms have been developed for this problem

are "boolean" answer:

one. Each story is compared, when it appears, with

"Clustering" already detected.

two. If the news can join this grouping

It is assumed to correspond to the same event.

3. If the news can not be incorporated is supposed to

It corresponds to a new event and a new grouping starts.

Classi fi caci'on text

b'asicas ideas

general model classi fi caci'on

- Let's consider *The vector model documents / t'erminos*

  Every document is represented as a vector: $d = (w_{one}, .. w_N)$

  where each w_i represents the weight that the t'ermino i in document d.

- Suppose we have M documents $\{ d_i, i =$ one.. $M\}$
- A set of H classes provided for documents.
- The problem of classifying documents (categorize) *It does not seem very different from the problem of classifying General caci'on*

Classi fi caci'on text

b'asicas ideas

t'ıpicos problems classified text caci'on

Document Categorizaci'on

We have a document colecci'on classi fi ed by t'opicos and we want to train a categorizer way that allows us to classify new documents. They tend to be "long" documents.

problems

- Car initially classified documents. Usually you must become an expert.
- The t'opicos documents often change over time.

Classi fi caci'on text

b'asicas ideas

t'ıpicos problems classified text caci'on

M'as problems with short texts that include a word catgorizador .

- Classi fi er news (Google News for example). Est'a detecci'on related to the event. In some cases it is pre-clustering and working with jerarqu'ıas
- Electr'onicos spam filters emails
- An'alisis feelings on social networks.
- Studies short texts categorized database (descriptive an'alisis):

  m'edicas descriptions. Diagn'osticos etc. Opini'on
  surveys with free text.
  Miner'ıa Web usage and content (an'alisis of t'opicos on Twiter yh'abitos of use etc ..)

Classi fi caci'on text

b'asicas ideas

general model classi fi caci'on

- We dataset:

items variables	t_{one}	t_{two}	· · · · · · ·	t_N	C
d_{one}	w_{eleven}	w_{12}	· · · · · · ·	$w_{one\ N}$	c_{one}
⋮	⋮	⋮	· · · · · · ·	⋮	⋮
d_M	$w_{M\,one}$	$w_{M\,two}$	· · · · · · ·	w_{MN} c_M	

- It is predicted, with the values of { w_{ij}} class c_i to which the document belongs.

- In principle the process of classi fi caci'on not vary:

one. We have a training set and a set of test

two. We apply a model classi fi caci'on the training set

3. We apply the model to predict the test set.

Four. We measure how well the model

Classi fi caci'on text

speci fi c problems classi fi caci' *text on*

issue

Vector REPRESENTATION documents:

one. In principle it contains many variables. It is impossible to think, principle in discrete t'ecnicas as spanning trees of decisi'on

two. The matrix documents / t'erminos is scattered.

Necessary reduce dr'asticamente *t'erminos space*

Solution

one. Reducci'on apply principal component.

two. Establishing processes classify binary caci'on and reduce dictionary the representative M'as t'erminos class to consider

Classi fi caci'on text

speci fi c problems classi fi caci' text on

Aplicaci'on problems of major components

- interpretaci'on the semantics of the classification caci'on is lost. All m'etodos podr'ıan be predictive in this case.
- Classi fi caci'on of each document involves a process prior tranformaci'on. *REPRESENTATI is necessary to use the document in the space factor*

Processes binary classi fi caci'on

- If there is class H H processes classified develop caci'on increasingly taking a class as positive examples and the rest as negative.

- Every time the dictionary of the process considering the relevant M'as t'erminos class we are studying is changed.

Classi fi caci'on text

speci fi c problems classi fi caci' text on

Processes binary classi fi caci'on

- The process can lead to classi fi ed each document belongs to a class M'as:

belonging to several classes remains (B' Index search t'opicos)

It is assigned somehow the "winning" class (probability measures, distance measurements etc.)

Usually working with classi fi cations binary. The adaptaci'on processes classified caci'on take into account this fact

*Processes classified caci'on word used M'as t'ecnicas taking the usual starting the vector model documents / t'erminos. Seg' one type is usually t'ecnica use the presence / absence or tf * idf model*

Classi fi caci'on text

Aplicaci' t'ecnicas on the classi fi caci' on a word

Using the KNN t'ecnica

- Overall similarity measures are used instead of distances:

  If it is a binary pattern similarity measures if it is a model tf *
  idf measure cosine.

- The main problem is computational aplicaci'on : For each document must calculate the similarity with everyone.

- Soluci' on : Aplicaci'on of t'ecnicas of recuperaci'on of informaci'on All documents to classify a "consultation" k-documents that best satisfy the process is greatly improved with the use of inverted'indices are obtained (t'ermino / documents) ordered t'erminos

Classi fi caci'on text

Aplicaci' t'ecnicas on the classi fi caci' on a word

Using rule-based t'ecnicas

- As explanatory m'etodos est'an M'as get oriented rules describing the classes to classify.
- Commonly used binary data
- You should try to work with simple rules that prevent overlearning. This involves using t'ecnicas pruning.
- Experimental data indicate that it is not necessary to work with a large n' umber of starting t'erminos
- It is also possible to work with spanning trees of decisi'on following the same fi losof'ıa above.

Using T'ecnicas Probabil'ısticas

Naive Bayes M'etodo

We aqu'ı predict a binary class variable N random binary (t_{one}, .. t_N), or N-dimensional binary ua varaible $\bar{t}$. The expressions the conditional probability, lead to:

$$Pr(C \mid \bar{t} = \bar{x}) = \frac{one}{Pr(\bar{t} = \bar{x}) \exp(\sum_j w_j x_j + b)}$$

where w_i Y b They can be calculated from the estimated probabilities. $Pr(t_j = 1 \mid C)$, $Pr(C)$ etc.

There are two models of computation: Bernoulli and multinomial

Classi fi caci'on text

Using T'ecnicas Probabil'ısticas

Regresi' on Log'ıstica

Since we are classi fi ed in binary form you can use tf * idf and work with models log'ıstica.Los regresi'on coefficients of regres'ı' log'ıstica on can give us an idea of the weight of the classified t'erminos caci'on. Especially if it has done in steps.

Classi fi caci'on text

GLM .

For each document *D* to classify its puntuaci'on is calculated to belong to a given class of linear form:

$$punt(D) = \sum_{j} w_j x_j + b$$

x_j is 0 or 1 in the document and w_j They are the "weights of each t'ermino in class". The process of classifying the documents caci'on is very r'apido with this m'etodo:

one. the t'erminos are arranged with their weights (inverted list)

two. D for each list is scrolled by adding the weights of t'erminos
It is in D, *sum (D)*

3. ¿ *sum (D) >* ↲ *>* 0? the document is assigned to the class.

Classi fi caci'on text

GLM .

- The process of classifying M'as r'apido caci'on is possible.
- They may extend lists weights using, sin'onimos, n-grams etc.

- Problem: learn the weights

  Direct optimizaci'on leads to a non-convex porblema with local m'ınimos.

  It works to find the hyperplane separating the positive and negative points

  It is normally used an SVM

Asociaci'on rules and TM

An interesting example

- *T: t'erminos in a document*
- *D: word set*
- *{ pain love}) {death}*

b'asica Idea

A set of documents associated with t'erminos can be viewed as a transactional database:

items variables	t_{one}	t_{two}	· · · · · ·	t_N
d_{one}	1	0	· · · · · ·	one
⋮	⋮	⋮	· · · · · ·	⋮
d_M	0	0	· · · · · ·	one

Asociaci'on rules and TM

Associations and co-occurrences between a text caracter'ısticas

Formulation

T = t1, t2, ..., tn set t'erminos

D = d1, d2, ..., dm set of documents indexed by t'erminos

Each document di generates a subset *di (t)* ✓ *T.*

Be *OR* ✓ *T.* The set of all documents D such that d

OR ✓ *d (T)* It is called *covering assembly* U and [U] shows. In its simplest form a rule of asociaci'on is a implicaci'on of the form:

U) or with *or* ✓ *YOU*

Asociaci'on rules and TM

There are several approaches to the problem depending on c'omo de fi na T d (T):

- T is the set of all documents t'erminos in d (T) is the document REPRESENTATION t as a "pocket" word. This approach is not very efficient and computationally expensive

- T is a set of keywords in documents d (T) is the document indexaci'on T through a set of keywords.

  In this approach the problem is to obtain T. t'ecnicas of co-occurrences based on natural language processing can be employed

Asociaci'on rules and TM

- The elements of T are grouped seg' classi fi caci'on one taxon'omica. (Thesaurus)
  This approach allows you to discover partnerships with different level of granularity

- T elements are phrases or full sentences. In this case it is discovered co-occurrences of sentences in documents colecci'on.

- T elements are "episodes" (complete sets of vector caracter'ısticas).

  In this case the so-called "rules of episodes" are obtained.

Extracci' on attributes of semantics in short texts

issue

There attributes database (relational or not) corresponding to short text with a restricted semantics.

Examples:

- Databases M'edicas (diagn'osticos, descripci'on interventions etc.)
- Fields "observations" in various situations (surveys, expert etc.)
- Document Abstract

It is desired to obtain a REPRESENTATION of the attribute (intemedia shape) such that it:

- Collect semantics
- allow the query as a model ROO (where applicable)
- allows performing TM on said structure

Extracci' on attributes of semantics in short texts

b'asicas ideas

- Once "clean" text, sets / sequences repeated words in n'

  UMBER sufficient to constitute tuples

  "Sentences" or "concepts" system.
- It is possible that some "sentences" are repeated incomplete, corresponding to less specialized concepts, so that we have a subset of t'erminos reticulum.

- This leads us to the concept of "itemset (seq) frequent" in transactional database of texts. Remember that every frequent itemset has the property "a priori" Then

If you get frequent itemset base texts:

- *The maximal itemset give us complete sentences*
- *The subret'ıculo gives us the full semantics structure of the database.*

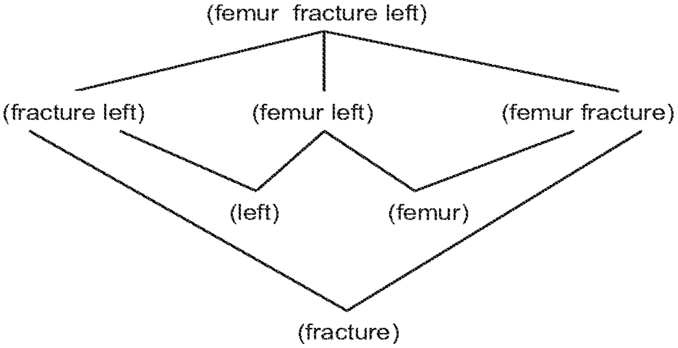
Extracci' on attributes of semantics in short texts

Example: database ED

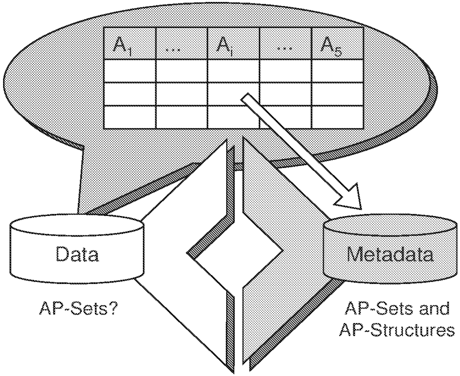
Field 1	Field 2	Field 3	...	Support
right	fracture			4.26
right	femur			1.826
back	sebaceous	cyst		1.304
back	cyst			1.304
right	femur	fracture		1.173
femur	left			1.043
femur	fracture	left		1.043
. . .	. . .			. . .

Extracci' on Semantics in short text attributes

Example: database ED



Extracci' on Semantics in short text attributes



Extracci' on Semantics in short text attributes

⇮ The overall semantics of a set of short texts is collected in a structure derived from whole frequent itemset

⇮ The value of each tuple seg' a text containing a subestructure global

⇮ Are de fi ned properties and operations ap-structures that allow:

- Get the value of each tuple in the attribute as substructure
- Approximate perform queries using this new data structure

⇮ This structure is implemented as a TDA in a BDROO

⇮ We have developed TM with these structures, studying models:

- Data Warehousing
- And obtaining ordered itemset GENERATION autom'atica tag-clouds