UNIVERSITÉ DE LYON

ICOM

INFO STAT

Applications in text mining

Master Data Mining

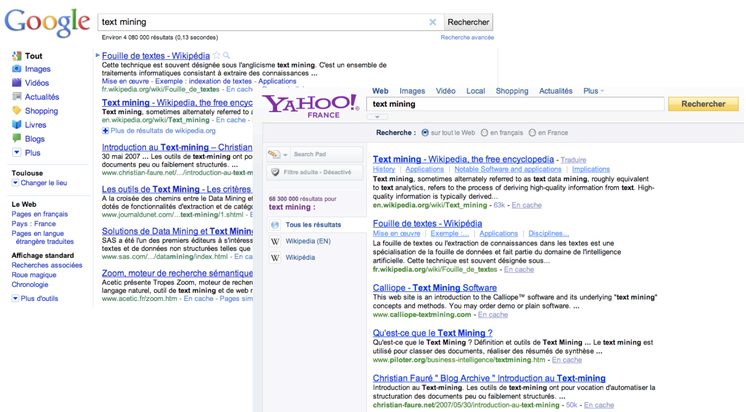
Julien Velcin

Applications in text mining

- Information Retrieval (IR)
- Web access filtering
- Document summarization
- Information Extraction (IE)
- Question-Answering (QA)
- Text classification
- Spam detection
- Technology, Economics watch

Etc.

Information Retrieval (IR)



Information Retrieval (IR)

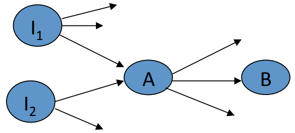
- Mainly based on document indexing
  = find the important **meanings** and create an internal representation of the websites
- Factors to consider:
  - Accuracy to represent meanings (semantics)
  - Exhaustiveness (cover all the contents)
  - Facility for computer to manipulate
- What is the best representation of contents?
  - Char. string (char trigrams): not precise enough
  - Word: good coverage, not precise
  - Phrase: poor coverage, more precise
  - Concept: poor coverage, precise

Information Retrieval (IR)

- Matching score model
  - Document D = a set of weighted keywords
  - Query Q = a set of non-weighted keywords
  - R(D, Q) = Σ_i w(t_i , D)
    where t_i is in Q.
- Boolean model
- VSM model
- Probabilistic models

Information Retrieval (IR)

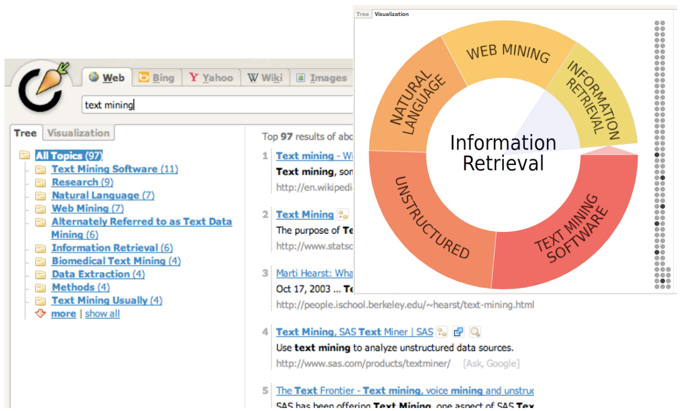
- PageRank in Google:



$$PR(A) = (1-d) + d\sum_i \frac{PR(I_i)}{C(I_i)}$$

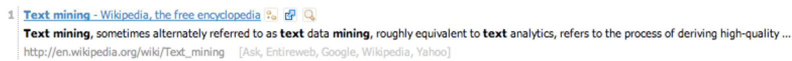
- Assign a numeric value to each page
- The more a page is referred to by important pages, the more this page is important
- *d:* damping factor (0.85)
- Many other criteria: e.g. proximity of query words
  - "…information retrieval …" better than "… information … retrieval …"

Result Clustering



Result Clustering

- Use any search engine to get snippets:



- Text clustering to organize snippets into a tree
- Attach meaningful labels to the categories
  - Frequent patterns
  - Named entities
- Some existing systems: Clusty, Carrot2, Kartoo…

Web access filtering

- Website classification for parental control
  - 2 European Resarch Projects: NetProtect I & II
  - EADS, OpteNet, MGN, Hypertech
- Developpement of a new text classifier
  - Based on combinations of Support Vector Machines (SVM)
  - Designed to deal with 4 classes (pornography, violence, bomb making, drugs), and 8 languages (english, french, italian, portuguese, dutch, german, spanish, greek)

Web access filtering

	English(En)	French(Fr)
B	40	113
¬B	182	223
D	145	205
¬D	203	183
P	393	292
¬P	332	155
V	190	187
¬V	110	150
E	453	440
G	563	470
Total	2611	2418

Table 1 : Netprotect II Database content

Bomb making
^B counter-examples
Drugs
^D counter-examples
Pornography
^P counter-examples
Violence
^V counter-examples
Child oriented websites (**E**)
Generic websites

[Grilheres et al., 2004]

Web access filtering

	B	D	P	V	Boolean	Score
TEX			X		X	
SVM1	X					X
SVM2		X				X
SVM3			X			X
SVM4				X		X
ADR	X	X	X	X	X	
IMG			X		X	

TEX: Artificial neuronal network of NetProtect I (text)

SVM*i*: (Biclass) Support Vector Machine (text)

(take the n strongest word given the TF-IDF score)

ADR: based on regular expression on the name/address

IMG: machine learning system using only picture's features

Web access filtering

- Results on the dedicated categories:

Classifier Treated Category	Blocking	OverBlocking	GCR	I(GCR,5%)
SVM1 Bomb-Making	0.7818	0.1282	0.8373	0.8135-0.8586
SVM2 Drug	0.8261	0.1850	0.8203	0.7956-0.8426
SVM3 Pornography	0.9407	0.1850	0.8885	0.8678-0.9063
SVM4 Violence	0.7027	0.2300	0.7302	0.7022-0.7565
TEX Pornography	0.8444	0.0485	0.8889	0.8682-0.9067
IMG Pornography	0.3407	0.0583	0.5903	0.5599-0.6200

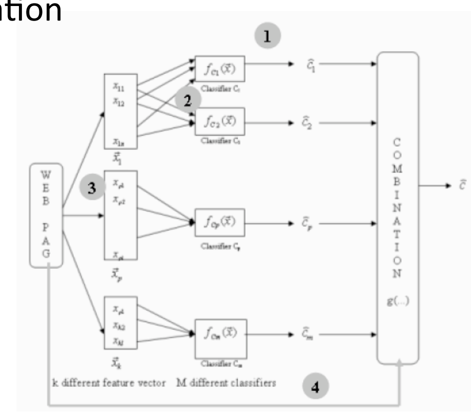
- Results on the 4 categories:

Classifier	Blocking	OverBlocking	GCR	I(GCR,5%)
ADR	0.6486	0.0895	**0.8255**	0.8011-0.8475
SVM1	0.3303	0.1400	0.6881	0.6591-0.7157
SVM2	0.3514	0.1457	0.6910	0.6621-0.7186
SMV3	0.5736	0.2165	0.7154	0.6870-0.7422
SVM4	0.5195	0.1544	0.7398	0.7121-0.7657
TEX	0.4384	0.0101	0.8109	0.7858-0.8337
IMG	0.2102	0.0101	0.7368	0.7090-0.7629

Web access filtering

combination

OR
AND
Majority Voting (MV)
Logical Rules (RUL)
kNN
DSkNN
NB
MLP
SVM



Web access filtering

Blocking
Rate of harmful pages correctly blocked

Overblocking
Rate of harmless pages blocked

GCR
Good Classification Rate

I(GCR,5%)
Confidence interva at 5%

Classifier	**Blocking**	**OverBlocking**	**GCR**	**I(GCR,5%)**
OR	0.9550	0.3304	0.7622	0.7352-0.7872
AND	0.0150	0	0.6803	0.6511-0.7872
MV	0.4655	0.1328	0.7368	0.7090-0.7629
RUL	0.5976	0.0188	0.8567	0.8340-0.8768
NB	0.7417	0.0404	0.8889	0.8682-0.9067
MLP	0.8018	0.0418	0.9074	0.8881-0.9237
kNN	0.7898	0.0404	0.9074	0.8881-0.9237
kNN+RUL	0.8048	0.0346	**0.9133**	0.8945-0.9290
DSkNN	0.7958	0.0303	**0.9133**	0.8945-0.9290
DSkNN+RUL	0.8048	0.0346	**0.9133**	0.8945-0.9290
PSVM	0.7958	0.0317	0.9123	0.8934-0.9281
PSVM+RUL	0.8108	0.0346	**0.9152** +	0.8966-0.9307
GSVM	0.7928	0.0289	**0.9133**	0.8945-0.9290
GSVM+RUL	0.8048	0.0346	**0.9133**	0.8945-0.9290

Document summarization

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

Document summarization

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

Example of summarization

Cracks Appear in U.N. Trade Embargo Against Iraq.

Human written summary

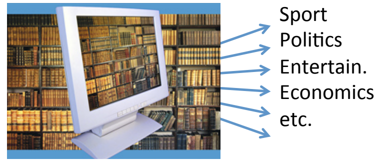
Cracks appeared in the U.N. trade embargo against Iraq. The State Department reports that **Cuba and Romania have struck oil deals with Iraq** as others attempt to trade with Baghdad in defiance of the sanctions. Iran has agreed to exchange food and medicine for Iraqi oil. **Saddam has offered developing nations free oil if they send their tankers to pick it up.** Thus far, none has accepted. Japan, accused of responding too slowly to the Gulf crisis, has promised $2 billion in aid to countries hit hardest by the Iraqi trade embargo. **President Bush has promised that Saddam's aggression will not succeed.**

7800 chars, 1300 words

17

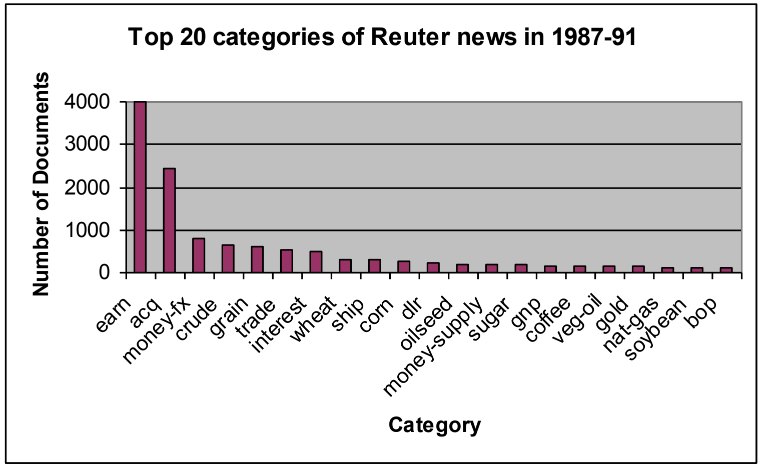
Text classification

- Automatically classify documents into **predefined** classes
  e.g., digital libraries



- Application areas:
  - Email SPAM filtering
  - Internet directory construction (ex.: Yahoo!)
  - Automatic indexing …

A very classical dataset



SPAM detection

- SPAM = Junk emails
- In 2009: 97% of emails are SPAM!
  http://news.bbc.co.uk/2/hi/technology/7988579.stm
- Anti-spam is about 95% accurate today, but can achieve about 99% if correctly trained
- Numerous systems:
  SpamAssassin (MessageLabs), Bitdefender AntiSpam, POP File, Spamihilator, Cactus Spam Filter…
- A lot of heuristics, partly using ML

Email Format

- **From:** The e-mail address, and optionally the name of the sender
- **To:** The e-mail address[es], and optionally name[s] of the message's recipient[s]
- **Subject:** A brief summary of the contents of the message
- **Date:** The local time and date when the message was written
- **Cc:** Carbon copy
- **Bcc:** Blind Carbon Copy
- **Received:** Tracking information generated by mail servers that have previously handled a message
- **Content-Type:** Information about how the message has to be displayed, usually a MIME type
- **Reply-To:** Address that should be used to reply to the sender.
- **References:** Message-ID of the message that this is a reply to, and the message-id of this message, etc.
- **In-Reply-To:** Message-ID of the message that this is a reply to.
- **X-Face:** Small icon.

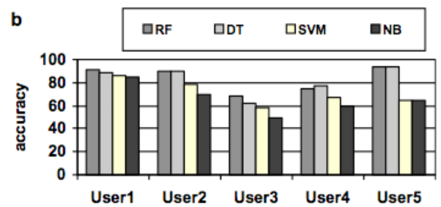
Image SPAM

Antispam technologies of Bitdefender

- Bayesian filters
- Heuristics filters
- Neuronal Networks (ART, ARTMAP, NeuNet, ART+)
- URL filters
- KNN
- ASSL (script language)
- SURBL
- Blacklists and whitelists
- Image filters

Naive Bayes for email classification

- Based on a simple (even simplistic) assumption, NB performs interesting performances for the task of email classification [Koprinska et al., 06]
- For the task of filing emails into folders (4 classes):



Information Extraction (IE)

- Identify phrases in language that refer to specific types of entities and relations in text.
- Named entity recognition is task of identifying names of people, places, organizations, etc. in text.

  people organizations places

  – Michael Dell is the CEO of Dell Computer Corporation and lives in Austin Texas.
- Relation extraction identifies specific relations between entities.

Question Answering

- Directly answer natural language questions based on information presented in a corpora of textual documents (e.g. the Web).
  – When was Barack Obama born?
    - August 4, 1961
  – Who was president when Barack Obama was born?
    - John F. Kennedy
  – How many presidents have there been since Barack Obama was born?
    - 9

Main events related to text mining

- Artificial Intelligence
  – IJCAI, AAAI, UAI, UM, ECAI
- Natural Language Processing
  – ACL, CoNLL, EACL, EMNLP, IJCNLP, LREC
- Machine Learning
  – ICML, ECML, ALT, COLT, NIPS, ICALT
- Data Mining / Database
  – ICDM, SIGKDD, PKDD, VLDB, ICDE, SDM, PAKDD, DAWAK
- Information Retrieval
  – SIGIR, TREC, ECIR, CIKM
- Others
  – DocEng, ICWSM…

National and international contest

- Continued development of corpora and competitions on shared data:

  TREC Q/A - SENSEVAL/SEMEVAL - CONLL Shared Tasks (NER, SRL…) - KDD contest - etc.
- For instance: SIAM TM Competition 2007
  – Objective: develop TM algo for doc classification
  – Aviation safety reports documenting one or more problems that occurred on certain flights
  – Validation measures: precision/recall