# Investigating word interactions in texts. Application to text categorization in genomics

Martine Cadot, Michel Zitt, Gabriel Meurin, Alain Lelu

# Investigating word interactions in texts. Application to text categorization in genomics.

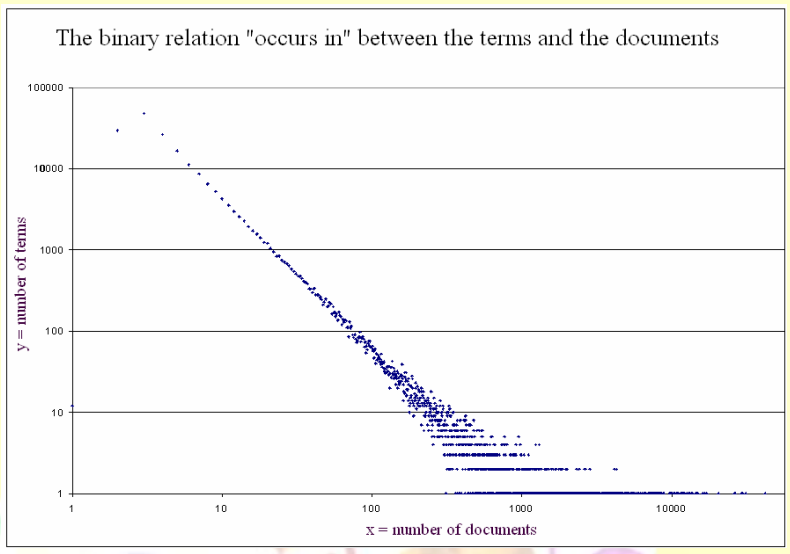
Martine Cadot (LORIA), Michel Zitt (INRA, OST), Gabriel Meurin (LORIA/INRA), Alain Lelu (LORIA, LASELDI)

Words interacting in a text may be compared, to a certain extent, to molecules interacting and building "complexes", i.e. phrases, named entities, or longer-range semantic or syntactic associations. We will call them "k-itemsets", k being their interaction level. We have shown (Cadot 06) that an adequately built subset of these k-itemsets is enough for describing the entirety of the relations at work in a corpus represented as a set of "bag-of-words" documents, whatever the level k of these relations.

Our objective is to reconstruct each category into which a corpus of scientific abstracts has been split, using a set of Boolean queries as a best compromise between conciseness and reproducibility of this categorization.

## I – The corpus, its 50 catégories.

- Scientific abstracts in genomics, pulled out from the Web of Science (Thomson Scientific ed.).
- This subset has been delineated using a hybrid method, based both on lexical queries and citation expansion/ shrinkage (Zitt et al. 2006) → 120,000 abstracts from 1999 to 2005.
- A vocabulary of 237,000 lemmatized words and phrases (>2 occurrences) has been pulled out and filtered (NeuroNav, www.diatopie.com). I.e.: *sequence, polymorphism, folded_structure, chromosome_4B, greenbug_resistence_gene,…*




This figure reads: e.g. 1047 terms occur each one exactly in 21 documents.

- 50 categories resulted from a clustering of the abstracts by the Axial K-means method (Bassecoulard et al. 2007).

M 1/Human_genome/Human_genome_project
M 2/Translocation/FISH/leukemia
M 3/Plant_genomics/Transgenic_plants
M 4/DNA_sequence/Satellite
M 5/Strain/Microbial_genomics
M 6/Cell identity & Gene expression
M 7/Enzyme/Escherichia_Coli
M 8/Alignment/Bioinformatics
M 9/Genome/
M10/Comparative_genomic_hybridization_/Tumor
M11/SNPs/Polymorphism
M12/Network/Biological_networks/Model
M13/Transcriptional/Saccharomyces_cerevisiae/Transcriptome
M14/Locus/Microsatellite_locus/Polymorphism
M15/Cell_line/Tumor/Genome_&_Cancer
M16/Spectrometry/Proteomics
M17/Map/Linkage_maps/Polymorphism
M18/Population_genomics
M19/Repair/DNA_damage
M20/Resistance/Resistance_genes/Plant & Trout_resistance
M21/Hybrid/Somatic_hybrids/Ferility
M22/Human/C-DNA/Gene_annotation
M23/Exon/Genomic_organization/Gene_annotation
M24/System/Systems biology/Bioinformatics
M25/Patient/Disease_genomics/Biomarkers/Pharmacogenomics
M26/Virus/Nucleotide_sequence
M27/Evolution/Evolutionary_genomics
M28/Cancer/Genome & cancer
M29/Promoter/Transcription
M30/Mutant/Mutagenesis
M31/LOH/Tumor_suppressor/Genome & Cancer
M32/Marker/RAPD/AFLP/Polymorphism
M33/RNA-/Virus
M34/PCR/Methods/applications
M35/C-DNA/Transcription/C-DNA_library
M36/Polymorphism
M37/Cell/DNA_damage
M38/Genome/Genome_sizes
M39/DNA/Arrays/Genomic techniques
M40/QTL/Trait/Mapping/Polymorphism
M41/Signaling/Kinase/MAPK
M42/Mutation/Missence_mutation
M43/Mouse/Murine_genomics
M44/Expression/Cell_identity&Gene_expression
M45/LOD/Linkage_analysis/Polymorphism
M46/Human/Primate/Gene_annotation/Comparative_genomics
M47/Species/Phylogeny/Evolutionary_genomics
M48/C57BL/Congenic_strains/Murine_genomics
M49/Residue/Amino_acid_sequence
M50/Virus/Virus_replication/Virus_recombin

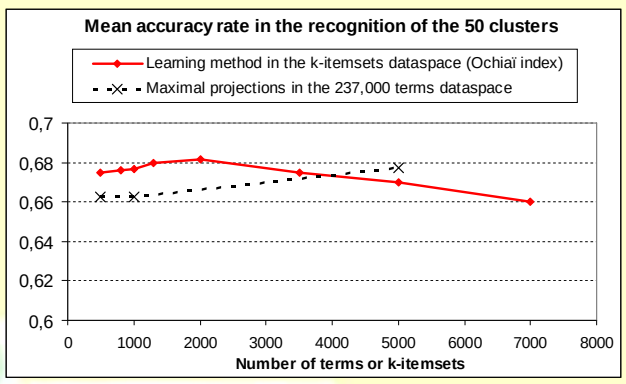
Ex. of the 14 first terms (words and phrases) most typical of the « Human Genome Project » cluster:

| No | Phrase | Ochiaï | No | Phrase | Ochiaï |
|---|---|---|---|---|---|
| 1 | human_genome | 0,467 | 8 | human_genome_sequence | 0,122 |
| 2 | human | 0,336 | 9 | primate | 0,107 |
| 3 | genome | 0,257 | 10 | sequence | 0,103 |
| 4 | human_genome_project | 0,238 | 11 | chimpanzee | 0,099 |
| 5 | project | 0,157 | 12 | completion | 0,096 |
| 6 | draft | 0,138 | 13 | disease | 0,094 |
| 7 | human_chromosome | 0,125 | 14 | genomics | 0,092 |
| | | | 15 | human_gene | 0,092 |

As most of data analysis methods do, this data partition takes into account the only « 2-itemsets » (a k-itemset of support s is an elementary association of k terms present in s documents).

**II – Concise and reproducible representation of the 50 categories:** using itemsets of order 1, 2, *and higher order ones,* which express complex interactions between terms in specific contexts.

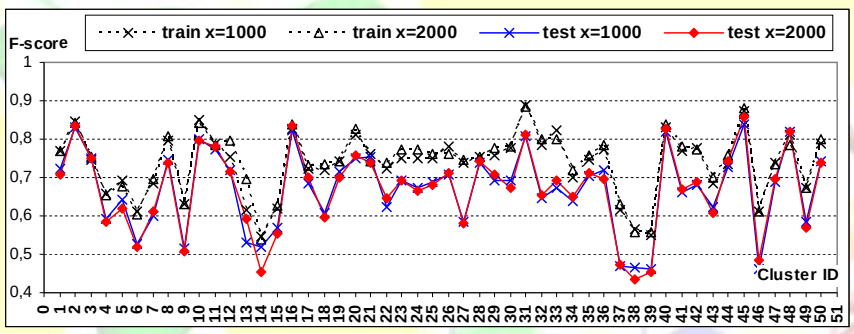
- MIDOVA method (Cadot 2006) for mining ordered lists of informative itemsets specific to each category (train set: 1/10th of the corpus, test set: 9/10 th) → ordered lists of simple Boolean queries for identifying the class of any document out of the corpus, and extending this categorization process to other databases (patents, …).
- Results
  - At the same time as a 100% intrinsically exact reconstitution rate results from using the whole 237,000 terms, a maximum 68% rate results from using about x = 2000 k-itemsets (and then decreases), with a statistically-controlled generalization ability:




  - 68% is a mean value, embedding many discrepancies:

>*80% :* clusters N°45 (LOD/ Linkage_analysis/ Polymorphism) and N°16 (Spectrometry/ Proteomics)

<*50% :* clusters N°14 (Locus/ Microsatellite_Locus/ Polymorphism) and N° 38 (genome/ genome_size)




  - example of class description: the 6 first k-itemsets of cluster 10:

*comparative_genomic_hybridization*,
*hybridization* AND *tumor* AND *genomic*,
*hybridization* AND *tumor*,
*CGH* AND *comparative_genomic_hybridization*,
*losses* AND *genomic*,
*tumor* AND *genomic*.

As may be observed *comparative_genomic_hybridization* and *CGH* acronym appear together, integrally or partly, at the top of the list.

## References

Bassecoulard E., Lelu A. and Zitt M. (2007). Mapping nanosciences by citation flows: a preliminary analysis, *Scientometrics,* vol 70, n°3, pp. 859-880.

Cadot M. (2006). *Extraire et valider les relations complexes en sciences humaines : statistiques, motifs et règles d'association*. Doctoral dissertation, University of Franche-Comté, France.

Laurens P., Zitt M. and Bassecoulard E. (to appear), "Delineation of the genomics field by hybrid citation-lexical methods: interaction with experts and validation process", *Scientometrics*.

Zitt M., Ramanana-Rahary S. and Bassecoulard E. (2006). Delineating complex scientific fields by a hybrid lexical-citation method: an application to nanosciences, *Information Processing and Management* Vol. 42-6, pp. 1513-1531.