Extraction de concepts par AFC dans des corpus

Stage effectué au laboratoire Er-Tim, Inalco

Valentin Taillandier, encadré par Damien Nouvel

École Normale Supérieure de Rennes

Définitions

Corpus Recueil de documents similaires.

Lemme Unité autonome constituant le lexique d'une langue.

Vector Space Model

Approche fréquentielle

	avoir	**être**	**gâteau**	**ordinateur**	**farine**
doc1	5	4	0	2	0
doc2	4	3	1	0	2

Vector Space Model

Approche fréquentielle

	avoir	**être**	**gâteau**	**ordinateur**	**farine**
doc1	5	4	0	2	0
doc2	4	3	1	0	2

Représentation vectorielle

$$\vec{c_i} = \frac{\sum_{\vec{u} \in \mathcal{C}_i} \vec{u}}{|\mathcal{C}_i|}$$

Vector Space Model

Approche fréquentielle

	avoir	**être**	**gâteau**	**ordinateur**	**farine**
doc1	5	4	0	2	0
doc2	4	3	1	0	2

Représentation vectorielle

$$\vec{c_i} = \frac{\sum_{\vec{u} \in C_i} \vec{u}}{|C_i|}$$

Similarité cosinus

$$\text{simcos}(\vec{u}, \vec{v}) = 1 - \frac{\vec{u}.\vec{v}}{\|\vec{u}\| \|\vec{v}\|}$$

Estimation de la catégorie associée à un document

Pour des vecteurs normalisés

$$\text{argmax}_{i \in [\![1,n]\!]} \{\vec{u}.\vec{c_i}\}$$

Algorithme TFIDF

- Repondération

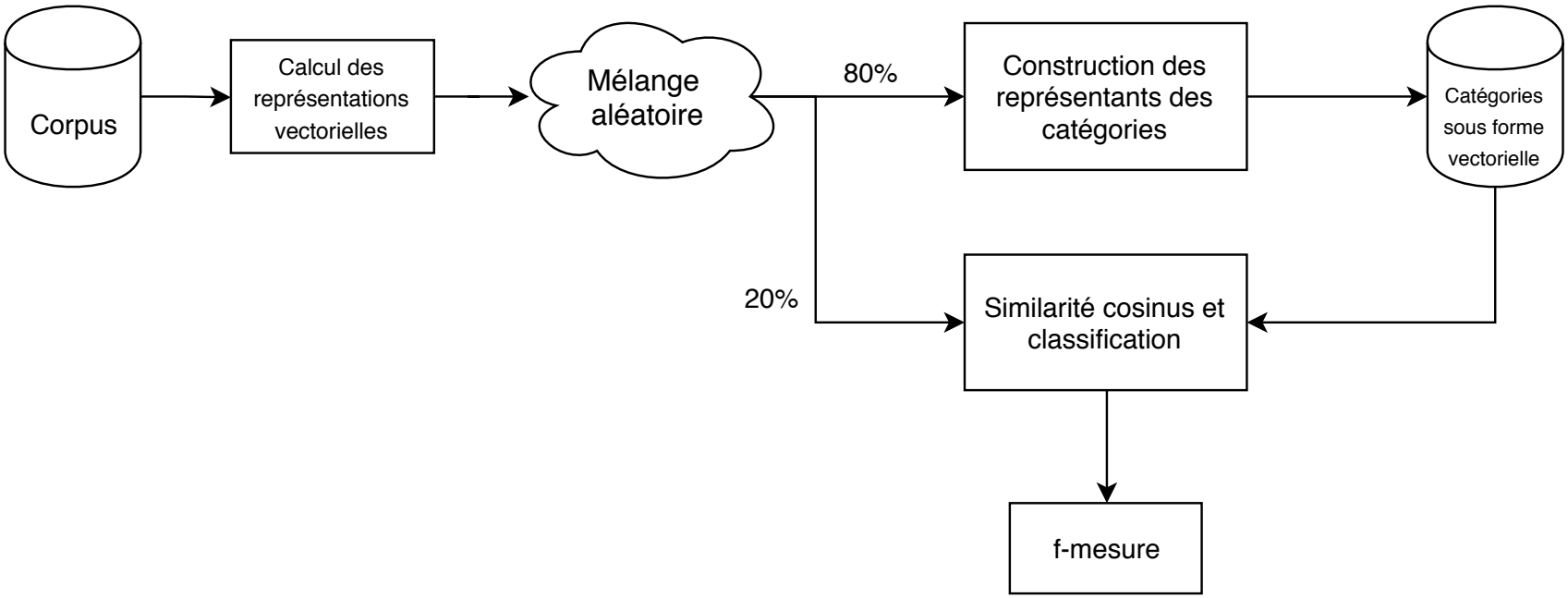
Fréquence inverse de document

$$\text{idf}_{mot} = \log \frac{1}{\text{proportion de document}}$$

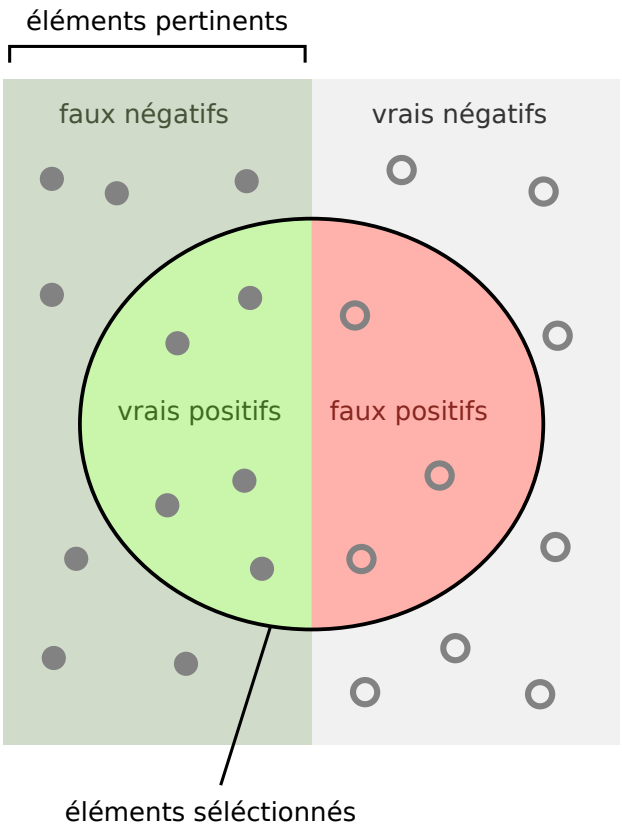
Analyse sémantique latente

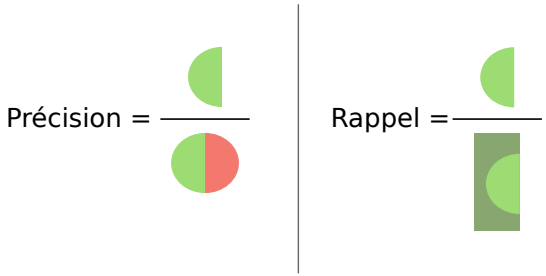
- Réduire la largeur des matrices.
- Décomposition en valeurs singulières.

Schéma de principe



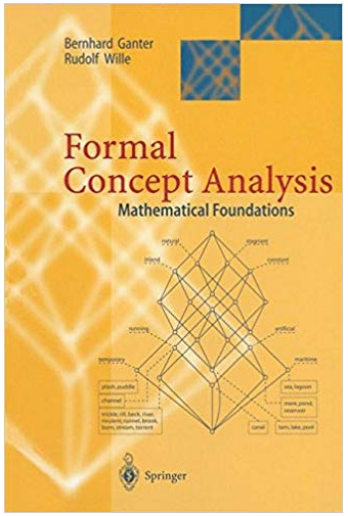
Évaluation





. Adapté de Wikipédia

Recherche sur l'AFC



Formal Concept Analysis : Mathematical Foundations Ganter et Wille, 2012.

Analyse formelle de concepts

	composé	pair	impair	premier	carré
1			X		X
2		X		X	
3			X	X	
4	X	X			X
5			X	X	
6	X	X			
7			X	X	
8	X	X			
9	X		X		X
10	X	X			

Analyse formelle de concepts

Contexte formel

- G est l'ensemble des objets.
- M est l'ensemble des attributs.
- $I \subseteq G \times M$, une relation.

On définit deux opérateurs de dérivation pour $A \subseteq G$ et $B \subseteq M$ par $A' = \{m \in M | \forall g \in A, gIm\}$ et $B' = \{g \in G | \forall m \in B, gIm\}$.

Concept formel

Un concept du contexte (G, M, I) est alors un couple (A, B) où $A \subseteq G$ et $B \subseteq M$ vérifient $A' = B$ et $B' = A$.

L'AFC appliquée au TAL

	mot1	mot2	mot3	mot4	mot5
doc1			x		x
doc2		x		x	
doc3			x	x	
doc4	x	x			x
doc5			x	x	
doc6	x	x			
doc7			x	x	
doc8	x	x			
doc9	x		x		x
doc10	x	x			

Logiciel d'extraction

PCBO (Parallel Cbo) Krajca, Outrata et Vychodil, 2008.

Choix des documents

- *Abstracts* français des articles de Wikipédia.
- Catégories : amphibiens, oiseaux, poissons, insectes, mammifères et reptiles.

Choix des documents

- *Abstracts* français des articles de Wikipédia.
- Catégories : amphibiens, oiseaux, poissons, insectes, mammifères et reptiles.

Corpus résultant

- 4000 abstracts
- 6 catégories
- 4500 mots

Contextes aléatoires

- Générer des contextes aléatoires.
- Définir un nombre d'objets et un nombre d'attributs.
- Probabilité un demi.
- 32go de mémoire vive et 16 processeurs Intel©Xeon©E5-2690 v3 à 2,60GHz.

Résultats

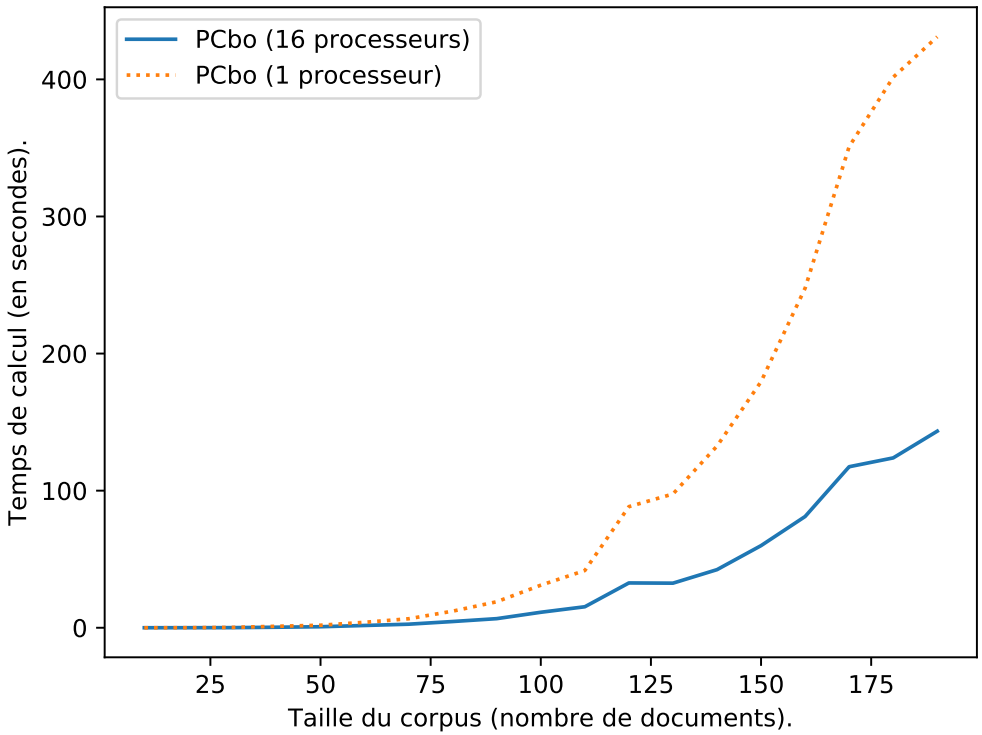


FIGURE – Évolution du temps de calcul en fonction du nombre de documents (taille du vocabulaire fixée à 200 lemmes).

Résultats

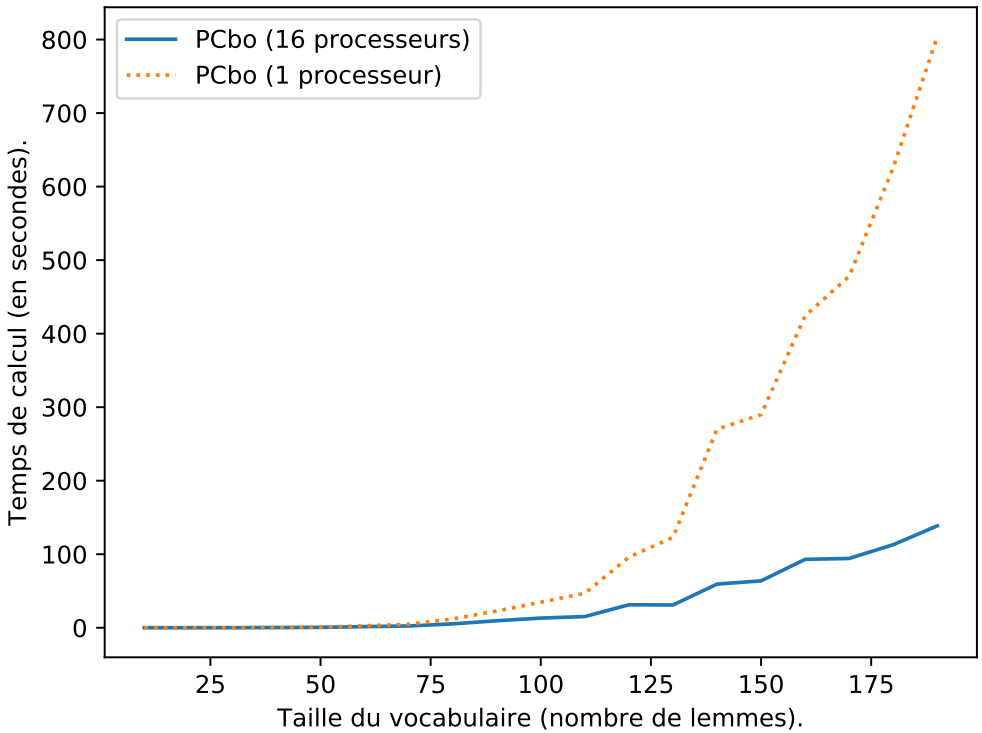


FIGURE – Évolution du temps de calcul en fonction de la taille du vocabulaire (nombre de documents fixé à 200).

Contextes basés sur un corpus réel

- Loi de Zipf.
- George Kingsley Zipf (Mandelbrot, 1957).

Résultats

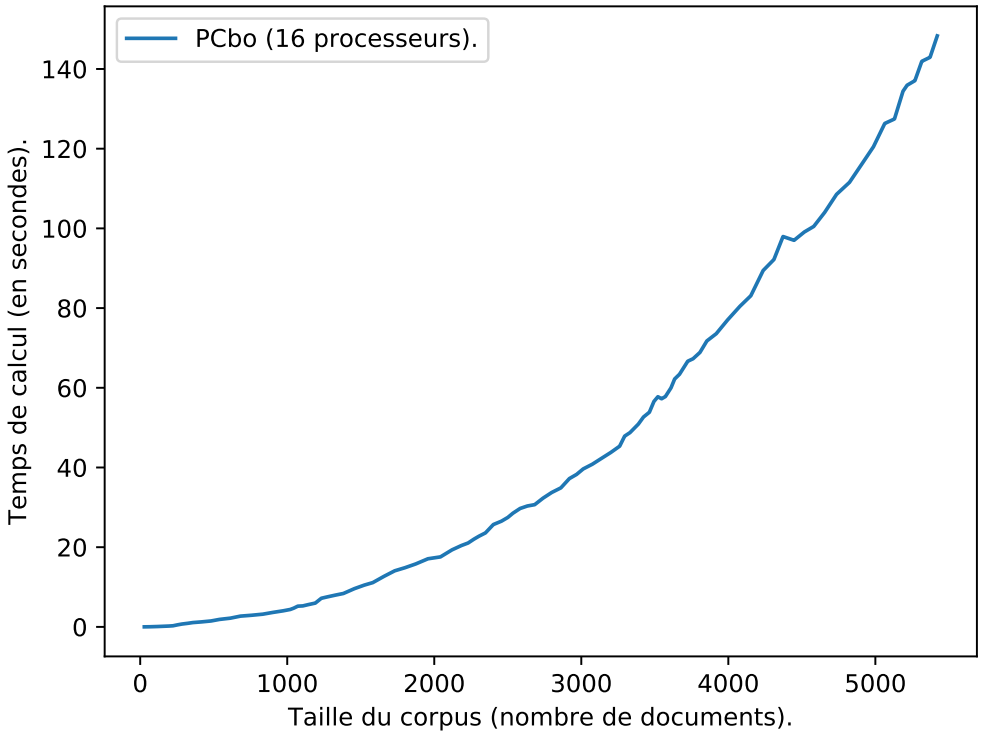


FIGURE – Évolution du temps de calcul en fonction du nombre de documents choisis dans l'ontologie *Animal*.

Résultats

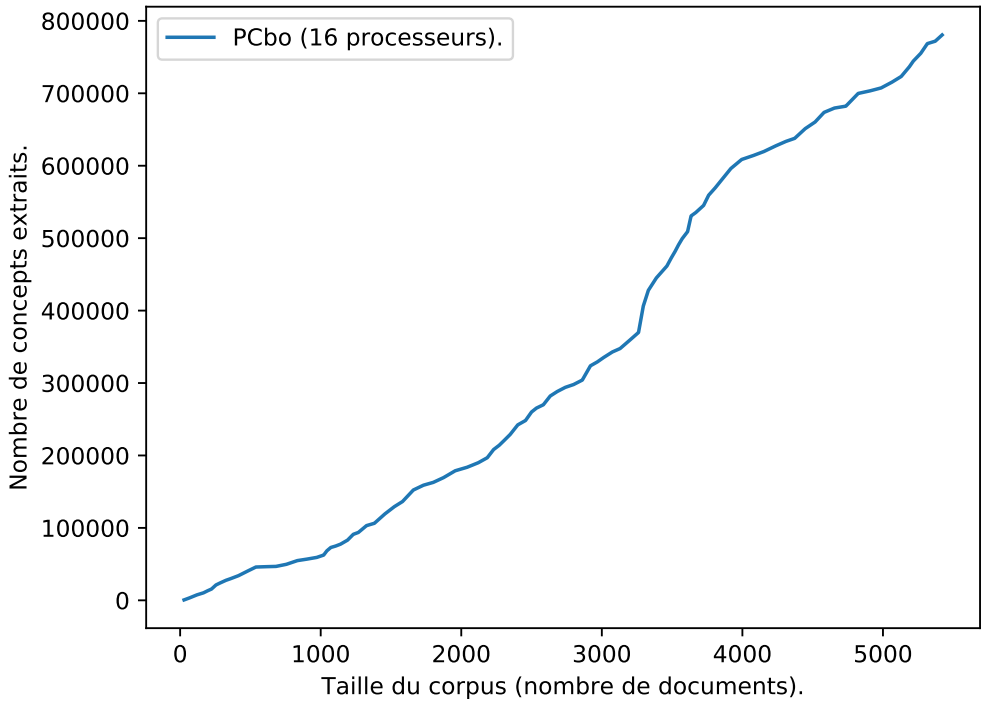


FIGURE – Évolution du nombre de concepts extraits fonction du nombre de documents choisis dans l'ontologie *Animal*.

La méthode classique

- Appliquer TFIDF puis LSA (300 composantes).
- 1000 essais en 22 minutes.

La méthode classique

- Appliquer TFIDF puis LSA (300 composantes).
- 1000 essais en 22 minutes.

Résultats

- F-mesure de 0,94.
- Écart type de 0,018.

Réel \ Estimé	**Amphibien**	**Oiseau**	**Poisson**	**Insecte**	**Mammifère**	**Reptile**
Amphibien	115	1	0	0	0	2
Oiseau	0	312	3	7	10	1
Poisson	0	1	55	2	4	1
Insecte	0	1	0	73	1	2
Mammifère	0	0	1	3	72	2
Reptile	0	0	0	0	0	82

FIGURE – Matrice de confusion lors d'un essai avec TFIDF et LSA.

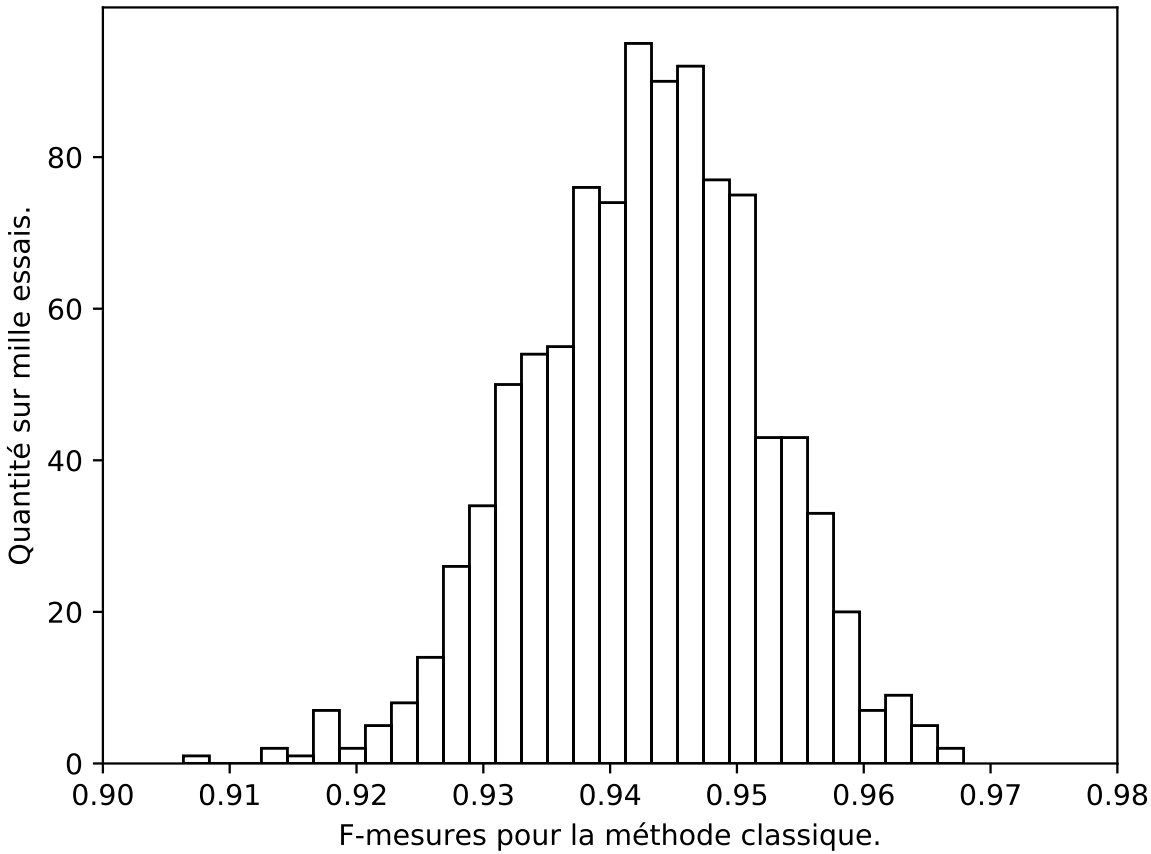


FIGURE – Histogramme de la répartition des f-mesures pour mille essais avec TFIDF et LSA.

L'analyse formelle de concepts

- Calcul des vecteur documents dans l'espace des concepts.
- Filtrage de certains concepts.
- 5 heures pour 1000 essais.

L'analyse formelle de concepts

- Calcul des vecteur documents dans l'espace des concepts.
- Filtrage de certains concepts.
- 5 heures pour 1000 essais.

Résultats

- F-mesure de 0,89.
- Écart type de 0,014.

Réel \ Estimé	**Amphibien**	**Oiseau**	**Poisson**	**Insecte**	**Mammifère**	**Reptile**
Amphibien	114	0	0	0	0	5
Oiseau	0	303	3	3	7	1
Poisson	0	0	50	0	1	0
Insecte	0	0	1	52	5	1
Mammifère	0	5	3	2	52	3
Reptile	0	0	2	1	0	69

FIGURE – Matrice de confusion lors d'un essai avec les concepts formels.

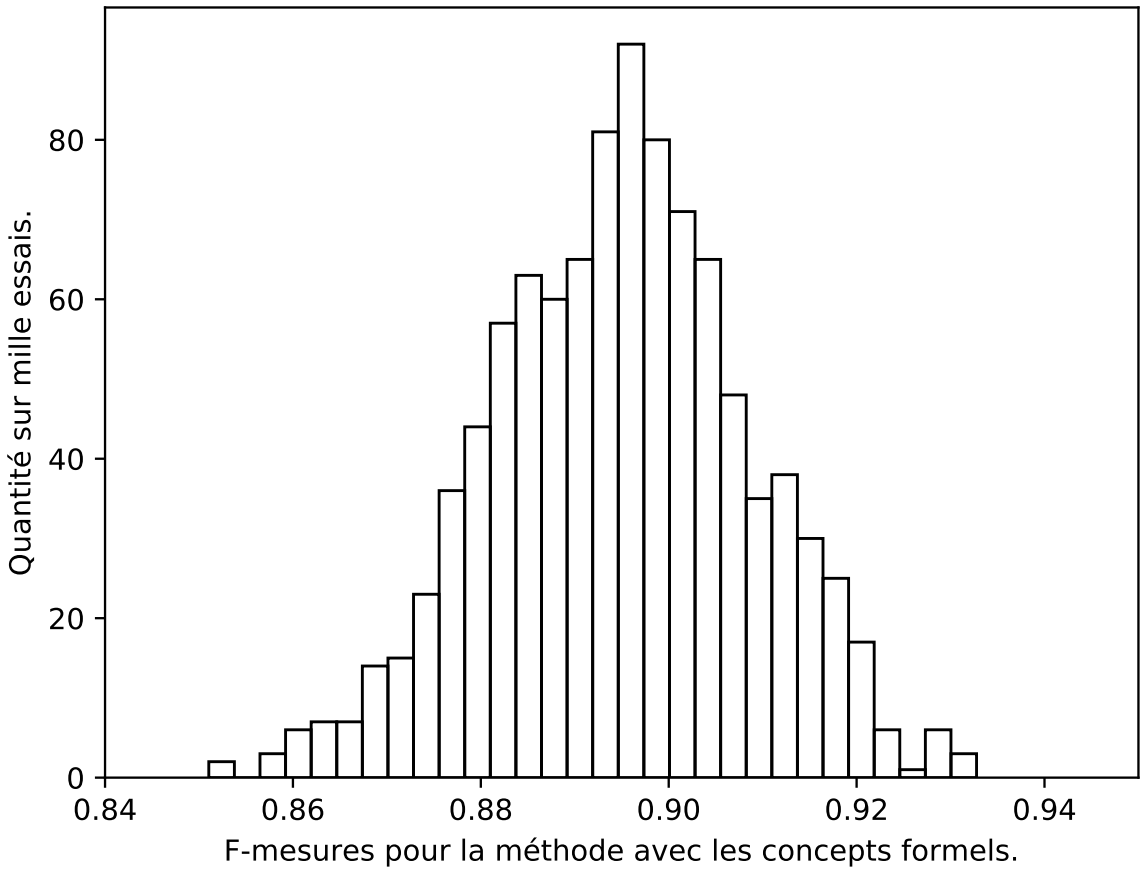


FIGURE – Histogramme de la répartition des f-mesures pour mille essais avec les concepts formels.

Observations

- La méthode avec les concepts est lente.
- De moins bons résultats.
- L'AFC apporte des informations discriminantes.

Conclusion

- Expérimentation sur un schéma particulier.
- Notre corpus est petit.

Conclusion

- Expérimentation sur un schéma particulier.
- Notre corpus est petit.

Idées

- Heuristique pour accélérer l'extraction des concepts.
- Assimiler les concepts à des variables de la logique des prédicats.

Bibliographie I


Ganter, Bernhard et Rudolf Wille (2012). *Formal concept analysis : mathematical foundations*. Springer Science & Business Media.


Krajca, Petr, Jan Outrata et Vilem Vychodil (2008). "Parallel recursive algorithm for FCA". In : *CLA*. T. 2008. Citeseer, p. 71–82.


Mandelbrot, Benoît (1957). "Etude de la loi dEstoup et de Zipf : fréquences des mots dans le discours". In : *Logique, langage et théorie de linformation*.