Rameau et l'automate : *que vaut l'indexation générée par une intelligence artificielle ?*

Les expérimentations menées par le Labo de l'Abes révèlent le potentiel de l'Intelligence artificielle comme outil d'indexation. Prochaine étape : tester en situation réelle.

D'un point de vue informatique, l'indexation automatique avec un vocabulaire tel que Rameau est une tâche complexe et ambitieuse. En termes techniques, on parlera d'une classification *multilabel* extrême. En effet, il s'agit bien de *classer,* c'est-à-dire de ranger les documents dans des cases prédéfinies, et non pas de regrouper les documents semblables (ce qu'on appelle « *clustering* »). En second lieu, cette classification est dite « multilabel » car un même document peut appartenir à différentes classes, c'est-à-dire être indexé par plusieurs concepts Rameau à la fois. Enfin, cette classification multilabel est dite *extrême* car le nombre de classes est très important, en l'occurrence autour de 100 000, ce qui complique considérablement l'affaire. Les caractéristiques de cette tâche en font un réel défi pour une machine, mais aussi pour les humains. Cette symétrie inhabituelle est un point important, que nous retrouverons plus loin.

ENTRAÎNER

Selon l'approche *machine learning,* nous essayons d'apprendre à la machine à indexer avec Rameau en lui soumettant de nombreux exemples. En l'occurrence, nous avons extrait du Sudoc un corpus d'environ 150 000 notices d'*ebooks* ayant une indexation Rameau et un résumé en français. Nous avons entraîné ces données avec deux techniques : le logiciel ANNIF (qui intègre différents algorithmes) et le calcul d'*embeddings* (vectorisation sémantique du texte). Avant de lancer le programme d'entraînement, nous avons mis de côté 30 % des notices extraites pour en faire un corpus de test. Après entraînement, nous demandons au programme de proposer une indexation Rameau pour ces notices, puis nous comparons ces propositions à l'indexation réelle (celle du Sudoc), pour savoir si la machine a « bien » travaillé ou pas. Toute la question est de savoir ce que signifie « bien » pour ce type de tâche.

ÉVALUER

En principe, le travail de la machine sera parfait si elle parvient à indexer ces notices de test exactement comme les catalogueurs du Sudoc :

- Tous les concepts du Sudoc ont été proposés par la machine (rappel = 1)
- Tous les concepts proposés par la machine sont présents dans les notices Sudoc (précision = 1).

Si cette approche de l'évaluation est pertinente pour des tâches de classification simple (binaire ou multiclasse), elle semble inadéquate pour notre classification multilabel, pour les raisons suivantes :

1. **Une indexation peut être en partie correcte.**
2. **Il n'y a pas qu'une seule manière de bien indexer une notice.**

1. Une indexation peut être en partie correcte

Si la machine propose un tiers des concepts Sudoc (rappel = 0,33) et qu'un tiers de ces propositions sont dans le Sudoc (précision = 0,33), on est loin d'un échec complet.
En l'occurrence, voici les scores obtenus avec l'un des algorithmes proposés par le logiciel ANNIF :

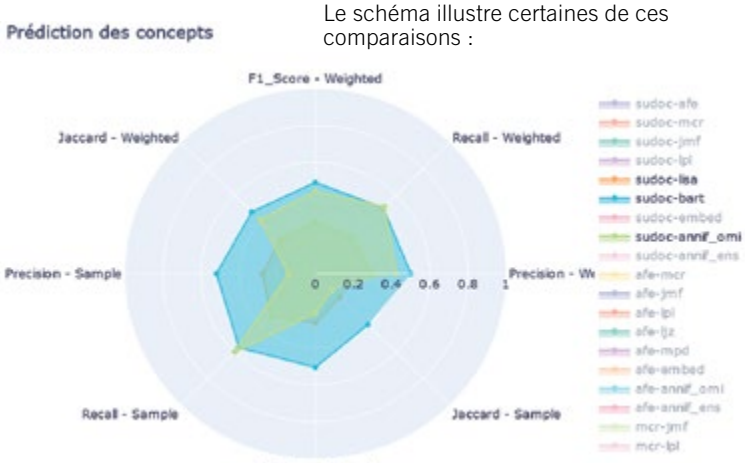
- Rappel = 0,25 (25 % des concepts Sudoc sont trouvés)
- Précision@5 = 0,35 (35 % des 5 premiers concepts proposés sont dans le Sudoc)

À ce stade, comme on le verra, il serait imprudent de conclure que ce résultat est décevant ou satisfaisant.

2. Il n'y a pas qu'une seule manière de bien indexer une notice

Qu'en est-il des concepts proposés par la machine et absents du Sudoc ? Sont-ils pour autant incorrects ? Il se peut que ces propositions originales couvrent une notion que l'indexation Sudoc couvre au moyen d'un autre concept Rameau proche, voire que la proposition machine exprime une notion négligée par l'indexation Rameau (incomplète dans ce cas). Bref, que l'indexation automatique soit meilleure que le laisse présager les chiffres ci-dessus.
Cette hypothèse est d'autant plus pertinente qu'on sait bien qu'il n'existe pas une seule manière d'indexer correctement un document. C'est vrai pour les humains. Il faut donc tenir compte de cette pluralité pour évaluer la machine.
Partant de cette intuition bien connue et documentée, nous avons décidé de « réindexer » une centaine de notices Sudoc. Six bibliothécaires de l'Abes, aux profils variés, ont été invités à proposer au maximum trois concepts (ou chaînes) Rameau, en ne disposant que du titre et du résumé.
À l'issue de ce travail d'autant plus fastidieux qu'il était contraint par les artifices inhérents à tout dispositif expérimental, nous disposions d'une pluralité d'indexations humaines et machine pour une centaine de notices permettant d'effectuer différentes comparaisons, plus riches et réalistes que la seule comparaison Sudoc/machine, et notamment de comparer chaque réindexation avec celle du Sudoc, avec l'algorithme machine, ou de comparer les réindexations entre elles.

Le schéma illustre certaines de ces comparaisons :



Même si la pluralité des méthodes de mesure complique la lecture, on peut voir que, pour les humains comme pour la machine, ni la précision ni le rappel ne dépassent 0.6. D'autre part, on constate que l'algorithme Omikuji d'ANNIF fait en général mieux que le réindexeur Lolo (le prénom a été modifié).

Compte tenu de la divergence entre les indexations produites par des bibliothécaires (Sudoc compris) à propos du même document, l'évaluation des propositions de la machine ne peut se limiter à mesurer la distance *absolue* entre celles-ci et les indexations Sudoc. Il est plus judicieux de se demander si la machine est beaucoup plus loin du Sudoc que les indexations humaines. Mais dans ce cas, pourquoi prendre le Sudoc comme point de référence ? Il est aussi légitime de privilégier n'importe quelle indexation humaine, voire une agrégation des indexations humaines.

Union ? Intersection ?

Le tableau ci-contre permet de comparer les propositions de deux algorithmes à l'union et l'intersection des réindexations pour l'ouvrage *Éthologie animale et humaine : communication et comportement* de Jacques Goldberg[1].

SUDOC + UNION DES 6 RÉINDEXEURS	INTERSECTION DES 6 RÉINDEXEURS	ANNIF	EMBEDDINGS V2
Sudoc	Éthologie comparée (4 fois)	Éthologie	***Comportement animal***
Comportement animal		Comportement animal	Manuels d'enseignement supérieur
Éthologie comparée	Comportement humain (2)	Relations homme-animal	Éthologie – Manuels d'enseignement supérieur
Éthologie humaine		Comportement humain	Neuroéthologie
Union des réindexations	Éthologie humaine (2)	Communication	
Éthologie comparée		Communication interpersonnelle	***Éthologie humaine***
Éthologie	Comportement animal (2)	Communication non verbale	Éthologie
Comportement animal		Éthologie humaine	Comportement social des animaux
Comportement humain	***Subdivisions***	Communication dans les organisations	
Éthologie humaine	Modèles animaux	Animaux	
Comportement humain Modèles animaux			
Communication Comportement social des animaux			
Sciences du comportement			

D'une manière générale, les indexations propres à la machine sont sans doute moins pertinentes que celles propres à tel ou tel bibliothécaire. Sans pour autant être systématiquement incorrectes ou incongrues. Comment le savoir ?

NOTER LES INDEXATIONS

Nous avons donc choisi d'ajouter une strate d'évaluation consistant à porter un jugement qualitatif sur les indexations proposées, celles du Sudoc, celles des « réindexeurs » et celles de la machine (plusieurs algorithmes). Pour autant, cette notation n'est pas arbitraire : elle suit un barème que nous avons défini pour noter à la fois la qualité de chaque proposition de concept et la qualité du groupe de concepts choisi par tel indexeur pour tel document.

Selon notre grille,

une proposition de concept est :

- exacte ou non (0 ou 1)
- plus ou moins précise (0 ou 1 ou 2)

Un groupe de concepts est :

- plus ou moins complet (0 ou 1 ou 2)
- redondant ou non (0 ou 1)

Le notateur s'appuie sur la lecture du titre et du résumé (comme la machine) pour donner une note sur chaque composante du barème. Il reste ensuite à effectuer la moyenne sur chaque composante, puis la moyenne globale.

	COMPLÉTUDE (0 OU 1 OU 2)	REDONDANCE (0 OU 1)
Sudoc	1,5	0,98
Union des réindexeurs	1,6	0,98
ANNIF (omikuji)	1,5	0,26
Embeddings v1	1,3	0,40

	EXACTITUDE (0 OU 1)	PRÉCISION (0 OU 1 OU 2)
Sudoc	0,99	1,8
Union des réindexeurs	0,98	1,9
ANNIF (omikuji)	0,60	1,5
ANNIF (omikuji). 2 premières propositions	0,86	1,6
Embeddings v1	0,66	1,6
Embeddings v1. 2 premières propositions	0,79	1,6

Même si des marges et des pistes d'amélioration existent, ces résultats nous semblent suffisamment bons pour envisager d'aller plus loin, à savoir expérimenter en situation, en intégrant un service de proposition d'indexation Rameau dans l'environnement de catalogage du Sudoc en tant qu'aide à la décision. Les modalités de cette expérimentation sont encore à préciser.

YANN NICOLAS
Responsable du Labo de l'Abes
nicolas@abes.fr

[1] https://www.sudoc.fr/147294509