Université Lumière Lyon 2 - ICOM

Projet Initiation à la recherche

Catégorisation d'articles scientifiques à partir des relations sémantiques des mots clés

Élèves :
Dorian LAMOTHE
Mame Libasse MBOUP
Saliou NDAO

11 avril 2021

Table des matières

1 Introduction

De nos jours, les grandes entreprises, les grandes administrations et les laboratoires de recherche font face à la croissance accélérée de production d'articles scientifiques. La gestion de cette masse de documents textuels a attiré l'attention de plusieurs chercheurs dans le domaine du traitement automatique de texte. Ainsi, ces derniers à travers leurs travaux essaient d'apporter des solutions pour répondre aux questions suivantes : Comment classer les documents ? Comment les stocker pour y retrouver rapidement les informations qu'ils contiennent ? Comment diffuser ces informations à ceux qui sauront les utiliser ? Comment filtrer une information pertinente parmi toutes les informations contenues dans les documents stockés ? Ainsi un besoin de les organiser de manière automatique et rapide s'impose pour une meilleure accessibilité. Cependant c'est une technique ancienne qui date des années 60 avec le travail acharné de Maron. [Mar61] qui a connu des progrès considérables notamment vers les années 90 avec l'apparition d'algorithmes beaucoup plus performants qu'auparavant et d'ordinateurs avec de meilleurs capacités de calcul. Beaucoup d'autres chercheurs se sont penchés aussi sur cette question tels que (Osborne Motta, 2012, 2014) qui ont par exemple travaillé sur l'analyse des communautés scientifiques en utilisant les liens de citations et des axinomies de topics de recherche dans le domaine informatique. C'est dans ce contexte que notre étude portera particulièrement sur la catégorisation d'articles scientifiques à partir des relations sémantiques des mots clés.

2 Résumé

En raison de la croissance exponentielle de données dans les publications scientifiques, il est aujourd'hui très difficile pour les chercheurs et lecteurs de trouver des articles ou documents scientifiques qui se rapportent à leur centre d'intérêts. Ainsi, mettre au point un système de catégorisation automatique de documents par thématique en se basant sur les relations sémantiques des mots clés pose un défi de taille. Nous avons fait une revue de la littérature afin d'étudier et de comprendre les différentes méthodologies de catégorisation automatique de textes proposées par les scientifiques. Une des possibilités ayant attirée notre attention est l'utilisation d'une encyclopédie en ligne, BabelNet, qui en fonction d'un ou plusieurs mots clés entrés en paramètre renvoie une catégorie ou domaine correspondant. Ainsi par ce document, nous comptons partir sur la même approche BabelNet, tout en proposant de nouvelles idées pour réaliser un tel outil. Pour y arriver nous avons décomposé le travail en deux parties : une première partie qui consiste à extraire des clusters d'articles grâce aux algorithmes de clustering (CAH ou k-means), puis une deuxième partie qui consistera à affecter aux articles d'un groupe une catégorie en se basant soit sur les mots clés les plus récurrents des articles du groupe soit sur la catégorie d'un article pris au hasard dans le cluster.

Mots clés : Sémantique, Automatique, Catéforisation, BabelNet, Clustering, CAH

3 État de l'art : Catégorisation automatique

3.1 Qu'est-ce que la catégorisation automatique ?

La catégorisation automatique de textes est une technique qui consiste à classer de manière automatisée des documents suivant certains critères comme le thème, les mots clés, le style etc. Autrement dit, elle consiste à associer un texte non-structuré à une étiquette, qui correspond à une classe bien précise.
Il existe plusieurs approches qui ont été abordées pour la catégorisation automatique de texte dans la littérature que nous étudierons en détails.

3.2 Technologies de catégorisation existantes

De manière générale, on distingue deux façons d'aborder le problème de la classification automatique de texte [Seb99] : soit sur la base de règles/mots-clés, soit sur la base d'un modèle de machine learning.

3.2.1 Ensemble Règles/Mots-clés

Jusqu'à la fin des années 80, la méthode dominante de résolution des problèmes de classification automatique de textes fut la construction manuelle d'ensemble de règles d'encodage des connaissances d'experts métiers du domaine. Ces règles étaient sous forme d'une implication logique où le texte à classifier (antécédent) portait, typiquement, sur la présence ou l'absence de certains mots, et où le conséquent désignait la catégorie d'appartenance du texte. Cette approche peut s'avérer très efficace. Entre autres, le système construis a démontré une précision de classification impressionnante, mais sur un corpus de textes relativement petit, ce qui ne permet pas de tirer des conclusions favorables [Yan99]. En fait, malgré de bons résultats sur des banques de textes bien précises, l'inconvénient de cette approche est que l'édition des règles de décision peut s'avérer très longue. Et surtout, si des catégories s'ajoutent ou si on désire utiliser le classificateur dans un autre domaine, on doit répéter l'exercice. C'est donc la pertinence de cet ensemble de règles, qui évolue dans le temps, qui mine l'intérêt envers cette façon de faire. En plus, c'est une méthode qui demande beaucoup de travail dont des tâches plutôt laborieuses, telles que la création de règles et de listes de mots-clés. Enfin, cette méthode est difficile à maintenir.

3.2.2 Un modèle de Machine Learning

Avec le développement rapide des algorithmes de Machine Learning, on peut appréhender le problème d'une autre façon. L'idée c'est de construire un système qui va apprendre par eux-mêmes à classifier les documents qu'on lui soumet. L'un des intérêts d'un modèle

de Machine Learning est que l'on n'a pas besoin de faire éditer des règles à des experts métiers car justement, le but de l'entraînement est de montrer des exemples au modèle afin qu'il décèle, de façon automatique, des règles qui permettront de classifier de nouvelles données. Les avantages sont une très bonne efficacité, une économie considérable en termes de main-d'œuvre experte et l'indépendance du domaine.
L'apprentissage automatique est principalement classé en trois types : l'apprentissage supervisé, l'apprentissage non supervisé et l'apprentissage par renforcement

Apprentissage Supervisé : L'apprentissage supervisé (ou classification) consiste à construire un modèle basé sur un jeu de données d'apprentissages et des labels ou étiquettes (nom des catégories ou des classes) et à l'utiliser pour classer de nouvelles données [SR09]. Cette technique est utilisée dans plusieurs applications telles que les systèmes de recommandation (youtube, e-commerce), la prédiction des pannes d'équipement, les systèmes de reconnaissance d'images.... D'une manière générale les algorithmes d'apprentissage supervisé sont utilisés pour résoudre :

— Classification : les problèmes de prédiction de variable qualitatives
— Régression : utilisée pour prédire une valeur réelle (variable quantitative). Par exemple, prédire la température en direct dans une pièce.

Quelques algorithmes de classification supervisée : Classification Bayésienne,Machine à vecteurs de support (SVM), Réseau de neurones, Forêts aléatoires etc.

Apprentissage non supervisé : Dans l'apprentissage non supervisé (en anglais clustering) on cherche à construire des groupes (classes ou clusters) d'objets (individus ou variables) similaires à partir d'un ensemble hétérogène d'objets. Chaque groupe résultant de ce processus doit vérifier les deux propriétés suivantes :

— les objets d'une classe doivent être le plus similaires possibles, on parle de cohésion interne.
— Association (comme le fait d'associer un produit à un client en fonctions des produits achetés précédemment)

Il existe plusieurs méthodes de classification, mais le point de similarité entre ces méthodes repose sur une mesure précise appelée distance ou métrique qui permet de juger de quantifier la similarité des objets à regrouper. Ces applications sont nombreuses notamment dans l'analyse de réseaux sociaux, l'analyse de documents textes, le marketing, la santé etc.

Apprentissage semi-supervisé : C'est une classe de techniques d'apprentissage automatique qui utilise un ensemble de données étiquetées et non étiquetées. Il se situe ainsi entre l'apprentissage supervisé qui n'utilise que des données étiquetées et l'apprentissage non supervisé qui n'utilise que des données non étiquetées. Il a été démontré que l'utilisation de données non étiquetées, en combinaison avec des

données étiquetées, permet d'améliorer significativement la qualité de l'apprentissage.

Un autre intérêt provient du fait que l'étiquetage de données nécessite souvent l'intervention d'un utilisateur humain. Lorsque les jeux de données deviennent très grands, cette opération peut s'avérer fastidieuse. Dans ce cas, l'apprentissage semi-supervisé, qui ne nécessite que quelques étiquettes, revêt un intérêt pratique évident.

3.3 Quelques applications de la catégorisation de textes

3.3.1 Organisation des documents

A l'heure ou les documents arrivent de façon très intensive , il est nécessaire de structurer ces données et de mettre en place une organisation adaptée. L'organisation de ces informations présente plusieurs avantages telles que la structuration de l'information, l'assurance de la cohérence et la fiabilité de l'information, et surtout pour avoir un accès aisé à l'information . Des exemples pratiques d'organisation de documents ; ranger sa bibliothèque, organiser ses dossiers ordinateur, astuces pour rangement décoration .
Toutes ces questions relatives à l'organisation peuvent être traiter par la catégorisation automatique . Différentes approches permettraient d'arriver à attribuer à chaque document une catégorie, de lui associer à un cluster comme la méthode du premier arrivé, premier servi.

3.3.2 Indexation des documents

Dans la littérature, s'y trouve différentes techniques de catégorisation automatique adressées à la pratique d'indexation de documents .
L'AIR/X, système à base de règles est une méthode permettant d'effectuer une indexation automatique avec des termes d'index (descripteurs) à partir d'un vocabulaire prescrit. Ce processus de plusieurs étapes , utilise des bases de règles qui seront développées à chacune de ces étapes . Principalement les bases de règles s'identifient à l'approche du thésaurus qui sont essentiellement dérivées du thésaurus du domaine du sujet. L'application phare de d'AIR/X est le système AIR/PHYS développé pour une grande base de données de physique. Ce système datant de 1985 , provient du développement du dictionnaire d'indexation PHYS/PILOT pour la base de données PHYS ayant un thésaurus avec 22 683 descripteurs. La base de données PHYS est composé de documents portant des numéros significatif des formules physiques ou chimiques ainsi que des données numériques . La méthode AIR/PHYS est bien détaillé sur cet article[Fuh+91].

4 Principe de notre approche

Dans le cadre de notre travail scientifique, nous comptions affecter des thèmes ou catégories à des articles scientifiques sans que ces mêmes thèmes ne soient appréhendés sur un échantillon d'apprentissage.

Dans ce cadre, nous avons étudié de près la littérature scientifique et conclu qu'une des possibilités était d'utiliser une encyclopédie en ligne qui nous renverrait une catégorie correspondant à un concept entré en paramètre (un ou plusieurs mots clés).

Nous avons fait le choix de BabelNet qui est un outil que nous détaillerons plus tard dans cet article.

Une des méthodologies que nous privilégions était d'affecter un domaine à chaque article en extrayant ces mêmes domaines depuis les mots-clés et ngrams.

4.1 Les données DBLP

Notre travail s'est articulé autour des données DBLP, qui est une encyclopédie bibliographique regroupant des articles scientifiques traitant principalement de machine learning.

Au sein de ces données nous avions à notre disposition les titres et les mots clés des différents articles.

Au cours de notre travail nous nous sommes rendus compte que cette base présentait plusieurs problèmes :

— les mots-clés sont en fait seulement les mots importants du titre de l'article ;
— le champ d'études à travers la base DBLP est restreint au seul Machine Learning.

Néanmoins, nous trouvions que c'était là un challenge jusque là non exploré, que nous voulions expérimenter.

Comme énoncé précédemment, nous privilégions l'approche individuelle pour chaque article en les catégorisant un par un. Cependant, nos articles présentent souvent plus de 10 mots-clés, et donc de multiples n-grams associés à ces derniers. Or l'API de BabelNet ne nous permet de faire que 1000 requêtes individuelles quotidiennes.

Si nous voulions améliorer notre outil, il nous fallait donc trouver une méthode pour éviter à avoir à faire ce grand nombre de requêtes.

Une idée qui nous est venu à l'esprit est d'extraire en amont des clusters d'articles, puis d'affecter la catégorie à l'ensemble des éléments du cluster en fonction des mots-clés qui reviennent le plus ou d'un article pris au hasard dans le cluster.

Pour cela nous avons extrait 300 articles au hasard de la base et procédé, en amont du clustering, au nettoyage de la base en enlevant les stopwords.

4.2 Clustering

Nous partons du principe selon lequel, certains de nos documents présentent des caractéristiques similaires, c'est-à-dire des centres d'intérêts communs. Nous jugeons ainsi pertinent d'appliquer une technique de Clustering dans un premier temps afin de regrouper ces derniers dans des clusters homogènes, puis procéder dans un deuxième temps à la labellisation de ces derniers en se concentrant uniquement sur les mots clés les plus importants d'un groupe donné.

4.2.1 Matrice termes-document (MDT)

L'idée consiste à considérer les mots-clés comme des descripteurs et d'utiliser une représentation en sac de mots de notre corpus d'articles. Nous obtenons ainsi un tableau individus-variables (individus = articles, variables = mots-clés) dont les valeurs représentent la présence ou non d'un mot clé dans un article. Il est dès lors propice au traitement à l'aide des algorithmes de data mining, en minimisant au possible la perte d'information.

La construction de cette matrice passe par un processus de tokenisation constitué de plusieurs étapes :

— repérer les mots (tokens) présents dans les documents en se concentrant uniquement sur ceux qui sont important pour notre étude. Chose facile grâce aux fonctionnalités offertes par NLP.
— ces derniers vont constituer le dictionnaire. Potentiellement, le nombre de mots est très important. Il peut y avoir des redondances dans le dictionnaire, il faudra pouvoir les traiter
— les mots deviennent des descripteurs(features, termes)
— on associe alors l'absence ou la présence des mots à chaque document

En suivant cette série d'étapes, nous obtenons la matrice Termes-Documents ci-dessous :

document	accurate	ace	acquisition	adapt	adaptive	ade	align	analyze	appearance	approach	...	variation	via	video	videos	view	viewpoints
0	0	0	0	0	0	0	0	0	0	1	...	0	0	0	0	0	0
10	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0
21	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0
28	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0
34	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0
44	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0
50	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0
60	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0
69	0	0	0	0	0	0	0	0	0	1	...	0	0	0	0	0	0
76	0	0	0	0	0	0	0	0	0	0	...	0	0	1	0	0	0
82	0	1	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0

FIGURE 1 – Matrice documents-termes

4.2.2 Application des algorithmes de clustering

Dans cette partie nous avons voulu mettre en évidence la similarité existante entre les différentes articles de notre jeu de données par application d'un algorithme de partitionnement tels que la classification ascendante hiérarchique (CAH) ou k-Means (méthodes plus adaptéds pour des données plus conséquentes). En effet plus deux documents sont similaires et plus ils auront tendance à utiliser les mêmes termes ; inversement, deux documents non similaires auront tendance a avoir des termes décorrélés. Il existe plusieurs mesures de ressemblance entre groupes d'individus : critère du saut minimal, lien complet, critère de ward etc. Nous avons utilisé la méthode de 'ward' qui permet une meilleure séparation des classes.

Ainsi l'algorithme va prendre en entrée notre matrice documents-termes. Comme la taille de notre matrice n'est pas très grande, nous avons opté pour la CAH qui ne nécessite pas une connaissance du nombre de classes ou groupes à constituer à priori. Ci-dessous une figure montrant le dendrogramme obtenu après application de la méthode :

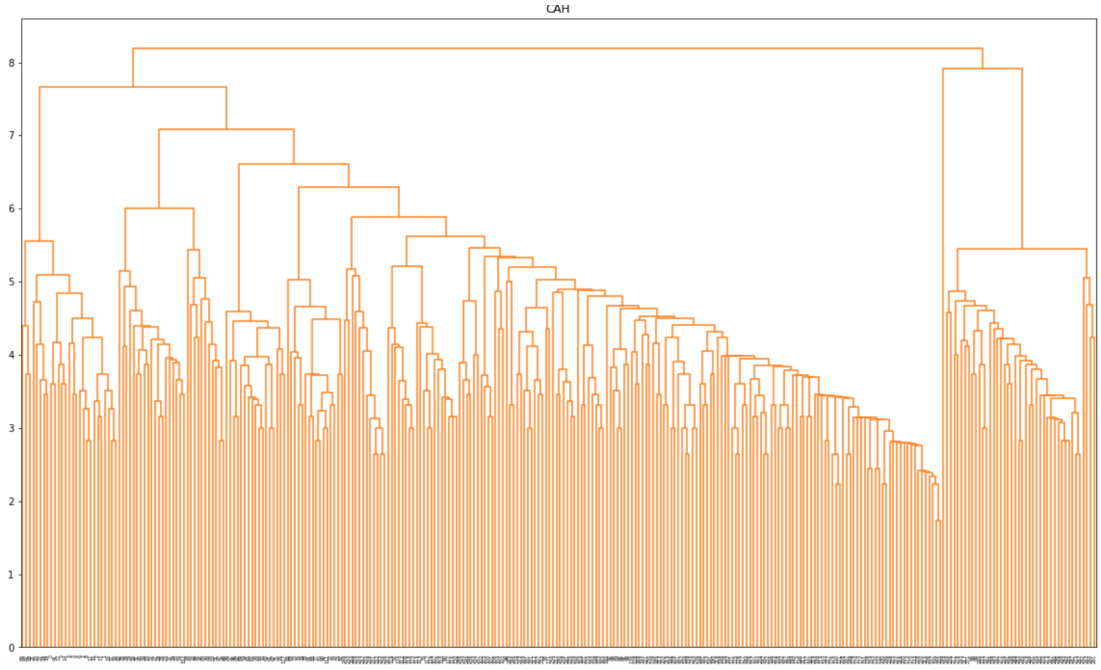


FIGURE 2 – Dendrogramme pour un jeux de 300 articles

.

En définissant un niveau de coupure, on obtient une partition c'est-à-dire le nombre de classes. Une partition est bonne si et seulement si les individus d'une même classe sont proches autrement dit les individus de deux classes différentes sont éloignés. Mathématiquement ça se traduit par :

— Une variabilité intra-classe petite : très peu de variabilité à l'intérieur d'une classe
— ou une variabilité inter-classe grande : une grande variabilité d'une classe à l'autre.

En effet ces deux critéres ne font qu'un. D'après le théoréme de Huygens, l'inertie totale que renferme nos données est égale à l'inertie intra-classe plus l'inertie inter classe. Et comme l'inertie totale est constante, minimiser l'inertie intra-classe revient à maximiser l'inertie inter-classe. Donc on peut se focaliser sur l'un des critères pour définir une meilleure partition.

La coupure au niveau 5 de notre dendrogramme nous permet d'obtenir ainsi 26 groupes d'articles. Ce niveau étant assez haut, les groupes constitués ne sont pas tous homogènes. Ce qui peut être minimisé en coupant plus bas.

Pour une bonne visualisation de nos groupes, nous avons pu implémenter une méthode factorielle (ACP) afin de pouvoir projeter les articles sur le plan factoriel à deux dimensions avec un code couleur selon le groupe d'appartenance. Nous avons choisit de mettre le numéro des documents comme labels pour une meilleure lisibilité.

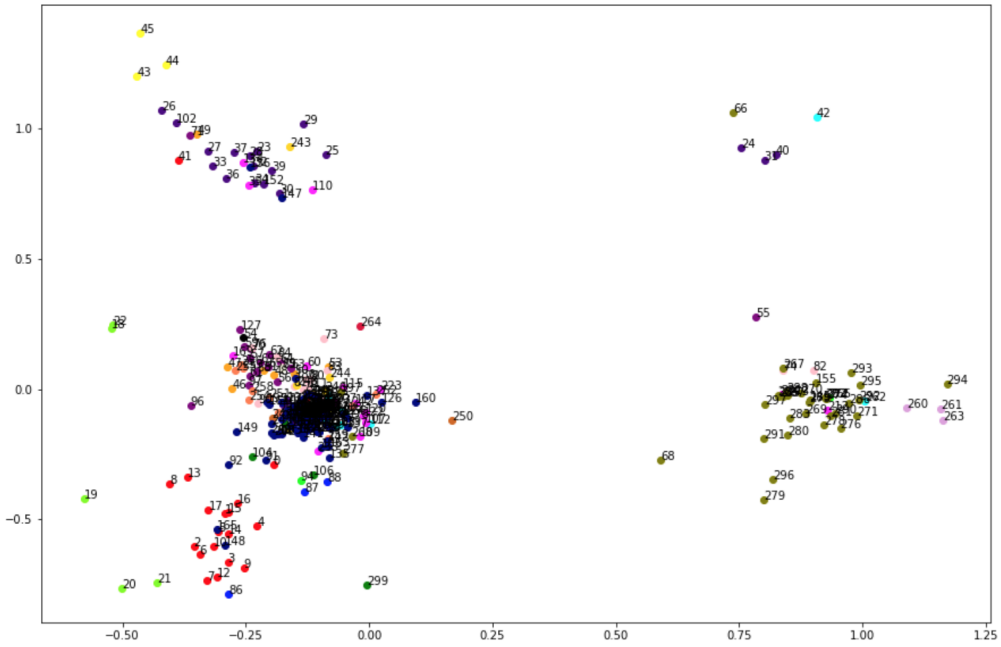


FIGURE 3 – Visualisation des groupes sur le plan factoriel

D'après la figure, on constate des couleurs qui se mélangent expliquant la non homogénéité de tous les groupes obtenus. Ainsi, en agissant sur le niveau de coupure du dendrogramme, on réduit cet effet. Ce qui sera fait dans la suite de nos travaux.

groupe	titre_article	keywords
20	Detect in RGB. Optimize in Edge: Accurate 6D Pose Estimation for Texture-less Industrial Parts.	detect rgb optimize edge accurate 6d pose estimation texture less industrial parts
21	A Semi-Blind Channel Estimation Algorithm for One-bit Massive MIMO Systems.	semi blind channel estimation algorithm one bit massive mimo systems
	Asynchronous testing of real-time systems.	asynchronous testing real time systems
	Comparative Characteristic Analysis of Circular and Double D Power Pads for Electric Vehicle Wireless Charging Systems.	comparative characteristic analysis circular double power pads electric vehicle wireless charging systems
	Differential Privacy Preservation for Smart Meter Systems.	differential privacy preservation smart meter systems
	Fuzzy-voting systems in smart cities.	fuzzy voting systems smart cities
	Literature review on smart lighting systems and their application in industrial settings.	literature review smart lighting systems application industrial settings
	On-Line State Estimation of 2-qubit Quantum Systems.	line state estimation qubit quantum systems
	Symposium: WISH - Workgroup on Interactive Systems in Healthcare.	symposium interactive systems healthcare
	Synchronization Tracking of a Class of Uncertain Nonidentical Networked Euler-Lagrange Systems on Directed Graphs.	synchronization tracking class uncertain nonidentical networked euler lagrange systems directed graphs
22	A fast and ultra low power time-based spiking neuromorphic architecture for embedded applications.	fast ultra low power time based spiking neuromorphic architecture embedded applications
	Impacts of Linear Controllers for Power Interfaces in Ideal Transformer Model Based Power Hardware-in-the-Loop.	impacts linear controllers power interfaces ideal transformer model based hardware loop
	Real-Time Basic Principles Nuclear Reactor Simulator Based on Client-Server Network Architecture with WebBrowser as User Interface.	real time basic principles nuclear simulator based client server network architecture user interface
	Semi-Partitioned Scheduling of Dynamic Real-Time Workload: A Practical Approach Based on Analysis-Driven Load Balancing.	semi partitioned scheduling dynamic real time workload practical approach based analysis driven load balancing
23	Time-varying formation control for general linear multi-agent systems with time-varying delays and switching topologies.	time varying formation control general linear multi agent systems time delays switching topologies
24	Fashion Is Taking Shape: Understanding Clothing Preference Based on Body Shape From Online Sources.	fashion taking shape understanding clothing preference based body online sources
	What Dress Fits Me Best?: Fashion Recommendation on the Clothing Style for Personal Body Shape.	fits best fashion recommendation clothing style personal body shape
25	A 20.5TOPS and 217.3GOPS/mm² Multicore SoC with DNN Accelerator and Image Signal Processor Complying with ISO26262 for Automotive Applications.	mm sup multicore soc dnn accelerator image signal processor automotive applications
	A Simulation Modeling Framework with Autonomous Vehicle Region-based Routing and Public Transit Diversion Integration.	simulation modeling framework autonomous vehicle region based routing public transit integration
	A document-based neural relevance model for effective clinical decision	document based neural relevance model effective clinical decision

FIGURE 4 – Extrait du jeux de données avec les groupes d'articles

Cette approche un peu fruste a le mérite de la simplicité et de la mécanisation. Mais elle ne règle pas le problème des synonymes.

Maintenant que nous avons obtenu nos différents groupes d'articles, chaque groupe traduisant la ressemblance entre les articles le constituant, nous pouvons aborder la partie catégorisation de ces derniers. L'idée c'est d'affecter une même catégorie aux articles d'un groupes (homogènes) en se basant soit :

— sur les mots clés les plus récurrents du groupes
— ou en considérant la catégorie d'un des articles du groupe.

Ainsi comme nous avons des groupes homogènes, la catégorie décelée pour un article du groupe pourra se généraliser sur l'ensemble des articles du même cluster. Ce que nous allons détailler dans la section suivante.

4.3 Étiquetage des groupes

4.3.1 L'outil BabelNet

BabelNet est un dictionnaire encyclopédique multilingue qui prend en entrée des lexiques sémantiques (WordNet) et d'autres bases de données encyclopédiques telles que

Wikipédia. Cet outil, disposant d'une API à la prise en main aisée, permet donc à l'utilisateur d'obtenir plusieurs renseignements sur un mot donné : ses catégories, ses domaines mais aussi ses synsets (concepts sémantiquement reliés au mot).

Dans la littérature scientifique, cet outil a été utilisé à plusieurs reprises, notamment dans le cadre de topic modelling.

Notre utilisation de l'outil se base donc sur l'utilisation de son API en deux étapes : on renseigne un concept/mot-clé au programme et ce dernier nous renvoit les synsets correspondants. Une fois ces synsets obtenus on peut extraire leurs catégories et leurs domaines correspondants.

Cependant plusieurs problèmes se sont mis en évidence au cours de la programmation :

— un mot/concept peut avoir plusieurs synsets (par exemple, le mot 'NATURAL' nous renvoit 44 définitions. Le programme ne peut pas savoir laquelle est la plus pertinente dans le cadre de la labellisation ;

— l'API BabelNet présente certains bugs qui ne nous permettent pas d'extraire la catégorie d'un concept alors que cette dernière est renseignée sur le site (par exemple, pour 'neural networks') ;

Notre méthodologie s'est axée autour de deux grandes méthodes d'étiquetage :

— la première était de prendre un article au hasard dans le cluster et d'assigner le domaine extrait à l'ensemble des articles du cluster ;

— la seconde était de prendre les 10 mots-clés dont l'occurrence est la plus grande dans le cluster et de chercher le domaine correpsondant le plus à ces derniers pour l'assigner à l'ensemble du cluster.

Méthode des mots-clés les plus redondants

On récupère dans un premier temps les n mots-clés qui reviennent le plus dans le cluster (n=10 dans notre exemple).

Une fois cette étape réalisée, on procède à la construction des trigrams et des bigrams. Cela consiste en l'assemblage des mots-clés deux à deux ou trois à trois. Ainsi, si l'on dispose des mots "machine" et "learning" dans notre liste ces derniers seront correctement interprétés.

Une fois cette étape remplie, il nous faut extraire les domaines correspondants aux 3-grams (trigrams), 2-grams (bigrams) et 1-grams (mots-clés uniques).

On reporte les résultats dans un dataframe et on compte ensuite quel domaine revient chez le plus grand nombre de concepts.

Par exemple, sur l'image ci-dessous, il s'agit des catégories 'COMPUTING' 'MEDIA' et 'MUSIC' car parmi les concepts existants sur BabelNet extraits des mots-clés ils reviennent chez 5 d'entre-eux.

Ainsi, nous avons du faire face à plusieurs problèmes :

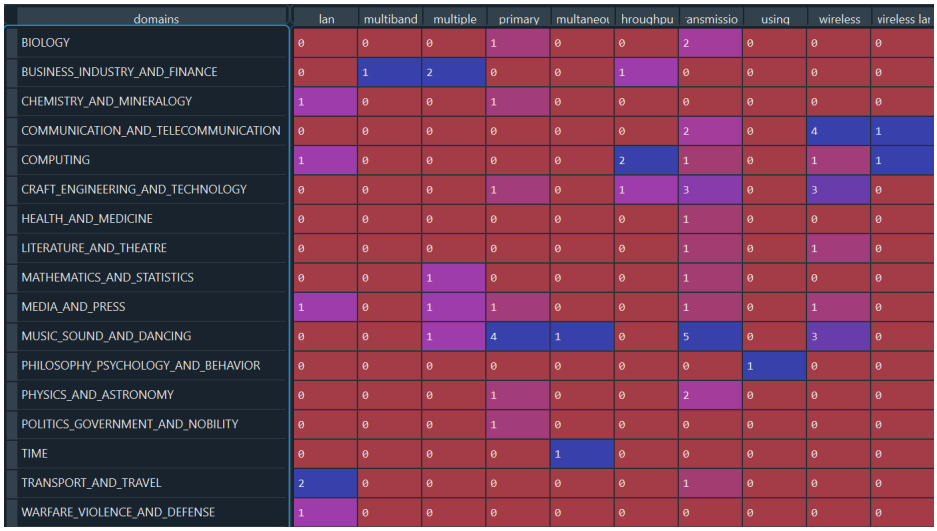
— certains clusters n'étaient pas assez peuplés. Ainsi, on se retrouvait souvent avec des mots-clés avec seulement une occurence.

— l'API BabelNet a un problème qui fait que certains concepts comme 'NEURAL NETWORKS' ont des domaines sur le site que l'API ne nous renvoit pas.

Au final, le but de notre démarche était de vérifier si les domaines renvoyés étaient principalement 'COMPUTING' et 'MATHEMATICS' car ce sont les sujets principaux de la base DBLP.

Nous aurions également pu nous intéresser aux catégories qui sont plus précises que les domaines. Cependant, ces dernières étaient beaucoup trop nombreuses et le fait de s'intéresser aux mots un par un nous renvoyait des catégories parfois absurdes.

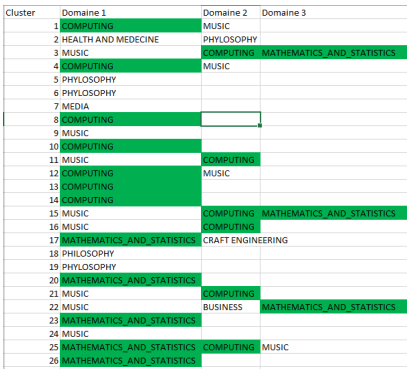
Nous nous sommes donc concentrés sur les 3 domaines qui revenaient le plus par cluster et voici une copie de nos résultats :



domains	lan	multiband	multiple	primary	multaneou	hroughpu	ansmissio	using	wireless	vireless lar
BIOLOGY	0	0	0	1	0	0	2	0	0	0
BUSINESS_INDUSTRY_AND_FINANCE	0	1	2	0	0	1	0	0	0	0
CHEMISTRY_AND_MINERALOGY	1	0	0	1	0	0	0	0	0	0
COMMUNICATION_AND_TELECOMMUNICATION	0	0	0	0	0	0	2	0	4	1
COMPUTING	1	0	0	0	0	2	1	0	1	1
CRAFT_ENGINEERING_AND_TECHNOLOGY	0	0	0	1	0	1	3	0	3	0
HEALTH_AND_MEDICINE	0	0	0	0	0	0	1	0	0	0
LITERATURE_AND_THEATRE	0	0	0	0	0	0	1	0	1	0
MATHEMATICS_AND_STATISTICS	0	0	1	0	0	0	1	0	0	0
MEDIA_AND_PRESS	1	0	1	1	0	0	1	0	1	0
MUSIC_SOUND_AND_DANCING	0	0	1	4	1	0	5	0	3	0
PHILOSOPHY_PSYCHOLOGY_AND_BEHAVIOR	0	0	0	0	0	0	0	1	0	0
PHYSICS_AND_ASTRONOMY	0	0	0	1	0	0	2	0	0	0
POLITICS_GOVERNMENT_AND_NOBILITY	0	0	0	1	0	0	0	0	0	0
TIME	0	0	0	0	1	0	0	0	0	0
TRANSPORT_AND_TRAVEL	2	0	0	0	0	0	1	0	0	0
WARFARE_VIOLENCE_AND_DEFENSE	1	0	0	0	0	0	0	0	0	0

FIGURE 5 – Domaines extraits d'un cluster

On répète ainsi le programme pour l'ensemble des clusters et obtenons les résultats suivants :



Cluster	Domaine 1	Domaine 2	Domaine 3
1	COMPUTING	MUSIC	
2	HEALTH AND MEDECINE	PHYLOSOPHY	
3	MUSIC	COMPUTING	MATHEMATICS_AND_STATISTICS
4	COMPUTING	MUSIC	
5	PHYLOSOPHY		
6	PHYLOSOPHY		
7	MEDIA		
8	COMPUTING		
9	MUSIC		
10	COMPUTING		
11	MUSIC	COMPUTING	
12	COMPUTING	MUSIC	
13	COMPUTING		
14	COMPUTING		
15	MUSIC	COMPUTING	MATHEMATICS_AND_STATISTICS
16	MUSIC	COMPUTING	
17	MATHEMATICS_AND_STATISTICS	CRAFT ENGINEERING	
18	PHILOSOPHY		
19	PHYLOSOPHY		
20	MATHEMATICS_AND_STATISTICS		
21	MUSIC	COMPUTING	
22	MUSIC	BUSINESS	MATHEMATICS_AND_STATISTICS
23	MATHEMATICS_AND_STATISTICS		
24	MUSIC		
25	MATHEMATICS_AND_STATISTICS	COMPUTING	MUSIC
26	MATHEMATICS_AND_STATISTICS		

FIGURE 6 – Domaines affiliés au cluster

Nos résultats sont stockés dans un Excel du fait que une même clé BabelNet ne permet pas de réasliser l'ensemble requêtes à l'API (cette dernière étant limitée à 1000). Ainsi, notre programme ne tourne pas sur tous les clusters mais bien sur un seul. On a donc ainsi un programme semi-automatisé.

On remarque ainsi que 13 clusters semblent affiliés à des bonnes catégories, donc la moitié. On trouve cependant le domaine 'MUSIC' qui revient souvent.

En effet les limites de cette méthode réside dans le fait que on dispose parfois de clusters peu peuplés et où les 10 premiers mots ne seront pas forcément représentatifs. Enfin comme nous l'avions souligné en amont un 1-gram va prendre plusieurs sens différents ce qui peut handicaper l'analyse.

Méthode de l'article pris au hasard

La méthode suivante réside elle dans le choix d'un article au hasard dans le cluster, du fait d'extraire le domaine et de l'affilier à l'ensemble des articles du groupe.

Pour cette dernière que nous avions développé vers la fin, nous avons juste testé sur quelques clusters et cela nous renvoyait globalement le même ratio que pour la méthode précédente.

De plus, ce choix de réaliser l'analyse seulement sur quelques clusters est aussi une contrainte. En effet, une clé BabelNet permet d'obtenir les domaines pour environ 4 clusters, suite à quoi on franchit généralement les 1000 requêtes maximales.

5 Conclusion

En conclusion, notre travail regroupe plusieurs idées inédites qui s'axent sur un jeu de données particulier et délicat. Nos résultats ne sont certes pas forcément exacts mais ouvrent à de nouveaux questionnements sur la catégorisation d'articles dans un domaine restreint comme le machine learning. Il y a plusieurs axes d'amélioration à notre programme. D'une part on pourrait pondérer l'importance des mots-clés lors de la labellisation pour donner plus de poids aux mots qui reviennent le plus. Pour la partie clustering elle est bien spécifique à cet article et intervient dans une optique de vulgarisation/facilitation. Cette partie pourrait être améliorée dans le sens où on pourrait faire une matrice termes-documents non pas sur les mots-clés mais sur les catégories renvoyées par BabelNet par exemple.

6 Remerciements

Ce travail a été réalisé dans le cadre de notre cours d'initiation à la recherche. Nous tenons à remercier Mme Fadila Bentayeb, professeur d'informatique à l'Université Lyon 2 et responsable du Master 1 Informatique, pour l'aide qu'elle nous a fournie et les connaissances qu'elle a su nous transmettre tout au long de ce cours. Nous lui adressons toute notre gratitude, pour sa disponibilité et la qualité de ses conseils, qui ont contribué à alimenter notre réflexion.

Références

[Mar61] M.E. Maron. "Automatic Indexing : An Experimental Inquiry". In : *Journal of the ACM (JACM)* 8.3 (1961), p. 404-417.

[Fuh+91] Norbert Fuhr et al. "AIR/X-a Rule Based Multistage Indexing System for Large Subject Fields". In : *RIAO*. T. 91. Citeseer. 1991, p. 606-623.

[Seb99] Fabrizio Sebastiani. "A tutorial on automated text categorisation". In : *Proceedings of ASAI-99, 1st Argentinian Symposium on Artificial Intelligence, pages 7–35, Buenos Aires, AR*. 1. 1999, p. 7-35.

[Yan99] Y. Yang. "An Evaluation of Statistical Approaches to Text Categorization. Information Retrieval". In : 1(1/2). 1999, p. 69-90.

[SR09] Silva et Ribeiro. "Inductive inference for large scale text classification". In : *kernel approaches and techniques*. T. 255. 2009.