Apprentissage d'une classification thématique générique et cross-langue à partir des catégories de la Wikipédia

François-Régis Chaumartin[1]
(1) Proxem, 19 boulevard de Magenta, 75010 Paris
frc@proxem.com

RESUME

La catégorisation de textes nécessite généralement un investissement important en amont, avec une adaptation de domaine. L'approche que nous proposons ici permet d'associer finement à un texte tout-venant écrit dans une langue donnée, un graphe de catégories de la Wikipédia dans cette langue. L'utilisation de l'index inter-langues de l'encyclopédie en ligne permet de plus d'obtenir un sous-ensemble de ce graphe dans la plupart des autres langues.

ABSTRACT

Cross-lingual and generic text categorization

Text categorization usually requires a significant investment, which must often be associated to a field adaptation. The approach we propose here allows to finely associate a graph of Wikipedia categories to any text written in a given language. Moreover, the inter-lingual index of the online encyclopedia allows to get a subset of this graph in most other languages.

MOTS-CLES : catégorisation, apprentissage, recherche d'information, Wikipédia, graphes
KEYWORDS: categorization, machine learning, information retrieval, Wikipedia, graphs

1 Présentation et objectifs

La catégorisation[1] est le processus qui consiste à associer à un document donné une ou plusieurs étiquettes prédéfinies. L'objectif d'une catégorisation automatique de textes est d'apprendre à la machine à effectuer cette classification en analysant son contenu. La nature même des catégories prédéfinies varie en fonction des objectifs ; il peut s'agir d'identifier la langue du texte, les thématiques abordées, mais aussi par exemple la priorisation souhaitée pour le traitement du document, ou encore les sentiments exprimés. La difficulté de la tâche varie selon le type et la longueur du document ; un tweet, un email, un article de presse, un document scientifique ou un avis de consommateur ne s'analysent généralement pas de la même façon.

Les étapes opérationnelle préalables à l'apprentissage d'une classification sont le plus souvent : i) la constitution du plan de classement, ii) l'annotation manuelle du corpus d'apprentissage, iii) la définition de caractéristiques linguistiques utilisées par l'algorithme d'apprentissage. Ces opérations peuvent être chronophages ; leur résultat n'est généralement applicable qu'au domaine particulier concerné par les catégories prédéfinies, et aux types de documents représentatifs du corpus d'apprentissage.

[1] On parle aussi de classification ; dans le milieu des sondages, on emploie plutôt le terme de codification.

L'approche que nous présentons ici concerne la **catégorisation thématique** ; notre objectif est d'identifier automatiquement les différents sujets dont parle un texte[2]. Notre ambition est double. D'une part, savoir traiter un document tout-venant (sous réserve d'une taille minimale) d'une façon **générique**, c'est-à-dire sans imposer préalablement une phase manuelle d'apprentissage spécifique au domaine ou à la langue du document. D'autre part, être capable de traduire (au moins certaines) des thématiques du document dans d'autres langues que celle du texte d'origine ; l'intérêt de ce point est d'autoriser alors une recherche **cross-langue** des documents associés à une thématique donnée.

2 Etat de l'art succinct

Si l'application de l'apprentissage automatique à la catégorisation de textes n'est pas nouvelle, son importance est grandissante. (Sebastiani, 2002) fournit un tableau comparatif des méthodes et applications possibles. (Dasari, Rao, 2012) complète cet état de l'art avec des approches plus récentes et mesure les progrès accomplis en 10 ans.

Une question se pose à propos des plans de classement, généralement définis pour un domaine particulier. Quel jeu d'étiquettes prédéfini serait suffisamment couvrant pour catégoriser d'une façon raisonnablement générique un texte tout venant ? Les catégories de la Wikipédia sont récemment apparues comme une possibilité de tel plan de classement universel. (Schönhofen, 2009) propose ainsi de les utiliser pour effectuer une catégorisation thématique avec un algorithme simple (dont l'implémentation met en œuvre un moteur de recherche) qui se contente d'exploiter les titres et les catégories des articles. Une idée proche est présentée dans (Yun *et al.*, 2011). Les catégories Wikipédia servent aussi de référence dans l'ontologie YAGO (Suchanek *et al.*, 2007).

3 Démarche

3.1 Utiliser les Wikipédia pour effectuer un apprentissage à large échelle

Les encyclopédies collaboratives Wikipédia figurent parmi les sources ayant de bonnes propriétés pour nous aider à atteindre notre objectif. En mars 2013, elles comptent 41 langues dotées de plus de 100 000 articles, et 70 autres avec au moins 10 000 articles. Ce volume permet de réaliser des apprentissages dans de nombreuses langues, dont certaines sont faiblement dotées en ressources lexicales.

Wikipédia propose différentes formes de structuration de l'information :

- Un article est classé dans une ou plusieurs catégories (en bas de chaque page) ;
- Les articles et catégories portant sur le même sujet, en différentes langues, sont reliés entre eux par l'intermédiaire d'un index interlingue (affiché à gauche) ;
- Les InfoBox présentent des données structurées sur un sujet sous forme de tables préformatées (encadrés placés en haut à droite ou en fin d'article) ;
- Les articles peuvent être rattachés à des portails, c'est-à-dire des regroupements thématiques offrant des points d'entrée dans l'encyclopédie ;
- Chaque article est organisé en sections et sous-sections.

[2] Par opposition à *ce que l'on en dit* ; nous ne chercherons pas ici à faire d'analyse d'opinions, par exemple.

Nous tirons parti notamment des deux premiers points. Les catégories sont organisées selon un graphe orienté au sein duquel une catégorie est reliée à d'autres, plus générales ou plus spécifiques. Par exemple[3], SCIENCE THEORIQUE et INFORMATIQUE sont les deux catégories mères de INFORMATIQUE THEORIQUE, qui possède 21 sous-catégories (ALGORITHMIQUE, CALCULABILITE …). Par ailleurs, 91 articles (Perceptron, Codage…) sont directement annotés avec (entre autres) la catégorie INFORMATIQUE THEORIQUE.

L'ensemble de ces graphes forme un plan de classement thématique cross-langue à large échelle. L'idée que présentons ici consiste à effectuer un apprentissage sur le contenu textuel des articles annotés par ces catégories. Nous allons mettre en œuvre pour cela des techniques –classiques et éprouvées– de recherche d'information, en stockant les résultats de cet apprentissage dans un moteur de recherche. La catégorisation d'un document revient alors simplement à effectuer une recherche dans l'index créé ; plus précisément, nous utiliserons le texte du document comme requête, et le moteur de recherche renverra comme résultat les catégories jugées les plus pertinentes.

3.2 Simplification des graphes de catégories de la Wikipédia

L'apprentissage est effectué sur chaque langue séparément. Nous commençons par charger en base de données la structure[4] fournie par le classique fichier XML d'import[5], pour en faciliter la manipulation ultérieure. Notre traitement commence par restructurer le graphe des catégories. Dans les versions que nous avons utilisées, celui de la Wikipédia en langue anglaise compte par exemple 438 251 sommets reliés par 949 017 arcs ; celui de la Wikipédia française contient 116 158 sommets et 230 217 arcs.

3.2.1 Détection et suppression des cycles

Chaque langue est organisée d'une façon spécifique, selon un graphe orienté de catégories qui possède une racine[6] (ou éventuellement plusieurs). La limite pratique des Wikipédia est la bonne volonté (ou la compétence) des internautes qui éditent les articles ; parfois, ils introduisent involontairement des cycles[7] entre catégories. La phase d'apprentissage devra explorer récursivement le graphe des racines jusqu'aux feuilles. Une opération préliminaire consiste donc à détecter puis supprimer ces cycles, de façon à travailler sur un graphe orienté acyclique (*directed acyclic graph* ou DAG en anglais) et éviter les boucles infinies. Nous appliquons pour cela l'algorithme décrit dans (Tarjan, 1972), qui détecte les zones fortement connexes d'un graphe orienté avec une exploration en profondeur à partir des racines.

[3] Nous noterons les catégories en petites majuscules plutôt que sous la forme « Catégorie: Libellé ».

[4] Les contenus textuels (balisés en syntaxe MediaWiki) ne sont pas importés pour des raisons de performance, mais l'empan permettant d'y accéder est créé en base de données lors de la lecture du fichier XML.

[5] Les fichiers XML compressés (« *dumps* ») sont téléchargeables sur http://dumps.wikimedia.org/

[6] C'est-à-dire une catégorie plus générale que toutes les autres. En français, elle est unique et s'appelle ARTICLE. Plusieurs racines coexistent pour l'anglais, proposant des organisations différentes ; nous avons choisi de partir de MAIN TOPIC CLASSIFICATIONS mais nous aurions aussi pu retenir FUNDAMENTAL CATEGORIES.

[7] En pratique, ces cycles existent dans les différentes Wikipédia, mais en nombre relativement faible.

La seconde étape consiste à enlever localement un arc jusqu'à supprimer tous les cycles. Le choix de l'arc à enlever a une dimension arbitraire ; nous privilégions ceux qui relient les sommets les plus bas dans la hiérarchie.

3.2.2 Suppression des catégories trop fines

Toutes les catégories ne sont pas pertinentes comme résultat d'un système de classification. En effet, beaucoup semblent avoir été créées pour pallier un déficit de structuration de la Wikipédia : par exemple, NAISSANCE PAR VILLE EN FRANCE compte plus de 1 000 sous-catégories correspondant à autant de villes ; CHRONOLOGIE PAR CONTINENT énumère des événements par année avec plus de 500 sous-catégories. Nous commençons par réduire ce graphe avec différentes heuristiques pour simplifier les manipulations informatiques ultérieures ; nous supprimons récursivement comme catégories trop fines :

- Les feuilles du graphe (les nœuds n'ayant pas de catégorie plus spécifique).
- Les catégories servant à annoter trop peu d'articles (10 dans notre expérience).

Lors de ces opérations, les articles directement reliés aux sommets supprimés sont alors annotés avec leur catégorie mère, de façon à préserver l'information correspondante. Au final, nous obtenons un DAG plus compact que celui d'origine (Cf. la table 1).

		En français	En anglais
Volumétrie initiale	Nombre de sommets	116 158	438 251
	Nombre d'arcs	230 217	949 017
Volumétrie après simplification	Nombre de sommets	59 267	229 626
	Nombre d'arcs	125 635	535 089

TABLE 1 – Volumétrie du graphe de catégories avant et après simplification.

Des heuristiques supplémentaires pourraient s'appliquer, par exemple à travers l'utilisation de patrons morphosyntaxiques pour détecter des noms de catégories particulières. Une catégorie comme NAISSANCE EN [ANNEE] n'est pas forcément pertinente dans notre problématique. Nous avons toutefois renoncé à cette approche, qui imposerait un paramétrage manuel particulier pour chaque langue, ce que nous souhaitons éviter.

3.3 Apprentissage par indexation dans un moteur de recherche

3.3.1 Principe général

Une fois le graphe simplifié, nous pouvons en indexer le contenu dans un moteur de recherche. L'objectif ici est d'associer à chaque catégorie un sac de mots (ou plus exactement un vecteur termes-fréquences) représentatif. La classification d'un document revient alors à utiliser ses termes comme critères de recherche ; le moteur renverra comme résultat les catégories les plus pertinentes, correspondant le mieux au document.

Notre implémentation met en œuvre le moteur de recherche *open source* Lucene[8]. Il permet d'indexer des textes selon une séquence d'opérations classique en recherche

[8] http://lucene.apache.org/

d'information : segmentation du texte en mots, normalisation de leur casse, suppression des diacritiques, suppression des mots grammaticaux (*stop words*), racinisation (*stemming*) et comptage des termes ; l'un des intérêts de Lucene est de proposer en standard ces opérations pour une trentaine de langues. Le résultat de ce processus est un vecteur des termes représentatifs des articles d'une catégorie, associés à leurs fréquences.

Le graphe simplifié des catégories est d'abord triée par ordre topologique inversé. L'indexation est effectuée avec une exploration récursive remontant des feuilles du DAG jusqu'à la racine. Le vecteur termes-fréquences d'une catégorie est calculé en fusionnant :

- Celui obtenu par le processus d'indexation décrit plus haut, appliqué au texte des articles directement annotés par la catégorie.
- Ceux déjà calculés sur ses k sous-catégories, pondérés par un facteur $1/(k+1)$.

On donne ainsi une importance prédominante aux termes des articles directement liées à la catégorie, tout en conservant la contribution due aux catégories plus spécifiques.

3.3.2 Amélioration du processus

Notre implémentation utilisant Lucene, les techniques classiques d'optimisation de moteur de recherche s'appliquent ici. En ce qui concerne la pertinence, une amélioration du mécanisme consiste à indexer aussi les termes composés ; leur utilisation lors de l'indexation et de la recherche améliore la pertinence des résultats, certes au prix d'une augmentation du temps de calcul. Nous utilisons des n-grammes[9] en plus des termes simples, avec n inférieur ou égal à 3 dans notre expérience[10].

Vus les volumes de texte manipulés, la taille de l'index Lucene peut devenir très importante (plusieurs giga-octets pour les langues les mieux dotées) ; elle s'accroît encore quand on indexe des n-grammes en plus des termes simples. Ce point a un impact direct sur les temps de recherche. De façon à limiter la taille de l'index et améliorer les performances, on peut choisir de ne pas indexer les hapax d'une encyclopédie en une langue donnée ; un examen manuel de l'index de la Wikipédia française montre par ailleurs que ce sont souvent des fautes d'orthographes, ce qui conforte ce choix. On peut aller plus loin dans cette démarche en enlevant les termes qui n'apparaissent que quelques fois dans le corpus. Nous avons retenu, dans notre expérience, les termes apparaissant 3 fois ou plus ; cela peut toutefois diminuer la qualité de l'apprentissage[11].

3.3.3 Stockage de l'information de structure du graphe

Chaque enregistrement indexé dans le moteur de recherche correspond à une catégorie Wikipédia donnée. Il contient son titre ainsi que le vecteur des termes qui lui sont associés directement (issus des articles de la catégorie) ou indirectement (via les sous-catégories). L'enregistrement stocke aussi des éléments de structure du graphe :

---

[9] Dans Lucene, les n-grammes sont appelés *shingles* (« bardeaux » en français : petites tuiles qui se recouvrent).

[10] Les termes composés les plus fréquents dans Wikipédia sont *championnat du monde*, *jeux olympiques*, *premier ministre* ou *guerre mondiale* en français (*United States*, *London borough*, *United Kingdom* ou *NHL league* en anglais)

[11] Par exemple, la banque grecque *Emporiki* n'apparaissait que deux fois dans la Wikipédia française avant 2008. Cela induisait de mauvaises catégorisations sur des textes courts récents parlant de la crise financière.

– La liste des catégories mères et des sous-catégories au sein d'une langue donnée. Cette information sera utilisée pour afficher le résultat sous forme graphique et aussi pour effectuer un filtrage améliorant la pertinence de la catégorisation.
– Les liens de l'index inter-langues, correspondant aux « traductions » de la catégorie vers les autres Wikipédia. Cette information servira à afficher les résultats d'une catégorisation d'une façon cross-langue.

La complétude de l'index inter-langues est aléatoire ; elle varie énormément en fonction des catégories[12]. Pour celles jugées importantes aux yeux des wikinautes, des liens sont fournis vers un grand nombre de langues ; en revanche, aucun lien n'existera parfois pour une catégorie trop fine ou d'intérêt secondaire.

3.4 Catégorisation d'un document

3.4.1 Principe

Une fois l'index Lucene constitué, la catégorisation d'un document devient trivialement simple, et revient à faire une recherche en utilisant le texte du document comme critère. Plus précisément, le texte est analysé avec le même processus qui a servi à l'indexation, y compris l'extraction des termes composés. Le vecteur termes-fréquences obtenu est alors utilisé par le moteur de recherche pour trouver les documents de l'index (correspondant aux catégories Wikipédia) les plus proches du texte, avec une pondération TF-IDF[13].

3.4.2 Exemple : catégorisation du présent article

Le résultat brut de la recherche est une liste à plat de catégories associées à un score de pertinence. La figure 1 illustre le résultat de la catégorisation thématique obtenue à partir du texte du présent article. Nous obtenons : RECHERCHE D'INFORMATION = 0,258 ; MOTEUR DE RECHERCHE = 0,203 ; MOT-VALISE = 0,198 ; INTELLIGENCE ARTIFICIELLE = 0,186 ; INFORMATIQUE THEORIQUE = 0,183 ; TRAITEMENT AUTOMATIQUE DU LANGAGE NATUREL = 0,173.

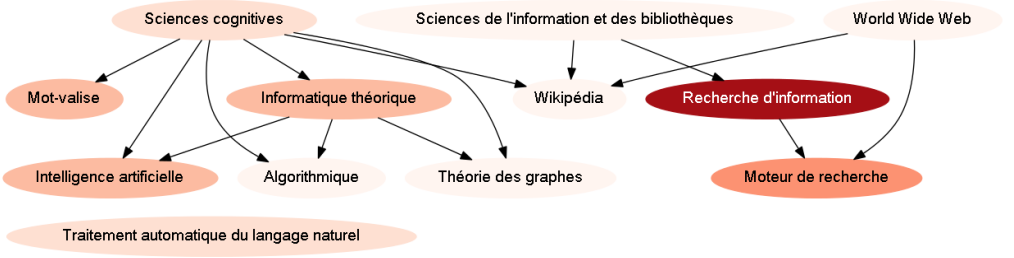

FIGURE 1 – Catégorisation thématique obtenue sur le texte du présent article.

[12] Notons que depuis mars 2013, la Wikipédia en français centralise les liens inter-langues dans Wikidata, une base de données structurée libre. Cela devrait contribuer à élargir et fiabiliser l'index inter-langues.

[13] TF-IDF (*term frequency-inverse document frequency*) est une méthode de pondération classique. Avec cette mesure statistique, le poids d'un terme augmente proportionnellement à son nombre d'occurrences dans le texte à catégoriser. Il varie également en fonction de la fréquence du terme dans l'index des catégories.

Dans les figures présentées ici, ces scores se traduisent visuellement par une couleur de fond d'autant plus sombre que la catégorie est pertinente ; pour les catégories reliées par des arcs, les plus générales s'affichent en haut et les plus spécifiques en bas. La structure locale du graphe (arcs entrants et sortants) est également stockée avec chaque catégorie (Cf. 3.3.3). Nous utilisons cette information pour reconstituer un graphe à partir de la liste à plat produite par le moteur de recherche. Un lecteur humain aura ainsi une visualisation plus riche qu'une simple liste. L'autre intérêt est de tenir compte de la géométrie locale du graphe de catégories pour établir une heuristique de filtrage supplémentaire. Si un sommet isolé ou de degré 1 présente aussi une pertinence trop faible, il est supprimé[14] ; ce filtrage augmente la précision de la catégorisation.

Enfin, les liens de l'index inter-langues permettent de passer sans effort de la figure 1 (en français) aux figures 2 (en anglais) et 3 (en allemand). On remarque que dans les deux cas, on n'obtient qu'un sous-graphe de celui en français. Les catégories MOT-VALISE et ALGORITHMIQUE n'ont pas d'équivalent exact en anglais ; de même, ALGORITHMIQUE manque en allemand, ainsi que TRAITEMENT AUTOMATIQUE DU LANGAGE NATUREL.

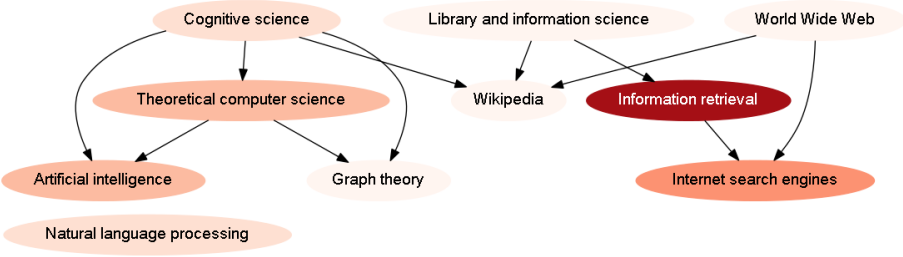

FIGURE 2 – Catégorisation thématique obtenue en anglais sur le texte du présent article.

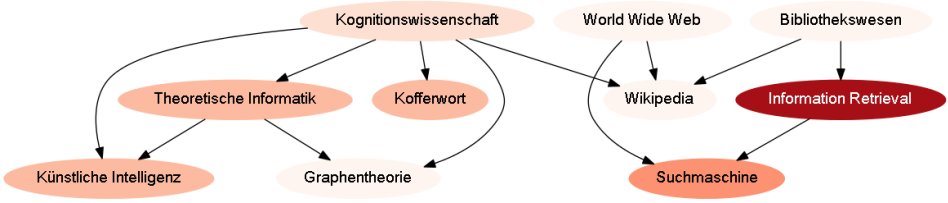

FIGURE 3 – Catégorisation thématique obtenue en allemand sur le texte du présent article.

4 Evaluation

Nous avons effectué une mesure préliminaire des résultats de notre algorithme pour prédire les bonnes catégories sur les articles de la Wikipédia française eux-mêmes, en utilisant uniquement leur contenu. Nous avons procédé à une validation croisée en

[14] Le seuil retenu ici est une pertinence inférieure à la moitié de celle de la catégorie la plus pertinente. Dans notre exemple, DEVELOPPEMENT LOGICIEL apparaissait initialement dans le graphe résultat, mais avec une pertinence inférieure au seuil (0,125) et un seul arc (vers ALGORITHMIQUE) ; elle a donc été enlevée.

divisant les articles de l'encyclopédie en 10 échantillons de même taille. Nous effectuons un apprentissage sur 9 d'entre eux, suivi d'un test sur le 10ème échantillon ; ce test est répété 10 fois en changeant à chaque fois l'échantillon de test. Chaque document testé est explicitement annoté par *k* catégories « officielles ». Le test lui-même consiste à prédire 10 catégories, puis à vérifier si on retrouve au moins l'une des catégories prédites dans les catégories officielles ; le résultat de chaque test élémentaire vaut donc 0 ou 1. Avec cette approche, la moyenne sur les 10 échantillons est de 91%. Ce protocole reste simpliste ; nous comptons l'améliorer en nous inspirant de celui utilisé pour le LSHTC challenge (Large Scale Hierarchical Text Classification, http://lshtc.iit.demokritos.gr).

5 Bilan et travaux futurs

Nous avons présenté une approche opérationnelle pour classifier du texte tout-venant écrit dans l'une des langues pour lesquelles il existe une encyclopédie Wikipédia. Le composant a été intégré à la plate-forme Antelope (Chaumartin, 2012) et utilisé avec un succès dans des projets de catégorisation de sites Web et de flux RSS en environnement multilingues. Notre démarche présente des similitudes avec (Schönhofen, 2009) ; nos contributions portent sur (i) l'amélioration de la qualité de l'indexation (processus d'exploration récursive partant des feuilles et utilisation des *n*-grammes sur l'intégralité du texte des articles) ; (ii) L'utilisation de la topologie du graphe résultat pour élaguer les catégories peu pertinentes ; (iii) l'exploitation de l'index inter-langue pour présenter une traduction (au moins partielle) du résultat dans les autres langues disposant aussi d'une Wikipédia. Les résultats sont encourageants mais peuvent encore être améliorés.

Remerciements

Je remercie les ingénieurs de Proxem pour leur soutien, notamment Fanny Parganin.

Références

CHAUMARTIN, F.-R. (2012). *Antelope, une plate-forme de TAL permettant d'extraire les sens du texte : théorie et applications de l'ISS*. Thèse de doctorat, Université Paris Diderot.

DASARI, D. B., RAO V. G. (2012). Text Categorization and Machine Learning Methods: Current State of the Art. In *GJCST*, Vol. 12, N°11.

SCHÖNHOFEN, P. (2009). Identifying document topics using the Wikipedia category network. In *Web Intelligence and Agent Systems*, Vol. 7, N°2, pages 195-207.

SEBASTIANI, F. (2002). Machine Learning in Automated Text Categorization. In *ACM Computing Surveys*, Vol. 34, N°1, pages 1-47.

SUCHANEK F., KASNECI G., WEIKUM G. (2007). Yago: a core of semantic knowledge. In *WWW 2007*, pp. 697-706.

TARJAN, R. E. (1972). Depth-first search and linear graph algorithms. In *SIAM Journal on Computing*, Vol. 1, N°2, p. 146–160.

YUN, J., JING, L., YU, J., HUANG, H., ZHANG, Y. (2011). Document Topic Extraction Based on Wikipedia Category. Actes de *Computational Sciences and Optimization (CSO)*.